

International Journal on Advances in Intelligent Systems



The *International Journal on Advances in Intelligent Systems* is Published by IARIA.

ISSN: 1942-2679

journals site: <http://www.iariajournals.org>

contact: petre@iaria.org

Responsibility for the contents rests upon the authors and not upon IARIA, nor on IARIA volunteers, staff, or contractors.

IARIA is the owner of the publication and of editorial aspects. IARIA reserves the right to update the content for quality improvements.

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy or print, providing the reference is mentioned and that the resulting material is made available at no cost.

Reference should mention:

International Journal on Advances in Intelligent Systems, issn 1942-2679
vol. 6, no. 3 & 4, year 2013, http://www.iariajournals.org/intelligent_systems/

The copyright for each included paper belongs to the authors. Republishing of same material, by authors or persons or organizations, is not allowed. Reprint rights can be granted by IARIA or by the authors, and must include proper reference.

Reference to an article in the journal is as follows:

<Author list>, "<Article title>"
International Journal on Advances in Intelligent Systems, issn 1942-2679
vol. 6, no. 3 & 4, year 2013, <start page>:<end page> , http://www.iariajournals.org/intelligent_systems/

IARIA journals are made available for free, proving the appropriate references are made when their content is used.

Sponsored by IARIA

www.iaria.org

Copyright © 2013 IARIA

Editor-in-Chief

Freimut Bodendorf, University of Erlangen-Nuernberg, Germany

Editorial Advisory Board

Dominic Greenwood, Whiteston Technologies AG, Switzerland

Josef Noll, UiO/UNIK, Norway

Said Tazi, LAAS-CNRS, Universite Toulouse 1, France

Radu Calinescu, Oxford University, UK

Editorial Board

Jemal Abawajy, Deakin University - Victoria, Australia

Sherif Abdelwahed, Mississippi State University, USA

Habtamu Abie, Norwegian Computing Center/Norsk Regnesentral-Blindern, Norway

Siby Abraham, University of Mumbai, India

Witold Abramowicz, Poznan University of Economics, Poland

Imad Abugessaisa, Karolinska Institutet, Sweden

Arden Agopyan, CloudArena, Turkey

Dana Al Kukhun, IRIT - University of Toulouse III, France

Leila Alem, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia

Panos Alexopoulos, ISOCO, Spain

Vincenzo Ambriola, Università di Pisa, Italy

Junia Anacleto, Federal University of Sao Carlos, Brazil

Razvan Andonie, Central Washington University, USA

Cosimo Anglano, DISIT - Computer Science Institute, Università del Piemonte Orientale, Italy

Richard Anthony, University of Greenwich, UK

Avi Arampatzis, Democritus University of Thrace, Greece

Sofia J. Athenikos, Amazon, USA

Isabel Azevedo, ISEP-IPP, Portugal

Costin Badica, University of Craiova, Romania

Ebrahim Bagheri, Athabasca University, Canada

Fernanda Baiao, Federal University of the state of Rio de Janeiro (UNIRIO), Brazil

Flavien Balbo, University of Paris Dauphine, France

Suliman Bani-Ahmad, School of Information Technology, Al-Balqa Applied University, Jordan

Ali Barati, Islamic Azad University, Dezful Branch, Iran

Henri Basson, University of Lille North of France (Littoral), France

Carlos Becker Westphall, Federal University of Santa Catarina, Brazil

Ali Beklen, Cloud Arena, Turkey

Helmi Ben Hmida, FH MAINZ, Germany

Petr Berka, University of Economics, Czech Republic

Julita Bermejo-Alonso, Universidad Politécnica de Madrid, Spain
Aurelio Bermúdez Marín, Universidad de Castilla-La Mancha, Spain
Lasse Berntzen, Vestfold University College - Tønsberg, Norway
Michela Bertolotto, University College Dublin, Ireland
Ateet Bhalla, Oriental Institute of Science & Technology, Bhopal, India
Freimut Bodendorf, Universität Erlangen-Nürnberg, Germany
Karsten Böhm, FH Kufstein Tirol - University of Applied Sciences, Austria
Pierre Borne, Ecole Centrale de Lille, France
Marko Bošković, Research Studios, Austria
Christos Bouras, University of Patras, Greece
Anne Boyer, LORIA - Nancy Université / KIWI Research team, France
Stainam Brandao, COPPE/Federal University of Rio de Janeiro, Brazil
Stefano Bromuri, University of Applied Sciences Western Switzerland, Switzerland
Vít Bršlica, University of Defence - Brno, Czech Republic
Dumitru Burdescu, University of Craiova, Romania
Diletta Romana Cacciagrano, University of Camerino, Italy
Kenneth P. Camilleri, University of Malta - Msida, Malta
Paolo Campegnani, University of Rome Tor Vergata, Italy
Marcelino Campos Oliveira Silva, Chemtech - A Siemens Business / Federal University of Rio de Janeiro, Brazil
Ozgu Can, Ege University, Turkey
José Manuel Cantera Fonseca, Telefónica Investigación y Desarrollo (R&D), Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Miriam A. M. Capretz, The University of Western Ontario, Canada
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Davide Carboni, CRS4 Research Center - Sardinia, Italy
Luis Carriço, University of Lisbon, Portugal
Rafael Casado Gonzalez, Universidad de Castilla - La Mancha, Spain
Michelangelo Ceci, University of Bari, Italy
Fernando Cerdan, Polytechnic University of Cartagena, Spain
Alexandra Suzana Cernian, University "Politehnica" of Bucharest, Romania
Carlos Cetina, Technical Universidad San Jorge, Spain
Sukalpa Chanda, Gjøvik University College, Norway
David Chen, University Bordeaux 1, France
Luke Chen, University of Ulster @ Jordanstown, UK
Ping Chen, University of Houston-Downtown, USA
Kong Cheng, Telcordia Research, USA
Po-Hsun Cheng, National Kaohsiung Normal University, Taiwan
Dickson Chiu, Dickson Computer Systems, Hong Kong
Sunil Choenni, Research & Documentation Centre, Ministry of Security and Justice / Rotterdam University of Applied Sciences, The Netherlands
Ryszard S. Choras, University of Technology & Life Sciences, Poland
Smitashree Choudhury, Knowledge Media Institute, The UK Open University, UK
William Cheng-Chung Chu, Tunghai University, Taiwan
Christophe Claramunt, Naval Academy Research Institute, France
Cesar A. Collazos, Universidad del Cauca, Colombia
Phan Cong-Vinh, NTT University, Vietnam

Christophe Cruz, University of Bourgogne, France
Beata Czarnacka-Chrobot, Warsaw School of Economics, Department of Business Informatics, Poland
Claudia d'Amato, University of Bari, Italy
Sérgio Roberto P. da Silva, Universidade Estadual de Maringá - Paraná, Brazil
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Dragos Datcu, Netherlands Defense Academy / Delft University of Technology , The Netherlands
Antonio De Nicola, ENEA, Italy
Claudio de Castro Monteiro, Federal Institute of Education, Science and Technology of Tocantins, Brazil
Noel De Palma, Joseph Fourier University, France
Jan Dedek, Charles University in Prague, Czech Republic
Zhi-Hong Deng, Peking University, China
Stojan Denic, Toshiba Research Europe Limited, UK
Vivek S. Deshpande, MIT College of Engineering - Pune, India
Sotirios Ch. Diamantas, Pusan National University, South Korea
Leandro Dias da Silva, Universidade Federal de Alagoas, Brazil
Jerome Dinet, Univeristé Paul Verlaine - Metz, France
Jianguo Ding, University of Luxembourg, Luxembourg
Yulin Ding, Defence Science & Technology Organisation Edinburgh, Australia
Alexiei Dingli, University of Malta, Malta
Mihaela Dinsoreanu, Technical University of Cluj-Napoca, Romania
Ioanna Dionysiou, University of Nicosia, Cyprus
Roland Dodd, CQUniversity, Australia
Nima Dokoochaki, Royal Institute of Technology (KTH)-Kista, Sweden
Suzana Dragicevic, Simon Fraser University- Burnaby, Canada
Mauro Dragone, University College Dublin (UCD), Ireland
Marek J. Druzdzel, University of Pittsburgh, USA
Carlos Duarte, University of Lisbon, Portugal
Raimund K. Ege, Northern Illinois University, USA
Jorge Ejarque, Barcelona Supercomputing Center, Spain
Larbi Esmahi, Athabasca University, Canada
Simon G. Fabri, University of Malta, Malta
Umar Farooq, Amazon.com, USA
Mehdi Farshbaf-Sahih-Sorkhabi, Azad University - Tehran / Fanavaran co., Tehran, Iran
Anna Fensel, Semantic Technology Institute (STI) Innsbruck and FTW Forschungszentrum Telekommunikation
Wien, Austria
Stenio Fernandes, Federal University of Pernambuco (CIn/UFPE), Brazil
Oscar Ferrandez Escamez, University of Utah, USA
Florin Filip, Romanian Academy, Romania
Agata Filipowska, Poznan University of Economics, Poland
Ziny Flikop, Scientist, USA
Adina Magda Florea, University "Politehnica" of Bucharest, Romania
Francesco Fontanella, University of Cassino and Southern Lazio, Italy
Panagiotis Fotaris, University of Macedonia, Greece
Enrico Francesconi, ITTIG - CNR / Institute of Legal Information Theory and Techniques / Italian National Research
Council, Italy
Rita Francese, Università di Salerno - Fisciano, Italy

Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria
Sören Frey, syscovery Business Solutions GmbH, Germany
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Somchart Fugkeaw, Thai Digital ID Co., Ltd., Thailand
Naoki Fukuta, Shizuoka University, Japan
Mathias Funk, Eindhoven University of Technology, The Netherlands
Adam M. Gadomski, Università degli Studi di Roma La Sapienza, Italy
Alex Galis, University College London (UCL), UK
Crescenzo Gallo, Department of Clinical and Experimental Medicine - University of Foggia, Italy
Matjaz Gams, Jozef Stefan Institute-Ljubljana, Slovenia
Raúl García Castro, Universidad Politécnica de Madrid, Spain
Fabio Gasparetti, Roma Tre University - Artificial Intelligence Lab, Italy
Joseph A. Giampapa, Carnegie Mellon University, USA
George Giannakopoulos, NCSR Demokritos, Greece
David Gil, University of Alicante, Spain
Harald Gjermundrod, University of Nicosia, Cyprus
Angelantonio Gnazzo, Telecom Italia - Torino, Italy
Luis Gomes, Universidade Nova Lisboa, Portugal
Nan-Wei Gong, MIT Media Laboratory, USA
Francisco Alejandro Gonzale-Horta, National Institute for Astrophysics, Optics, and Electronics (INAOE), Mexico
Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece
Victor Govindaswamy, Texas A&M University-Texarkana, USA
Gregor Grambow, University of Ulm, Germany
Fabio Grandi, University of Bologna, Italy
Andrina Granić, University of Split, Croatia
Carmine Gravino, Università degli Studi di Salerno, Italy
Dominic Greenwood, Whitestep Technologies, Switzerland
Michael Grottko, University of Erlangen-Nuremberg, Germany
Vic Grout, Glyndŵr University, UK
Maik Günther, Stadtwerke München GmbH, Germany
Francesco Guerra, University of Modena and Reggio Emilia, Italy
Alessio Gugliotta, Innova SPA, Italy
Richard Gunstone, Bournemouth University, UK
Fikret Gurgen, Bogazici University, Turkey
Ivan Habernal, University of West Bohemia, Czech Republic
Maki Habib, The American University in Cairo, Egypt
Till Halbach Røssvoll, Norwegian Computing Center, Norway
Jameleddine Hassine, King Fahd University of Petroleum & Mineral (KFUPM), Saudi Arabia
Ourania Hatzis, Harokopio University of Athens, Greece
Yulan He, Aston University, UK
Kari Heikkinen, Lappeenranta University of Technology, Finland
Cory Henson, Wright State University / Kno.e.sis Center, USA
Arthur Herzog, Technische Universität Darmstadt, Germany
Rattikorn Hewett, Whitacre College of Engineering, Texas Tech University, USA
Celso Massaki Hirata, Instituto Tecnológico de Aeronáutica - São José dos Campos, Brazil
Jochen Hirth, University of Kaiserslautern, Germany

Bernhard Hollunder, Hochschule Furtwangen University, Germany
Thomas Holz, University College Dublin, Ireland
Władysław Homenda, Warsaw University of Technology, Poland
Carolina Howard Felicissimo, Schlumberger Brazil Research and Geoengineering Center, Brazil
Jingwei Huang, University of Illinois at Urbana-Champaign, USA
Weidong (Tony) Huang, CSIRO ICT Centre, Australia
Xiaodi Huang, Charles Sturt University - Albury, Australia
Eduardo Huedo, Universidad Complutense de Madrid, Spain
Marc-Philippe Huget, University of Savoie, France
Chi Hung, Tsinghua University, China
Chih-Cheng Hung, Southern Polytechnic State University - Marietta, USA
Edward Hung, Hong Kong Polytechnic University, Hong Kong
Muhammad Iftikhar, Universiti Malaysia Sabah (UMS), Malaysia
Prateek Jain, Ohio Center of Excellence in Knowledge-enabled Computing, Kno.e.sis, USA
Wassim Jaziri, Miracl Laboratory, ISIM Sfax, Tunisia
Hoyoung Jeung, SAP Research Brisbane, Australia
Yiming Ji, University of South Carolina Beaufort, USA
Jinlei Jiang, Department of Computer Science and Technology, Tsinghua University, China
Weirong Jiang, Juniper Networks Inc., USA
Hanmin Jung, Korea Institute of Science & Technology Information, Korea
Ilya S. Kabak, "Stankin" Moscow State Technological University, Russia
Eleanna Kafeza, Athens University of Economics and Business, Greece
Hermann Kaindl, Vienna University of Technology, Austria
Ahmed Kamel, Concordia College, Moorhead, Minnesota, USA
Faouzi Kamoun, University of Dubai, UAE
Rajkumar Kannan, Bishop Heber College(Autonomous), India
Teemu Kanstrén, VTT, Finland
Fazal Wahab Karam, Norwegian University of Science and Technology (NTNU), Norway
Dimitrios A. Karras, Chalkis Institute of Technology, Hellas
Koji Kashiwara, The University of Tokushima, Japan
Nittaya Kerdprasop, Suranaree University of Technology, Thailand
Katia Kermanidis, Ionian University, Greece
Serge Kernbach, University of Stuttgart, Germany
Nhien An Le Khac, University College Dublin, Ireland
Malik Jahan Khan, Lahore University of Management Sciences (LUMS), Lahore, Pakistan
Reinhard Klemm, Avaya Labs Research, USA
Ah-Lian Kor, Leeds Metropolitan University, UK
Arne Koschel, Applied University of Sciences and Arts, Hannover, Germany
George Kousiouris, NTUA, Greece
Philipp Kremer, German Aerospace Center (DLR), Germany
Dalia Kriksciuniene, Vilnius University, Lithuania
Dariusz Król, AGH University of Science and Technology, ACC Cyfronet AGH, Poland
Roland Kübert, Höchstleistungsrechenzentrum Stuttgart, Germany
Markus Kunde, German Aerospace Center, Germany
Dharmender Singh Kushwaha, Motilal Nehru National Institute of Technology, India
Andrew Kusiak, The University of Iowa, USA

Dimosthenis Kyriazis, National Technical University of Athens, Greece
Vitaveska Lanfranchi, Research Fellow, OAK Group, University of Sheffield, UK
Mikel Larrea, University of the Basque Country UPV/EHU, Spain
Angelos Lazaris, University of Southern California, USA
Philippe Le Parc, University of Brest, France
Gyu Myoung Lee, Institut Telecom, Telecom SudParis, France
Kyu-Chul Lee, Chungnam National University, South Korea
Tracey Kah Mein Lee, Singapore Polytechnic, Republic of Singapore
Daniel Lemire, LICEF Research Center, Canada
Haim Levkowitz, University of Massachusetts Lowell, USA
Kuan-Ching Li, Providence University, Taiwan
Tsai-Yen Li, National Chengchi University, Taiwan
Yangmin Li, University of Macau, Macao SAR
Jian Liang, Nimbus Centre, Cork Institute of Technology, Ireland
Haibin Liu, China Aerospace Science and Technology Corporation, China
Lu Liu, University of Derby, UK
Qing Liu, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Shih-Hsi "Alex" Liu, California State University - Fresno, USA
Xiaoqing (Frank) Liu, Missouri University of Science and Technology, USA
David Lizcano, Universidad a Distancia de Madrid, Spain
Henrique Lopes Cardoso, LIACC / Faculty of Engineering, University of Porto, Portugal
Sandra Lovrencic, University of Zagreb, Croatia
Jun Luo, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
Prabhat K. Mahanti, University of New Brunswick, Canada
Jacek Mandziuk, Warsaw University of Technology, Poland
Herwig Mannaert, University of Antwerp, Belgium
Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece
Antonio Maria Rinaldi, Università di Napoli Federico II, Italy
Ali Masoudi-Nejad, University of Tehran, Iran
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Gerrit Meixner, German Research Center for Artificial Intelligence (DFKI) / Innovative Factory Systems (IFS) / Center for Human-Machine-Interaction (ZMMI), Germany
Zulfiqar Ali Memon, Sukkur Institute of Business Administration, Pakistan
Andreas Merentitis, AGT Group (R&D) GmbH, Germany
Jose Merseguer, Universidad de Zaragoza, Spain
Frederic Migeon, IRIT/Toulouse University, France
Harald Milchrahm, Technical University Graz, Institute for Software Technology, Austria
Fatma Mili, Oakland University, USA
Les Miller, Iowa State University, USA
Marius Minea, University POLITEHNICA of Bucharest, Romania
Yasser F. O. Mohammad, Assiut University, Egypt
Shahab Mokarizadeh, Royal Institute of Technology (KTH) - Stockholm, Sweden
Martin Molhanec, Czech Technical University in Prague, Czech Republic
Dorothy Monekosso, University of Ulster at Jordanstown, UK
Charalampos Moschopoulos, KU Leuven, Belgium
Mary Luz Mouronte López, Ericsson S.A., Spain

Henning Müller, University of Applied Sciences Western Switzerland - Sierre (HES SO), Switzerland
Susana Munoz Hernández, Universidad Politécnica de Madrid, Spain
Adrian Muscat, University of Malta, Malta
Peter Mutschke, GESIS - Leibniz Institute for the Social Sciences - Bonn, Germany
Bela Mutschler, Hochschule Ravensburg-Weingarten, Germany
Deok Hee Nam, Wilberforce University, USA
Fazel Naghdy, University of Wollongong, Australia
Joan Navarro, Research Group in Distributed Systems (La Salle - Ramon Llull University), Spain
Saša Nešić, University of Lugano, Switzerland
Rui Neves Madeira, Instituto Politécnico de Setúbal / Universidade Nova de Lisboa, Portugal
Toàn Nguyễn, INRIA Grenoble Rhone-Alpes/ Montbonnot, France
Andrzej Niesler, Institute of Business Informatics, Wrocław University of Economics, Poland
Michael P. Oakes, University of Sunderland, UK
John O'Donovan, University of California - Santa Barbara, USA
Kouzou Ohara, Aoyama Gakuin University, Japan
Jonice Oliveira, Universidade Federal do Rio de Janeiro, Brazil
Ian Oliver, Nokia Location & Commerce, Finland / University of Brighton, UK
Michael Adeyeye Oluwasegun, University of Cape Town, South Africa
Sigeru Omatu, Osaka Institute of Technology, Japan
Sascha Opletal, University of Stuttgart, Germany
Flavio Oquendo, European University of Brittany/IRISA-UBS, France
Fakri Othman, Cardiff Metropolitan University, UK
Enn Õunapuu, Tallinn University of Technology, Estonia
Jeffrey Junfeng Pan, Facebook Inc., USA
Hervé Panetto, University of Lorraine, France
Malgorzata Pankowska, University of Economics, Poland
Harris Papadopoulos, Frederick University, Cyprus
Laura Papaleo, ICT Department - Province of Genoa & University of Genoa, Italy
Agis Papantoniou, National Technical University of Athens, Greece
Thanasis G. Papaioannou, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Andreas Papasalouros, University of the Aegean, Greece
Eric Paquet, National Research Council / University of Ottawa, Canada
Kunal Patel, Ingenuity Systems, USA
Carlos Pedrinaci, Knowledge Media Institute, The Open University, UK
Juan C Pelaez, Defense Information Systems Agency, USA
Yoseba Peña, University of Deusto - DeustoTech (Basque Country), Spain
Cathryn Peoples, University of Ulster, UK
Asier Perallos, University of Deusto, Spain
Christian Percebois, Université Paul Sabatier - IRIT, France
Andrea Perego, European Commission, Joint Research Centre, Italy
Mark Perry, University of Western Ontario/Faculty of Law/ Faculty of Science - London, Canada
Willy Picard, Poznań University of Economics, Poland
Meikel Poess, Oracle, USA
Agostino Poggi, Università degli Studi di Parma, Italy
R. Ponnusamy, Madha Engineering College-Anna University, India
Dorin Popescu, University of Craiova, Romania

Stefan Poslad, Queen Mary University of London, UK
Wendy Powley, Queen's University, Canada
Radu-Emil Precup, "Politehnica" University of Timisoara, Romania
Jerzy Prekurat, Canadian Bank Note Co. Ltd., Canada
Didier Puzenat, Université des Antilles et de la Guyane, France
Sita Ramakrishnan, Monash University, Australia
Elmano Ramalho Cavalcanti, Federal University of Campina Grande, Brazil
Juwel Rana, Luleå University of Technology, Sweden
Martin Randles, School of Computing and Mathematical Sciences, Liverpool John Moores University, UK
Christoph Rasche, University of Paderborn, Germany
Ann Reddipogu, ManyWorlds UK Ltd, UK
Ramana Reddy, West Virginia University, USA
René Reiners, Fraunhofer FIT - Sankt Augustin, Germany
Paolo Remagnino, Kingston University - Surrey, UK
Sebastian Rieger, University of Applied Sciences Fulda, Germany
Andreas Riener, Johannes Kepler University Linz, Austria
Ivan Rodero, NSF Center for Autonomic Computing, Rutgers University - Piscataway, USA
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Alejandro Rodríguez González, University Carlos III of Madrid, Spain
Aitor Rodríguez-Alsina, University Autònoma of Barcelona (UAB), Spain
Paolo Romano, INESC-ID Lisbon, Portugal
Vicente-Arturo Romero-Zaldivar, Atos Origin SAE, Spain
Agostinho Rosa, Instituto de Sistemas e Robótica, Portugal
José Rouillard, University of Lille, France
Paweł Różycki, University of Information Technology and Management (UITM) in Rzeszów, Poland
Igor Ruiz-Agundez, DeustoTech, University of Deusto, Spain
Michele Ruta, Politecnico di Bari, Italy
Melike Sah, Trinity College Dublin, Ireland
Francesc Saigi Rubió, Universitat Oberta de Catalunya, Spain
Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Yacine Sam, Université François-Rabelais Tours, France
Ismael Sanz, Universitat Jaume I, Spain
Ricardo Sanz, Universidad Politécnica de Madrid, Spain
Marcello Sarini, Università degli Studi Milano-Bicocca - Milano, Italy
Munehiko Sasajima, I.S.I.R., Osaka University, Japan
Minoru Sasaki, Ibaraki University, Japan
Hiroyuki Sato, University of Tokyo, Japan
Jürgen Sauer, Universität Oldenburg, Germany
Patrick Sayd, CEA List, France
Dominique Scapin, INRIA - Le Chesnay, France
Kenneth Scerri, University of Malta, Malta
Adriana Schiopoiu Burlea, University of Craiova, Romania
Rainer Schmidt, Austrian Institute of Technology, Austria
Bruno Schulze, National Laboratory for Scientific Computing - LNCC, Brazil
Wieland Schwinger, Johannes Kepler University Linz, Austria
Hans-Werner Sehring, T-Systems Multimedia Solutions GmbH, Germany

Paulo Jorge Sequeira Gonçalves, Polytechnic Institute of Castelo Branco, Portugal
Sandra Sendra Compte, Polytechnic University of Valencia, Spain
Kewei Sha, Oklahoma City University, USA
Hossein Sharif, University of Portsmouth, UK
Roman Y. Shtykh, Rakuten, Inc., Japan
Kwang Mong Sim, Gwangju Institute of Science & Technology, South Korea
Robin JS Sloan, University of Abertay Dundee, UK
Vasco N. G. J. Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
Don Sofge, Naval Research Laboratory, USA
Christoph Sondermann-Woelke, Universitaet Paderborn, Germany
George Spanoudakis, City University London, UK
Vladimir Stantchev, SRH University Berlin, Germany
Claudius Stern, University of Paderborn, Germany
Mari Carmen Suárez-Figueroa, Universidad Politécnica de Madrid (UPM), Spain
Kåre Synnes, Luleå University of Technology, Sweden
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Yehia Taher, ERISS - Tilburg University, The Netherlands
Yutaka Takahashi, Senshu University, Japan
Azzelarabe Taleb-Bendiab, Liverpool John Moores University, UK
Dan Tamir, Texas State University, USA
Jinhui Tang, Nanjing University of Science and Technology, P.R. China
Yi Tang, Chinese Academy of Sciences, China
Said Tazi, LAAS-CNRS, Université Toulouse 1, France
John Terzakis, Intel, USA
Sotirios Terzis, University of Strathclyde, UK
Vagan Terziyan, University of Jyväskylä, Finland
Ioan Toma, STI Innsbruck/University Innsbruck, Austria
Lucio Tommaso De Paolis, Department of Innovation Engineering - University of Salento, Italy
Davide Tosi, Università degli Studi dell'Insubria, Italy
Raquel Trillo Lado, University of Zaragoza, Spain
Tuan Anh Trinh, Budapest University of Technology and Economics, Hungary
Simon Tsang, Applied Communication Sciences, USA
Theodore Tsiligiridis, Agricultural University of Athens, Greece
Antonios Tsourdos, Cranfield University, UK
José Valente de Oliveira, University of Algarve, Portugal
Cristián Felipe Varas Schuda, NIC Chile Research Labs, Chile
Eugen Volk, University of Stuttgart, Germany
Mihaela Vranić, University of Zagreb, Croatia
Chieh-Yih Wan, Intel Labs, Intel Corporation, USA
Jue Wang, Washington University in St. Louis, USA
Shenghui Wang, OCLC Leiden, The Netherlands
Zhonglei Wang, Karlsruhe Institute of Technology (KIT), Germany
Laurent Wendling, University Descartes (Paris 5), France
Maarten Weyn, University of Antwerp, Belgium
Nancy Wiegand, University of Wisconsin-Madison, USA

Alexander Wijesinha, Towson University, USA
Eric B. Wolf, US Geological Survey, Center for Excellence in GIScience, USA
Ouri Wolfson, University of Illinois at Chicago, USA
Yingcai Xiao, The University of Akron, USA
Reuven Yagel, The Jerusalem College of Engineering, Israel
Fan Yang, Nuance Communications, Inc., USA
Maribel Yasmina Santos, University of Minho, Portugal
Zhenzhen Ye, Systems & Technology Group, IBM, US A
Jong P. Yoon, MATH/CIS Dept, Mercy College, USA
Shigang Yue, School of Computer Science, University of Lincoln, UK
Constantin-Bala Zamfirescu, "Lucian Blaga" Univ. of Sibiu, Romania
Claudia Zapata, Pontificia Universidad Católica del Perú, Peru
Marek Zaremba, University of Quebec, Canada
Filip Zavoral, Charles University Prague, Czech Republic
Yuting Zhao, University of Aberdeen, UK
Hai-Tao Zheng, Graduate School at Shenzhen, Tsinghua University, China
Yu Zheng, Microsoft Research Asia, China
Zibin (Ben) Zheng, Shenzhen Research Institute, The Chinese University of Hong Kong, Hong Kong
Bin Zhou, University of Maryland, Baltimore County, USA
Alfred Zimmermann, Reutlingen University - Faculty of Informatics, Germany
Wolf Zimmermann, Martin-Luther-University Halle-Wittenberg, Germany

CONTENTS

pages: 165 - 176

Exact Logic Minimization and Multiplicative Complexity of Concrete Algebraic and Cryptographic Circuits

Nicolas Courtois, UCL, United Kingdom
Theodosios Mourouzis, UCL, United Kingdom
Daniel Hulme, UCL, United Kingdom

pages: 177 - 187

KLocator: An Ontology-Based Framework for Scenario-Driven Geographical Scope Resolution

Panos Alexopoulos, iSOCO S.A., Spain
Carlos Ruiz, iSOCO S.A., Spain
Boris Villazon-Terrazas, iSOCO S.A., Spain
José-Manuél Gómez-Pérez, iSOCO S.A., Spain

pages: 188 - 198

Three Principles for the Design of Energy Feedback Visualizations

Robert S. Brewer, University of Hawaii at Manoa, USA
Yongwen Xu, University of Hawaii at Manoa, USA
George E. Lee, University of Hawaii at Manoa, USA
Michelle Katchuck, University of Hawaii at Manoa, USA
Carleton A. Moore, University of Hawaii at Manoa, USA
Philip M. Johnson, University of Hawaii at Manoa, USA

pages: 199 - 212

Regular Polysemy in WordNet and Pattern based Approach

Abed Alhakim Freihat, Dept. of Information Engineering and Computer Science University of Trento,, Italy
Fausto Giunchiglia, Dept. of Information Engineering and Computer Science University of Trento,, Italy
Biswanath Dutta, Documentation Research and Training Centre Indian Statistical Institute (ISI), India

pages: 213 - 222

A Human Surface Prediction Model Based on Linear Anthropometry

Ameersing Luximon, The Hong Kong Polytechnic University, Hong Kong
Yan Luximon, The Hong Kong Polytechnic University, Hong Kong
Huang Chao, The Hong Kong Polytechnic University, Hong Kong

pages: 223 - 234

Kinematic Description of Bimanual Performance in Unpredictable Virtual Environments: A Lifespan Study

Andrea H Mason, University of Wisconsin - Madison, USA
Patrick J Grabowski, University of Wisconsin - La Crosse, USA
Drew N Rutherford, University of Wisconsin - Madison, USA
Andrew R Minkley, University of Wisconsin - Madison, USA

pages: 235 - 255

High End Computing Using Advanced Archaeology and Geoscience Objects

Claus-Peter Rückemann, Westfälische Wilhelms-Universität Münster (WWU) and Leibniz Universität Hannover and North-German Supercomputing Alliance (HLRN), Germany

pages: 256 - 265

Using Semantic Web Technologies to Follow the Evolution of Entities in Time and Space

Benjamin Harbelot, Laboratoire Le2i, UMR-6302 CNRS, Département Informatique, University of Burgundy, France
Helbert Arenas, Laboratoire Le2i, UMR-6302 CNRS, Département Informatique, University of Burgundy, France
Christophe Cruz, Laboratoire Le2i, UMR-6302 CNRS, Département Informatique, University of Burgundy, France

pages: 266 - 278

Designing for 3D User Experience in Tablet Context Design and Early Phase User Evaluation of Four 3D GUIs

Minna Pakanen, Department of Information Processing Science, University of Oulu, Finland
Leena Arhippainen, Center for Internet Excellence, Finland
Seamus Hickey, Department of Information Processing Science, University of Oulu, Finland

pages: 279 - 288

Constructing Autonomous Systems: Major Development Phases

Nikola Serbedzija, Fraunhofer FOKUS, Germany
Annabelle Klarl, LMU Munich, Germany
Philip Mayer, LMU Munich, Germany

pages: 289 - 299

Interactive Rigid-Body Dynamics and Deformable Surface Simulations with Co-Located Maglev Haptic and 3D Graphic Display

Peter Berkelman, University of Hawaii at Manoa, USA
Sebastian Bozlee, University of Portland, USA
Muneaki Miyasaka, University of Washington, USA

pages: 300 - 317

Autonomous Load Balancing of Data Stream Processing and Mobile Communications in Scalable Data Distribution Systems

Rafael Oliveira Vasconcelos, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil
Markus Endler, Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil
Berto Gomes, Federal University of Maranhão (UFMA), Brazil
Francisco Silva, Federal University of Maranhão (UFMA), Brazil

pages: 318 - 328

Comparison of Simultaneous Measurement of Lens Accommodation and Convergence in Stereoscopic Target with Sine Curve Movement

Takehito Kojima, Graduate School of Information Science Nagoya University, Japan
Tomoki Shiomi, Tokushima Labour Standards Inspection Office, Japan
Kazuki Yoshikawa, Graduate School of Information Science Nagoya University, Japan
Masaru Miyao, Graduate School of Information Science Nagoya University, Japan

pages: 329 - 341

Automobile Driving Interface Using Gesture Operations for Disabled People

Yoshitoshi Murata, Iwate Prefectural University, Japan
Kazuhiro Yoshida, Iwate Prefectural University, Japan

pages: 342 - 355

Business Process Modelling for Measuring Quality

Farideh Heidari, Delft University of Technology, The Netherlands
Pericles Loucopoulos, University of Manchester, The United Kingdom
Frances Brazier, Delft University of Technology, The Netherlands

pages: 356 - 366

Machine Learning Methods Applied on Long Term Data Analysis for Rain Detection in a Partial Discharge Sensor Network

Leandro H. S. Silva, Polytechnic School of Pernambuco - University of Pernambuco, Brazil
Sérgio C. Oliveira, Polytechnic School of Pernambuco - University of Pernambuco, Brazil
Eduardo Fontana, Department of Electronics and Systems - Federal University of Pernambuco, Brazil

pages: 367 - 375

Implementation of a Map-Reduce based Context-Aware Recommendation Engine for Social Music Events

Wolfgang Beer, Software Competence Center Hagenberg GmbH, Austria
Sandor Herramhof, Evtogram Labs GmbH, Austria
Christian Derwein, Evtogram Labs GmbH, Austria

pages: 376 - 393

An Error Detection Strategy for Improving Web Accessibility for Older Adults

Alfred Taylor Sr., Iowa State University, USA
Les Miller, Iowa State University, USA
Sree Nilakanta, Iowa State University, USA
Jeffrey Sander, Iowa State University, USA
Saayan Mitra, Iowa State University, USA
Anurag Sharda, Iowa State University, USA
Bachar Chama, Iowa State University, USA

Exact Logic Minimization and Multiplicative Complexity of Concrete Algebraic and Cryptographic Circuits

Nicolas T. Courtois
University College London,
Gower Street, London, UK
n.courtois@cs.ucl.ac.uk

Theodosios Mourouzis
University College London,
Gower Street, London, UK
theodosios.mourouzis.09@ucl.ac.uk

Daniel Hulme
University College London,
Gower Street, London, UK
d.hulme@cs.ucl.ac.uk

Abstract—Two very important NP-hard problems in the area of computational complexity are the problems of Matrix Multiplication (MM) and Circuit Optimization. Solving particular cases of such problems yield to improvements in many other problems as they are core sub-routines implemented in many other algorithms. However, obtaining optimal solutions is an intractable problem since the space to explore for each problem is exponentially large. All suggested methodologies rely on well-chosen heuristics, selected according to the topology of the specific problem. Such heuristics may yield to efficient and acceptable solutions but they do not guarantee that no better can be done. In this paper, we suggest a general framework for obtaining solutions to such problems. We have developed a 2-step methodology, where in the first place we describe algebraically the problem and then we convert it to a SAT-CNF problem, which we solve using SAT-solvers. By running the same procedure for different values of k we can obtain optimal solutions and prove that no better can be done. We decrease the k until "UNSAT" is obtained. Using the suitable encoding step for each problem we have been able to obtain exact and optimal solutions for different problems which are sufficiently small, allowing us to solve them on an average PC. We have been able to prove the exact number of multiplications needed for multiplying two non-square matrices of sufficiently small dimensions, as well as obtaining optimal representations with respect to meaningful metrics for several S-boxes used in prominent ultra-lightweight ciphers such as GOST, PRESENT and CTC2.

Index Terms—Linear Algebra, Fast Matrix Multiplication, Complex Numbers, quaternions, Strassen's algorithm, Multiplicative Complexity, Asynchronous Circuits, Logic Minimization, Automated Theorem Provers, Block Ciphers, CTC2, PRESENT, GOST, SAT solvers

I. INTRODUCTION

Optimization of arbitrary algebraic computations over rings in the general non-commutative setting is considered as one of the most interesting topics in theoretical computer science and mathematics. In general such optimization problems are expected to be computationally very hard [1], [2].

In this paper, we study two fundamental problems. We study the problem of minimizing the Multiplicative Complexity (MC) of algebraic computations, such as the Matrix Multiplication (MM) [1]. MC is the minimum number of AND gates that are needed, if we allow an unlimited number of *NOT* and *XOR* gates. Informally, we are interested in reducing the

number of multiplications involved in an arbitrary algebraic computation to the lowest possible bound, allowing unlimited number of additions. Initially, we study the optimization problem over small fields such as $GF(2)$. However, in some cases solutions found do not yet yield a general solution for a larger ring, and there can be additional lifting steps [1].

The second problem we study is the combinatorial logic optimization of general Boolean circuits, with respect to a given set of elementary operations. Logic optimization is also a well-known hard problem which interests the chip maker industry and researchers in complexity. Good optimizations are particularly important in industrial hardware implementations of standard cryptographic algorithms [3], [4]. This is because cryptography is computationally very costly and the improved designs can be used in hundreds of millions of integrated circuits and produce important savings. These ciphers have very small S-boxes, yet, nobody knows how to implement them in an optimal way, and new cryptographic implementations with less gates are obtained almost each year [5], [6].

In practice, there are no known analytic techniques nor direct prescriptive algorithms, which can construct such optimal circuits. Developing an optimal circuit representation for a small-size Boolean function of the form $GF(2)^8 \rightarrow GF(2)$ with respect to AND gates remains still an open problem. Is it possible to determine once for all, what is the minimum possible number of gates? Exact bounds are very hard to be obtained in these areas, as the problem is mainly solved by heuristic techniques and is known to be computationally very hard.

In this paper, we view these problems as constraint satisfaction problems which we attempt to solve them by methods of formal coding [7] and later solve with software such as SAT solvers [8]. The striking feature of this type of methods is that, if we use a "complete" SAT solver (and we have enough CPUs), as opposed to a "stochastic" one, and if it is fast enough to complete, and it outputs UNSAT, we obtain a proven lower bound, a very rare thing in complexity.

Our method consists of three basic steps. In the first step, we formally encode the problem by writing a system of equations which describes our problem as a system of polynomial

equations over the finite field of two elements GF(2). In the case of the MM problem and some other of our algebraic optimizations [1], we use the Brent Equations [9] in the encoding step. Circuit minimization problems are encoded formally as a form of straight-line representation problem, which we encode them as a quantified set of multivariate relations that need to be satisfied. Then, we proceed by converting our defined modulo 2 problem to a SAT problem using the Courtois-Bard-Jefferson method [7] and then (only if required) we may add additional steps such as lifting the solution to larger fields or rings [1] or re-optimize for circuit depth, or many other [5], [6].

This type of methodology was recently applied with success to optimize linear circuits [10] and bi-linear circuits [11]. We have developed a method to do this also for non-linear circuits. We have been greatly influenced by the work of Boyar and Peralta on the AES cipher S-box and its MC, however, we can also produce many optimizations from scratch with arbitrary gates and without the Boyar-Peralta heuristic. Though this type of exact optimizations is computationally very intensive and therefore currently only possible for fairly small circuits, the preliminary results obtained are very encouraging and allow for direct applications in cryptographic hardware synthesis, systematic synthesis of implementations resistant to side-channel attacks, and also in cryptanalysis [12], [3], [13]. In this paper, we also report our results on PRESENT and GOST, two block ciphers known for their exceptionally low hardware cost.

A. Structure of the Paper

The organization of this paper is as follows:

Section II: We refer to several reasons highlighting the importance of solving such problems. We outline several improvements which yield in many other applications, as a consequence of improving the state-of-art algorithms for solving MM and combinatorial circuit optimization problems. Obtaining even solutions to small problems may yield significant improvements to general problems as these solutions to smaller instances can be recursively used to handle the general problem.

Section III: We describe all technical details of our 2-step methodology. Initially, we describe the encoding step which we employ for each problem. For example in case of MM as a tool of encoding we use the Brent Equations, while for optimizing circuits we invented a general framework of encoding. Then we briefly analyze how to obtain the corresponding CNF-SAT problem of a given problem, which is algebraically encoded. Additionally, we describe provably aspects of our methodology and we highlight how powerful are the SAT solvers for solving exactly such NP-hard problems.

Section IV: We apply our methodology for obtaining new formulae for multiplying two non-necessarily square matrices. We have been able to solve exactly with respect to the number of multiplications needed to multiply such matrices, for sufficiently small matrices. Several small instances are solved and presented.

Section V: We optimize arbitrary non-linear digital circuits for silicon implementation and cryptanalysis. We apply our methodology for obtaining optimal circuit representations for the S-boxes used in many prominent ultra-lightweight block ciphers such as PRESENT and GOST. For experimentation we have been able to optimize the 3-bit to 3-bit S-box of CTC cipher with respect to different meaningful metrics.

II. MOTIVATION FOR LOW MC AND LOW GATE COUNT OPTIMIZATIONS

We briefly outline what will be the benefits in both academic and industrial world if some better optimizations are found for the problems of MM and gate-efficient implementation.

A. Matrix Multiplication Problem

Obtaining the minimum number of 2-input multiplications needed for computing the product of two matrices A, B , is considered among the most difficult optimization problems in the area of computer science and mathematics. Given two matrices A, B the MM is defined as follows (Def. 1).

Definition 1: (Matrix Multiplication [MM])

Let A and B two $n \times n$ matrices, $n \in \mathbb{N}$, with entries in a ring \mathcal{R} (not necessarily commutative), such that

$$A_{m,n} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \text{ and } B_{m,n} = \begin{pmatrix} b_{1,1} & b_{1,2} & \cdots & b_{1,n} \\ b_{2,1} & b_{2,2} & \cdots & b_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \cdots & b_{n,n} \end{pmatrix}.$$

Then, the entries of the product matrix $C = AB$ are given by

$$C_{p,q} = [AB]_{p,q} = \sum_{i=1}^n a_{p,i} b_{i,q}. \quad (1)$$

The multiplication is defined as in the ring \mathcal{R} .

We are interested in obtaining new formulae for computing the product matrix C involving as few number of 2-input multiplications as possible. Fast linear algebra for large matrices leads to significant improvements in many other areas as follows [14]:

- Commercial software such as MATLAB, MATHEMATICA and GAUSS
- Economic Modeling
- Weather prediction
- Signal processing
- Gauss Elimination algorithm for solving a system of linear polynomial equations
- Algorithms for solving non-linear polynomial equations
- Recognizing if a word of length n belongs to a context-free language
- Transitive closure of a graph or a relation on a finite set
- Statistical analysis of large data sets

- Integer factorization
- Cryptanalysis

B. Circuit Complexity

There are many reasons why circuits of low MC are very important especially for industrial applications and for cryptography. More analytic explanations can be found in [3].

- Develop certain so called Bitslice parallel-SIMD software implementations of block ciphers such as in [15]
- Lower the hardware implementation cost of encryption algorithms in silicon chips
- Prevent Side Channel Attacks (SCA) on smart cards such as Differential Power Analysis (DPA) [16]

C. Cryptanalysis Applications:

In addition, more or less all optimizations in this paper have direct applications in software cryptanalysis, cf. [12], [3], [4], [13]. One reason for that is the fact that dense linear algebra is very frequently the last step in many algebraic attacks. Another family of applications are algebraic attacks on symmetric cryptography. A more compact representation of a cipher is known to improve the running time of many such attacks [12], [13].

III. A SAT-SOLVER BASED METHODOLOGY

A. General Methodological Framework

In this paper, we formally encode optimization problems as systems of multivariate equations over $GF(2)$, then we convert them to SAT problems, and then we solve them.

This sort of encoding is subject to continuous improvement, and it can be viewed itself as a hard but solvable optimization problem. Therefore, our current timings (current methods are quite slow) and results are very likely to be improved in the future in very substantial ways.

We describe all the main steps in our approach. Our main scientific contributions are the methodology and concrete results that are mathematical theorems about lower complexity bounds and concrete optimizations, which are reusable building blocks for algorithms, scientific computing and the industry. For those results which we show to be optimal they can no longer be improved, however, the timing of obtaining these results (computations that need to be done only once) can still certainly be improved.

In what follows, we present two major encoding methodologies. The Coding Methodology 1 is designed to address the efficient MM problem and other algebraic optimizations, We use the Brent Equations [9] in the encoding step.

The Coding Methodology 2 is designed to address more general problems of MC and other optimizations of arbitrary circuits which are no longer initially described by multivariate polynomial expressions, but by a truth table. Circuit minimization problems are encoded formally as a form or straight-line representation problem, then we describe it as a quantified set of multivariate relations [3], and then it has to hold simultaneously for different input values in the truth table.

B. Coding Methodology 1

Our algorithm for solving for MM problems is as follows:

- 1) Form the Brent Equations (or write a quantified set of multivariate relations that describes the problem)
- 2) Consider only solutions in $0,1$ =integers modulo 2
- 3) Convert to SAT with Courtois-Bard-Jefferson method [7]
- 4) Lift the solution from $GF(2)$ to the general bigger fields by another constraint satisfaction algorithm

C. Brent Equations

Brent Equations are used for encoding problems in a more formal algebraic language. After the encoding we convert this problem to a SAT problem and then we try to obtain a solution using several SAT solvers.

Brent's Method: Suppose we want to multiply a $M \times N$ matrix A by a $N \times P$ matrix B using T 2-input multiplications. We solve the above problem by solving the following system of $(MNP)^2$ equations in $T(MN + NP + MP)$ unknowns, cf. [9]:

$$\{\forall i \forall j \forall k \forall L \forall m \forall n, \sum_{p=1}^T \alpha_{ijp} \beta_{kLp} \gamma_{mnp} = \delta_{ni} \delta_{jk} \delta_{Lm}\}$$

A solution to this set of equations implies that the coefficient entries c_{ij} of the product matrix $C = AB$ can be written as

$$c_{nm} = \sum_{p=1}^T \gamma_{mnp} q_p,$$

where the products q_1, q_2, \dots, q_T are given by the expression $q_p = (\sum \alpha_{ijp} a_{ij})(\sum \beta_{kLp} b_{kL})$.

This form of encoding can be generalized for describing other problems such as complex number multiplication and quaternion multiplication.

D. Coding Methodology 2

This methodology is very different, and in fact it is possible to see that for many circuits both methodologies could be applied. The motivation for this second method is that not every circuit is very algebraic and it can be described efficiently by sparse multivariate polynomial expressions. Especially, in industrial cryptographic primitives we expect that the resulting circuits will have a very low gate count since the efficient hardware implementation is one of the main priorities of designers. Thus, it is very realistic that a system of equation describing such a cryptographic primitive will be very sparse.

Therefore, the Brent-like approach could lead to a very large problem to solve. However, we can also describe the initial problem as a substitution box with a truth table. We proceed as follows.

First, we write a certain system of equations $\mathcal{C}1$ in which variables will be divided in several disjoint categories:

- 1) We will have the " x " variables which will be inputs of the truth table.
- 2) The " y " variables which will be outputs of the truth table.
- 3) The " t " variables which will be inputs of gates.
- 4) The " q " variables which are outputs of gates.
- 5) The " b " variables will define the function of each gate. (For example, one gate could be AND, OR, XOR and

the model is $b(uv) + b'(u + v)$, and when $b = 1, b' = 0$ this will be AND gate.)

- 6) The "a" variables which will be the unknown connections between different gates.

A crucial problem is how these connections are described in $C1$. The purpose of the "a" variables is to make each new "t" variable depended on some combinations of past variables of type "x", "t", "q". These "a" variables will encode ALL the unknown connections between different gates, their inputs, their outputs, our inputs "x" and our outputs "y".

For example, in MC optimizations we can say that each variable is an affine or linear combination of previous variables. In other optimizations we can furthermore add constraints of type $a_i a_j = 0$ which say that in a certain set of a_i only at most one of these variables is at 1. We provide a toy example of our algebraic description $C1$ below:

```
t1 = a1 * x1 + a2 * x2
a1 * a2 = 0
t2 = a3 * x1 + a4 * x2
a3 * a4 = 0
q1 = t1 * t2
y1 = a5 * q1 + a6 * x1 + a7 * x2
y1 = x1 * x2
```

Another problem is how to describe the relation between the inputs "x" and the outputs "y" efficiently. In the toy example above this is done in the last line. Several methods for coding small size I/O relations are described in [12]. It is the open problem to see what is the best method and the best method will depend a lot on the type of the circuit. One very good method is to use a previous circuit with more gates (!).

Now our circuit minimization problem is encoded formally as a straight-line computation problem. Then we expand this system of equations $C1$ into another system of equations $C2$ as follows:

- 1) Our problem $C1$ is a quantified system of constraints. We need to determine the variables of types "a" and "b".
- 2) Our circuit $C1$ (see toy example above) must be true for every "x". We can privilege values of small Hamming weight (for some circuits we do NOT need to put all possible values of "x").
- 3) We make several copies of the circuits and we rename the "x" and "y" and "q" and "t" variables, however, the "a" and "b" variables remain common in all circuits.
- 4) We write all these circuits as systems of multivariate relations [3] and concatenate them. We call $C2$ the resulting system of equations.
- 5) We convert $C2$ to SAT and solve for the "a" and "b" variables.
- 6) We take the values of the "a" and "b" variables, we ignore all the other assignments, and substitute in the original (single) circuit model $C1$.
- 7) We check if our solution is correct, and potentially optimize for XORs, to decrease the number of intermediate

variables, etc.

E. SAT Solver Step

Satisfiability (SAT) is the problem of determining if the variables of a given Boolean formula can be assigned in a way as to make the formula evaluate to TRUE [17]. SAT was the first known example of an NP-complete problem. A wide range of other decision and optimization problems can be transformed into instances of SAT and a class of algorithms called SAT solvers can efficiently solve a large enough subset of SAT instances such as MiniSAT solver [8]. Our aim is to transform problems like MM into SAT problems.

SAT solvers had both theoretical and practical improvements and have made a lot of progress in recent years. The basis of most SAT solvers is the Davis-Putnam backtrack search, which searches for a solution by recursively choosing a variable and trying to assign to it one value and then the other. At each stage of search a propagation step is performed which attempts to imply the assignments to as many unassigned variables as possible based on previous assignments. As a result of this it may uncover a clause which cannot be satisfied, so search backtracks.

In the major SAT competition every year, almost all the previous years winners are beaten by new competitors who design more efficient solvers. Thus SAT solvers are carefully designed to run on a large range of problems with no tuning required by users.

Problems arising either from academia or from industry can be solved by SAT solvers if are converted to Conjunctive Normal Form (CNF). In Boolean logic, a formula is in CNF if it is a conjunction of clauses, where a clause is a disjunction of literals.

At a first glance, this seems to be inefficient as conversion to CNF can skip extra structural information of the original problem. However, the performance of SAT solvers is often able to offset this structural information loss.

F. On the Complexity of SAT-solvers

Unfortunately, the time complexity of a SAT solver is not easy to determine. A very large system in CNF can be easily solved by a SAT solver on an average PC, but beyond some point the probability of solving such a system from 1 becomes 0.

In cryptanalysis, that implies we can derive the key of a reduced version. As the number of rounds grows, the time complexity of such an algebraic attack becomes infinite. Unfortunately, there is no clear indication when the problem becomes infeasible.

G. Conversion to SAT

In order to solve a problem using a SAT solver we need to convert this problem to its CNF (cf. Def. 2).

Definition 2: (Conjunctive Normal Form)

A Boolean function f is said to be in conjunctive normal form, if it is a conjunction of clauses, where each clause is a disjunction of literals, i.e., f can be expressed in the form

$$\wedge_{I \subseteq M} (\vee_{i \in I} x_i), M = \{1, \dots, n\}$$

We have been using three major methods to convert a system of multivariate polynomial equations over $GF(2)$ to a SAT problem. This idea has been pioneered by Bard and Courtois see [12] and has become a very important tool in modern cryptanalysis and automated problem solving.

As these methods are quite slow, it is too early to say which one is better for the purpose of our optimizations.

- 1) We can use the Courtois-Bard-Jefferson tool [7] which is available for download.
- 2) Another method is local approximation, it has been frequently used in cryptanalysis, see [12], [13].
- 3) Yet another method is to use a SAT solver which accepts native XORs, such as CryptoMiniSat by Soos [18], and therefore new conversion methods can be proposed, see [13] for some applications of these very promising new encodings which seem to be really excellent in cryptanalysis applications however, they have not yet been tested in the setting of this paper.

A very basic approach to map a given problem to SAT-CNF is firstly to derive a 2-degree system of equations from the algebraic description of the problem using the fact that [7]:

$$\{m = wxyz\} \Leftrightarrow \{a = wx, b = yz, m = ab\}$$

In general, CNF expressions describe instances of SAT problems, thus we need to obtain the CNF of this multivariate system of quadratic equations. This conversion proceeds by three major steps as follows [7]:

STEP 0: The CNF form must not contain any constants. Since all clauses must be true in a solution, we introduce constants by adding clauses of the form $T \vee T \vee \dots \vee T$, which implies that variable T is true in any satisfying solution. T encodes constant 1, while \bar{T} encodes 0.

STEP 1: (Polynomial System to Linear System)

Every polynomial is a sum of linear and higher degree terms. Given a monomial $a = wxyz$ over \mathbb{F}_2 , then this is tautological equivalent to

$$a \Leftrightarrow (w \wedge x \wedge y \wedge z)$$

$$(w \vee \bar{a})(x \vee \bar{a})(y \vee \bar{a})(z \vee \bar{a})(a \vee \bar{w} \vee \bar{x} \vee \bar{y} \vee \bar{z}).$$

Thus, for each monomial of degree d we have $d+1$ clauses, while the total length of clauses is $3d+1$.

STEP 2: (Linear System to CNF expression)

After expressing each monomial involved the next step is to express the logical XORs. The sum $a \oplus b \oplus c \oplus d \oplus 0$ is equivalent to:

$$(\bar{a} \vee b \vee c \vee d)(a \vee \bar{b} \vee c \vee d)(a \vee b \vee \bar{c} \vee d)(a \vee b \vee c \vee \bar{d})$$

$$(\bar{a} \vee \bar{b} \vee \bar{c} \vee d)(\bar{a} \vee \bar{b} \vee c \vee \bar{d})(\bar{a} \vee b \vee \bar{c} \vee \bar{d})(a \vee \bar{b} \vee \bar{c} \vee \bar{d})$$

However, handling long XORs is a hard problems for SAT solvers. For example, given a sum of length h we split it into different sub-sums and encode each sum separately. More details can be found in [7] since the scope of this paper to contribute towards the encoding step and both conversion and

solving techniques can be considered as black-box procedures. Note that the conversion procedure in this section is polynomial in time and more details are found in [7].

H. Provably Optimal Aspects of Our Methods:

All the optimizations which are claimed EXACT in this paper are optimal: they have been proven impossible to further improve. This is achieved with an automated software proof with UNSAT and would be PROVABLY OPTIMAL if we had a proof of correctness of the SAT solver software and of course if there is no bug in the SAT solver software.

For example, a SAT solver could claim UNSAT for a certain problem or even output an incorrect proof of UNSAT. However, we can overcome this problem as we have a portfolio of around 500 different SAT solvers software and we can re-check our results with other SAT solvers. Even if we assume the presence of bugs in this software, one can consider that our proofs are *probabilistic proofs*.

Possibly the probability of error could be very small and under some additional assumptions we could have better confidence that our automated proof is indeed correct. We also claim that what we do could be extended to produce fully verifiable mathematical proofs written in a formal language, which prove these optimality results. Some SAT solvers already have the ability to output such proofs.

In order to obtain optimal solutions with respect to a count k for a problem X we proceed as follows in *Algorithm 1*:

Algorithm 1: Given a decision problem X and a count k for the metric of our interest proceed as follows:

- 1) Convert this to SAT-CNF
- 2) Obtain "SAT" and a solution
- 3) Set $k := k - 1$
- 4) Repeat Until "UNSAT"
- 5) Output: k_{\min} such that is "SAT"

IV. ON SOLVING THE MM PROBLEM

A very common approach for tackling the MM problem is to work by solving fixed-size problems and then apply the solution recursively to higher dimensions. The general framework for gluing together solutions for smaller instances and obtain solutions to the general problem is provided by the *divide-and-conquer* paradigm [19].

The complexity of solving the general problem depends on the complexity of solving the underlying smaller sub-problems. Thus, even a slight improvement in such a sub-problem may lead to a huge improvement in the general problem. This general concept can be seen as a pure combinatorial optimization problem with fixed size, which have been studied by many authors since Strassen [2], [20].

In this section, we provide a short description regarding the complexity of existing techniques for solving the MM up-to-date. Additionally, we apply our SAT-based methodology for solving smaller instances of the MM problem. We present new formulaes for multiplying sufficiently small matrices and in some cases we are able to prove that these formulaes are optimal with respect to the number of 2-input multiplications required.

A. On the Complexity of MM

The complexity of the naive algorithm for computing the product of two $n \times n$ matrices is $\mathcal{O}(n^3)$ and similarly the complexity for multiplying a $m \times p$ matrix by a $p \times n$ matrix is $\mathcal{O}(mpn)$. Clearly, as the computation of the product matrix of two $n \times n$ matrices contains n^2 entries, that implies at least n^2 operations are needed and that a proven lower bound for the complexity is $\mathcal{O}(n^2)$.

Thus, the exponent of MM problem over a general non-commutative ring \mathcal{R} defined as

$$\omega(\mathcal{R}) := \inf \tau \in \mathbb{R} | \mathbb{M}_{\mathcal{R}} = \mathcal{O}(n^\tau)$$

Improving the exponent τ of the complexity $\mathcal{O}(n^\tau)$ of MM problem is one of the main interests of the academic community. The first attempt was in 1969 by Volker Strassen who has been able to decrease the complexity of square MM to $\mathcal{O}(n^{2.807})$, by applying recursively the optimal solution he obtained for multiplying two 2×2 matrices with 7 multiplications [20] (cf. *Theorem 1*).

Theorem 1: (Strassen's Algorithm using 7 Multiplications)

Given two 2×2 matrices A, B over a ring \mathcal{R} , with entries $a_{i,j}, b_{i,j} \in \mathcal{R} \ 1 \leq i, j \leq 2$, then the entries $c_{i,j}$ of the product matrix $C = AB$ can be computed by the following formulae,

$$\begin{aligned} P_1 &= (a_{1,1} + a_{2,2})(b_{1,1} + b_{2,2}) \\ P_2 &= (a_{2,1} + a_{2,2})b_{1,1} \\ P_3 &= a_{1,1}(b_{1,2} + b_{2,2}) \\ P_4 &= a_{2,2}(-b_{1,1} + b_{2,1}) \\ P_5 &= (a_{1,1} + a_{1,2})b_{2,2} \\ P_6 &= (-a_{1,1} + a_{2,1})(b_{1,1} + b_{1,2}) \\ P_7 &= (a_{1,2} - a_{2,2})(b_{2,1} + b_{2,2}), \\ c_{1,1} &= P_1 + P_4 - P_5 + P_7 \\ c_{1,2} &= P_2 + P_4 \\ c_{2,1} &= P_3 + P_5 \\ c_{2,2} &= P_1 + P_3 - P_2 + P_6 \end{aligned}$$

Afterwards, Coppersmith and Winograd developed an algorithm to perform MM of square matrices of complexity $\mathcal{O}(n^{2.376})$ [21]. They achieved such a complexity reduction by proving new formulas for computing the inner product of two n -dimensional vectors using fewer 2-input multiplications.

Later in 1975, Laderman published a solution for multiplying 3×3 matrices with 23 multiplications [22]. Since then, this topic generated very considerable interest and yet to this day it is not clear if Laderman's solution in case of 3×3 multiplication can be further improved. For example, we cannot prove if 23 is optimal and no formulas exist using 22 multiplications or even less.

In 2005, a team of scientists from Microsoft Research and two US universities established a new method for finding such algorithms based on group theory, and their best method so far gives an exponents of 2.41 [23], close to Coppersmith-Winograd result and subject to further improvement. However, exponent τ is quite low and it is conjectured that one should be able to do MM in so called *soft quadratic time*, with possibly some poly-logarithmic overheads, which could even

be sub-exponential in the logarithm. This in fact would be nearly linear in the size of the input. Amazingly enough, many scientists conjecture that it could be nearly quadratic like $\mathcal{O}(n^2(\log_2(n))^a)$, for some a .

Our Contribution: In this paper, we proceed by solving the corresponding Brent equations for a given MM problem[9], by converting it into a SAT-CNF problem. This approach has been tried many times before, cf. [9], [17]. Note that this methodology is more generic and it is also applied to multiplication of non-square matrices.

B. New Formulaes for MM Problem

Using our SAT-based methodology as described in previous chapters, we have been able to obtain the following results as presented in Table I.

TABLE I
THE OUTPUT OF APPLYING OUR METHODOLOGY FOR SOLVING THE MM PROBLEM USING A FIXED NUMBER OF MULTIPLICATIONS

| Inputs | No.Mults. | SAT | Av.Time(s) |
|--------|-----------|-----|------------|
| 2,2,2 | 7 | YES | 0.55 |
| 2,2,2 | 6 | NO | 1062.7 |
| 2,2,3 | 11 | YES | 474.5 |
| 2,2,3 | 10 | NO | 4032.2 |
| 2,2,4 | 16 | YES | 0.63 |
| 2,2,4 | 15 | YES | 3152.8 |

As we see from the same table we can prove that multiplying two 2×2 matrices can not be done using less than 7 multiplications and thus Strassen's formulae are optimal.

Using stochastic SAT solvers, we can solve exactly the decision problem: "Can we multiply two matrices A, B using exactly k 2-input multiplications?". We have tried to solve all these underlying decision problems for small problems and we have been able to prove that no better can be done. A new exact result is as formulated below in *Theorem 2*.

Theorem 2: Given two matrices $A \in M_{2 \times 2}(\mathcal{R})$ and $B \in M_{2 \times 3}(\mathcal{R})$ where \mathcal{R} an arbitrary non-commutative ring, then we can compute the product matrix $C = AB$ using at most 11 multiplications

Proof: An upper bound for solving this problem is by naive MM and it is 12 multiplications in total over a general non-commutative ring \mathcal{R} .

First, we consider the Brent Equations corresponding to 11 multiplications. Thus, we obtain 144 equations in 176 unknowns (12098 right clauses).

Then, we convert it to a SAT problem, which we solve using CryptoMiniSat in approximately 474.54s=0.132h. We have obtained the following set of equations for solving the MM problem using 11 multiplications.

$$\begin{aligned} P01 &:= (-a_{11} - a_{12} + a_{21} + a_{22}) * (b_{23}); \\ P02 &:= (-a_{11} - a_{12} + a_{21}) * (b_{12} - b_{23}); \\ P03 &:= (a_{11} - a_{21}) * (-b_{13} + b_{23}); \\ P04 &:= (a_{11}) * (b_{11}); \\ P05 &:= (a_{22}) * (-b_{21} + b_{23}); \end{aligned}$$

$$\begin{aligned}
P06 &:= (-a_{11} - a_{12}) * (b_{12}); \\
P07 &:= (a_{21}) * (b_{12} - b_{13}); \\
P08 &:= (a_{22}) * (b_{21} - b_{22}); \\
P09 &:= (a_{12}) * (-b_{12} + b_{22}); \\
P10 &:= (a_{21}) * (-b_{11} + b_{12}); \\
P11 &:= (a_{12}) * (b_{21}); \\
c_{11} &= (P04 + P11); \\
c_{12} &= (-P06 + P09); \\
c_{13} &= (P02 - P03 - P06 - P07); \\
c_{21} &= (P01 + P02 - P05 - P06 - P10); \\
c_{22} &= (P01 + P02 - P05 - P06 - P08); \\
c_{23} &= (P01 + P02 - P06 - P07);
\end{aligned}$$

Initially, only 117 out of 144 equations were also true over $\mathbb{Z}/4\mathbb{Z}$. After we applied our heuristic lifting technique we lifted all solutions over this ring and the solution was true over an arbitrary ring.

Then, we have obtained the corresponding Brent Equations for 10 multiplications, 144 equations in 160 unknowns and proceeded similarly. The output of our algorithm is UNSAT, implying that there is no solution for this problem. We have verified the UNSAT result using several other SAT-solvers for minimizing the errors due to bugs in software.

Hence, 11 multiplications is the minimal number of required multiplications for solving this problem. ■

In addition, we have applied our methodology for solving the Laderman's problem for multiplying $2 \times 3 \times 3$ matrices using 23 multiplications. Amazingly, we have obtained a new non-isomorphic solution to the same problem and we present it below in *Theorem 3*.

Theorem 3: Given two square matrices matrices $A, B \in M(R, 3)$ where R an arbitrary non-commutative ring, then we can compute the product matrix $C = A.B$ using at most 23 multiplications

Proof: An upper bound for solving this problem is by naive MM and it is 27 multiplications in total over a general non-commutative ring R .

Firstly, we write down the Brent Equations corresponding to 23 multiplications. Thus, we obtain 729 equations in 621 unknowns. Then we convert them to a SAT-CNF problem, which we solve using CryptoMiniSat. The following set of equations is obtained.

$$\begin{aligned}
P01 &:= (a_{23}) * (-b_{12} + b_{13} - b_{32} + b_{33}); \\
P02 &:= (-a_{11} + a_{13} + a_{31} + a_{32}) * (b_{21} + b_{22}); \\
P03 &:= (a_{13} + a_{23} - a_{33}) * (b_{31} + b_{32} - b_{33}); \\
P04 &:= (-a_{11} + a_{13}) * (-b_{21} - b_{22} + b_{31}); \\
P05 &:= (a_{11} - a_{13} + a_{33}) * (b_{31}); \\
P06 &:= (-a_{21} + a_{23} + a_{31}) * (b_{12} - b_{13}); \\
P07 &:= (-a_{31} - a_{32}) * (b_{22}); \\
P08 &:= (a_{31}) * (b_{11} - b_{21}); \\
P09 &:= (-a_{21} - a_{22} + a_{23}) * (b_{33}); \\
P10 &:= (a_{11} + a_{21} - a_{31}) * (b_{11} + b_{12} + b_{33}); \\
P11 &:= (-a_{12} - a_{22} + a_{32}) * (-b_{22} + b_{23}); \\
P12 &:= (a_{33}) * (b_{32}); \\
P13 &:= (a_{22}) * (b_{13} - b_{23}); \\
P14 &:= (a_{21} + a_{22}) * (b_{13} + b_{33});
\end{aligned}$$

$$\begin{aligned}
P15 &:= (a_{11}) * (-b_{11} + b_{21} - b_{31}); \\
P16 &:= (a_{31}) * (b_{12} - b_{22}); \\
P17 &:= (a_{12}) * (-b_{22} + b_{23} - b_{33}); \\
P18 &:= (-a_{11} + a_{12} + a_{13} + a_{22} + a_{31}) * (b_{21} + b_{22} + b_{33}); \\
P19 &:= (-a_{11} + a_{22} + a_{31}) * (b_{13} + b_{21} + b_{33}); \\
P20 &:= (-a_{12} + a_{21} + a_{22} - a_{23} - a_{33}) * (-b_{33}); \\
P21 &:= (-a_{22} - a_{31}) * (b_{13} - b_{22}); \\
P22 &:= (-a_{11} - a_{12} + a_{31} + a_{32}) * (b_{21}); \\
P23 &:= (a_{11} + a_{23}) * (b_{12} - b_{13} - b_{31}); \\
c_{11} &= P02 + P04 + P07 - P15 - P22; \\
c_{12} &= P01 - P02 + P03 + P05 - P07 + P09 + P12 \\
&\quad + P18 - P19 - P20 - P21 + P22 + P23; \\
c_{13} &= -P02 - P07 + P17 + P18 - P19 - P21 + P22; \\
c_{21} &= P06 + P08 + P10 - P14 + P15 + P19 - P23; \\
c_{22} &= -P01 - P06 + P09 + P14 + P16 + P21; \\
c_{23} &= P09 - P13 + P14; \\
c_{31} &= P02 + P04 + P05 + P07 + P08; \\
c_{32} &= -P07 + P12 + P16; \\
c_{33} &= -P07 - P09 + P11 - P13 + P17 + P20 - P21; ■
\end{aligned}$$

This new set of equations for multiplying two 3×3 matrices is non-isomorphic to the system of equations obtained by Ladermann. A full explanation and proof of this is found in [1]. This embraces the conjecture that maybe it can be done with fewer multiplications. We will try to investigate even more this in the future by either seeking for further improvements in our encoding step or running our algorithms on more CPUs working in parallel.

V. EXACT COMBINATORIAL CIRCUIT OPTIMIZATION

In this section, we apply our methodology for obtaining optimal circuit representations for sufficiently small digital circuits with respect to various meaningful metrics. We study circuit representations with respect to the following metrics:

1. *Multiplicative Complexity (MC)*: is the minimum number of AND gates (infinite number of XORs is allowed).

2. *Bitslice Gate Complexity (BGC)*: is the minimum number of 2-input gates of types XOR, OR, AND, NOT needed. This model is relevant in so called *bitslice parallel-SIMD* implementations of block ciphers, e.g. in [15].

3. *Gate Complexity (GC)*: is the minimum number of 2-input gates of types XOR, AND, OR, NAND, NOR, NXOR.

4. *NAND Complexity (NC)*: is defined by the minimum number of 2-input NAND gates.

In order to compute such circuits, we apply the heuristic methodology suggested by Boyar and Peralta [24] based on the notion of MC as follows:

Step 1: First compute the MC.

Step 2: Optimize the number of XORs separately, cf. [25], [10].

Step 3: (Optional Step) At the end do additional optimizations to decrease the circuit depth, and possibly additional software optimizations, cf. [5], [24], [6].

We apply Coding Methodology 2 and we encode the problem formally as a straight-line representation problem,

described by a quantified set of multivariate relations and we convert it to SAT with the Courtois-Bard-Jefferson tool [7] or other methods. Earlier work on computing the MC can be found in [3].

In the next section we apply our methodology for obtaining optimal representations with respect to all these circuit notions for the CTC S-box.

A. Optimal Representations of CTC S-box

More generally, Coding Methodology 2 allows to optimize for arbitrary gates, not only for MC. As a proof of concept we consider the following S-box with 3 inputs and 3 outputs, which have been generated at random for the CTC2 cipher [3] and is defined as,

$$\{7, 6, 0, 4, 2, 5, 1\}.$$

We have tried to optimize this S-box with the well known software Logic Friday (based on Espresso min-term optimization developed at Berkeley) and we obtained 13 gates, which obviously can be further improved. With our software and in a few seconds we obtained several interesting results, each coming with a proof that it is an optimal result. All our theorems are presented in the *Lemmas* below.

Lemma 4: The MC of CTC S-box is exactly 3 (we allow 3 AND gates and an unlimited number of XOR gates) (cf. Figure 1)

Proof: We have obtained the following straight-line program for this problem:

$$\begin{aligned} t_{00} &= x_{01} + x_{02} + 1 & y_0 &= q_{00} + q_{01} + x_{02} \\ t_{01} &= x_{00} + x_{02} + 1 & y_1 &= q_{00} + q_{01} + q_{02} \\ q_{00} &= t_{00} \times t_{01} & y_2 &= q_{00} + x_{00} \\ t_{02} &= x_{02} \\ t_{03} &= x_{00} + x_{01} + x_{02} \\ q_{01} &= t_{02} \times t_{03} \\ t_{04} &= x_{01} + x_{02} + 1 \\ t_{05} &= q_{00} + x_{01} + 1 \\ q_{02} &= t_{04} \times t_{05} \end{aligned}$$

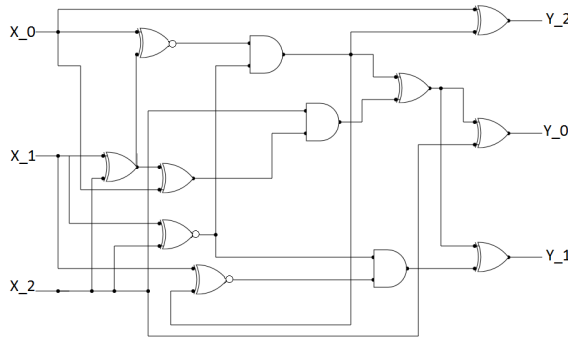


Fig. 1. Our provably optimal implementation of CTC2 S-box [3] with MC 3.

Lemma 5: The Bitslice Gate Complexity (BGC) of CTC S-box is exactly 8 (allowed are XOR,OR,AND,NOT) (we allow

3 AND gates and an unlimited number of XOR gates) (cf. Figure 2)

Proof: Straight-Line Program for CTC2 S-box w.r.t Bit-slice Complexity

$$\begin{aligned} t_{00} &= x_{01} & q_{03} &= t_{06} + t_{07} & t_{15} &= x_{00} \\ t_{01} &= x_{00} & t_{08} &= q_{01} & q_{07} &= t_{14} + t_{15} \\ q_{00} &= t_{00} \times t_{01} + t_{00} + t_{01} & t_{09} &= x_{02} & y_0 &= q_{04} \\ t_{02} &= q_{00} & q_{04} &= t_{08} + t_{09} & y_1 &= q_{05} \\ t_{03} &= 1 & t_{10} &= q_{00} & y_2 &= q_{07} \\ q_{01} &= t_{02} + t_{03} & t_{11} &= q_{03} \\ t_{04} &= x_{00} & q_{05} &= t_{10} + t_{11} \\ t_{05} &= x_{02} & t_{12} &= q_{04} \\ q_{02} &= t_{04} \times t_{05} & t_{13} &= q_{05} \\ t_{06} &= x_{01} & q_{06} &= t_{12} \times t_{13} \\ t_{07} &= q_{02} & t_{14} &= q_{06} \end{aligned}$$

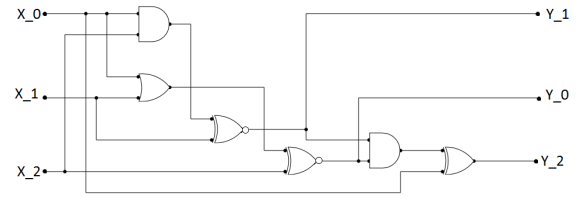


Fig. 2. Our provably optimal implementation of CTC2 S-box with Gate Complexity 6.

Lemma 6: The Gate Complexity (GC) of CTC S-box is exactly 6 (allowing XOR,OR,AND,NOT,NAND,NOR,NXOR) (cf. Figure 3)

Proof: Straight-Line Program for CTC2 S-box w.r.t Gate Complexity

$$\begin{aligned} t_{00} &= x_{01} & q_{03} &= t_{06} + t_{07} \\ t_{01} &= x_{00} & t_{08} &= q_{03} \\ q_{00} &= t_{00} \times t_{01} + t_{00} + t_{01} + 1 & t_{09} &= q_{01} \\ t_{02} &= x_{02} & q_{04} &= t_{08} \times t_{09} \\ t_{03} &= q_{00} & t_{10} &= q_{00} + q_{04} \\ q_{01} &= t_{02} + t_{03} & t_{11} &= x_{00} \\ t_{04} &= q_{01} & q_{05} &= t_{10} + t_{11} \\ t_{05} &= x_{00} & y_0 &= q_{01} \\ q_{02} &= t_{04} \times t_{05} + 1 & y_1 &= q_{03} \\ t_{06} &= q_{02} & y_2 &= q_{05} \\ t_{07} &= x_{01} \end{aligned}$$

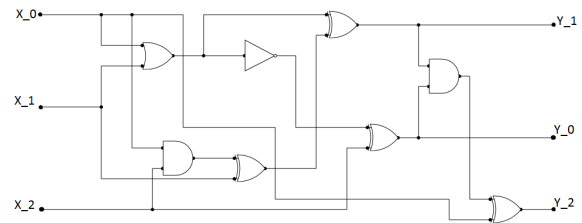


Fig. 3. Our provably optimal implementation of CTC2 S-box with Bitslice Gate Complexity 8.

Lemma 7: The NAND Complexity (NC) of CTC S-box is exactly 12 (only NAND gates and constants) (cf. Figure 4)

Proof: Straight-Line Program for CTC2 S-box w.r.t Bit-slice Complexity

$$\begin{aligned}
 m_{00} &= x_{00} & m_{10} &= q_{05} \\
 m_{01} &= x_{01} & t_{12} &= x_{02} \\
 m_{02} &= x_{02} & t_{13} &= m_{10} \\
 m_{03} &= x_{00} & q_{06} &= t_{12} \times t_{13} + 1 \\
 m_{04} &= x_{01} & m_{11} &= q_{06} \\
 t_{00} &= m_{03} & t_{14} &= x_{00} \\
 t_{01} &= m_{02} & t_{15} &= m_{09} \\
 q_{00} &= t_{00} \times t_{01} + 1 & q_{07} &= t_{14} \times t_{15} + 1 \\
 m_{05} &= q_{00} & m_{12} &= q_{07} \\
 t_{02} &= m_{05} & t_{16} &= m_{08} \\
 t_{03} &= m_{01} & t_{17} &= m_{12} \\
 q_{01} &= t_{02} \times t_{03} + 1 & q_{08} &= t_{16} \times t_{17} + 1 \\
 m_{06} &= q_{01} & m_{13} &= q_{08} \\
 t_{04} &= m_{06} & t_{18} &= m_{10} \\
 t_{05} &= x_{02} & t_{19} &= m_{09} \\
 q_{02} &= t_{04} \times t_{05} + 1 & q_{09} &= t_{18} \times t_{19} + 1 \\
 m_{07} &= q_{02} & m_{14} &= q_{09} \\
 t_{06} &= m_{06} & t_{20} &= m_{12} \\
 t_{07} &= m_{07} & t_{21} &= m_{13} \\
 q_{03} &= t_{06} \times t_{07} + 1 & q_{10} &= t_{20} \times t_{21} + 1 \\
 m_{08} &= q_{03} & m_{15} &= q_{10} \\
 t_{08} &= m_{06} & t_{22} &= m_{15} \\
 t_{09} &= m_{04} & t_{23} &= m_{11} \\
 q_{04} &= t_{08} \times t_{09} + 1 & q_{11} &= t_{22} \times t_{23} + 1 \\
 m_{09} &= q_{04} & m_{16} &= q_{11} \\
 t_{10} &= m_{05} & y_0 &= m_{16} \\
 t_{11} &= m_{06} & y_1 &= m_{14} \\
 q_{05} &= t_{10} \times t_{11} + 1 & y_2 &= m_{13}
 \end{aligned}$$

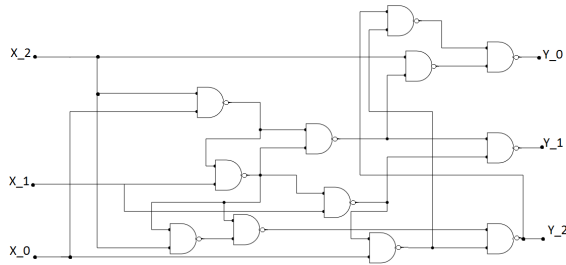


Fig. 4. Our provably optimal implementation of CTC2 S-box with NAND Complexity 12.

Proof: (Complimentary Optimality Proof)

Unlike the great majority of circuit optimizations, needed each time a given cipher is implemented in hardware, our results are exact. They are obtained by solving the problem at a given gate count k , the SAT solver outputs SAT and a solution, and if for $k-1$ gates the SAT solver is good enough and fast enough, it will output UNSAT and we obtain a proven lower bound, a rare thing in complexity.

B. Optimizing the PRESENT S-box

In this section we apply our methodology for optimal circuit representations with respect to bitslice implementation metric for the PRESENT S-box defined as,

$$\{12, 5, 6, 11, 9, 0, 10, 13, 3, 14, 15, 8, 4, 7, 1, 2\} \text{ [26].}$$

In Figure 5, we present a circuit implementation of it with MC 4 (i.e., using only 4 AND gates) and we prove that this circuit implementation is optimal with respect to the AND gates.

Lemma 8: The MC of the PRESENT S-box is exactly 4.

Proof: Initially, we encoded the problem using 3 AND gates and our thoroughly designed and tested system outputs UNSAT. This could be converted to a formal proof that the MC is at least 3. For 4 AND gates, our system outputs SAT and a solution. This can be seen as a software proof.

Further optimization of the linear part, which is also optimal as we also obtained UNSAT for lower numbers, allowed us to minimize the number of XORs to the strict minimum possible (prove by additional UNSAT results).

As a result, for we have obtained an implementation of the PRESENT S-box with 25 gates in total: 4 AND, 20 XOR, 1 NOT, which is optimal w.r.t our Boyar-Peralta 2-step methodology, which is as follows: In overall gate complexity since 25 gates are still not satisfactory.

Straight-Line Program for PRESENT S-box w.r.t Bit-slice Complexity

$$\begin{aligned}
 u_{00} &= x_3 + x_1 & t_{00} &= x_4 & y_1 &= p_{02} + u_{02} \\
 u_{01} &= u_{00} + 1 & t_{01} &= u_{05} & y_2 &= v_{01} + u_{06} \\
 u_{02} &= x_1 + x_4 & p_{00} &= t_{00} \times t_{01} & y_3 &= v_{03} + u_{07} \\
 u_{03} &= u_{01} + x_4 & t_{02} &= p_{00} + u_3 & y_4 &= v_{00} + x_2 \\
 u_{04} &= x_3 + x_2 & t_{03} &= p_{00} + x_2 \\
 u_{05} &= u_{04} + x_4 & p_{01} &= t_{02} \times t_{03} \\
 u_{06} &= u_{03} + x_2 & t_{04} &= u_{04} \\
 u_{07} &= u_{06} + u_{00} & t_{05} &= x_3 \\
 u_{08} &= x_1 + u_{03} & p_{02} &= t_{04} \times t_{05} \\
 & & t_{06} &= u_{02} + u_{08} \\
 & & t_{07} &= p_{02} + u_{01} \\
 & & p_{03} &= t_{06} \times t_{07} \\
 & & v_{00} &= p_{03} + p_{00} \\
 & & v_{01} &= p_{03} + p_{02} \\
 & & v_{02} &= p_{00} + p_{02} \\
 & & v_{03} &= v_{02} + p_{01}
 \end{aligned}$$

A better result in terms of gate complexity can be achieved by the following method: we observe that AND gates and OR gates are affine equivalents, and it is likely that if we implement certain AND gates with OR gates, we might be able to further reduce the overall complexity of the linear parts. We may try all possible 2^4 cases where some AND gates are implemented with OR gates. Even better results can be obtained if we consider also NOR and NAND gates. By this method, starting with the right optimization with MC=4, as several such optimizations may exist, we can obtain the following new implementation of the PRESENT S-box which requires only 14 gates total.

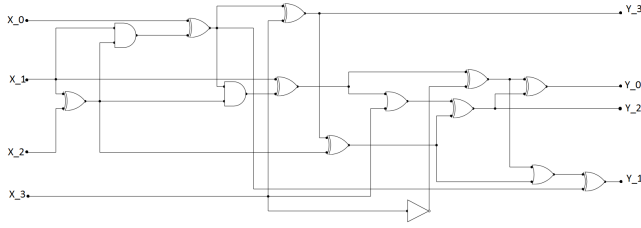


Fig. 5. Our provably optimal Bitslice-type implementation of PRESENT S-box with 14 gates.

Applications. This implementation is used in our recent bit-slice implementation of PRESENT, see [15]. In addition, we postulate that this implementation of the PRESENT S-box is in certain sense optimal for DPA-protected hardware implementations with linear masking, as it minimizes the number of non-linear gates (there are only 4 such gates).

Discussion. Our best optimisation of the PRESENT S-box seems to confirm the Boyar-Peralta heuristic to the effect that some of the best possible gate-efficient implementations are very closely related to the notion of MC. However, the most recent implementations of the AES S-box, in the second paper by Boyar and Peralta, show that further improvements, and also circuit depth improvements, can be achieved also by relaxing the number of ANDs used as in the latest optimization of the 4-bit inverse in $GF(2^4)$ for AES given on Figure 1 in [6].

C. Optimal Representations of GOST S-boxes

In this section, we apply our automated methodology encode-and-then-solve for obtaining optimal circuit representations of the 8 S-boxes $S1 - S8$ of GOST block cipher with respect to the number of AND gates.

GOST block cipher has a simple 32-round Feistel structure, which encrypts a 64-bit block using a 256-bit key defined in the standard GOST 28147-89 [27]. We consider the main standard and most widely known version of the GOST block cipher, known as "GostR3411 94 TestParamSet", also known as the one used by the Central Bank of the Russian Federation, and 7 additional versions which are found in the OpenSSL source code.

We have obtained for all these versions optimal representations with respect to the number of AND gates and these results are summarized in Table II.

Lemma 9: The MC of the eight S-boxes of GOST cipher $S1-S8$ and for 8 principal known version of GOST as specified in OpenSSL are EXACTLY equal to the values given on Table II.

Further work. With suitable encoding for other components, and in particular for the addition modulo 2^{32} in GOST cipher, we are potentially able to provably minimize the number of non-linear gates in a whole cipher to a (proven) lower bound. We can obtain very compact algebraic encodings of GOST which can be used for algebraic cryptanalysis, see [4], [13].

TABLE II
MC FOR ALL KNOWN GOST S-BOXES

| S-box Set Name | | | | | | | | |
|-----------------------------------|------|------|------|------|------|------|------|--|
| $S1$ | $S2$ | $S3$ | $S4$ | $S5$ | $S6$ | $S7$ | $S8$ | |
| GostR3411_94_TestParamSet | | | | | | | | |
| 4 | 5 | 5 | 5 | 5 | 5 | 4 | 5 | |
| GostR3411_94_CryptoProParamSet | | | | | | | | |
| 4 | 5 | 5 | 4 | 5 | 5 | 4 | 5 | |
| Gost28147_TestParamSet | | | | | | | | |
| 4 | 4 | 4 | 4 | 4 | 5 | 5 | 5 | |
| Gost28147_CryptoProParamSetA | | | | | | | | |
| 5 | 4 | 5 | 4 | 4 | 4 | 5 | 5 | |
| Gost28147_CryptoProParamSetB | | | | | | | | |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | |
| Gost28147_CryptoProParamSetC | | | | | | | | |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | |
| Gost28147_CryptoProParamSetD | | | | | | | | |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | |
| GostR3411_94_SberbankHashParamset | | | | | | | | |
| 4 | 4 | 4 | 5 | 5 | 4 | 4 | 4 | |

The better the optimizations obtained, the more compact the representation we get for a given cipher, and heuristically this leads to better algebraic attacks. This provides additional motivation for our work. Such compact representations can be combined with several complexity reduction and differential attacks to obtain attacks against full rounds of GOST.

D. Optimization of the Majority Function

The Majority function is a function of the form $\mathbb{F}_2^n \rightarrow \mathbb{F}_2$, which is False when $\frac{n}{2}$ of its inputs are false and vice versa for True. It is a highly non-trivial task to obtain circuit representations with optimal MC in the case when n is odd.

Using our 2-step automated procedure we have been able to find optimal circuit representation for the Majority function in cases when $n = 3, 5, 7$.

Lemma 10: The MC for the Majority Function when $n = 3$ is 1 (cf. Figure 6)

Proof: Using our methodology we have obtained the following circuit representation.

$$t_0 = x_0 \oplus x_2, t_1 = x_0 \oplus x_1 \\ q_0 = t_0 \wedge t_1, q_1 = q_0 \oplus x_1$$

Lemma 11: The MC for the Majority Function when $n = 5$ is 3 (cf. Figure 7)

Proof: Using our methodology we have obtained the following circuit representation.

$$k_0 = x_0 \oplus x_1, k_1 = x_3 \oplus x_4, t_0 = k_0 \oplus k_1, \\ t_1 = x_1 \oplus x_2, q_0 = t_0 \wedge t_1, t_2 = x_0 \oplus x_3, \\ q_1 = k_1 \wedge t_2, k_2 = k_1 \oplus t_2, t_3 = q_1 \oplus k_2, \\ k_3 = x_2 \oplus x_4, t_4 = q_0 \oplus k_3, q_2 = t_3 \wedge t_4, \\ o_0 = q_2 \oplus x_4$$

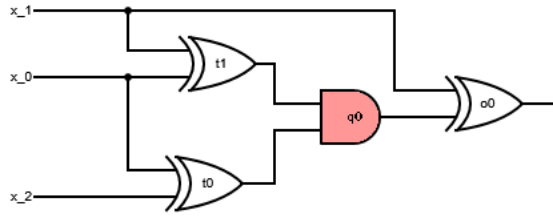


Fig. 6. Circuit Representation of Majority function on 3 inputs with optimal MC=1

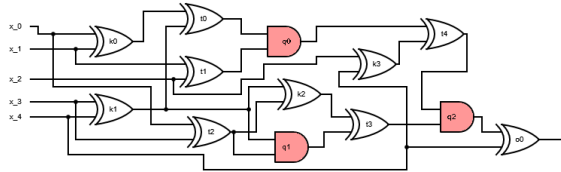


Fig. 7. Circuit Representation of Majority function on 5 inputs with optimal MC=3

Lemma 12: The MC for the Majority Function when $n = 7$ is 4 (cf. Figure 8)

Proof: Using our methodology we have obtained the following circuit representation. We have obtained a circuit with 23 gates in total: 4 AND gates, 1 NOT gate and 18 XOR gates.

$$\begin{aligned}
 t_0 &= x_0 \oplus x_1, t_1 = x_0 \oplus x_2, q_0 = t_0 \wedge t_1, \\
 t_2 &= x_4 \oplus x_5, t_3 = x_3 \oplus x_4, q_1 = t_2 \wedge t_3, \\
 p_0 &= q_0 \oplus q_1, k_0 = x_1 \oplus x_2, k_1 = x_4 \oplus x_6, \\
 k_2 &= x_3 \oplus x_5, l_0 = k_0 \oplus k_1, l_1 = p_0 \oplus l_0, \\
 t_4 &= l_1 \oplus 1, t_5 = k_1 \oplus k_2, q_2 = t_4 \wedge t_5, \\
 k_3 &= x_0 \oplus x_4, t_6 = p_0 \oplus k_3, p_1 = q_0 \oplus q_2, \\
 k_4 &= x_0 \oplus x_6, t_7 = p_1 \oplus k_4, q_3 = t_6 \wedge t_7, \\
 p_2 &= q_0 \oplus q_3, o_0 = p_2 \oplus x_0
 \end{aligned}$$

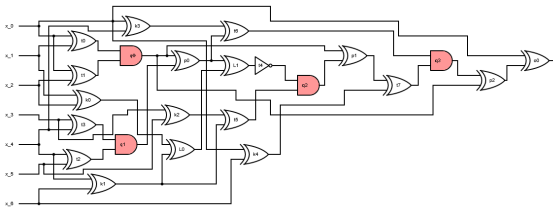


Fig. 8. Circuit Representation of Majority function on 7 inputs with optimal MC=4

VI. CONCLUSION

The construction of efficient computational circuits and the optimization of arbitrary algebraic computations over fields and rings in the general non-commutative setting is considered as one of most important problems in computer science, applied mathematics and the industry. It has numerous

applications in improving various linear/polynomial/graph/language/cryptographic algorithms and an important footprint in many current applications. For example, in improving basic all-purpose linear algebra routines or in efficient implementation of important cryptographic algorithms, which are used in countless software (and hardware) systems.

In this paper, we studied two fundamental problems in the area of computational complexity. The notion of MC which minimizes the number of elementary non-linear operations (AND gates), and more general problems of gate complexity under almost any circuit complexity "metric", for example MC, gate count w.r.t. a specific set of gates, circuit depth, circuit width, etc. Following the heuristic of Boyar-Peralta, we used MC as an essential tool for optimizing potentially arbitrary algebraic computations over fields and rings and in particular for binary circuits. Additionally, we focused on the combinatorial logic optimization of general digital circuits with particular attention to small substitution boxes (S-boxes), which are massively used in the industrial cryptographic schemes and which are small enough for some very advanced methods to be invented. Thus, in many cases we managed to obtain optimal results which can no longer be improved, and we have a formal mathematical proof of these claims by an automated software proof technique with a SAT solver.

We have developed a fully automated procedure for obtaining new formulas for Matrix Multiplication (MM) problems, complex multiplication problems, multiplication of quaternions and for construction of optimal circuit representations with any given Boolean functions, with respect to any given set of basic gates. Our methodology consists of three main steps. In the first step, we formally encode these problems as polynomial equations, then convert them into a SAT problem using the Courtois-Bard-Jefferson [7] or other methods and then we solve these problems using SAT solver software. Thus, we have been able to find new formulas for multiplying two 3×3 matrices using 23 2-input multiplications [11] and also multiplying a 2×2 matrix by a 2×3 matrix using only 11 multiplications, naive multiplication needs 12. We have been able to construct several optimal circuit representations for the S-box of CTC2 cipher [3] with respect to its MC, Bitslice Gate Complexity, Gate Complexity and NAND Gate Complexity. Additionally, we constructed an optimal circuit representation with 14 gates for the PRESENT S-box, which is the best currently known [15] and we computed the exact MC of the 8 S-boxes S1-S8 used in the GOST cipher. The amazing thing is that our methodology can find EXACT circuit representations (which is very hard to obtain in the area of computational complexity). This is if our SAT solvers used in the final solving stage are complete (in a sense that they are fast enough and output UNSAT if the problem has no solution). In future works, we need to address the questions of what form or language such automated proofs could be output, shared and published.

Cryptography is always very costly, and a lot of effort is always done in order to improve the implementation of any given cipher [5], [6]. To the best of our knowledge

provably optimal circuits have never been found before for any cryptographic algorithm, and optimal digital circuits have never been yet used in industrial applications. Interestingly, if a cryptographic algorithm such as AES which tends to be present nowadays in more or less every major CPU [5] could be implemented in a provably optimal way (or at least it would be optimal for smaller components like in the present paper), people in the industry would no longer need to make their designs proprietary. The best digital designs could be developed at universities and licensed or offered to the worldwide industry or jointly developed by cooperative industrial consortia to benefit every company small and large. Thus, the (large) computational cost and research effort needed to develop such top-end (excessively good) implementations and optimizations could be amortized and justified.

ACKNOWLEDGMENT

This research was supported by the UK Technology Strategy Board in the United Kingdom under project 9626-58525.

REFERENCES

- [1] N.T. Courtois, D. Hulme, and T. Mourouzis, "Multiplicative Complexity And Solving Generalized Brent Equations With SAT Solvers," in COMPUTATION TOOLS 2012, in the Third International Conference on Computational Logics, Algebras, Programming, Tools, and Benchmarking, pp. 22-27, 2012.
- [2] J. Patarin, N. Courtois, and L. Goubin, "Improved Algorithms for Isomorphism of Polynomials," in Advances in CryptologyEUROCRYPT'98, pp. 184-200, Springer Berlin Heidelberg, 1998.
- [3] N.T. Courtois, D. Hulme, and T. Mourouzis, "Solving Circuit Optimisation Problems in Cryptography and Cryptanalysis," in Proceedings of 2nd IMA Conference Mathematics in Defence, UK, Swindon, 2011.
- [4] N.T. Courtois, D. Hulme, and T. Mourouzis, "Solving Circuit Optimisation Problems in Cryptography and Cryptanalysis," in IACR Cryptology ePrint Archive 2011, Report 475,2012.
- [5] J. Boyar, P. Matthews, and R. Peralta, "Logic Minimization Techniques with Applications to Cryptology," Journal of Cryptology, vol. 26, p. 280-312, 2013.
- [6] J. Boyar, and R.Peralta, "A depth-16 circuit for the AES S-box," in IACR Cryptology ePrint Archive, Report 33,2011.
- [7] G.V. Bard, N.T. Courtois, and C. Jefferson, "Efficient Methods for Conversion and Solution of Sparse Systems of Low-Degree Multivariate Polynomials over GF(2) via SAT-Solvers," ECRYPT workshop Tools for Cryptanalysis, 2007.
- [8] N. Sorensson, and N. Een, "Minisat v1. 13-a sat solver with conflict-clause minimization," SAT journal pp. 53-59, 2005.
- [9] R. Brent, "Algorithms for matrix multiplication," Tech. Report Report TR-CS-70-157, Department of Computer Science, Stanford, 1970.
- [10] C. Fuhs, and P. Schneider-Kamp, "Synthesizing Shortest Linear Straight-Line Programs over GF(2) Using SAT," in SAT 2010, Theory and Applications of Satisfiability Testing, Springer LNCS 6175, pp. 71-84, 2010.
- [11] N.T. Courtois, G.V. Bard, and D. Hulme, "A New General-Purpose Method to Multiply 3x3 Matrices Using Only 23 Multiplications," in arXiv preprint arXiv:1108.2830, 2011.
- [12] N. T. Courtois, and G. Bard, "Algebraic Cryptanalysis of the Data Encryption Standard," in Cryptography and Coding, 11-th IMA Conference, pp. 152-169, LNCS 4887, Springer, 2007.
- [13] N.T. Courtois, "Algebraic Complexity Reduction and Cryptanalysis of GOST," in IACR Cryptology ePrint Archive, Report 626, 2011.
- [14] A. Edelman, "Large Dense Numerical Linear Algebra in 1994 (survey)," Journal of Supercomputer Applications. Vol. 7, p. 113128, 1993.
- [15] M. Albrecht, N.T. Courtois, D. Hulme, and G. Song, "Bit-Slice Implementation of PRESENT in pure standard C," Available online at www.nicolascourtois.com, 2011.
- [16] E. Prouff, C. Giraud, and S. Aumonnier, "Provably Secure S-Box Implementation Based on Fourier Transform," in CHES 2006, Springer LNCS 4249, pp. 216-230, 2006.
- [17] G. Bard, "Algorithms for Solving Linear and Polynomial Systems of Equations over Finite Fields with Applications to Cryptanalysis," Submitted in Partial Fulfillment for the degree of Doctor of Philosophy of Applied Mathematics and Scientific Computation, 2007.
- [18] M. Soos, "CryptoMiniSat 2.5.0," in SAT Race competitive event booklet, 2010.
- [19] S. Dasgupta, C. Papadimitriou, and U. Vazirani, "Algorithms," 2nd Edition, 2006.
- [20] V. Strassen, "Gaussian elimination is not optimal," in Numerische Mathematik Vol 13 pp. 354-356, 1969.
- [21] D. Coppersmith, and S.Winograd, "On the asymptotic complexity of matrix multiplication," SIAM Journal Comp., Vol 11, pp 472-492, 1980.
- [22] J.D. Laderman, "A Non-Commutative Algorithm for Multiplying 3x3 Matrices Using 23 Multiplications," in Amer. Math. Soc. Vol. 82, Number 1, 1976.
- [23] H. Cohn, R. Kleinberg, B. Szegedyz, and C. Umans, "Grouptheoretic Algorithms for Matrix Multiplication," in FOCS05, 46th Annual IEEE Symposium on Foundations of Computer Science, pp. 379-390, 2005.
- [24] J. Boyar, and R. Peralta, "A New Combinational Logic Minimization Technique with Applications to Cryptology," in SEA 2010, pp. 178-189, 2009.
- [25] J. Boyar, P. Matthews, and R. Penalta, "On the Shortest Linear Straight-Line Program for Computing Linear Forms," in Mathematical Foundations of Computer Science 2008, pp. 168-179, Springer Berlin Heidelberg, 2008.
- [26] A. Bogdanov, L.R. Knudsen, G. Leander, C. Paar, A. Poschmann, and M.J.B. Robshaw,"PRESENT: An Ultra-Lightweight Block Cipher," in CHES 2007, LNCS 4727, pp. 450-466, Springer, 2007.
- [27] A. Poschmann, S. Ling, and H. Wang, "256 Bit Standardized Crypto for 650 GE GOST Revisited," in CHES 2010, LNCS 6225, pp. 219-233, 2010.

KLocator: An Ontology-Based Framework for Scenario-Driven Geographical Scope Resolution

Panos Alexopoulos, Carlos Ruiz, Boris Villazon-Terrazas, and José-Manuel Gómez-Pérez
iSOCO, Intelligent Software Components S.A.
 Av. del Partenon, 16-18, 28042, Madrid, Spain
 Email: {palexopoulos, cruiz, bvillazon, jmgomez}@isoco.com

Abstract—The automatic extraction of geographical information from textual pieces of information is a challenging task that has been getting increasing attention from application and research areas that need to incorporate location-awareness in their methods and services. In this paper, we present KLocator, a novel ontology-based system for correctly identifying geographical entity references within texts and mapping them to knowledge sources, as well as determining the geographical scope of texts, namely the areas and regions to which the texts are geographically relevant. Compared to other similar approaches, KLocator has two important novelties: i) It does not utilize only background geographical information for performing the above tasks but allows the exploitation of any kind of semantic information that is explicitly or implicitly related to geographical entities in the given domain and application scenario. ii) It is highly customizable, allowing users to define and apply custom geographical resolution models that best fit to the domain(s) and expected content of the texts to be analyzed. Both these features, according to our experiments, manage to substantially improve the effectiveness of the geographical entity and scope resolution tasks, especially in scenarios where explicit geographical information is scarce.

Keywords—Geographical Entity Resolution; Geographical Scope Resolution; Ontologies; Semantic Data.

I. INTRODUCTION

In this paper, we present KLocator a novel ontology-based framework for performing geographical semantic analysis of textual information, in the form of geographical entity and scope resolution. An initial version of the framework has already been presented in [1]; in this paper we extend this work by providing i) a more comprehensive positioning of it in the semantic information processing research landscape, ii) a detailed technical description of the system's implementation and way of use and iii) enhanced experiments with more input data.

In general, our work is related to Geographical Intention Retrieval [2], an area that covers techniques related to the retrieval of information involving some kind of spatial awareness. The goal is to improve services and applications that rely on geographical information, ranging from its quite straightforward use in map services, to more advanced personalization techniques. The main idea is that a text or a query has a geographic scope. For instance, a query for cheap flights from London to Paris would include both

London and Paris in the geographic scope, but not locations in between. Similarly, a text describing the Eiffel tower will have the geographic scope of Paris, rather than of France.

Current geo-location services retrieve likely geographical locations for given keywords or text [3] by mostly applying data mining and statistical techniques on large-scale Web data. Nevertheless, the analysis they perform is primarily a syntactic one, without any exploitation of the text's semantics. The result of this are problems like ambiguity where locations with the same name (Paris, France vs. Paris, Texas) or locations named somehow similar to non-geographic concepts (such as Reading, UK) are not correctly resolved. Thus, semantic analysis, either built on top of statistical analysis or as a standalone approach, can improve current approaches by extracting not only geographical entities from a text, but also other types of entities (people, companies, etc.) that can, via reasoning or inference techniques, improve the accuracy and completeness of the extracted geographical information.

Of course, a bottleneck in applying semantic approaches is the need for geographical knowledge bases as input to the system. Previous approaches have tried to build geographic knowledge on top of different kind of resources, including ad hoc ontologies, geo-gazetteers or more generic knowledge hubs such as Wikipedia. A more promising approach, however, for avoiding or at least limiting the initial entry barriers for geographical semantic analysis is the reuse of Open Data. In particular, the Linked Data initiative [4] provides a crucial starting point for building a large and reliable geographical centered knowledge base, with enough information from other type of entities to allow for a comprehensive coverage of most domains. Moreover, there are some Linked Data initiatives, such as GeoLinkedData [5] and LinkedGeoData [6], that aim to enrich the Web of Data with geographical data.

Given the above, in this paper, we focus on geographical analysis of textual information and we present KLocator, a novel ontology-based framework that focuses on tackling two problems:

- 1) The problem of **geographical entity resolution**, namely the detection within a text of geographical entity references and their correct mapping to ontological

uris that represent them.

- 2) The problem of **geographical scope resolution**, namely the determination of areas and regions to which the text is geographically relevant.

The proposed framework has two distinguishing characteristics. First, unlike other ontology-based approaches which utilize only geographical information for performing the above tasks, it allows the exploitation of any kind of semantic information that is explicitly or implicitly related to geographical entities in the given domain and application scenario. In that way, it manages to significantly improve the accuracy of the above tasks in domains and scenarios where explicit geographical information is scarce.

Second, it is highly customizable as it allows users to define and apply **Geographical Resolution Evidence Models**, based on their knowledge about the domain(s) and expected content of the texts to be analyzed. This allows KLocator to adapt to the particular characteristics of different domains and scenarios and be more effective than other similar systems primarily designed to work in open domain and unconstrained scenarios.

The rest of the paper is as follows. Section II presents related works. Section III describes in detail KLocator's geographical resolution framework while Section IV provides implementation details of the system as well as guidelines on how to use it in practice. Section V presents and discusses experimental results regarding the evaluation of the framework's effectiveness. Finally, a critical discussion of the overall framework is given in Section VI, while conclusions and lines of future work are outlined in Section VII.

II. RELATED WORK

In this section, we describe relevant to our work approaches, both from the area of geographical information processing and from the semantic content analysis one.

A. Geographical-Specific Approaches

The majority of related approaches to our work are found in the area of geographical information retrieval [2], where several approaches based on information retrieval, machine learning or semantic techniques attempt to resolve geographic entities and scope.

Andogah et al. [7] describe an approach to place ambiguity resolution in text consisting of three components; a geographical tagger, a geographical scope resolver, and a placename referent resolver. The same authors, in [8], also propose determining the geographical scope as means to improve the accuracy in relevance ranking and query expansion in search applications. However, these processes only rely on limited geographical information rather than using some other data available.

Following a strict semantic approach, Kauppinen et al. [9] present an approach using two ontologies (SUO - a large Finnish place ontology, and SAPO - a historical and

geographical ontology) and logic rules to deal with heritage information where modern and historical information is available (e.g., new name for a place, new borders in a country). This method is combined with some faceted search functionalities, but they do not propose any method for disambiguating texts.

More related to the fact that the disambiguation of a location depends on the context (such as in "London, England" vs. "London, Ontario"), Peng et al. [10] propose an ontology-based method based on local context and sense profiles combining evidence (location sense context in training documents, local neighbor context, and the popularity of individual location sense) for such disambiguation.

B. Generic Entity and Scope Resolution Approaches

Since geographical entities are just a subset of named entities that are typically considered in the Information Extraction literature (persons and organizations are other examples), more generic named entity resolution approaches may be applied to them.

A recent ontology-based entity resolution approach is described in [11] where an algorithm for entity reference resolution via Spreading Activation on RDF Graphs is proposed. The algorithm takes as input a set of terms associated with one or more ontology elements and uses the ontology graph and spreading activation in order to compute Steiner graphs, namely graphs that contain at least one ontology element for each entity. These graphs are then ranked according to some quality measures and the highest ranking graph is expected to contain the elements that correctly correspond to the entities.

Several approaches utilize Wikipedia as a highly structured knowledge source that combines annotated text information (articles) and semantic knowledge (through the DBpedia [12] and YAGO [13] ontologies). For example, DBpedia Spotlight [14] is a tool for automatically annotating mentions of DBpedia resources in text by using i) a lexicon that associates multiples resources to an ambiguous label and which is constructed from the graph of labels, redirects and disambiguations that DBpedia ontology has and ii) a set of textual references to DBpedia resources in the form of Wikilinks. These references are used to gather textual contexts for the candidate entities from wikipedia articles and use them as disambiguation evidence.

A similar approach that uses the YAGO ontology is the AIDA system [15], which combines three entity disambiguation measures: the prior probability of an entity being mentioned, the similarity between the contexts of a mention and a candidate entity, and the semantic coherence among candidate entities for all mentions together. The latter is calculated based on the distance between two entities in terms of type and subclass of edges as well as the number of incoming links that their Wikipedia articles share.

Other related approaches utilize ontological information for semantically characterizing documents [16] [17] [18]. The first two frameworks assume a categorized ontology, i.e., an ontology whose concepts belong to particular predefined categories (e.g., education, sports, politics, etc.) and, based on the entities found in the document, they compute the categories it belong to through graph similarity measures. On the other hand, the framework of [18] annotates particular segments of the documents with entities derived from a database.

The difference between the above approaches and KLocator is detected in the way they treat the available semantic data. For example, Spotlight uses the DBpedia ontology only as an entity lexicon without really utilizing any of its relations, apart from the redirect and disambiguation ones. Thus, it is more text-based than ontology-based. On the other hand, AIDA builds an entity relation graph by considering only the type and subclass of relations as well as “assumed” relations inferred by the links within the articles. The problem with this approach is that important semantic relations that are available in the ontology are not utilized and, of course, there is no control over which edges of the derived ontology graph should be utilized in the given scenario. Such control is not provided either in [11] or any of the rest aforementioned approaches.

III. GEOGRAPHICAL SCOPE RESOLUTION FRAMEWORK

KLocator facilitates geographical entity and scope resolution in application scenarios where:

- The documents’ domain(s) and content nature are a priori known or can be predicted.
- Comprehensive ontologies covering these domain(s) are available (either purposely built or from existing sources such as Linked Data).

By content nature, we practically mean the types of semantic entities and relations that are expected to be found in the documents. For example, in film reviews one can expect to find films along with directors and actors that have directed them or played in them, respectively. Similarly, in texts describing historical events one will probably find, among others, military conflicts, locations where these conflicts took place and people and groups that participated in them. Documents with known content nature, like the above, can be found in many application scenarios where content is specialized and focused (e.g., reviews, scientific publications, textbooks, reports, etc).

Given such scenarios, our proposed framework targets the two tasks of geographical entity and scope resolution based on a common assumption: that the existence of both geographical and non-geographical entities within a text may be used as **evidence** that indicate which is the most probable meaning of an ambiguous location term as well as which locations constitute the geographical scope of the whole text.

To see why this assumption makes sense, assume a historical text containing the term “Tripoli”. If this term is collocated with terms like “*Siege of Tripolitsa*” and “*Theodoros Kolokotronis*” (the commander of the Greeks in this siege) then it is fair to assume that this term refers to the city of Tripoli in Greece rather than the capital of Libya. Also, in a historical text like “*The victory of Greece in the Siege of Tripolitsa under the command of Kolokotronis was decisive for the liberation from Turkey*”, the evidence provided by “*Siege of Tripolitsa*” and “*Kolokotronis*” and “*Greece*” indicates that Tripoli is more likely to be the location the text is about rather than Turkey.

Now, which entities and to what extent are potential evidence in a given application scenario depends on the domain and expected content of the texts that are to be analyzed. For example, in the case of historical texts we expect to use as evidence historical events and persons that have participated in them. For that reason, our approach is based on the a priori determination and acquisition of the optimal evidential knowledge for the scenario in hand. This knowledge is expected to be available in the form of an ontological knowledge base and it is used within the framework to perform geographical entity and scope resolution. The framework components that enable this are the following:

- A **Geographical Resolution Evidence Model** that contains both geographical and non-geographical semantic entities that may serve as location-related evidence for the application scenario and domain at hand. Each entity is assigned evidential power degrees, which denote its usefulness as evidence for the two resolution tasks.
- A **Geographical Entity Resolution Process** that uses the evidence model to detect and extract from a given text terms that refer to locations. Each term is linked to one or more possible location uris along with a confidence score calculated for each of them. The uri with the highest confidence should be the correct location the term refers to.
- A **Geographical Scope Resolution Process** that uses the evidence model to determine, for a given text, the location uris that potentially fall within its geographical scope. A confidence score for each uri is used to denote the most probable locations.

In the following paragraphs, we elaborate on each of the above components.

A. Geographical Resolution Evidence Model

For the purpose of this paper, we define an ontology as a tuple $O = \{C, R, I, i_C, i_R\}$ where

- C is a set of concepts.
- I is a set of instances.
- R is a set of binary relations that may link pairs of concept instances.

- i_C is a concept instantiation function $C \rightarrow I$.
- i_R is a relation instantiation function $R \rightarrow I \times I$.

Given an ontology, the **Geographical Resolution Evidence Model** defines which ontological instances and to what extent should be used as evidence towards i) the correct meaning interpretation of a location term to be found within the text and ii) the correct geographical scope resolution of the whole text. More formally, given a domain ontology O and a set of locations $L \subseteq I$, a geographical resolution evidence model consists of two functions:

- A **location meaning evidence function** $lme_f : L \times I \rightarrow [0, 1]$. If $l \in L$ and $i \in I$ then $lme_f(l, i)$ is the degree to which the existence, within the text, of i should be considered an indication that l is the correct meaning of any text term that has l within its possible interpretations.
- A **geographical scope evidence function** $gse_f : L \times I \rightarrow [0, 1]$. If $l \in L$ and $i \in I$ then $gse_f(l, i)$ is the degree to which the existence, within the text, of i should be considered an indication that l represents the geographical scope of the text.

It is important to note that, though similar in form, these two functions have different meaning and use which, as we show in subsequent sections, is reflected in the way they are calculated and applied. Function lme_f is to be used for disambiguation purposes and its values depend primarily on the ambiguity of the evidential entities. On the other hand, gse_f is to be used for geographical scope resolution and its values have mostly to do with the number and of the evidential entities.

Both functions are expected to be constructed prior to the execution of the resolution process through a semi-automatic process. To do that, for a given domain and scenario, we need to consider the concepts whose instances are directly or indirectly related to locations and which are expected to be present in the texts to be analyzed. The more domain specific the texts are, the smaller the ontology needs to be and the more effective and efficient the whole resolution process is expected to be. In fact, it might be that using a larger ontology than necessary could reduce the effectiveness of the resolution process.

For example, assume that the texts to be analyzed are about American History. This would mean that the locations mentioned within these texts are normally related to events that are part of this history and, consequently, locations that had nothing to do with these events need not be considered. In that way, the range of possible meanings for location terms within the texts as well as the latter's potential scope is considerably reduced. Therefore, a strategy for selecting the minimum required instances that should be included in the location evidence model would be the following:

- First, identify the concepts whose instances may act as location evidence in the given domain and texts.

- Then, identify the subset of these concepts, which constitute the central meaning of the texts and thus “determine” mostly their location scope.
- Finally, use these concepts in order to limit the number of possible locations that may appear within the text as well as the number of instances of the other evidential concepts.

The result of the above process should be a location evidence mapping function $lem : C \rightarrow R^n$ which given an evidential concept $c \in C$ returns the relations $\{r_1, r_2, \dots, r_n\} \in R^n$ whose composition links c 's instances to locations.

Table I shows such a mapping for the history domain and in particular about that of military conflicts where, for instance, military conflicts provide scope related evidence for the locations they have taken place in and military persons provide evidence for locations they have fought in. The latter mapping, shown in the third row of the table, is facilitated by the chain of two relations: i) the inverse of the relation **dbpprop:commander** that relates persons with battles they have commanded and ii) the relation **dbpprop:place** that relates battles to their locations). In a similar way, one may define a location evidence mapping for the same scenario by, for example, considering the military conflicts mentioned in the text as evidence for the disambiguation of the military persons.

Table I
LOCATION EVIDENCE MAPPING FUNCTION FOR MILITARY CONFLICTS DOMAIN

| Evidence Concept | Location Linking Relation(s) |
|-------------------|---|
| Military Conflict | <i>dbpprop:place</i> |
| Military Conflict | <i>dbpprop:place, dbpedia-owl:isPartOf</i> |
| Military Person | <i>is dbpprop:commander of, dbpprop:place</i> |
| Location | <i>dbpedia-owl:isPartOf</i> |

Using this mapping function, we can calculate the location meaning evidence function lme_f as follows. Given a location $l \in L$ and an instance $i \in I$, which belongs to some concept $c \in C$ and is related to l through the composition of relations $\{r_1, r_2, \dots, r_n\} \in lem(c)$, we derive the set of locations $L_{amb} \subseteq L$ which share common names with l and are also related to i through $\{r_1, r_2, \dots, r_n\} \in lem(c)$. Then the value of the function lme_f for this location and this instance is:

$$lme_f(l, i) = \frac{1}{|L_{amb}|} \quad (1)$$

The intuition behind this formula is that the evidential power of a given instance is inversely proportional to the number of different target locations it provides evidence for. If, for example, a given military person has fought in 2 different locations with the same name, then its evidential power for this name is 0.5.

Using the same equation we can also calculate the geographical scope evidence function $gsef$, the only difference being that we consider the set L'_{amb} that contains all the locations related to i , not just the ones with the same name as l :

$$gsef(l, i) = \frac{1}{|L'_{amb}|} \quad (2)$$

Again, the intuition here is that the geographical scope-related evidential power of a given instance is inversely proportional to the number of different locations it is related to. Thus, if the military person of the above example has fought battles in 4 locations in total (independently of whether they share the same name), then its scope-related evidential power would be 0.25.

B. Geographical Entity Resolution

Given a text document and a location meaning evidence function, the detection and disambiguation of the text's locations is performed as follows. First, we extract from the text the set of terms T that match to some $i \in I$ along with a term-meaning mapping function $m : T \rightarrow I$ that returns for a given term $t \in T$ the instances it may refer to. We also consider I_{text} to be the superset of these instances.

Then, we consider the set of potential locations found within the text $L_{text} \subseteq I_{text}$ and for each $l \in L_{text}$ we derive all the instances from I_{text} that belong to some concept $c \in C$ for which $lem(c) \neq \emptyset$. Subsequently, by combining the location evidence model function $lmef$ with the term meaning function m we are able to derive a location-term meaning support function $sup_m : L_{text} \times T \rightarrow [0, 1]$ that returns for a location $l \in L_{text}$ and a term $t \in T$ the degree to which t supports l . If $l \in L_{text}$, $t \in T$ then

$$sup_m(l, t) = \frac{1}{|m(t)|} \cdot \sum_{i \in m(t)} lmef(l, i) \quad (3)$$

Using this function, we are able to calculate for a given term $t \in T$ in the text the confidence that it refers to location $l \in m(t)$:

$$c_{ref}(t, l) = \frac{\sum_{t_j \in T} K(l, t_j)}{\sum_{l' \in m(t)} \sum_{t_j \in T} K(l', t_j)} \cdot \sum_{t_j \in T} sup_m(l, t_j) \quad (4)$$

where $K(l, t) = 1$ if $sup_m(l, t) > 0$ and 0 otherwise.

In other words, the overall support score for a given candidate location is equal to the sum of the location's partial supports (i.e., function sup_m) weighted by the relative number of terms that support it. It should be noted that in the above process, we adopt the one referent per discourse approach, which assumes one and only one meaning for a location in a discourse.

C. Geographical Scope Resolution

The process of geographical scope resolution is similar to the entity resolution one, the difference being that we consider as candidate scope locations not only those found within the text but practically all those that are related to instances of the evidential concepts in the ontology. In that way, even if a location is not explicitly mentioned within the text, it still can be part of the latter's scope.

More specifically, given a text document and a geographical scope evidence function $gsef$ we first consider as candidate locations all those for which there is evidence within the text, that is all those for which $gsef(l, i) > 0$, $l \in L$, $i \in I_{text}$. We call this set L_{cand} . Then, for a given $l \in L_{cand}$ we compute the scope related support it receives from the terms found within the text as follows:

$$sup_s(l, t) = \frac{1}{|m(t)|} \cdot \sum_{i \in m(t)} gsef(l, i) \quad (5)$$

Finally, we compute the confidence that l belongs to the geographical scope of the text in the same way as Equation (4) but with sup_s substituting sup_m :

$$c_{scope}(l) = \frac{\sum_{t_j \in T} K(l, t_j)}{\sum_{l' \in L_{cand}} \sum_{t_j \in T} K(l', t_j)} \cdot \sum_{t_j \in T} sup_s(l, t_j) \quad (6)$$

where $K(l, t) = 1$ if $sup_s(l, t) > 0$ and 0 otherwise.

IV. SYSTEM IMPLEMENTATION AND USAGE

In this section, we provide details on the technical realization of KLocator and illustrate the way it is meant to be used.

A. System Architecture

The main components of KLocator's architecture, depicted in Figure 1, are the following.

- **Geographical Resolution User Interface:** This interface, depicted in Figure 2 allows users to define and manage their own geographical resolution evidence models and use them to geographically resolve texts.
- **Geographical Resolution Service:** This service layer implements and exposes the required functionality for performing the geographical entity and scope resolution tasks, as described in Section III.
- **Evidence Model Management Service:** This service layer implements and exposes the required functionality for defining, storing and editing geographical resolution evidence models.
- **Evidence Model Repository:** This repository stores all the created evidence models.
- **Evidence Model Manager** This is a low level API for retrieving and manipulating information from the Evidence Model repository.

- **Semantic Data Repository:** This repository stores all the domain and scenario-related ontologies and semantic data that are meant to be used by the system.
- **Semantic Data Manager:** This is a low level API for retrieving and manipulating information from the available ontologies and semantic data that are to be used by the Geographical Resolution Service. At the moment, it is designed to work with locally stored data but, in the future, it will be able to query directly the Linked Open Data Cloud [19].

B. System Usage

The definition and usage of geographical resolution evidence models is performed through the user interface of KLocator. The whole process comprises three steps:

- 1) The user (manually) defines the scenarios's location evidence mapping function by determining the location-related concepts whose instances may serve as contextual disambiguation and scope evidence within his/her scenario's texts.
- 2) The system automatically generates the functions *lmef* and *gsef* and stores them for future use.
- 3) The user is then able to apply the model to relevant texts and perform geographical entity and scope resolution.

The execution of the first step starts by pressing the "Create New Evidence Model" button to reveal the model creation form (Figure 3). Then, a name should be given for the new model (e.g., "Locations in Military Conflict Texts") and the table form to be filled with information like that of Table I. In doing that, the user first selects the target concept (e.g., "PopulatedPlace"), then the one to be used as evidence (e.g., "MilitaryConflict") and then the (automatically calculated) relation path between them that we want to consider.

When the model is complete the "Generate Model" button is used to store the model in the server and generate location-evidence entity pairs as in Table II. Depending on the size of the underlying ontology, the generation of these pairs can take a while but it is a process that will need to be performed only once.

When the generation process is finished, the new model appears as an option in the list of defined evidence models and can be used to perform location disambiguation and scope resolution. To do that one needs to select the model and then use the "Input Text" form and the "Perform Geographical Resolution" button to analyze texts relevant to the scenario the model has been defined for. Figure 4 shows the results of executing this process on an example text.

V. EXPERIMENTAL EVALUATION

To evaluate the effectiveness of KLocator, we used it to perform geographical entity and scope resolution on

historical texts describing military conflicts. In particular, we performed two experiments. In the first, we focused on correctly resolving ambiguous location references within the texts while in the second, on correctly determining the texts' geographical scope.

In both cases, we considered DBPedia as a source of semantic information, utilizing a subset of it comprising about 4120 military conflicts, 1660 military persons, 4270 locations and, of course, the relations between them (conflicts with locations, conflicts with persons, etc.). Using this semantic data, we defined the location evidence mapping function of Table I and we used it to automatically calculate the evidential functions *lmef* and *gsef* for all pairs of locations and evidential entities (other locations, conflicts and persons).

Table II shows a small sample of these pairs where, for example, James Montgomery acts as evidence for the disambiguation of Beaufort County, South Carolina because he has fought a battle there. Moreover, his evidential power for that location is 0.5, practically because he has fought a battle in another location called Beaufort County. Similarly, Pancho Villa acts as evidence for the consideration of Columbus, New Mexico as the scope of a text (because he has fought a battle there) and his evidential power for that is 0.2 since, according to the ontology, has fought battles in 4 other locations as well.

Table II
EXAMPLES OF LOCATION EVIDENTIAL ENTITIES

| Location | Evidential Entity | lmef | gsef |
|---------------------------------|-------------------|------|------|
| Columbus, Georgia | James H. Wilson | 1.0 | 0.17 |
| Columbus, New Mexico | Pancho Villa | 1.0 | 0.2 |
| Beaufort County, South Carolina | James Montgomery | 0.5 | 0.5 |

Using this model, we first applied our proposed geographic entity resolution process in a dataset of 150 short texts describing military conflicts like the following: "The Siege of Augusta was a significant battle of the American Revolution. Fought for control of Fort Cornwallis, a British fort near Augusta, the battle was a major victory for the Patriot forces of Lighthorse Harry Lee and a stunning reverse to the British and Loyalist forces in the South". The choice of this domain and scenario was driven from the fact that it has the key characteristics of the application scenarios our framework is designed for, namely predictability of text content and available background ontological knowledge.

The texts were manually compiled from web resources, including Wikipedia and other history-related pages. They were, in average, 2-4 sentences long, all contained ambiguous location entities but little other geographical information and, in average, each ambiguous location reference had 2.5 possible interpretations. For each such reference, 2 human

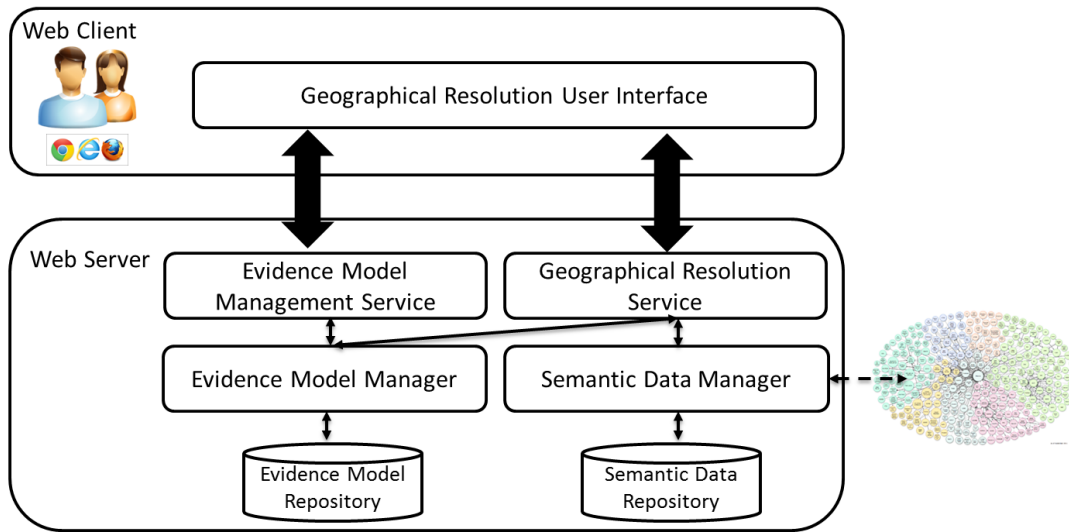


Figure 1. High Level Overview of KLocator Architecture

judges identified the correct location it referred to, with a very high inter-agreement of 0.9.

Then, we used KLocator to perform the same task automatically and we measured the precision and recall of the process. Precision was determined by the fraction of correctly interpreted locations (i.e., locations for which the interpretation with the highest confidence was the correct one) to the total number of interpreted locations (i.e., terms with at least one interpretation). Recall was determined by the fraction of correctly interpreted locations to the total number of annotated locations in the input texts. It should be noted that all target locations for disambiguation in the input texts were known to the knowledge base (i.e., DBPedia).

Table III shows results achieved by our approach compared to those achieved by some well-known publicly available semantic annotation and disambiguation services, namely DBPedia Spotlight [20], AIDA [21] [15], Wikimeta [22], Zemanta [23], AlchemyAPI [24] and Yahoo! [25]. As one can see, the consideration of non-geographical semantic information that our approach enables, manages to significantly improve the effectiveness of the geographical entity resolution task.

Of particular significance is the improvement achieved over DBPedia Spotlight and AIDA as these two systems i) also use DBPedia as a knowledge source and ii) they provide some basic mechanisms for constraining the types of entities to be disambiguated, though not in the same methodical way as our framework does. Practically, the two systems merely provide the users the capability to select the classes whose instances are to be included in the process.

In all cases, it should be made clear that the goal of this comparison was not to disprove the effectiveness and value of these systems as tools for open domain and unconstrained situations but rather to illustrate the importance of

customization and verify our claim that our approach is more appropriate for disambiguation in “controlled” scenarios, i.e., scenarios in which a priori knowledge about what entities and relations are expected to be present in the text is available. Of course, the availability of comprehensive background semantic knowledge about the domain is also an important effectiveness factor, but this is a requirement for any relevant system that follows a knowledge-based approach. A useful evaluation of popular semantic entity recognition systems for open scenarios may be found at [26].

Table III
GEOGRAPHICAL ENTITY RESOLUTION EVALUATION RESULTS

| System/Approach | Precision | Recall | F_1 Measure |
|-------------------|-----------|--------|---------------|
| Proposed Approach | 88% | 83% | 85% |
| DBPedia Spotlight | 71% | 69% | 70% |
| AIDA | 44% | 40% | 42% |
| Wikimeta | 33% | 30% | 31% |
| Zemanta | 26% | 30% | 28% |
| AlchemyAPI | 26% | 28% | 27% |
| Yahoo! | 24% | 26% | 25% |

As a second experiment, we applied our proposed geographic scope resolution process in two different datasets, all comprising 150 short military conflict related texts but with different characteristics. The first dataset comprised texts whose geographical scope was not explicitly mentioned within them and which contained little other geographical information. The second dataset comprised texts whose geographical scope related locations were explicitly and unambiguously mentioned within them but along with other geographical entities that were not part of this scope.

In both cases, we used again 2 human judges who decided the scope location of the texts, with an inter-agreement of 0.85. The we used KLocator to automatically determined for each text the possible locations that comprised its

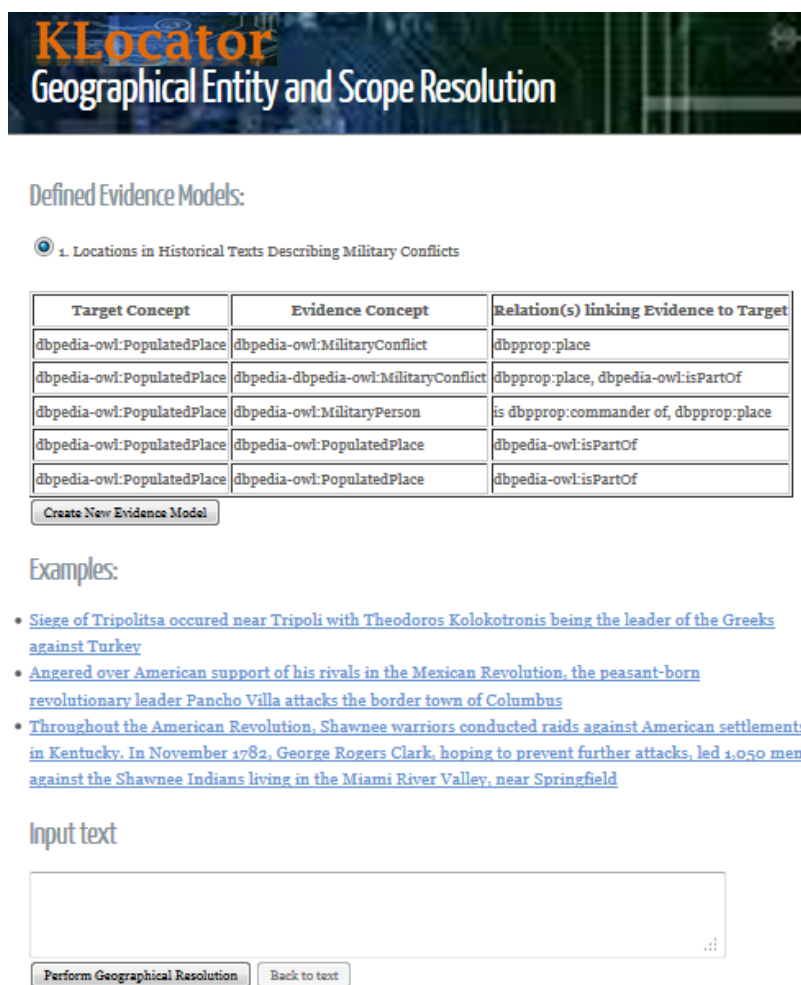


Figure 2. KLocator User Interface

geographical scope and ranked them using the confidence score derived from Equation (6). We then measured the effectiveness of the process by determining the number of correctly scope resolved texts, namely texts whose highest ranked scope locations were the correct ones. As a baseline, we compared our results to the ones derived from Yahoo! Placemaker [27] geoparsing web service.

The results of the above process are shown in Table IV. As one can see, the improvement our method achieves in the effectiveness of the scope resolution task is quite significant in both datasets and especially in the first one where the scope-related locations are not explicitly mentioned within the texts. This verifies the central idea of our approach that non-geographical semantic information can significantly improve the geographical scope resolution process and in particular the subtasks of:

- 1) Inferring relevant to the text's geographical scope locations even in the absence of explicit reference of them within the text (first dataset).

- 2) Distinguishing between relevant and non-relevant to the text's geographical scope locations, even in the presence of non-relevant location references within the text (second dataset).

Table IV
GEOGRAPHICAL SCOPE RESOLUTION EVALUATION RESULTS

| System/Approach | Dataset 1 | Dataset 2 |
|-------------------|-----------|-----------|
| Proposed Approach | 70% | 85% |
| Yahoo! Placemaker | 18% | 30% |

VI. DISCUSSION

It should have been made clear from the previous that our KLocator is not independent of the content or domain of the input texts but rather adaptable to them. That is exactly its main differentiating feature as our purpose was not to build another generic geographical resolution system but rather a reusable framework that can i) be relatively

New Evidence Model Creation

Evidence Model Name:

| Target Concept | Evidence Concept | Relation(s) linking Evidence to Target | |
|---|---|--|---|
| http://dbpedia.org/ontology/PopulatedPlace | http://dbpedia.org/ontology/MilitaryConflict | http://dbpedia.org/ontology/place | |
| http://dbpedia.org/ontology/PopulatedPlace | http://dbpedia.org/ontology/MilitaryConflict | http://dbpedia.org/ontology/place , http://dbpedia.org/ontology/isPartOf | <input type="button" value="Delete row"/> |
| http://dbpedia.org/ontology/PopulatedPlace | http://dbpedia.org/ontology/MilitaryPerson | http://dbpedia.org/ontology/commander (inverse), http://dbpedia.org/ontology/place | <input type="button" value="Delete row"/> |
| http://dbpedia.org/ontology/PopulatedPlace | http://dbpedia.org/ontology/PopulatedPlace | http://dbpedia.org/ontology/isPartOf | <input type="button" value="Delete row"/> |

Figure 3. New Evidence Model Creation Form

Examples:

- [Siege of Tripolitsa occurred near Tripoli with Theodoros Kolokotronis being the leader of the Greeks against Turkey](#)
- [Angered over American support of his rivals in the Mexican Revolution, the peasant-born revolutionary leader Pancho Villa attacks the border town of Columbus](#)
- [Throughout the American Revolution, British forces burned down settlements in Kentucky. In November 1782, George Rogers Clark, hoping to prevent further attacks, burned the town of Maysville in the Miami River Valley, near Springfield](#)

Input text

Siege of Tripolitsa occurred near [Tripoli](#) with Theodoros Kolokotronis being the leader of the Greeks against [Turkey](#)

Figure 4. Semantic Entity Resolution Example

easily adapted to the particular characteristics of the domain and application scenario at hand and ii) exploit these characteristics in order to increase the effectiveness of the disambiguation process. Our motivation for that was that, as the comparative evaluation showed, the scenario adaptation capabilities of existing generic systems can be inadequate in certain scenarios (like the ones described in this paper), thus limiting their applicability and effectiveness.

In that sense, our proposed framework is not meant as a substitute or rival of other geographical resolution approaches (that operate in open domains, use geographical information and relevant heuristics and apply machine learning and statistical methods) but rather as a complement of them in application scenarios where text domain and content are a priori known and comprehensive domain ontological knowledge is available (as in the case of historical texts used in our experiments).

Of course, the usability and effectiveness of our approach is directly proportional to the content specificity of the texts to be disambiguated and the availability of a priori knowledge about their content. The greater these two parameters are, the more applicable is our approach and the more effective the disambiguation is expected to be. The opposite is true as the texts become more generic and the information we have out about them more scarce. A method that could

a priori assess how suitable is our framework for a given scenario would be useful, but it falls outside the scope of this paper.

Also, the framework's approach is not completely automatic as it requires some knowledge engineer or domain expert to manually define the scenario's geographical resolution evidence mapping function. Nevertheless, this function is defined at the schema level thus making the number of required mappings for most scenarios rather small and manageable.

As far as the scalability of our approach is concerned, the main computational burden of the process is the building of the evidence index which takes place offline. In our experiments with the history knowledge base, the index building took about 2 minutes, in a standard server. On the other hand, the online location identification process took 1-2 seconds, depending of course on the size of the text. More generally, although we have not yet formally evaluated scalability, the fact that our framework is based on the constraining of the semantic data to be used makes us expect that it will perform faster than traditional approaches that use the whole amount of data. Furthermore, as the resolution evidence model is constructed offline and stored in some index, the most probable bottleneck of the process will be the phase of determining the candidate entities for

the extracted terms rather than the resolution process.

Finally, the typical errors our system is prone to, are related to two steps of the process, namely text entity detection and entity and scope resolution. In the text entity detection step, it can be the case that tokens are matched to wrong entities for reasons having to do with the linguistic analysis subsystem and/or the quality of the semantic data (entity coverage labeling, etc). On the other hand, entity and scope resolution typically fails when available evidence in the text is either too poor or too ambiguous.

VII. CONCLUSION

In this paper, we proposed KLocator, a novel framework for optimizing geographical entity and scope resolution in texts by means of domain and application scenario specific non-geographical semantic information. First, we described how, given a priori knowledge about the domain(s) and expected content of the texts that are to be analyzed, one can define a model that defines which and to what extent semantic entities (especially non-geographical ones) can be used as contextual evidence indicating two things:

- Which is the most probable meaning of an ambiguous location reference within a text (geographical entity resolution task).
- Which locations constitute the geographical scope of the whole text (geographical scope resolution task).

Then, we described how such a model can be used for the two tasks of geographical entity and scope resolution by providing corresponding processes. The effectiveness of these processes was experimentally evaluated in a comprehensive and comparative to other systems way. The evaluation results verified the ability of our framework to significantly improve the effectiveness of the two resolution tasks by exploiting non-geographical semantic information.

Given the semi-automatic nature of our framework and its dependence on the availability of comprehensive semantic data, future work will focus on investigating how statistical and machine learning approaches may be used, in conjunction with our approach, in order to i) automatically build geographical resolution evidence models based on text corpora and ii) deal with cases where available domain semantic information is incomplete.

ACKNOWLEDGMENT

This work was supported by the European Commission under contracts FP7- 248984 GLOCAL and FP7-287615 PARLANCE.

REFERENCES

- [1] P. Alexopoulos, C. Ruiz, and J. M. Gomez-Perez, "Optimizing geographical entity and scope resolution in texts using non-geographical semantic information," in *Proceedings of the Sixth International Conference on Advances in Semantic Processing (SEMAPRO)*, 2012, pp. 65–70.
- [2] C. B. Jones and R. S. Purves, "Geographical information retrieval," *International Journal of Geographical Information Science*, vol. 22, no. 3, pp. 219–228, Jan. 2008.
- [3] J. Raper, G. Gartner, H. Karimi, and C. Rizos, "Applications of location-based services: a selected review," *J. Locat. Based Serv.*, vol. 1, no. 2, pp. 89–111, Jun. 2007.
- [4] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," *Int. J. Semantic Web Inf. Syst.*, vol. 5, no. 3, p. 122, 2009.
- [5] L. M. V. Blázquez, B. Villazón-Terrazas, V. Saquicela, A. de León, Ó. Corcho, and A. Gómez-Pérez, "Geolinked data and inspire through an application case," in *GIS*, 2010, pp. 446–449.
- [6] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, "Linked-geodata: A core for a web of spatial open data," *Semantic Web Journal*, vol. 3, no. 4, pp. 333–354, 2012.
- [7] G. Andogah, G. Bouma, J. Nerbonne, and E. Koster, "Place-name ambiguity resolution," in *Methodologies and Resources for Processing Spatial Language (Workshop at LREC 2008)*, 2008.
- [8] G. Andogah, G. Bouma, and J. Nerbonne, "Every document has a geographical scope," *Data and Knowledge Engineering*, 2012.
- [9] T. Kauppinen, R. Henriksson, R. Sinkkilä, R. Lindroos, J. Vtinen, and E. Hyvnen, "Ontology-based disambiguation of spatiotemporal locations," in *Proceedings of the 1st international workshop on Identity and Reference on the Semantic Web (IRSW2008)*, 5th European Semantic Web Conference 2008 (ESWC 2008). Tenerife, Spain: CEUR Workshop Proceedings, ISSN 1613-0073, June 1-5 2008.
- [10] Y. Peng, D. He, and M. Mao, "Geographic named entity disambiguation with automatic profile generation," in *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, ser. WI '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 522–525.
- [11] J. Kleb and A. Abecker, "Entity reference resolution via spreading activation on rdf-graphs," in *Proceedings of the 7th international conference on The Semantic Web: research and Applications - Volume Part I*, ser. ESWC'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 152–166.
- [12] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: a nucleus for a web of open data," in *Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference*, ser. ISWC'07/ASWC'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 722–735.
- [13] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge," in *16th international World Wide Web conference (WWW 2007)*. New York, NY, USA: ACM Press, 2007.
- [14] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "Dbpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th International Conference on Semantic Systems*, ser. I-Semantics '11. New York, NY, USA: ACM, 2011, pp. 1–8.

- [15] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum, "Robust disambiguation of named entities in text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 782–792.
- [16] M. Janik and K. Kochut, "Wikipedia in action: Ontological knowledge in text categorization," in *ICSC*. IEEE Computer Society, 2008, pp. 268–275.
- [17] M. Wallace, P. Mylonas, G. Akrivas, Y. Avrithis, and S. Kollias, *Automatic thematic categorization of multimedia documents using ontological information and fuzzy algebra*. Studies in Fuzziness and Soft Computing, Soft Computing in Ontologies and Semantic Web, Springer, Ma, Z. (Ed.), Vol. 204, 2006.
- [18] V. T. Chakaravarthy, H. Gupta, P. Roy, and M. K. Mohania, "Efficiently linking text documents with relevant structured information," in *VLDB*, U. Dayal, K.-Y. Whang, D. B. Lomet, G. Alonso, G. M. Lohman, M. L. Kersten, S. K. Cha, and Y.-K. Kim, Eds. ACM, 2006, pp. 667–678.
- [19] "Linked open data cloud," <http://www.lod-cloud.net/>, accessed: 15/12/2013.
- [20] "Dbpedia spotlight," <http://spotlight.dbpedia.org/>, accessed: 15/12/2013.
- [21] "Aida," <https://gate.d5.mpi-inf.mpg.de/webaida/>, accessed: 15/12/2013.
- [22] "Wikimeta," <http://www.wikimeta.com>, accessed: 15/12/2013.
- [23] "Zemanta," <http://www.zemanta.com>, accessed: 13/07/2013.
- [24] "Alchemy api," <http://www.alchemyapi.com>, accessed: 15/12/2013.
- [25] "Yahoo!" <http://developer.yahoo.com/search/content/V2/contentAnalysis.html>, accessed: 15/12/2013.
- [26] G. Rizzo and R. Troncy, "NERD: A framework for evaluating named entity recognition tools in the Web of data," in *ISWC 2011, 10th International Semantic Web Conference, October 23-27, Bonn, Germany, 2011*.
- [27] "Yahoo! placemaker," <http://developer.yahoo.com/boss/geo/>, accessed: 15/12/2013.

Three Principles for the Design of Energy Feedback Visualizations

Robert S. Brewer, Yongwen Xu, George E. Lee, Michelle Katchuck, Carleton A. Moore, and Philip M. Johnson

Department of Information and Computer Sciences

University of Hawai'i at Mānoa

Honolulu, HI, USA

{rbrewer, yxu, gelee, katchuck, cmoore, johnson}@hawaii.edu

Abstract—To achieve the full benefits of the Smart Grid, end users must become active participants in the energy ecosystem. This paper presents the Kukui Cup challenge, a multifaceted serious game designed around the topic of energy conservation that incorporates a variety of energy feedback visualizations, online educational activities, and real-world activities such as workshops and excursions. We describe our experiences developing energy feedback visualizations in the Kukui Cup based on in-lab evaluations and field studies in college residence halls. We learned that energy feedback systems should address these three factors: 1) they should be actionable, 2) domain knowledge should go hand in hand with feedback systems, and 3) feedback must be “sticky” if it is to lead to changes in behaviors and attitudes. We provide examples of both successful and unsuccessful visualizations, and discuss how they address the three factors we have identified.

Keywords—Visualization, serious games, energy feedback, energy, energy literacy, smart grid.

I. INTRODUCTION

The development of the Smart Grid and the two-way communication that it provides, have enabled a variety of new customer-facing possibilities including real-time feedback on electricity usage, real-time pricing, and demand response. However, to make full use of this potential, end-users of the Smart Grid will need to be engaged about their electricity use, and become more energy literate. We believe that in addition to a Smart Grid, we need Smart Consumers. One common theme among customer-related aspects of the Smart Grid is the development of energy feedback visualizations [1].

In this context, we have developed the Kukui Cup Challenge, a serious game [2] (a game with additional goals beyond just entertainment), designed around the topic of energy. The Kukui Cup includes a variety of energy feedback visualizations [3] designed to inform and engage the players about their energy use. The Kukui Cup also includes a multifaceted online game with educational activities, and real-world activities such as workshops and excursions [4], [5].

The Kukui Cup is designed to provide players with insight into how their behaviors affect energy consumption and production. Such behaviors occur on a spectrum, from the short-term, immediate impact behaviors such as turning off lights, to the longer-term, collective impact of behaviors such as

considering the energy policies of political candidates when deciding how to vote. Creating a challenge that helps players understand energy from this wide scope sets the Kukui Cup apart from other similar “energy game” initiatives. It also impacts on our understanding of effective feedback for smart grid customer-facing applications.

Our work is also influenced by where we live. Even within the United States of America, the State of Hawai'i faces a number of unique challenges in the pursuit of sustainability for its citizens, compared to other states. Hawai'i has fertile agricultural land, and a variety of renewable energy sources (wind, solar, geothermal, wave), but it imports 85% of its food, and over 90% of its energy, in the form of oil and coal. In fact, Hawai'i is the most fossil fuel-dependent state in the United States. The Kukui Cup is designed in the context of these challenges, and as a remote archipelago, the issues are felt more keenly here than in many other parts of the world. While we believe the Kukui Cup can be a useful tool for addressing global energy challenges, the world will also need to dramatically increase the amount of energy coming from renewable sources, and invest in grid infrastructure.

Based on our experiences designing and evaluating energy feedback in the Kukui Cup with students living in residence halls, we have three recommendations for designing energy feedback systems for smart grid consumers:

- 1) they should be actionable,
- 2) domain knowledge should go hand in hand with feedback systems, and
- 3) feedback must be “sticky” if it is to lead to changes in behaviors and attitudes.

This paper explores how we came to these conclusions, and what evidence we have collected that supports these conclusions.

We first describe the Kukui Cup system, followed by an explanation of how energy goals and baselines are used in the Kukui Cup. With that foundation, we discuss our results from developing and deploying the Kukui Cup in the field over two years in the areas of designing energy feedback visualizations, the importance of energy literacy in understanding energy feedback, and our use of a serious game to encourage users to engage with the energy feedback information. Finally, we end with sections describing our plans for future work and our conclusions.

II. THE KUKUI CUP

College residence hall energy competitions have become a widespread mechanism for engaging students in energy issues, with more than 160 taking place or being planned for the 2010–2011 academic year in North America [6]. Residence hall energy competitions are events where residence halls, or floors within a residence hall, compete to see which team will use the least energy over a period of time. The competitions tap into both the residents' competitive urges, and their interest in environmental issues. However, unlike home residents, the dormitory residents typically do not financially benefit from any reduction in electricity use resulting from their behavior changes. There is no discount on their room and board charges even if consumption is reduced, because the residence hall fees are flat-rate. Since they lack even a monthly bill as feedback, residents are largely unaware of their energy usage.

Residence hall energy competition technologies range in complexity from simple web pages with weekly electricity data to complicated web applications [7, pp. 6–11]. An early adopter of the residence hall energy competition, Oberlin College, developed a real-time electricity consumption feedback system as described by Petersen et al. [8]

To build on this area of active energy work, we decided to target our serious game to college students living in residence halls. The Kukui Cup extends the typical college energy competition into a broader energy *challenge* where electricity consumption feedback is only one part of a larger game experience for players. The challenge is named after the kukui nut (also known as candlenut), which was burned by Native Hawaiians to provide light, making it an early form of stored energy in Hawai'i.

In a Kukui Cup challenge, residents are grouped into teams based on where they live. Different floors of a building or entire buildings can be formed into teams. The electricity usage of teams is measured either through manual meter readings or through automated meter data collection. In addition to the energy competition, the Kukui Cup has an energy literacy competition where players can earn points by engaging in educational and social activities on the challenge website. The point system provides a way to motivate players to explore and use the system, as the Kukui Cup is currently deployed as an extracurricular activity.

Much of the point competition revolves around a section of the challenge website called the Smart Grid Game (SGG). The Smart Grid Game consists of rows of actions arranged into columns based on a particular topic (similar to the popular game show "Jeopardy"), shown in Fig. 1. Clicking on a square in the SGG shows details about the action and explains how players can complete the action to earn points. There are several types of actions: short YouTube videos on energy and sustainability topics, activities like measuring the flow rate of a shower, excursions such as visiting a farm

| Level 1 | Level 2 | Level 3 | Level 4 | Level 5 | Level 6 | Level 7 |
|---------------|--------------------|----------------|---------------|-----------------|---------|---------|
| Electrimental | Let's Get Physical | Power To Burn? | Watts Up? | Water Cycle | | |
| Audit Video | Check energy | Energy Issues | Pull the plug | Sink Flow | | |
| 40 | Computer Sleep | Energy Now | Go meatless | Turn off sink | | |
| 🔒 | 20 | HCEI | 100 | Shower flow | | |
| 🔒 | 50 | Lighting video | Write Poem | Shorter showers | | |

Legend:
■ activity ■ commitment ■ event ■ excursion ■ filler

Figure 1. The Smart Grid Game widget, displaying level 2 actions.

that produces all its own electricity, and commitments such as public declarations of the intent to carpool or not eat meat. There are also creative actions such as writing a poem about energy or community engagement actions such as writing a letter to the editor on a sustainability topic. The flexibility of the SGG allows us to provide a wide variety of interesting actions for players to take part in.

The completion of each action (with the exception of commitments) is verified through the challenge website before points are awarded. For activities, players are usually asked a randomly-selected question, and their answer is placed in a queue for challenge administrators to review. The administrator can approve or reject the submission, award points, and provide feedback on the players' answers. The game also supports activities that are verified by submission of an uploaded image such as a photo or screenshot.

To encourage players to interact with each other during the challenge, the Kukui Cup provides a number of social features. The *social bonus* provides a way for players to earn additional points by completing certain actions with other players. To earn the social bonus, players include the email address of another player when they complete the verification step for an action. If the submitted email address corresponds to a player that has completed that same action, then the submitting player is awarded a small, configurable number of points. In addition to incentivizing joint play, the social bonus can provide a pretext for a player to initiate contact with another player, such as arranging to attend a workshop. In the college residence hall setting, it can be helpful to have such a pretext for meeting new friends. Social media (in particular Facebook) is integrated into the Kukui Cup as well. Players can share their game accomplishments directly on Facebook, and the Kukui Cup Facebook page is used to share information about the challenge, including upcoming events and short videos of events that have taken

place.

As part of the initial log on process to the Kukui Cup website, new players can enter in the email address of a referring player to earn a *referral bonus*. To ensure that the new player starts actually playing the game, they must earn a certain number of points before both the referred and referring player receive the referral bonus points. In the 2012 Kukui Cups, the referral bonus was variable based on the degree of the new player's team participation. Therefore, both new and referring players received more points if the new player was from a team with few participants. The variable referral bonus encourages players to reach out to teams that were not yet involved in the Kukui Cup and gave a small boost to those new players.

Kukui Cup challenges can also be configured to provide incentives to players in the form of prizes for the top scores both at the individual (points) and team level (points and energy use). One possible downside with the prizes provided in the competition as incentives is that they only go to the top performers in each competition. For those participants so far behind the leaders that they know they will not win the point competitions, the prizes provide little incentive, or possibly a disincentive: why play if there is no way to win a prize? Another challenge with the prizes is that to be effective, they had to appeal to all participants, which limits the options for prizes.

To address the two problems with prizes, we developed the Raffle Game, inspired by Prabhakar's work incentivizing road congestion reduction [9]. In the Raffle Game, there are a variety of raffle prizes available in each round of the competition. For each 25 points a participant earns in the challenge, they receive a virtual raffle ticket. Participants can allocate their raffle tickets among the prizes available, and they can change their allocations at any time until the end of the round. Then a winning ticket is "drawn" from those allocated to each raffle prize, and the owner of that ticket wins the prize. Since the winner of each prize is determined randomly among the allocated tickets, even a player who has allocated a single ticket can win a prize, thereby providing less active players with the potential to win prizes. However, because tickets are earned by earning points through playing the game, the more participants play the game, the better their chances of winning raffle prizes. We obtained many raffle prizes through donations by local businesses, which resulted in a wide assortment of types of prizes, including clothing, gift certificates to restaurants, coupons for outdoor activities, not all of which were of universal appeal to participants. The raffle format is especially advantageous when there are a variety of prizes because participants are able to pick and choose which prizes they are interested in. Further, there is opportunity to engage more people with different interests as a result.

A. Running a Kukui Cup

A Kukui Cup challenge consists of multiple components working together to provide the entire game experience. For challenges using real-time energy data, the open source WattDepot [10] system is used to collect, store, and analyze the data. The challenge website and associated game mechanics are provided by the open source Makahiki system [7], [11]. The current educational content is tailored to the needs of college students living in residence halls in Hawai'i, but can be tailored to suit other audiences or goals.

The final component in a Kukui Cup challenge is the administrators. Challenge administrators need to plan out the parameters of the competition (duration, number of teams), make game design choices such as point rubrics, customize the educational content for their organization, organize workshops, review player verification submissions, and distribute prizes. Kukui Cup challenges are labor intensive, but that labor provides the opportunity to interact with the players more fully and provide them with an expansive game experience.

B. Field Studies

In addition to in-lab evaluations and beta tests, there have been two sets of field studies of Kukui Cup challenges. The first Kukui Cup challenge took place over 3 weeks starting in October 2011 in four residence halls for first-year students on the University of Hawai'i at Mānoa campus containing a total of approximately 1070 residents. Pairs of floors, referred to as *lounges*, were the team unit in the 2011 Kukui Cup.

The second set of challenges started in September 2012. The University of Hawai'i (UH) 2012 Kukui Cup took place in the same four residence halls with approximately the same number of residents, but over the entire nine month academic year. The first month of the competition was an intensive period with multiple real-world events taking place each week, while the remaining months were less intensive. The goal of the much longer time frame is to discourage short-term and unsustainable behaviors (such as forgoing all electronic device use).

In addition to the 2012 UH Kukui Cup, two other educational institutions within the State of Hawaii, Hawaii Pacific University (approximately 200 residents) and the East-West Center (approximately 130 residents), have run their own challenges using the Kukui Cup system with our support.

III. BASELINES AND GOALS

Goal setting has been shown to be an effective tool in changing energy consumption behavior [12], [13] and are a common component of energy feedback mechanisms. Setting achievable goals is important from a game play perspective, so goals must typically be based on previous energy use. The most common way to generate a goal is to calculate a *baseline* of energy usage based on past energy

usage, and then set the goal as some percentage reduction from the baseline.

Two of the most common ways to calculate the electricity baseline are to average recent prior usage (such as the last two weeks), or to average usage from previous years. Both of these methods are problematic because they assume that the previous usage is representative of future usage, even though there are many factors that can significantly alter electricity use over time including: occupancy, weather, activities (e.g., studying for a big midterm exam), and changes to the building infrastructure such as efficiency upgrades. Any of these factors can lead to the baseline being an inaccurate predictor of future usage in an energy competition, as described by Johnson et al. [14].

Because baselines can be poor predictors of future electricity use, comparing actual electricity use to the baseline in order to determine how much electricity was “saved” by an intervention is misleading and can tempt designers to make claims about energy saved that cannot be substantiated. However, comparison of actual electricity usage to a goal generated from a baseline can be helpful as a game mechanic to motivate players to conserve energy.

In the 2011 Kukui Cup, we used a baseline that was derived from an average of the two weeks prior to the challenge. In the 2012 Kukui Cups, we switched to a dynamic baseline [14] that consists of the average electricity usage for the two previous weeks, but the baseline is recomputed every day throughout the challenge. The dynamic baseline means that as the challenge progresses, the baseline will include usage during the challenge. In essence, a goal generated from a dynamic baseline requires a team to reduce their energy usage compared to the recent past. Because the baseline is not a static value picked once before the challenge, anomalous conditions during the period before the challenge will soon be replaced with new, more representative data.

We incorporate the energy goal into an energy feedback game called the Daily Energy Goal Game (DEGG). Each team has a daily energy goal determined by the baseline. When a team’s energy usage at the end of the day is equal to or below the goal, they win the DEGG for that day. In the 2012 Kukui Cups, the energy competition is scored by the number of daily energy goals that each team has achieved, rather than an absolute measure of how much energy has been used or how much a team has reduced their usage below the baseline. By counting goals rather than measures of absolute or relative energy reductions, we hope to incentivize sustainable longer-term behavior changes rather than radical short-term changes such as moving out of the residence hall and into tents (as has been reported anecdotally in some other energy competitions). In an effort to link the energy and point competitions, when a team meets their daily energy goal, each team member receives a configurable number of points.

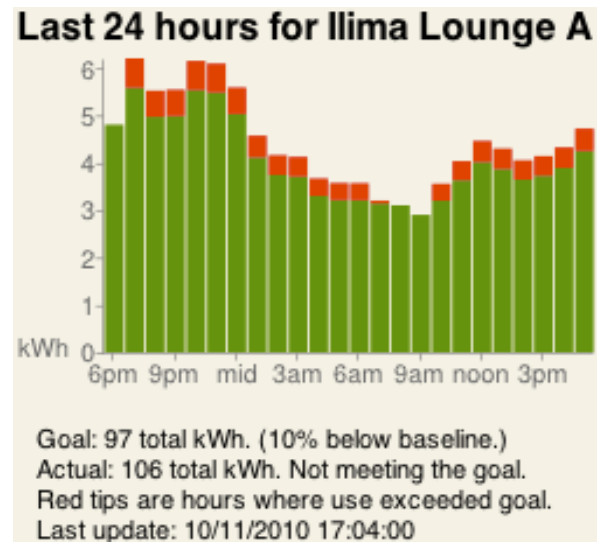


Figure 2. A bar chart visualization of energy use as compared to a goal.

IV. ENERGY FEEDBACK DESIGN EXPERIENCES

Feedback on electricity consumption has been used as a means for facilitating energy conservation by researchers in the human-computer interaction community [3] as well as in the broader energy efficiency [15]–[17] and environmental psychology [12], [13] communities. One reason for this focus on feedback is undoubtedly the hidden nature of electricity, so feedback provides an awareness that is otherwise unavailable.

There are three main decisions to be considered when designing an energy feedback visualization:

- 1) what type of data is to be displayed (power, energy, etc)?
- 2) how should the data be displayed to users?
- 3) on what time scale should the data be displayed?

Decisions on each of these factors will influence the success or failure of the visualization.

One of the fundamental principles of energy feedback in the Kukui Cup is that it be *actionable*. While any energy feedback may implicitly encourage energy conservation behaviors simply by making energy use visible, this level of feedback does not meet our definition of actionable. A feedback display that shows that a home has used 20 kWh so far on a particular day leaves the viewer with natural questions: is that a lot? What should I do if I wanted to reduce my energy use?

A. The Energy Bar Chart Visualization

An early attempt at energy feedback for the Kukui Cup is shown in Fig. 2. This “Energy Bar Chart” shows hourly energy use for a team participating in the Kukui Cup over 24 hours as compared to an energy goal. Note that the data shown in this particular figure are simulated. Bars that are

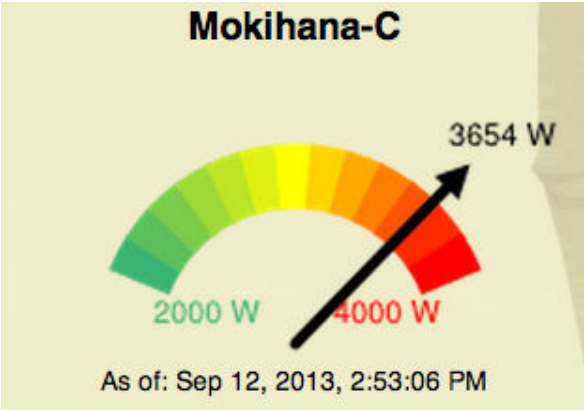


Figure 3. The Power Meter energy feedback visualization.

entirely green show the actual energy usage for that hour and indicate that the energy use was below the hourly goal. For mixed red and green bars, the main green portion represents the energy goal for that hour of the day, while the red tips of the bars represent the actual usage in excess of the goal.

This form of energy feedback shows the variation in energy use over the course of a day, which is an important energy literacy concept. It also shows in what parts of the day energy use is exceeding the goal, and by how much. By displaying the times during the day when the hourly goals are not being met, residents could focus on understanding what activities are going on during those periods.

As (naive) designers, we felt that this visualization provided a great deal of useful feedback both clearly and concisely. However, results of an in-lab evaluation were unequivocal: the visualization provided too much information, the meaning of its components was not obvious, and the “actionable” aspects were not obvious. This energy feedback visualization was a failure, and we began a redesign to address its deficiencies.

B. The Power Meter Energy Feedback Visualization

Another early energy feedback visualization developed for the Kukui Cup was the Power Meter, shown in Fig. 3. The Power Meter is an example of one of the most common types of energy feedback visualizations: a near real-time representation of how much power is being used. In the UH Kukui Cups, the Power Meter reflects the power use of a team consisting of 54 residents. The meter is calibrated so that the middle reading (needle pointing straight up) corresponds to the average team power use computed from the hourly baseline data, and the range of the meter dial is set to account for the largest swings in power use based on historical data. This calibration means that if the needle is on the right side of the dial it reflects more team power use than “normal” for this hour and day, and conversely if the needle is on the left side, it corresponds to less power usage

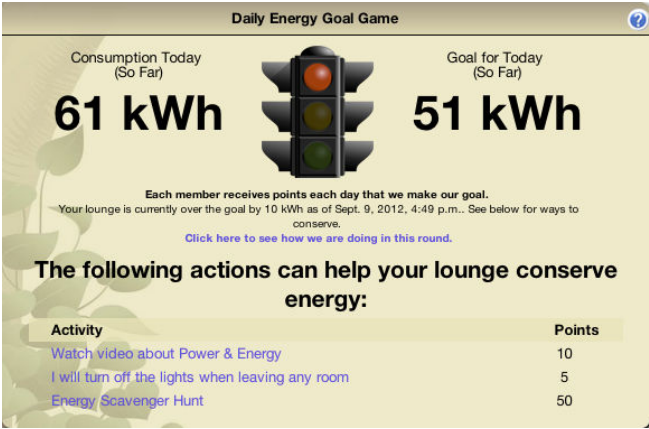


Figure 4. The Daily Energy Goal Game feedback visualization.

than normal. The calibration is updated each hour to match the new hourly baseline.

The motivation for the Power Meter is to enable players to explore their power use in real time by turning devices in their room and floor on and off to see how it impacted their power use, a classic use of high-frequency feedback. However, we have no evidence that players actually used the Power Meter visualization in this way. One difference between the Power Meter as used in the UH Kukui Cup challenges and similar energy feedback devices in a home setting is that there are many more people living in the metered space in the UH Kukui Cups. The large number of people whose electricity use is being metered could make it more difficult to see changes due to an individual device, unless that device uses a large amount of power (such as a microwave or hair dryer). As a result, we do not currently know if the Power Meter is successful as a visualization in our initial settings for the Kukui Cup. However, we have left the Power Meter as part of all Kukui Cups using real-time meter data as a counterpoint to the energy-based Daily Energy Goal Game visualization, in the hope of reinforcing the difference between power and energy. Understanding the different between power and energy was one of the energy literacy topics some participants seemed to have trouble grasping, so we felt any additional reinforcement might be helpful.

C. The Daily Energy Goal Game Feedback Visualization

To make our energy feedback easier to understand and also more actionable, we developed the Daily Energy Goal Game (DEGG) visualization shown in Fig. 4. The three most prominent components of the DEGG are:

- the energy consumption so far during the current day,
- the energy goal so far for the current day, and
- a traffic light that shows in the most straightforward way whether the team is meeting their energy goal.

The display updates once every 10 minutes with new energy data.

We picked the daily time frame for the game for two reasons. First, having a daily goal makes behavior changes more visible and feedback more immediate than a longer time frame such as weekly or monthly. Second, by concentrating on a daily goal, teams that are performing poorly on a particular day can redouble their efforts to do better the next day. Similarly, a team that does particularly well for one day cannot rest on their laurels, as they must make an effort to conserve every day. This game design reflects our belief that changing energy behaviors is a marathon and not a sprint: radical short-term changes made to win an energy competition are unlikely to be sustainable, and therefore are of very limited utility in achieving long-term energy conservation.

Residential energy use varies in intensity over the course of a day: typically low when people are sleeping and much higher during evening hours. For the students in our studies living in residence halls, the energy usage peak occurs at approximately midnight, and the lowest usage is between 8 and 9 AM, which is considerably different than an average single-family home. There is also daily variation between days of the week, as the activities taking place on a Monday night are different than those on a Saturday night. To account for the hourly and daily variation in energy use, we compute hourly and daily baselines for energy use, and the goal value is a percentage reduction from the baseline. The energy consumption and goal values displayed in the DEGG are computed over the time period from midnight to the current time. This choice of time frame is particularly important for the goal value, because if a daily goal value were simply spread evenly over the course of a day, players would see their energy use as always under the goal during low-usage periods, and going above the goal during the high-usage periods, possibly to a degree that makes it impossible to meet the goal for that day.

The DEGG also links the energy conservation competition with the point competition. When a team meets their daily energy goal, each team member is awarded an administrator-configured number of points. This linkage provides an additional incentive for players to pay attention to the energy competition, because successfully reducing energy use below the goal can significantly increase team point totals.

Below the traffic light display of the DEGG is a list of actions from the Smart Grid Game that players can take to either learn more about energy, or directly reduce their energy use. The actions displayed depend on what actions the player has already completed in the rest of the system. The DEGG is therefore highly actionable because it provides direct links to actions that players can take to reduce their energy usage, tailored to the opportunities available in their residence hall.

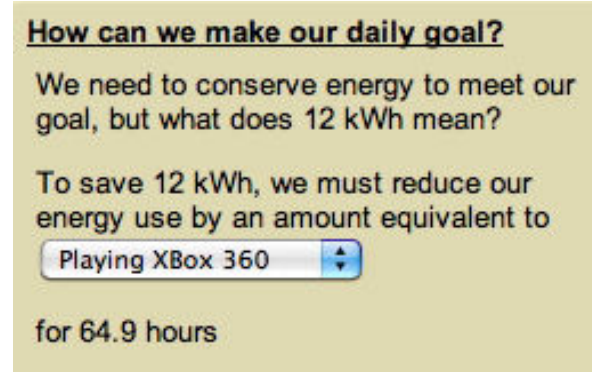


Figure 5. The “How can we make our daily goal?” widget.

Evaluation of the DEGG visualization during actual game play indicated that players do not have a problem understanding this visualization. The stoplight image provides a clear, unambiguous signal, and the actual/goal numbers provide further context. In addition, the visualization is explicitly paired with links to descriptions of appropriate actions for that player in the context of the game and the team’s current energy use. Log data indicated that players did click on these links in order to understand how to take action based on the energy feedback. This energy feedback visualization is thus a success and has been included in the current version of the Kukui Cup.

D. The “Wii Hours” Energy Feedback Visualization

In another energy feedback design effort, we created a small widget below the DEGG titled “How can we make our daily goal?”, which was intended to give players ideas on how to improve their team’s energy performance. This widget, shown in Fig. 5, shows how much the player’s team energy usage was above the goal, and provides a drop-down menu of electrical devices commonly present in student rooms: laptops, Xbox 360, Wii, etc. When a device is selected from the menu, the system displays the approximate number of hours of device use that would equal the amount of team energy use over the goal value. The time value is intended to show players how much device use they would need to *forego* in order to get back on track to their energy goal, and develop their intuition about the relative power use of different devices (i.e., plasma TVs use much more power than Wii game consoles). Therefore, a short time value could point out an easy way to make the goal, and a long time value would indicate less significant energy conservation.

However, during in-lab evaluations of the system, we found that multiple subjects misinterpreted the time value, thinking that high time values were bad rather than good. Since the Wii was the device on the list with the smallest power use (20 W) compared to an Xbox 360 or Playstation 3, it led to the highest time values. Some subjects drew the conclusion that using a Wii was worse than using an Xbox

360 or Playstation 3, which was precisely the opposite goal of this widget. One subject even took the time to use our in-game team discussion forum to post the message “don’t play wii” after using the widget! Because of this example, we dubbed this confusion the “Wii problem.”

Clearly, energy feedback that can lead some players to the opposite conclusion than intended is a failure. The “Wii Hours” visualization never made it into production, and we are still searching for a design variant that can convey this information in an unambiguous fashion to players with minimal energy literacy.

E. Canopy Energy Feedback

The 2011 Kukui Cup took place over three weeks, and as part of the game experience we wanted to create a more advanced level of experience for the top participants in the competition. The background of the competition website features a forest theme, so the Canopy is named to convey that it exists “above” the rest of the website. The Canopy is conceived as a way to keep the top participants engaged even if they complete most of the activities available in the Smart Grid Game. A primary way to retain their engagement is to offer more sophisticated visualizations that demand more thought and analysis from the top players.

The Canopy provides a series of “missions” that players can undertake. Some Canopy missions are to be accomplished individually, while other missions require 2 or 3 participants to work together. Players can indicate that they are “up” for a group mission to find other interested participants. Missions include looking at more advanced energy visualizations, and also activities such as seeking out places on campus that are wasting energy.

Canopy missions are like actions available in the rest of the game, but instead of earning points upon completion, Canopy activities earn players *Canopy Karma*, which is a separate point system for the Canopy. Canopy Karma is used instead of the standard points to ensure that the Canopy itself does not unbalance the point competition by providing a way for the top players to earn more points that are not available to the rest of the participants.

The energy data and visualizations available in the forest (the main game area), such as the Daily Energy Goal Game described earlier, are deliberately simple to avoid confusing participants with low energy literacy. The requirement for simplicity means that the energy data shown to players comes only from the participant’s team. This constraint can be relaxed in the Canopy, as players are more sophisticated and presumed to be interested in visualizations that compared teams.

Fig. 6 shows “Energy Hot Spots”, one of several forms of advanced energy feedback available in the Canopy. This visualization shows hourly electricity use for a selected team over the course of a selected week. This breakdown allows players to examine their usage patterns and see how they

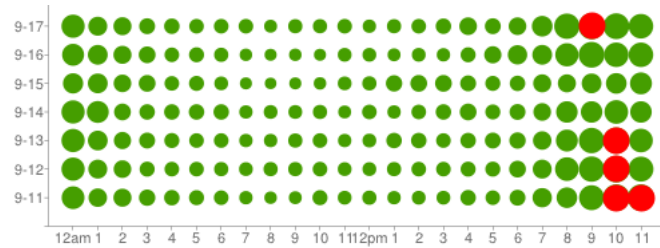


Figure 6. Energy Hot Spots visualization from the Canopy.

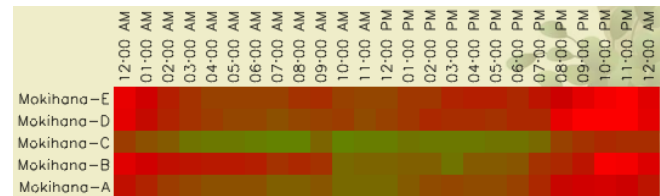


Figure 7. Heat Map visualization from the Canopy.

change throughout the day, between days of the week, and between different teams. The Canopy mission based around the Energy Hot Spots energy feedback asks players the following questions:

- What hours of the day seem to have the highest energy use? How does that compare to your own energy use patterns?
- What differences in energy use do you notice between teams?
- What are your thoughts on this visualization? What are its strengths and weaknesses?

Another Canopy visualization is the Heat Map, shown in Fig. 7. This visualization shows the energy use over time for all the teams in one residence hall, arranged spatially. This type of visualization can potentially show patterns related to the position of teams in a residence hall, and shows how energy use differs between teams over time. The Canopy mission based on the Heat Map visualization asks similar questions as the Hot Spots visualization, with the intent to make players investigate the visualization and introspect about their own energy use.

The Canopy was introduced in the third week of the 2011 UHM Kukui Cup, and the top 42 players were invited to enter the Canopy out of a total of 401 players. Although use of the Canopy was quite limited, players’ feedback about these advanced forms of energy feedback was uniformly positive: they reported enjoying the ability to compare energy use between teams and seeing what times different teams energy use peaked.

This positive feedback indicates that the Canopy energy feedback was a success, and also indicates the critical role of context and scaffolding in energy feedback design. To build on the Canopy’s role as another “level” of the game, in the 2012 UHM Kukui Cup, we removed the Canopy, but

the Smart Grid Game was segmented into explicit levels as shown earlier in Fig. 1. The advanced visualizations from the Canopy missions were turned into activities accessible in the more advanced levels of the Smart Grid Game.

V. DOMAIN KNOWLEDGE AND ENERGY FEEDBACK

Energy feedback systems provide data on some aspect of behavior with the goal of reducing negative environmental impact [3]. However, they often assume users possess some level of domain knowledge about the environmental topic they hope to address. The term *energy literacy* has been used to describe the understanding of energy concepts as they relate both on the individual level and on the national/global level.

Some examples of energy literacy are: understanding the difference between power and energy; knowing that a microwave uses much more power than a refrigerator, but that the refrigerator will use much more energy over the course of the day; and knowing how electricity is generated in one's community.

Unfortunately, all indications are that energy literacy is low in the United States. DeWaters and Powers have developed an energy literacy survey instrument for middle and high school students. They found that the student mean attitude scores were 73%, but that knowledge scores lagged far behind (42% correct) [18]. Based on their findings, they make some recommendations, such as energy curricula be "hands on, inquiry based, experiential, engaging, and real-world problem solving...", and using the campus as a "learning laboratory". Similarly, a nationwide survey in the United States of adults on energy by Southwell et al. found that the average respondent answered fewer than 60% of the energy knowledge questions correctly [19]. A online survey of 505 people in the United States regarding perceptions of energy consumption and savings conducted by Attari et al. also found significant problems with energy literacy [20].

A. Energy Literacy in Action

One energy literacy topic that we emphasize in the Kukui Cup is the difference between power and energy, power being the rate at which energy is being consumed or produced (measured in watts) and energy is the quantity of work that can be performed by a system (measured somewhat confusingly for electricity in kilowatt-hours). In the Kukui Cup we explain this relationship as being analogous to speedometer and odometer in a car.

Through answers submitted to the online activities in the Kukui Cup, we can see that many players have trouble understanding the concepts of power, energy and their interrelationship. Players often confuse the two concepts and often fail to grasp the time sensitivity of power, and thereby considering devices that consume a lot of power as "bad" irrespective of how long they are actually used. When the users of visualizations do not understand the concepts

that are being visualized, understanding of the visualizations becomes much more difficult. It is for this reason that we claim that energy feedback systems should incorporate educational components, or risk being unintelligible to users. However, we reject the notion that power and energy, watts and kilowatt-hours are too complicated and that users should be provided instead with analogies to cars driven or hamburgers eaten. These energy concepts are important for effective customer participation in the smart grid, and should not be reduced to analogies alone.

Domain knowledge can also be provided to participants in real-world contexts, rather than online contexts. We have witnessed improvements in players' energy literacy in the course of a single workshop. In the energy scavenger hunt workshop, attendees are grouped into teams of 3–4 people and provided with a plug load meter that shows the amount of electrical power consumed by whatever device is plugged into the meter. Each team is given 30 minutes record the power use of devices in their rooms and residence halls, looking for devices with the highest power use they can find, and also the most number of devices with distinct power use in 10 W intervals. The goals of the workshop (beyond entertainment) are to train players to measure device power use, and also to build their intuition about how much power different devices use. In the 2011 Kukui Cup, after the teams completed their hunt, each team was asked to present their results. One team reported that the microwave they measured used 200 W, and immediately several players from other teams shouted out that the first team must have done something wrong because microwaves use over 1000 W, as they started to form their energy intuition based on their own hunt.

B. Moving Beyond Individual Actions

In addition to shorter activities available in the SGG, the Kukui Cup also provides creative activities to encourage more in-depth explorations of energy and sustainability. These creative activities run the gamut from writing a haiku about a sustainability topic, to conducting an interview, or making a video. Creative activity submissions from players are assigned a point value by administrators based on the quality of the work and the effort it required to make. We hope that the creative activities provide a different outlet for players, and encourage them to think beyond their individual actions and bring a sense of ownership to the cause of sustainability.

As mentioned previously, the 2012 UH Kukui Cup took place over an entire academic year. Much of the educational content we developed was made available to players in the first intensive month of the challenge, leaving later rounds with less content, and thereby fewer reasons to continue playing. To address this problem, and to draw players into deeper play and understanding of sustainability issues, players were able to suggest new additions to the

SGG as part of the game. Players that provided useful new activities and events for the game earned points, and once the new actions were placed into the SGG, they provided additional educational content for other players. We also explored ways to tie class work into the Kukui Cup challenge. With the longer time frame afforded by the nine month competition, players were able to earn points in the competition by registering for sustainability-related classes, and picking sustainability-themed class projects. The goal of these changes was macro behavior changes like selecting a sustainability degree program or choosing a more efficient vehicle at some point in the future. We plan to follow up with the 2012 Kukui Cup players in future years to assess whether or not these macro behavioral changes did, in fact, occur.

There is a proposed renewable energy project in Hawai'i called "Big Wind" that would generate as much as 400 MW of electricity from wind farms covering substantial portions of two more rural islands (Moloka'i and Lāna'i) with excellent wind resources. The power would be transmitted via a new undersea cable to O'ahu, which has the majority of the state's population, but with inferior wind resources. There are advocates both for and against Big Wind, but to make an informed decision one should understand how O'ahu's electricity is generated now, and the characteristics and challenges of wind energy. We hope that the educational activities that are part of the Kukui Cup will equip players to make informed choices on these types of thorny policy questions.

VI. ENERGY FEEDBACK, STICKINESS, AND SERIOUS GAMES

A meta issue for all energy feedback systems is how to ensure that they continue to be "sticky" for users, as a feedback system that users do not view will be unable to accomplish anything. There are indications that the long-term impact of energy feedback may be diminished due to habituation. Froehlich suggests that the average user will spend less than one minute per day exploring their energy consumption behaviors [21]. A study by Houde et al. of households using Google PowerMeter found an "immediate decrease in electricity consumption, but in the long term these electricity savings decrease and disappear" [22]. This finding suggests that a primary concern for any energy feedback system is ensuring that users continue to interact with it over the long term. Put another way, energy feedback alone is not enough to accomplish the goal of long-range customer engagement with their energy consumption.

One solution to the lack of stickiness of energy feedback systems is the incorporation of game play. Serious games like the Kukui Cup provide an alternative route to promote both learning and engagement with energy feedback. It is for this reason that we designed the Kukui Cup as a serious game that incorporates electricity consumption feedback as

one aspect of the game experience, rather than an energy feedback system that has been "gamified" [23]. Gee has examined how learning takes place in the world of video games, and contrasted it with the way learning typically takes place in schools [24]. He points out that good games are both deep and complicated, but large numbers of players manage to learn how to play them. Players can spend enormous amounts of time playing games, and in the process of playing learn skills in a way not possible in a classroom.

One issue with the Kukui Cup is that the educational content is largely of interest only until its content has been assimilated. We do not anticipate that players would want to revisit most actions unless they were able to earn additional points. This limited engagement is in contrast to games that players enjoy playing over and over. Some serious games, such as the protein folding game Foldit, do manage to attract repeat players and meet their serious goals [25]. While the Kukui Cup may stop being of interest to players once they have completed all the content available, we hope that our attempts to engage players in broader sustainability opportunities such as taking classes and becoming involved in campus and community organizations make the Kukui Cup no longer necessary for them.

While games are not the only way to promote long-term engagement with energy issues, we submit that any normal energy feedback system will quickly be abandoned by users once the novelty wears off. There must be a continuing reason for users to revisit the system that even the most novel and interesting energy feedback systems lack.

VII. FUTURE WORK

The Kukui Cup is an ongoing project and we continue to build upon the four challenges that have been run to date. One interesting area of future work is moving beyond just visualizations of energy consumption by incorporating visualizations of energy generation in the grid. Ecotricity, a renewable energy utility in the UK, has developed such a grid-level feedback system [26]. The Ecotricity website provides a real-time dashboard that displays the types of energy sources used to power the UK grid (fossil fuels, nuclear, and renewable) and the overall carbon intensity of the grid in gCO₂ per kWh as shown in Fig. 8. The carbon intensity display is made actionable through a traffic light visualization that is green when the grid is emitting less carbon and red when it emits more carbon, with the intent that consumers could defer energy use when the grid was emitting more CO₂.

During California's energy crisis in 2000 and 2001, Lawrence Berkeley National Laboratory created a web site that graphed data from utility organizations [27]. The graphs showed consumer demand for electricity (actual and forecast), and the utilities' generation capacity. Darby reports anecdotal evidence that people viewing the graphs changed

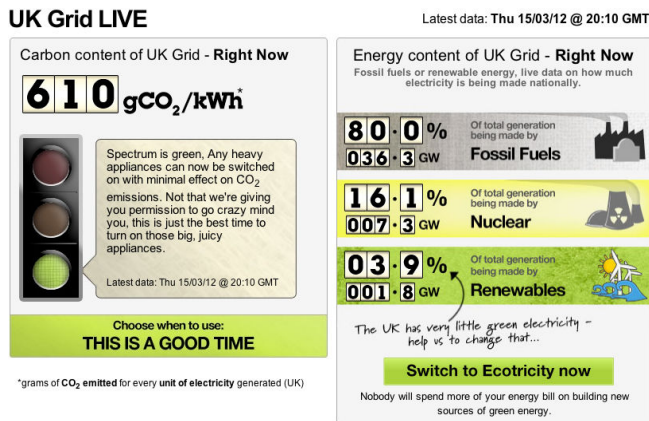


Figure 8. Ecotricity's live UK Grid dashboard.

their electricity usage based on the data [15]. In buildings generating their own electricity (such as through solar panels), there are additional opportunities to visualize the building's energy consumption along side the generation.

Beyond college campuses, we are planning to expand the Kukui Cup to other organizations in Hawai'i including middle and high schools. Kukui Cups run in schools will require changes to the educational content to make it relevant to those students, but may provide opportunities for integration with the curriculum.

A longer range goal is to integrate the Kukui Cup with Hawai'i's smart grid efforts. The Kukui Cup is currently an effort-intensive program, so scaling to hundreds of thousands of players will require scaling the management of the challenge, finding a means of funding, and a way for players to incorporate household energy data fairly, in a completely heterogeneous environment.

One final area of research is longitudinal studies of players after the game is over and they have moved out of the residence halls. We want to find out whether the Kukui Cup experience actually had lasting impacts on players, and whether they were able to continue any new behaviors after leaving the context of the residence hall.

VIII. CONCLUSION

We have described the Kukui Cup serious game, and our results from field trials of the system. We have discussed some of the energy feedback visualizations we developed, including both those that succeeded and those that failed. Based on our experiences, we provide three areas that energy feedback systems for the smart grid should address: they should be actionable, they must address users lack of domain knowledge, and they must find ways to be sticky.

ACKNOWLEDGMENTS

This research is supported in part by grant IIS-1017126 from the National Science Foundation; the HEI Charitable Foundation; Hawaiian Electric Company; the State of

Hawaii Department of Business, Economic Development and Tourism. We are also thankful for the support from the following organizations at the University of Hawai'i: the Center for Renewable Energy and Island Sustainability, Student Housing Services, Facilities Management, and the Department of Information and Computer Sciences.

We gratefully acknowledge the players of the 2011 and 2012 Kukui Cups and the members of the Kukui Cup team in addition to the authors who made the vision a reality: Kaveh Abhari, Hana Bowers, Greg Burgess, Caterina Desiato, Risa Khamsi, Amanda Pacholok, Morgan de Partee, Alyse Rutherford, Alex Young, and Chris Zorn.

REFERENCES

- [1] R. S. Brewer, Y. Xu, G. E. Lee, M. Katchuck, C. A. Moore, and P. M. Johnson, "Energy feedback for smart grid consumers: Lessons learned from the Kukui Cup," in *Proceedings of Energy 2013*, March 2013, pp. 120–126.
- [2] M. Zyda, "From visual simulation to virtual reality to games," *IEEE Computer*, vol. 38, no. 9, pp. 25 – 32, Sep 2005.
- [3] J. Froehlich, L. Findlater, and J. Landay, "The design of eco-feedback technology," in *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*. New York, New York, USA: ACM Press, Apr. 2010, pp. 1999–2008.
- [4] R. S. Brewer, G. E. Lee, and P. M. Johnson, "The Kukui Cup: a dorm energy competition focused on sustainable behavior change and energy literacy," in *Proceedings of the 44th Hawaii International Conference on System Sciences*, January 2011, pp. 1–10.
- [5] R. S. Brewer, "Fostering sustained energy behavior change and increasing energy literacy in a student housing energy challenge," Ph.D. dissertation, University of Hawaii, Department of Information and Computer Sciences, March 2013, [retrieved: December, 2013]. [Online]. Available: <http://csdl.ics.hawaii.edu/techreports/2010/10-08/10-08.pdf>
- [6] C. Hodge, "Dorm energy competitions: Passing fad or powerful behavior modification tool?" Presentation at the 2010 Behavior Energy and Climate Change conference, November 2010. [Online]. Available: http://www.stanford.edu/group/peec/cgi-bin/docs/events/2010/becc/presentations/2C_ChelseaHodge.pdf [retrieved: December, 2013].
- [7] G. E. Lee, "Makahiki: An extensible open-source platform for creating energy competitions," Master's thesis, University of Hawaii, June 2012. [Online]. Available: <http://csdl.ics.hawaii.edu/techreports/2011/11-01/11-01.pdf> [retrieved: December, 2013].
- [8] J. E. Petersen, V. Shunturov, K. Janda, G. Platt, and K. Weinberger, "Dormitory residents reduce electricity consumption when exposed to real-time visual feedback and incentives," *International Journal of Sustainability in Higher Education*, vol. 8, no. 1, pp. 16–33, 2007.
- [9] D. Merugu, B. S. Prabhakar, and N. S. Rama, "An incentive mechanism for decongesting the roads: a pilot program in Bangalore," in *Proceedings of NetEcon '09, ACM Workshop on the Economics of Networked Systems*, July 2009.

- [10] R. S. Brewer and P. M. Johnson, "WattDepot: An open source software ecosystem for enterprise-scale energy data collection, storage, analysis, and visualization," in Proceedings of the First International Conference on Smart Grid Communications, Gaithersburg, MD, October 2010, pp. 91–95.
- [11] P. M. Johnson, Y. Xu, R. S. Brewer, C. A. Moore, G. E. Lee, and A. Connell, "Makahiki+WattDepot: An open source software stack for next generation energy research and education," in Proceedings of the 2013 Conference on Information and Communication Technologies for Sustainability (ICT4S), February 2013.
- [12] L. J. Becker, "Joint effect of feedback and goal setting on performance: A field study of residential energy conservation," *Journal of Applied Psychology*, vol. 63, no. 4, pp. 428–433, 1978.
- [13] J. H. van Houwelingen and W. F. van Raaij, "The effect of goal-setting and daily electronic feedback on in-home energy use," *The Journal of Consumer Research*, vol. 16, no. 1, pp. 98–105, June 1989.
- [14] P. M. Johnson, Y. Xu, R. S. Brewer, G. E. Lee, M. Katchuck, and C. A. Moore, "Beyond kWh: Myths and fixes for energy competition game design," in Proceedings of Meaningful Play 2012, October 2012, pp. 1–10.
- [15] S. Darby, "The effectiveness of feedback on energy consumption," Environmental Change Institute, University of Oxford, Tech. Rep., 2006. [Online]. Available: <http://www.eci.ox.ac.uk/research/energy/downloads/smart-metering-report.pdf> [retrieved: December, 2013].
- [16] A. Faruqui, S. Sergici, and A. Sharif, "The impact of informational feedback on energy consumption—a survey of the experimental evidence," *Energy*, vol. 35, no. 4, pp. 1598–1608, 2010.
- [17] B. Foster and S. Mazur-Stommen, "Results from recent real-time feedback studies," American Council for an Energy-Efficient Economy (ACEEE), Tech. Rep. B122, February 2012. [Online]. Available: <http://aceee.org/research-report/b122> [retrieved: December, 2013].
- [18] J. E. DeWaters and S. E. Powers, "Energy literacy of secondary students in New York State (USA): A measure of knowledge, affect, and behavior," *Energy Policy*, vol. 39, no. 3, pp. 1699–1710, 2011.
- [19] B. G. Southwell, J. J. Murphy, J. E. DeWaters, and P. A. LeBaron, "Americans' perceived and actual understanding of energy," RTI Press, Tech. Rep. RR-0018-1208, 2012. [Online]. Available: <http://www.rti.org/pubs/rr-0018-1208-southwell.pdf> [retrieved: December, 2013].
- [20] S. Z. Attari, M. L. DeKay, C. I. Davidson, and W. B. de Bruin, "Public perceptions of energy consumption and savings," *Proceedings of the National Academy of Sciences*, vol. 107, no. 37, pp. 16 054–16 059, 2010.
- [21] J. Froehlich, "Moving beyond line graphs: The history and future of eco-feedback design," Presentation at the 2010 Behavior Energy and Climate Change conference, 2010. [Online]. Available: http://peec.stanford.edu/docs/events/2010/becc/presentations/3D_JonFroehlich.pdf [retrieved: December, 2013].
- [22] S. Houde, A. Todd, A. Sudarshan, J. A. Flora, and K. C. Armel, "Real-time feedback and electricity consumption: A field experiment assessing the potential for savings and persistence," *The Energy Journal*, vol. 34, no. 1, pp. 87–102, 2013.
- [23] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining "gamification"," in *Mindtrek 2011 Proceedings*. ACM Press, 2011, pp. 9–15.
- [24] J. P. Gee, *What Video Games Have to Teach Us About Learning and Literacy*. Palgrave Macmillan, 2007.
- [25] F. Khatib, F. DiMaio, S. Cooper, M. Kazmierczyk, M. Gilski, S. Krzywda, H. Zabranska, I. Pichova, J. Thompson, Z. Popović, M. Jaskolski, and D. Baker, "Crystal structure of a monomeric retroviral protease solved by protein folding game players," *Nat Struct Mol Biol*, vol. 18, no. 10, pp. 1175–1177, 10 2011.
- [26] "Ecotricity UK grid live website," <http://www.ecotricity.co.uk/our-green-energy/energy-independence/uk-grid-live>, [retrieved: December, 2013].
- [27] E. Bartholomew, C. Bolduc, K. Coughlin, B. Hill, A. Meier, and R. V. Buskirk, "Current energy website," <http://web.archive.org/web/20030507025649/http://currentenergy.lbl.gov/>, [retrieved: December, 2013].

Regular Polysemy in WordNet and Pattern based Approach

Abed Alhakim Freihat, Fausto Giunchiglia

Dept. of Information Engineering and Computer Science
University of Trento,
Trento, Italy
e-mail: {fraihat,fausto}@disi.unitn.it

Biswanath Dutta

Documentation Research and Training Centre
Indian Statistical Institute (ISI)
Bangalore, India
e-mail: bisu@drtc.isibang.ac.in

Abstract— WordNet represents the polysemous terms by capturing the different meanings of them at lexical level implicitly without giving emphasis on the polysemy types they belong to. This problem affects the usability of WordNet as a suitable knowledge representation resource for Natural Language Processing applications. The current work presents pattern based approach for solving the polysemy problem by transforming the implicit relations between the synsets at lexical level into explicit relations at the semantic level.

Keywords— *lexical databases; WordNet; specialization polysemy; metaphoric polysemy; homonymy; regular polysemy; polysemy reduction; lexical semantics; semantic search; knowledge engineering*

I. INTRODUCTION

Solving the polysemy problem is very crucial in many research fields including machine translation, information retrieval and semantic search [1] since polysemy in WordNet [2] is considered to be the main reason that makes it hard to use for natural language processing (NLP) and semantic applications.

Polysemy corresponds to various kinds of linguistic phenomena and belongs to different polysemy classes. Recognizing the polysemy class of a given polysemous term is essential in NLP since different polysemy phenomena require different processing strategies. Differentiating between the polysemy classes should be possible through explicit semantic relations between the senses of polysemous terms. Unfortunately, relations between polysemous terms are not provided in WordNet [3]. For instance, WordNet does not provide the distinction between homographs, and complementary terms [4].

In the last decades, many approaches have been introduced to solve the polysemy problem through merging the similar meanings of polysemous terms [5]. These approaches are sometimes helpful in cases, where terms have meanings that are similar enough to be merged. However, polysemous terms with similar meanings are a sub-case of the solution of specialization polysemy [6]. They represent only a small portion of the polysemy problem. In fact, a significant portion of the polysemous senses should not be merged, as they are just similar in meaning [7] and not redundant. In another approach, CORELEX [4] has been introduced as an ontology of systematic polysemous nouns extracted from WordNet.

However, CORELEX deals only with the upper level ontology of WordNet that corresponds mainly to the metonymy cases and does not provide a solution for other polysemy types [6].

In this paper, we introduce a pattern based approach that combines several ideas to solve the polysemy problem. Our approach follows the idea that the polysemy problem is a problem of semantic organization [9]. Thus, the goal of our approach is to reorganize the semantic structure of the polysemous terms in wordNet, where we transform the implicit relations between the polysemous terms at lexical level to explicit relations at the semantic level. This includes extending WordNet by adding new hierarchical and associative relations between the synsets to explicitly denote the polysemy type occurring between the meanings of each polysemous term, as suggested in [3]. To achieve this goal, our approach deals with all polysemy types at all ontological levels of WordNet. It deals with the lower level ontology of WordNet and it extends the merge operation suggested by the polysemy reduction approaches [5][10] by providing new operations that organize the relations between the meanings of polysemous terms. Our approach also deals with polysemy in the middle level, as it is the case in regular polysemy approaches [11] and also in the upper level ontology as in systematic polysemy approaches [4].

This paper is organized as follows. In Section II, we discuss the polysemy problem in wordNet. In Section III, we describe the current approaches for solving the polysemy problem in WordNet. In Section IV, we present the semantic relations that denote polysemy types and the operations that reorganize the structure of polysemous terms in WordNet. In Section V, we introduce a pattern based approach for solving the polysemy problem in the case of polysemous nouns. In Section VI, we discuss the rules that we use by reorganizing the ontological structure of polysemous terms. In Section VII, we discuss the results and evaluation of our approach. In Section VIII, we conclude the paper and describe our future research work.

II. POLYSEMY IN WORDNET

WordNet is a lexical database that organizes synonyms of English words into sets called synsets, where each synset is described through a gloss. For example, the words *happiness* and *felicity* are considered to be synonyms and are grouped into a synset {*happiness, felicity*} that is described through the gloss: *state of well-being*

characterized by emotions ranging from contentment to intense joy.

WordNet organizes the relations between synsets through semantic relations, where each word category has a number of relations that are used to organize the relations between the synsets of that grammatical category. For example, the hyponymy relation (X is a type of Y) is used to organize the ontological structure of nouns. WordNet 2.1 contains 147,257 words, 117,597 synsets and 207,019 word-sense pairs. Among these words there are 27,006 polysemous words, where 15,776 of them are nouns.

From linguistics, a term is polysemous if it has more than one meaning[19]. Linguists differentiate between contrastive polysemy, i.e., terms with completely different and unrelated meanings - also called homonyms or homographs; and complementary polysemy, i.e., terms with different but related meanings. Complementary polysemy is classified in three sub types: metonymy, specialization polysemy and metaphors. Following the above, we can classify the various forms of polysemy as follows:

1) Complementary polysemy: terms that have the same spelling and related meanings. Complementary polysemy can be:

a. Metonymy: substituting the name of an attribute or feature for the name of the thing itself, such as in the following example the term *chicken*:

Peter caught *a chicken* in his garden.

Peter prepared *chicken* for the dinner.

b. Specialization polysemy: a term is used to refer to a more general meaning and another more specific meaning, such as in the following example the term *methodology*:

#1 **methodology**, methodological analysis: the branch of philosophy.

#2 **methodology**: the system of methods followed in a particular discipline.

c. Metaphors: terms that have the same spelling and have literal and figurative meanings. Consider, for instance, the term *parasite*:

#1 **parasite**: an animal or plant that lives in or on a host (another animal or plant).

#2 leech, **parasite**, sponge, sponger: a follower who hangs around a host (without benefit to the host) in hope of gain or advantage.

2) Homographs: terms that have the same spelling and different unrelated meanings, such as in the following example the term *bank*:

Peter sat on the *bank* of the river.

Peter deposited money in the *bank*.

In WordNet, the number of senses for a polysemous term may range from 2 to more than 30. In some rare cases, the number of senses is even more. For instance, the noun *head* has 33 senses. Nevertheless, 90% of the polysemous terms are nouns. Table I shows the distribution of these polysemous nouns according to the number of senses they have. Notice that, in this paper, we are concerned with

polysemous nouns only, and not the verbs, adverbs and adjectives.

The fact that a term has more than two senses implies that the meanings of the term belong to more than one type of polysemy. For example, the term *food* has 3 senses as mentioned below, where the polysemy type between the first and the second meanings is specialization polysemy, while the third meaning is metaphoric.

TABLE I. POLYSEMIOUS NOUNS IN WORDNET

| # of senses | # of nouns (in percentage) |
|-------------|----------------------------|
| 2 | 9328 ($\approx 64.2\%$) |
| 3 | 2762 ($\approx 19\%$) |
| 4 | 1083 ($\approx 7.4\%$) |
| 5 | 555 ($\approx 3.8\%$) |
| 6 | 277 ($\approx 1.9\%$) |
| 7 | 194 ($\approx 1.3\%$) |
| 8 | 90 ($\approx 0.7\%$) |
| 9 | 88 ($\approx 0.6\%$) |
| 10 | 54 ($\approx 0.37\%$) |
| >10 | 94 ($\approx 0.64\%$) |
| Total | 14525 (=100%) |

#1 **food**, nutrient: any substance that can be metabolized by an organism to give energy and build tissue.

#2 **food**, solid food: any solid substance that is used as a source of nourishment.

#3 **food**, food for thought: anything that provides mental stimulus for thinking.

III. APPROACHES FOR SOLVING POLYSEMY IN WORDNET

The approaches of polysemy can be classified in three main approaches. The first is polysemy reduction, where the focus is on complementary polysemy to produce more coarse-grained lexical resources of existing fine-grained ones, such as WordNet. The second type of polysemy approaches focuses on classifying polysemy into systematic or regular polysemy and homographs. Based on this classification, CORELEX was introduced as ontology of systematic polysemous nouns extracted from WordNet. The third type of polysemy approaches is semantic relations extraction approaches. These approaches propose to enrich wordNet with semantic relations that correspond to the implicit relations between the complementary polysemous terms in WordNet.

In the following, we summarize polysemy reduction approaches, CORELEX, and the most prominent semantic relations extraction approaches. Notice that neither polysemy reduction approaches nor systematic polysemy approaches could solve the polysemy problem in WordNet. In general, polysemy reduction approaches could not solve the problem of the upper level ontology, while systematic

polysemy approaches did not provide a solution for the polysemy problem in the middle and lower level ontology of WordNet.

A. Polysemy Reduction Approaches

In polysemy reduction, the senses are clustered such that each group contains related polysemous words [10][14]. These groups are called homograph clusters. Once the clusters have been identified, the senses in each cluster are merged. To achieve this task, several strategies have been introduced. These strategies can be mainly categorized in semantic-based and statistics-based strategies [15]. Some approaches combine both strategies [10]. Although results of applications of these approaches are reported, these results are taken usually from applying them on sample data sets, and there is no way to verify these results independently. Polysemy reduction approaches typically rely on the application of some detection rules such as: *if S1 and S2 are two synsets containing at least two words, and if S1 and S2 contain the same words, then S1 and S2 can be collapsed together into one single synset* [10]. However, applying this rule may wrongly result in merging two different senses as in the following example:

```
#1 smoke, smoking: a hot vapor containing fine
particles of carbon being produced by combustion.
#2 smoke, smoking: the act of smoking tobacco or
other substances.
```

In general, polysemy reduction can neither predict the polysemy type occurring between the senses of polysemous words nor can deal with metonymy or metaphors. Polysemy reduction does not solve the polysemy problem in linguistic resource. Nevertheless, it can be potentially used to solve part of the problem, namely, the identification and merging of genuine redundant synsets.

B. CORELEX

J. Apresjan defined regular polysemy as follows: “*a polysemous Term T is considered to be regular if there exists at least another polysemous T' that is semantically distinguished in the same way as T*” [16]. CORELEX and regular polysemy approaches in general rely on this definition. These approaches follow two different methods to solve the polysemy problem in WordNet:

CORELEX, the first systematic polysemy lexical database, follows the generative lexicon theory [9] that distinguishes between systematic (also known as regular or logic) polysemy and homographs. Systematic polysemous words are systematic and predictable while homonyms are not regular and not predictable. The type of polysemy of the word *fish*, for example, is systematic since the meaning *food* can be predicted from the *animal* meaning, and so the word *fish* belongs to the systematic class *animal food*. The two meanings of *fish* describe two related aspects of *fish*: *fish* is an animal and *fish* is a food. A word is systematic polysemous means that the meanings of this word are not homonyms and they describe different aspects of the same

term. Following this distinction, CORELEX organizes the polysemous nouns of WordNet 1.5 into 126 systematic polysemy classes. The systematic polysemy classes in CORELEX have been determined in a top-down fashion considering the patterns in the upper level ontology of wordNet only [11]. The high level basic types in CORELEX patterns make them too coarse grained to extract useful semantic relations [4][11][18]. At the same time, there are hundreds of regular structural patterns that reside in the middle level and lower level ontology that are not covered by the high level basic types. These patterns correspond to metaphoric and specialization polysemy [1][4]. The underspecification method is not appropriate to CORELEX patterns that correspond to metaphoric polysemy. CORELEX patterns contain too many false positives [11][18]. Another important point is related to the fine grained nature of WordNet, where the meanings of some CORELEX classes are very difficult to disambiguate, and indistinguishable even for humans [13].

C. Semantic Relations Extraction Approaches

The semantic relations extraction approaches are regular polysemy approaches that attempt to extract implicit semantic relations between the polysemous senses via regular structural patterns. The basic idea in these approaches is that the implicit relatedness between the polysemous terms corresponds to variety of semantic relations. Extracting these relations and making them explicitly should improve wordNet [11][18]. These approaches refine and extend CORELEX patterns to extract the semantic relations. Beside the structural regularity, these approaches exploit also the synset gloss [4][11] and the cousin relationship [11][18] in WordNet. For example, the approach described in [4] exploits synset glosses to extract *auto-referent candidates*. The approach described in [18] uses several strategies, such as *ontological bridging* to detect relations between the sense pairs. In general, the extracted relations in these approaches are similar. For example, we find the relations *similar to*, *color of* in the results of the approach in [4]. The result in [11] contains relations such as *contained in*, *obtain from*. Similarly, the result in [18] contains relations such as *fruit of*, *tree of*.

The semantic relations extraction approaches are in general better than CORELEX in the following aspects. First of all, the discovered patterns in these approaches are more fine grained and enable to capture meaningful relations. These approaches classified the complementary polysemy in three sub classes: metonymy, metaphoric, and specialization polysemy. Another important point in these approaches is that these approaches considered the problem of false positives. Anyway, these approaches did cover only few patterns of the specialization polysemy and metaphoric cases. They did not address the problem of too fine grained senses or discourse dependent polysemy.

In this section, we have discussed some of the state of the art approaches for solving the polysemy problem in

WordNet. Our primary observation here is that these approaches complement each other. The same holds for the approach presented in this paper. Our approach is not an alternative for the presented solutions. Instead, we see our approach as a complementary solution for the state of the art solutions that focuses on the specialization polysemy problem. Solving the specialization polysemy problem is important complementary step for the presented solutions, because solving the specialization polysemy problem enhances the usability of WordNet as a knowledge representation resource. Solving the specialization polysemy problem includes solving the following problems in WordNet:

- **The problem of implicit relatedness:** The implicit relatedness between specialization polysemy cases is a hierarchical. For example, representing the hierarchical relatedness between the senses of the term *white croaker* from knowledge representation point of view is more appropriate than representing at the lexical level only.
 #1 *white croaker*, queenfish, *Seriphus politus*: silvery and bluish fish of California.
 #2 *white croaker*, kingfish, *Genyonemus lineatus*: silvery fish of California.
- **The problem of too fine grained senses:** WordNet contains a reasonable amount of too fine grained senses. For example, the first sense of the term *optimism* corresponds to optimistic feeling and the second meaning corresponds to disposition.
 #1 *optimism*: the optimistic feeling that all is going to turn out well.
 #2 *optimism*: a general disposition to expect the best in all things.
 Capturing the difference between the meanings of such cases is very difficult. In some cases, the too fine grained distinction between senses may result in redundancy as in the following example.
 #1 *lullaby*, *cradlesong*, *berceuse*: a quiet song intended to lull a child to sleep.
 #2 *lullaby*, *cradlesong*: a quiet song that lulls a child to sleep.
- **The problem of discourse dependent polysemy:** WordNet contains a reasonable amount of discourse dependent polysemy cases. For example, using the term *center* to refer to the following meanings cannot be understood without a proper context.
 #2 *center field*, *centerfield*, *center*: the piece of ground in the outfield directly ahead of the catcher.
 #6 *center*, *center of attention*: the object upon which interest and attention focuses.
 #7 *center*, *centre*, *nerve center*, *nerve centre*: a cluster of nerve cells governing a specific bodily process.
 #15 *plaza*, *mall*, *center*, *shopping mall*, *shopping center*, *shopping centre*: mercantile establishment consisting of a carefully landscaped complex of shops representing leading merchandisers.

It is not clear, which rule wordNet is following by adding such discourse dependent terms in the synset synonyms. In this example, it is not clear, why wordNet considers the term *center* to be a synonym in the previous cases and it does not consider it a synonym of the terms *city center*, *medical center*, or *research center*. In addition to these problems, our approach is able to discover homonymy and metaphoric cases that reside in the middle level or lower level ontology of wordNet. We provide a solution for the discovered homonymy and metaphoric cases as explained in the next section.

IV. DENOTING POLYSEMY TYPES AND ORGANIZING POLYSEMY IN WORDNET

A. Polysemy Type Relations

In the following, we explain the suggested relations to denote the polysemy types:

Homographs: There is no relation between the senses of a homograph term. Nevertheless, differentiating homographs from other polysemy types is very important improvement in wordNet. We use the relation *is_homograph* to denote that two synsets of a polysemous term are homographs. For example, this relation holds between the synsets *{saki as alcoholic drink}* and *{saki as a monkey}*.

Metonymy: In metonymy cases, there is always a *base meaning* of the term and other *derived meanings* that express different aspects of the base meaning [17]. For example, the term *chicken* has the base meaning *{a domestic fowl bred for flesh or eggs}* and a derived meaning *{the flesh of a chicken used for food}*. To denote the relation between the senses of a metonymy term, we use the relation *has_aspect*, where this relation holds between the base meaning of a term and the derived meanings of that term. To set up the relation we need to determine the base meaning and then relate the other derived meanings to it.

Metaphors: In metaphoric cases, we use the relation *Is_metaphor* to denote the metaphoric relation between the metaphoric meaning and literal meaning of a metaphoric term. For example, this relation is used to denote that *{cool as great coolness and composure under strain}* is metaphoric meaning of the literal meaning *{cool as the quality of being at a refreshingly low temperature}*. In the cases where this relation is applicable, we need to specify the literal meaning and the metaphoric meaning.

B. Operations for Specialization polysemy

Analysis of specialization polysemy cases shows that such cases can be classified based on the synset synonyms into the following three groups. To explain our idea, we have chosen cases, where the synsets of each term share the same common parent.

Let *T* be a polysemous term that occurs in two synsets *S₁* and *S₂*. We consider *T* in the following three cases:

Case 1: *T* has synonyms in *S₁* and has synonyms in *S₂* as in the case of *kestrel*:

#1 sparrow hawk, American kestrel, kestrel, Falco sparverius: small American falcon.
 #2. kestrel, Falco tinnunculus: small Old World falcon.

Case 2: T has synonyms in S_1 or in S_2 but not in both as in the case of *dorsum*:

#1 back, dorsum: the posterior part of a human (or animal) body from the neck to the end of the spine.

#2 dorsum: the back of the body of a vertebrate or any analogous surface.

Case 3: T has no synonyms in S_1 or S_2 as the in the case of *compatible software*:

#1 compatible software: application software programs that share common conventions.

#2 compatible software: software that can run on different computers without modification.

In case 1, T has synonyms in S_1 . This means that T is exchangeable with the other synonyms of S_1 and at the same time is also exchangeable with the synonyms of S_2 . Let T_1 , T_2 be non polysemous synonyms of T in S_1 and S_2 respectively. T_1 is synonymous with T but not with T_2 . Otherwise, T_1 and T_2 should appear in the same synset. The fact that T_1 and T_2 appear in two different sibling synsets indicates that they are not the same. We think that the semantic relatedness between S_1 and S_2 is encoded at lexical level rather than at semantic level. We have the same observation in case 2. The fact that one synset contains T only and the other synset contains additional terms indicates that the synset that contains T only is a more general meaning of the synset that have additional terms. We consider the terms in case 3 as candidates to be merged. Accordingly, we suggest the following operations to organize the relations between the senses in specialization polysemy cases.

Solution for Case 1: We add a new (missing) parent in cases where the polysemous meanings of a term T can be seen more specific meanings of an absent more general meaning. In such cases, we create the missing parent, which is a more general meaning and connect the more specific meanings to this newly created parent. This operation is schematized in Figure 1.

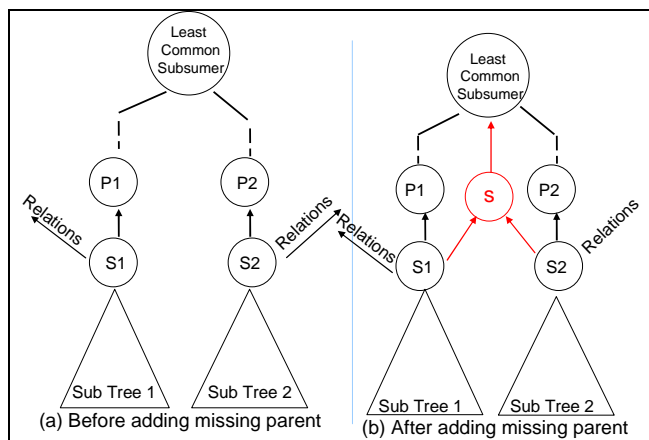


Figure 1. Adding missing parent

Solution for Case 2: We establish a new (missing) *is_a* relation to denote that a sense of a polysemous term T is more specific than another more general meaning of T. We schematize this operation as illustrated in Figure 2.

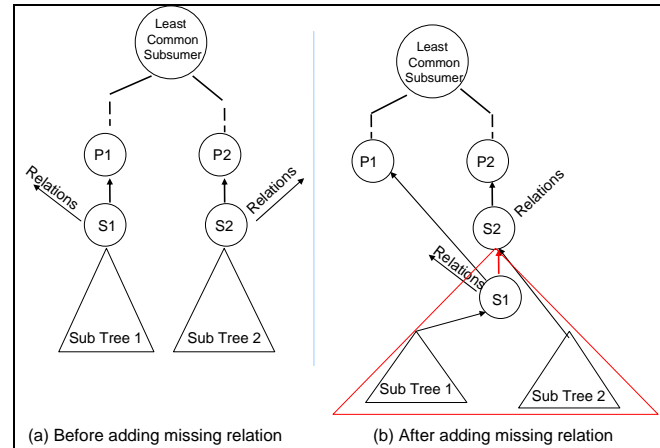


Figure 2. Adding missing relation

Solution for case 3: We merge the meanings. The merge operation is schematized as in Figure 3.

At the term level, we disambiguate the polysemous terms as follows: in case 1, we remove the polysemous terms from both child synsets and keep the polysemous words in the new added parent synset only. In case 2, we remove the polysemous term from the synset with the more specific meaning and keep it in the synset with the more generic meaning. The Merge operation in case 3 unifies the terms of both synsets in one synset. Thus, applying the three operations results in reducing the number of polysemous words in WordNet.

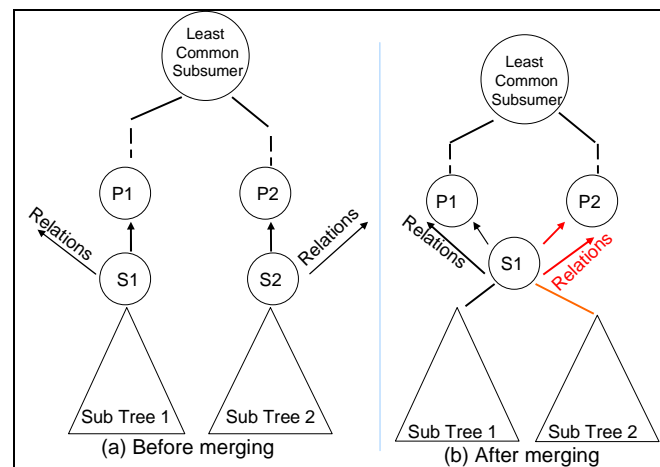


Figure 3. Merge operation

V. PATTERN BASED APPROACH FOR SOLVING POLYSEMY

In this section, we describe our approach for solving polysemy in WordNet. The approach has the following five

phases. Phases A, C, and E are automatic, while B and D are manual.

- A. Patterns Identification
- B. Patterns Classification
- C. Polysemy type Assignment
- D. Validation
- E. Applying the polysemy relations and operations

A. Patterns Identification

In this phase, we discuss the algorithm that is used to identify the regular type compatible patterns. Before describing the algorithm, we illustrate the definitions we used in the algorithm.

Definition 1 (Term). A term T is a triple $\langle \text{Label}, \text{Cat}, \text{Rank} \rangle$, where

- a) Label is the term label, i.e., a word which is the orthographic string representation of the term;
- b) Cat is the grammatical category of the term;
- c) Rank is the term rank, i.e., a natural number > 0 .

Wordnet organizes terms into synsets, where each synset contains an ordered list of synonymous terms. We define wordNet synsets as follows.

Definition 2 (wordNet synset) A synset S is a quintuple $\langle \text{Terms}, \text{Label}, \text{Gloss}, \text{Rels}, \text{Rank} \rangle$, where

- a) Terms is an ordered list of synonomous terms that have the same grammatical category, called synset synonyms;
- b) The grammatical category of the synset is the grammatical category of its terms;
- c) The term rank of the synset terms corresoponds to its position in the terms list;
- d) $\text{Label} \in \text{Ts}$ is the synset label, i.e., the prefferd term of the synset is the first term in the terms list;
- e) Gloss is a natural language text that describes the synset;
- f) Rels is a set of semantic relations that hold between the synsets;
- g) Rank is the synset rank, i.e., a natural number > 0 that reflects the familiarity of the synset.

The set Rels in the previous definition correspond to the semantic relations used by WordNet to organize the relations between the synsets. In the following, we define the relation hypernym and hyponym, the counter relation of hypernym.

Definition 3 (hypernym/hyponym relation). The relations hypernym/hyponym are hierarchical relations in WordNet that denote the superordinate/subordinate

relationship between synsets. For two synsets s_1, s_2 : s_1 is a hypernym of s_2 is equivalent to s_2 is a hyponym of s_1 .

Using the hypernym relation, wordNet organizes synsets in the case of nouns in a hierarchy. We define the hierarchy of WordNet in the case of nouns as follows:

Definition 4 (wordNet hierarchy). WordNet hierarchy WH is a rooted diagram $\langle N, E \rangle$, where

- a) N is a set of synsets that belong to the grammatical category noun;
- b) $\text{Entity} \in N$ is the root of WH ;
- c) $E \subseteq N^2$;
- d) $(s_1, s_2) \in E$ if s_1 is a hypernym of s_2 ;
- e) $\forall s ((s \in N \wedge s \neq \text{entity}) \Rightarrow \exists s' ((s', s) \in E))$.

A term is considered to be polysemous if it is found in the terms of more than one synset. We call such synsets polysemous synsets. It is possible for two polysemous synsets to share more than one term. Two polysemous synsets and their shared terms constitute a polysemy case. In the following, we define a polysemy case as follows.

Definition 5 (polysemy case) A polysemy case is a triple $\langle \text{Ts}, s_1, \dots, s_n \rangle$, where s_1, \dots, s_n are polysemous synsets that have the terms Ts in common.

Note that the polysemy cases $c_1 = \langle \text{Ts}, s_1, s_2 \rangle$ and $c_2 = \langle \text{Ts}, s_2, s_1 \rangle$ are considered to be one polysemy case. We exploit the hypernym relation to discover the terms that are “*semantically distinguished in the same way*” as stated in Apresjan’s definition. We consider polysemy cases to be semantically distinguished in the same way if they have the same structural pattern. In the following, we define structural patterns.

Definition 6 (Structural Pattern) Let $c = \langle \text{Ts}, s_1, \dots, s_n \rangle$ be a polysemy case. Let R be a subset of $\{s_1, \dots, s_n\}$. Let Q an ordered sequence of R , where $|R| = m, 2 \leq m \leq n$, and $Q = \langle s_1, \dots, s_m \rangle, s_i \in R, s_i \neq s_j$, for $i \neq j$. A structural pattern

is defined as $p \# \langle p_1, \dots, p_m \rangle$, such that each p_i is a direct hyponym of p and subsumes $s_i, 1 \leq i \leq m$. We call p the pattern head and p_i the pattern parts.

Definition 7 (Regular Structural Pattern) A pattern is regular if there are at least two terms that belong to it.

For example, the pattern *passerine* # $\langle \text{oscine}, \text{tyrannid} \rangle$ is regular since there are 3 terms that belong to it.

Definition 8 (Sub pattern) For a regular pattern $ptrn = p \# \langle p_1, \dots, p_m \rangle$. A pattern $ptrn'$ is a sub pattern of $ptrn$ if $ptrn' = p \# \langle p'_1, \dots, p'_k \rangle$ and $\exists p_i, p'_j (p_i = p'_j)$.

Sub patterns are important, since it is possible that the elements of a pattern and its sub patterns have the same polysemy type. For example, the pattern *passerine* # $\langle \text{oscine}$,

tyrannid> and its sub pattern *passerine*#<*oscine,wren*> both belong to the specialization polysemy patterns.

Definition 9 (Common parent class) A term belongs to the common parent class if it has at least two synsets that share the same hypernym.

For example, the synsets of the term *kestrel* in the previous section share the same hypernym. In polysemy reduction approaches, senses that have the common parent property are candidates to be merged. In our approach, such terms are candidates for specialization polysemy. Note that there are many terms that have this property, but they are not considered to be regular according to definition 1, since they have different hierarchical structures.

In the following, we explain the pattern extraction algorithm.

Algorithm : Regular Polysemy Patterns Extraction

Input:

PNOUNS = Polysemous nouns in WordNet

UR = the list of the unique beginners in WordNet

SNR = the number of the term synsets,

Output:

N = an associative array to store the regular patterns.

M = an associative array to store the sub patterns

P = a list to store the elements of the common parent class

O = a list of singleton patterns

1. *poly_nouns* = retrieve_polysemous_nouns(SNR);

2. **For each** *noun* **in** *poly_nouns*

3. *S* = retrieve_synsets(*noun*);

4. *ptrns* = construct_patterns(*S*);

5. **For each** *Q* \subseteq *S*

6. **If** *Q* \in *Common Parent*

7. add <*noun*, *Q*> to *P*;

8. **For each pattern** *ptrn* = *p*#<*p*₁,...*p*_{*m*}> **in** *ptrns*

9. **If** *p* \notin *UR*

10. Add *noun* to the list under *ptrn* in *N*;

11. **For each** *ptrn* **in** *N*

12. **If** |*N*[*ptrn*] > 1

13. *M*[*ptrn*] = sub_patterns(*ptrn*)

14. Remove sub_patterns(*ptrn*) from *N*

15. **For each** *ptrn* **in** *N*

16. **If** |*N*[*ptrn*] < 2

17. Add *ptrn* to *O*;

18. Remove *ptrn* from *N*;

19. **return** <*N*,*M*,*P*,*O*>;

The presented algorithm works in three phases:

1. **Patterns and common parent terms identification (line 1 to 10):** We retrieve the list of all nouns that have the sense number given in the algorithm input. We check, whether the term belongs to the common parent class and also whether it has regular patterns. We exclude the top level ontology patterns such as *physical entity*#<*physical object*, *physical process*>. Such patterns correspond usually to CORELEX patterns and

they are not specialization polysemy patterns. Notice also that it is possible for terms that have more than 2 senses to have more than one pattern. The function **construct_patterns** is explained below.

2. **Sub patterns identification (lines 11 to 14):** If more than one term belongs to a pattern, it is a regular pattern, and then we search all singleton patterns to identify possible sub patterns of that pattern. Identified sub patterns are removed from the patterns list and added to the sub patterns list.
3. **Singleton patterns identification (lines 15 to 18):** After identifying the sub patterns, the remaining singleton patterns are removed from the patterns list and added to the list of the singleton patterns.

In the following, we explain the algorithm **construct_patterns** that is used for constructing patterns.

Algorithm: construct_patterns

Input:

Synsets : a list of synsets

Output: a list of patterns

1. *parents* := a list of synsets;

2. **For each** synset *s*₁ **in** synsets

3. **For each** synset *s*₂ \neq *s*₁ **in** synsets

4. *parent* = getLeastCommonSubsumer(*s*₁, *s*₂);

5. **If** *parent* \notin *parents*

6. add *parent* to *parents*;

7. *patterns* : a list of pattern strings;

8. **For each** *parent* **in** *parents*

9. *ptrnSynsets* = an empty list of synsets;

10. **For each** synset *syn* **in** *Synsets*

11. **If** isHypernym(*parent*, *syn*)

12. add *syn* to *ptrnSynsets*;

13. **If** |*ptrnSynsets*| > 1

14. *ptrnLabel* = constructPtrnLabel(*parent*, *ptrnSynsets*);

15. add *ptrnLabel* to *patterns*;

16. **return** *patterns*;

The algorithm **construct_patterns** works as follows.

1. **Computing possible pattern heads (line 2 to 6):** We compute the possible pattern heads by computing the least common subsumer of the synsets.
2. **Grouping the pattern synsets (line 7 to 15):** We compute the synsets that belong to the patterns according to the pattern heads constructed in the previous step.
3. **Constructing the pattern label (line 14):** We use the function **constructPtrnLabel** that constructs the pattern label of a pattern.

The algorithm **constructPtrnLabel** is explained below.

Algorithm: constructPtrnLabel

Input:

ptrnSynsets: a list of synsets that belong to a pattern

parent: the pattern head of the synsets in ptrnSynsets

Output: patternLabel

```

1. ptrnLabel : a string;
2. ptrnLabel = parent. "#<";
3. sort ptrnSynsets;
4. For i = 0; i < |ptrnSynsets| -1
5.     Synset s = ptrnSynsets[i];
6.     p = getRootHypynomRelativeToSynset(parent, s);
8.     ptrnLabel .= p. ",";
9. Synset s = ptrnSynsets[|ptrnSynsets|-1];
10. p = getRootHypynomRelativeToSynset(parent, s);
11. ptrnLabel .= p. ">";
12. return ptrnLabel;
```

The algorithm constructs the pattern label as defined in definition 1. The synsets are sorted alphabetically to ensure that the pattern label is a unique identifier of the pattern.

The results of applying the algorithm on the polysemous terms in WordNet are as follows: the total number of the nouns in WordNet is 14525 nouns. The algorithm identified 12565 polysemy cases to belong to type compatible patterns. The algorithm returned four lists: a pattern list that contains 1169 patterns, a sub patterns list that contains 2855 sub patterns, the list of the common parents that contains 1002 cases, and a list that contains 700 singleton patterns. The average time to generate the patterns on a single computer is about 45 minutes.

The algorithm returns the following lists:

1. **a list of regular patterns:** contains the regular patterns, where at least two terms belong to each pattern.
2. **a list of sub patterns:** contains the sub patterns of the patterns identified in the regular patterns list.
3. **a list of common parent terms:** contains the terms, where the synsets or part of the synsets of these terms share the same hypernym.
4. **a list of singleton patterns:** this list contains the patterns that have less than two terms and are not sub patterns of any regular pattern. Notice that it is possible for terms that have more than 2 senses to have more than one pattern.

B. Patterns Classification

Our task in this phase is to classify the patterns in specialization polysemy and metaphoric polysemy. First of all, the terms that belong to the common parent are considered as specialization polysemy candidates. We consider also the polysemy type of the sub patterns as the polysemy type of the pattern, they belong to. To classify the patterns, we have arranged them into hierarchies. Some examples for the roots of the hierarchies are shown in Table II. The numbers rights to the types correspond to the number of patterns that belong to that type.

Analyzing the patterns under these types shows that these patterns can be classified into four groups:

1. Specialization polysemy patterns

2. Metaphoric patterns
3. Homonymy patterns
4. Mixed patterns

TABLE II. EXAMPLES FOR TYPE COMPATIBLE PATTERNS ROOTS IN WORDNET

| Patterns under physical entity | | Patterns under abstract entity | |
|--------------------------------|-----------|--------------------------------|-----------|
| Type | #patterns | Type | #patterns |
| substance | 19 | psychological feature | 1 |
| organism | 9 | cognition | 19 |
| person | 148 | attribute | 13 |
| animal | 5 | communication | 27 |
| plant | 6 | measure | 11 |
| artifact | 90 | group | 17 |
| process | 10 | time period | 6 |
| location | 7 | relation | 11 |
| thing | 5 | act | 70 |

In the following, we explain our criteria by classifying the patterns.

1. Specialization Polysemy patterns: the type of some specialization polysemy patterns can be determined directly by considering the type of the pattern only. For example, it is clear that the patterns whose type belongs to *animal* and the types under *animal* are specialization polysemy, or at least, it is not common at all to find a metaphoric link between the types under *animal*. The criterion for determining other specialization polysemy patterns is the *consistency* of the pattern subtypes.

2. Metaphoric patterns: to determine metaphoric patterns, we followed the idea that metaphors are human centric in the sense that we use metaphors to express our feelings, judgments, situations, irony, and so on. For example, when we use *sponger* to refer to some one, we are making a judgment upon that person. This gives us a hint, where to search for metaphoric patterns, namely, under the person type, or the types whose subtypes indicate meaning transfer from their literal meaning to a (metaphoric) human centric meaning as discussed below. Here, the type attribute is an example of such cases.

a. Metaphoric patterns under person: we found under the type person 106 patterns. Some of these patterns are specialization polysemy patterns and others are metaphoric. To determine metaphoric patterns under the type person, we searched for inconsistency between the subtypes of the patterns. For example, we find such inconsistency in the pattern person#<bad person, worker>. The subtype bad person is not consistent with the type worker and, therefore a specialization polysemy is totally excluded in this pattern. The term iceman is an example of terms that belong to this pattern:

```
#1 iceman: someone who cuts and delivers ice.
#2 hatchet man, iceman: a professional killer.
```

On the other hand, the subtypes of the pattern person#<expert, worker> are consistent and is considered as a specialization polysemy pattern. The term technician is an example for this pattern:

#1 **technician**: someone whose occupation involves training in a technical process.
 #2 **technician**: someone known for high skill in some intellectual or artistic technique.

b. Metaphoric patterns under attribute: our criteria here were to find meaning transfer between the subtypes. Attribute has the following four patterns: attribute#<property, trait>, attribute#<property, state>, attribute#<property, quality>, and attribute#<quality, trait>, with the following meanings:

Property: a basic or essential attribute shared by all members of a class.

State: a state of depression or agitation.

Quality: an essential and distinguishing attribute of something or someone.

Trait: a distinguishing feature of your personal nature.

The meaning transfer from property to human centric meaning is clear in the first three patterns. For example, in the term chilliness:

#1 **chilliness**, coolness, nip: the property of being moderately cold.

#2 coldness, frigidness, iciness, **chilliness**: a lack of affection or enthusiasm.

In the fourth pattern, the relation between quality and trait depends on whether the term under the quality subtype refers to *an attribute of something*, or *an attribute of someone*. The first case corresponds to metaphoric polysemy, while the second corresponds to specialization polysemy.

3. Homonymy Patterns: in general, homonymy cannot be considered as a type of regular polysemy. Nevertheless, we cannot exclude the existence of homonymy patterns. WordNet contains few homonymy patterns such as the following pattern: *organism#<animal, plant>*, where we find type mismatch between the subtypes. Specialization or metaphoric polysemy in such patterns is totally excluded.

4. Mixed patterns: this group contains the patterns that were identified to have more than one polysemy type. For example, the pattern attribute#<quality, trait> belongs to this group.

In summary, there are some patterns whose sub types indicate type inconsistency. After excluding these patterns, all patterns under the physical *entity* are candidates for specialization polysemy except the patterns under *person*, which contains both polysemy types. In the case of *abstract entity*, most of the patterns under attribute are candidates for metaphoric polysemy. The patterns under *cognition* and *communication* contain both polysemy types, and the rest types are candidates for specialization polysemy.

C. Polysemy type Assignment

In this phase, the terms are assigned to the polysemy type of the pattern they belong to. The terms that belong to singleton and mixed patterns are not assigned and they are subject to manual treatment in the validation phase. The terms that belong to specialization polysemy patterns are

assigned to the polysemy operation based on the synset synonyms as described in Section IV. Formally, we classify the specialization polysemy cases into three sub classes as follows.

Definition 10 (missing parent synsets sub class) Let $\langle Ts, s_1, s_2 \rangle$ be a specialization polysemy case. Let $Terms_1 = s_1.Terms$, $Terms_2 = s_2.Terms$. The synsets s_1, s_2 are considered to belong to the missing parent sub class, if the following holds: $Terms_1 \setminus (Terms_1 \cap Terms_2) \neq \emptyset \wedge Terms_2 \setminus (Terms_1 \cap Terms_2) \neq \emptyset$.

Definition 11 (missing relation synsets sub class) Let $\langle Ts, s_1, s_2 \rangle$ be a specialization polysemy case. Let $Terms_1 = s_1.Terms$, $Terms_2 = s_2.Terms$. The synsets s_1, s_2 are considered to belong to the missing relation synsets sub class, if the following holds: $(Terms_1 \subset Terms_2 \wedge Terms_1 \neq Terms_2) \vee (Terms_2 \subset Terms_1 \wedge Terms_1 \neq Terms_2)$.

Definition 12 (merge synsets sub class) Let $\langle Ts, s_1, s_2 \rangle$ be a specialization polysemy case. Let $Terms_1 = s_1.Terms$, $Terms_2 = s_2.Terms$. The synsets s_1, s_2 are considered to belong to the merge synsets sub class, if the following holds: $Terms_1 = Terms_2$.

D. Validation

In this phase, we manually validate the assigned polysemy type. This phase includes three tasks:

- 1. Validation of the assigned polysemy types**: we check whether each of the nouns belong to its assigned polysemy type.
- 2. Assigning the polysemy type**: for the terms that belong to the singleton and mixed patterns.
- 3. Excluding of false positives**: we exclude the false positives from the terms of the automatic assigned groups. Our judgments during the validation are based on knowledge organization. Word etymology and linguistic relatedness have secondary role.

In Table III, we show the results of our validation for sample patterns. An example for false positives that we found in the common parent group: the meanings of term *apprehender* are homographs.

#1 *knower, apprehender*: a person who knows or apprehends.

#2 *apprehender*: a person who seizes or arrests.

TABLE III. SAMPLE PATTERNS VALIDATION

| # of instances | Pattern | Assigned polysemy Type | # of False positives |
|----------------|---------------------------------|------------------------|----------------------|
| 2025 | Common Parent | Spec. polysemy | 93 |
| 164 | attribute#property,quality | Metaphoric | 7 |
| 88 | attribute#quality,trait | Metaphoric | 4 |
| 45 | vascular plant#herb,woody plant | Spec. polysemy | 1 |

| | | | |
|-----|-----------------------------|-------------------|----|
| 408 | event#act,happening | Metonymy | 29 |
| 328 | act#action,activity | Metaphoric | 0 |
| 21 | artifact#commodity,covering | Spec. polysemy | 10 |
| 56 | attribute#property,trait | Metaphoric | 0 |
| 19 | animal#invertebrate,larva | Spec. polysemy | 0 |
| 26 | woody plant#shrub,tree | Spec. polysemy | 0 |

E. Applying the polysemy relations and operations

In this phase, we annotate the resulting metaphoric and homonymy cases explicitly as described in Section IV. For resulting specialization polysemy cases, we apply for each case one of the following three specialization polysemy operations according to the specialization polysemy sub group it belongs to.

1. **Adding missing parent operation:** Let s_1, s_2 be a missing parent case. Let $T_1 = \{t_{11}, \dots, t_{1n}\}$, $T_2 = \{t_{21}, \dots, t_{2m}\}$ be the synonyms of s_1 and s_2 respectively. Let $T_p = T_1 \cap T_2$, $T_{np1} = T_1 \setminus T_p$, $T_{np2} = T_2 \setminus T_p$. Let t be the preferred term of T_p . Let r be the common root of s_1 and s_2 . Let t' be the preferred term in r . We create a common parent S_p of s_1 and s_2 as follows:
 - i) Create a new synset S_p such that:
The lemmas $S_p = T_p$;
The gloss of $S_p = t$ is a " ". t' .
 - ii) The lemmas of $s_1 = T_{np1}$
 - iii) The lemmas of $s_2 = T_{np2}$
 - iv) Connect S_p to r via the *is-a* relation
 - v) Connect s_1 to S_p via the *is-a* relation
 - vi) Connect s_2 to S_p via the *is-a* relation
 - vii) Remove redundant relations.

In Figure 4, we show an example for adding a missing parent. In this figure, the hierarchical relations between the two synsets and their hypernym synset in (a) are considered to be redundant and removed in (b).

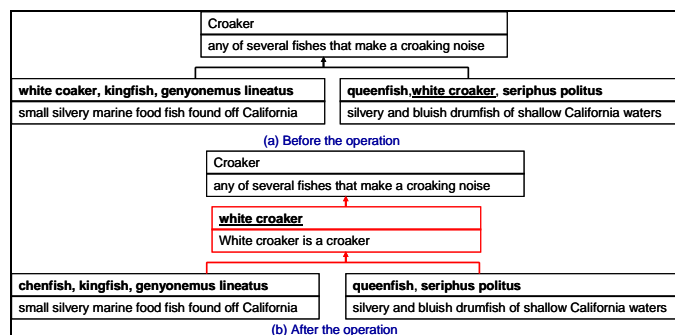


Figure 4. Example for adding a new missing parent

Adding missing relation operation: Let s_1, s_2 be a missing relation case. Let $T_1 = \{t_{11}, \dots, t_{1n}\}$, $T_2 = \{t_{21}, \dots, t_{2m}\}$ be the synonyms of s_1 and s_2 respectively. Let $T_p = T_1 \cap T_2$, $T_{np1} =$

$T_1 \setminus T_p$, $T_{np2} = T_2 \setminus T_p$. Let s_1 be the synset such that $T_1 \setminus T_p = \phi$.

- i) Connect s_1 to s_2 such that s_2 is $_a$ s_1 .
- ii) The lemmas of $s_2 = T_{np2}$.
- iii) Remove redundant relations.

In Figure 5, we show an example for adding a missing relation.

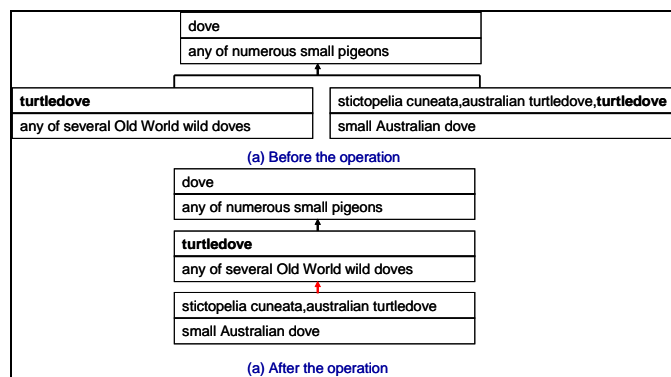


Figure 5. Example for adding missing relation

2. **Merge operation:** Let s_1, s_2 be two synsets of a merge case. We keep the synset with senses rank 1 as follows.
 - i) The gloss of s_1 = the gloss of s_1 " ". the gloss of s_2 .
 - ii) The relations of s_1 are the union of the relations of both synsets.
 - iii) Remove redundant relations.

In Figure 6, we show an example for merge operation.

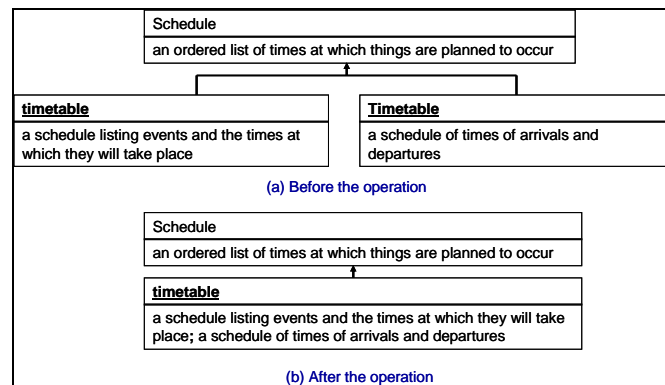


Figure 6. Example for merge operation

VI. OVERLAPPED TERMS SYNSETS

The relation between terms and synsets in WordNet is many to many. This means that it is possible for a term, or a synset, participates in more than one polysemy relation or operation of the same type (e.g., specialization polysemy operation). Considering such cases is very important, since the specialization polysemy operations make changes in the ontological structure and the synset synonyms. Changes in the ontological structure affect the structural patterns, and

changes in the synset synonyms affect the criteria for determining the polysemy operations between the synsets in specialization polysemy cases. The relation between specialization polysemy synsets is a binary relation and the specialization polysemy operations are applied pair wise.

To guarantee the correctness of the operations in cases of overlapped terms and synsets, we need rules for structural patterns construction and rules that control the order in which the operations are applied. In the following, we explain the strategy we are using to avoid multiple solutions and enforce the correct organization of such cases. The following figure represents an extreme case of overlapped terms and synsets.

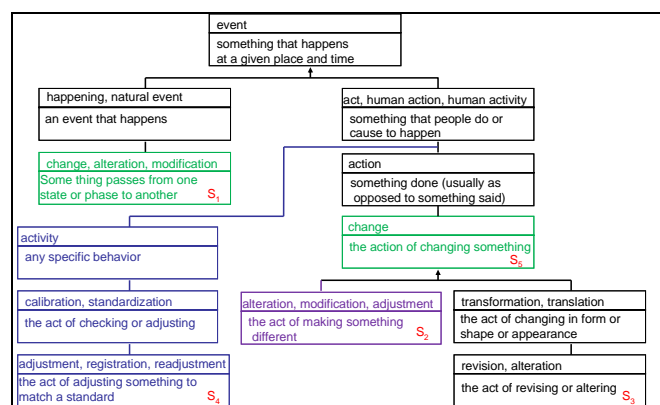


Figure 7. Example for Overlapped terms and synsets

In this example, we can see the following overlapped terms: The terms *alteration* and *modification* are found in s_1 and s_2 . The term *alteration* in s_1 , s_2 , and s_3 . At the same time we find the term *change* in s_1 and s_5 and the term *adjustment* in s_2 and s_4 . The synset s_2 participates in two operations: a missing parent operation in s_1 , s_2 and another missing parent operation in s_2 , s_4 .

Another important issue in the case of synset overlapping is that it is possible for a synset to follow more than one structural pattern. For example, the synset s_3 may follow the following patterns:

1. s_1 , s_3 belong to the pattern $\text{event} \# \langle \text{happening, act} \rangle$
2. s_2 , s_3 belong to the common parent *change*
3. s_4 , s_3 belong to the pattern $\text{act} \# \langle \text{action, activity} \rangle$

To handle this case and similar cases, we propose the following rules:

- **Patterns constructing rule:** construct the patterns bottom up and consider the first possible pattern. Applying this strategy, s_3 has only one pattern, which is the common parent with s_2 . At the same time s_2 belongs to two patterns, because the shared terms in s_2 , s_3 and s_2 , s_1 are different.
- **Operation Priority rules:** the operations are applied according to the following priority rules:

Synset level rule: apply the operations in a top down manner. For example, following this rule, we apply the operation on s_1 , s_5 before the operation on s_1 , s_2 .

Number of shared Terms rule: order the operations according to the number of shared terms. Following this rule, we apply the operation on s_1 , s_2 before the operation on s_2 , s_3 . The operations on s_2 , s_3 and the operation on s_2 , s_5 have the same priority.

Resulting changes rule: in case a synset is participating in more than one operation, the type of operation may change according to resulting changes from previous operations. For example, the operation on synset s_2 , s_4 is missing parent operation. The result of the operation on s_1 , s_2 , which is applied before the operation on s_2 , s_4 leads to changing the operation on s_2 , s_4 from missing parent operation to a missing relation operation.

Through the patterns construction rule, we guarantee two important aspects of specialization polysemy relatedness:

- **Specificity of terms:** the terms whose patterns can be constructed in the lower level are more specific than the terms who have other overlapping patterns in a higher level in the ontology.
- **Meaning relatedness:** through the bottom-up construction and discarding the higher level overlapped patterns, we guarantee that the captured meanings belong to the same, or very near ontology level and thus the meanings are more related than those of the discarded patterns.

We explain the importance of the operation priority rules as follows:

- **Synset level rule:** the synsets whose pattern is found at higher level of the ontology are usually a more general meanings. Thus applying the rules in a top level fashion guarantees that the results do not conflict with the original ontological structure. For example, it is clear that s_5 is a more general meaning of s_1 . Applying the missing parent operation on s_1 , s_2 before the missing relation operation on s_1 , s_5 , makes it impossible to keep the original relatedness between s_1 and s_5 .
- **Number of shared terms rule:** the rationale behind this rule is that synsets that share more terms are more related. We find evidence for this rule in polysemy reduction approaches. In polysemy reduction approaches, they consider synsets that share more than one term to be merged.
- **Resulting changes rule:** this rule is a consequence of the previous rules. We sort the rules according to the synset level rule and number of shared terms rule. These rules guarantee the correctness of the resulting

changes. Since we apply the operations pair wise, each of these operations affects the ontological structure and the terms of the operated synsets. Thus the subsequent operations should be applied on the resulting structure of the previous operations.

In Figure 8, we show the final result of applying the operations on the synsets in Figure 7.

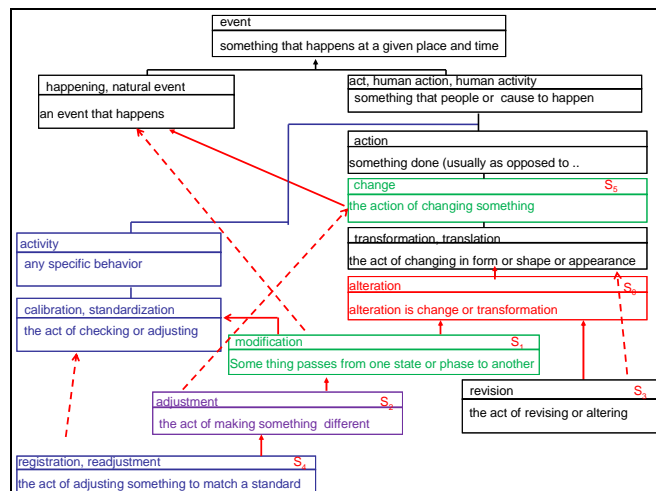


Figure 8. Solving overlapped terms and synsets

In Figure 8, the red colored lines and synsets are newly added. The dashed red lines are the removed relations. We apply the operations in the following order:

1. Missing relation operation on s_1 , s_5 (according to the synset level rule). This affects s_1 and s_5 in the following way. We connect s_1 to the synset *happening*. The synset s_1 now is a subtype of s_5 and the term *change* is removed from s_1 .
2. The operation on the synsets s_1 and s_2 has changed now to a missing relation instead of the original operation missing parent.
3. We apply the missing relation operation on s_1 , s_2 (according to the number of shared terms rule). The synset s_1 is connected to the synset *calibration* and s_2 is connected to s_1 . The relation between s_2 and the synset *change* is removed due to the relation redundancy rule. The terms *alteration* and *modification* are removed from s_2 .
4. The operation on s_2 , s_4 has changed to missing relation instead of the original missing parent. There is no change in the operation s_2 , s_3 .
5. Missing parent operation on s_2 , s_3 . This leads to creating a new synset s_0 . The synset s_0 has the term *alteration* only. The synsets s_2 and s_3 are connected to s_0 . The relation between s_3 and *transformation* is removed due to redundancy rule.
6. Missing relation operation on s_2 , s_4 . The term *adjustment* has been removed from s_4 .

VII. RESULTS AND EVALUATION

For the manual validation described in Section V and the evaluation process described in this section, we have developed a special user interface (see Figure 9). This user interface provides the local view of the polysemy cases. For each polysemy case, the user can view also the polysemy type of the displayed polysemy case and the polysemy operation (applicable for specialization polysemy cases). The user can then agree with the suggested polysemy type/polysemy operation or he can choose one of the provided alternative polysemy types. If the user can not decide, he can choose "No decision".

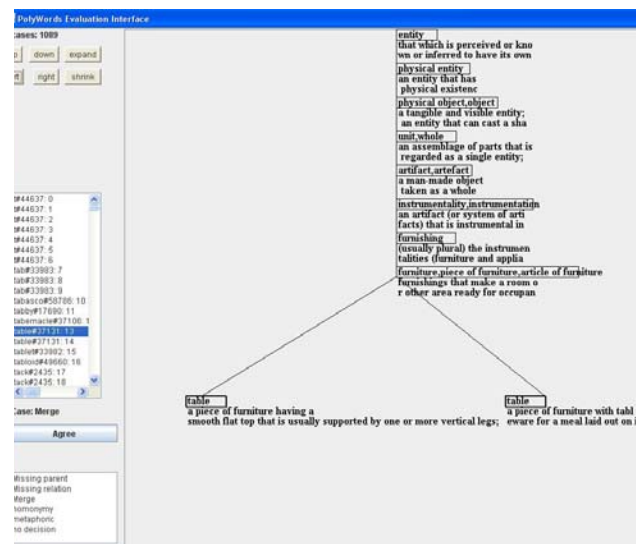


Figure 9. Polysemy Evaluation Interface

In Table IV, we present the results of our approach after the manual validation.

TABLE IV. VALIDATED RESULTS OF THE ALGORITHM

| Polysemy type | # of cases | # of cases in percentage (%) |
|-----------------------|------------|------------------------------|
| Metaphor | 1040 | 8.2 |
| Homograph | 1054 | 8.3 |
| Spec. Polysemy | 10200 | 80.7 |
| Systematic and Others | 361 | 2.8 |
| Total | 12655 | 100 |

The cases in the column systematic and others are the cases that we think that they should be processed in a subsequent phase of our approach in the framework of approaching CORELEX systematic polysemy or cases, where the presence of the polysemous term in one of the synsets is inappropriate and should be removed from one of the synsets. An example for such cases is the term *senate* that appears in the synset and its direct hypernym:

United States Senate, U.S. Senate, US Senate, Senate: the upper house of the United States Congress.

=> senate: assembly possessing high legislative powers

In Table V, we present the classification of specialization polysemy. The total number of reduced polysemous cases is 10200. The total number of merged synsets represents about

18% of the total processed cases. At the same time we have added 4212 new synsets and 8293 new *is a* relations, while have deleted 1907 synsets and 1907 *is_a* relations. This means that in our approach we have increased knowledge rather than decreasing knowledge to solve the polysemy problem.

TABLE V. SPECIALIZATION POLYSEMY RESULTS

| | # of words | # of words in percentage (%) |
|------------------|------------|------------------------------|
| Missing parent | 4212 | 41.3 |
| Missing relation | 4081 | 40 |
| Merge | 1907 | 18.7 |

To evaluate our approach, 1020 cases have been evaluated by two evaluators. In Table VI, we report the statistics of the evaluation, where the column polysemy type refers to homonymy, metaphoric, metonymy, or specialization polysemy and the column polysemy operation refers to creating missing parent, adding missing relation, or merging operation. Note that, polysemy operation is applicable in case of specialization polysemy. The table presents the agreement between the evaluators and our approach. The third row represents the number of cases, where at least one evaluator agrees with our approach.

TABLE VI. EVALUATION RESULTS

| | Polysemy type agreement | Polysemy operation agreement |
|-------------------|-------------------------|------------------------------|
| Evaluator 1 | 979 \approx 96% | 924 \approx 90.5% |
| Evaluator 2 | 945 \approx 92.5% | 855 \approx 84% |
| Partial agreement | 1006 \approx 98.5% | 978 \approx 96% |

As we can see from the results above, although the agreement with the approach is high, in many cases, the evaluators agree on the specialization polysemy type but disagree on the operation type. The explanation for this is that the operation is decided according to the nature of lemmas in both synsets as explained in Section IV.

VIII. CONCLUSION AND FUTURE WORK

In this paper, we introduced a pattern based approach for solving the polysemy problem in WordNet. Our approach deals and covers all polysemy cases at all ontological levels of wordNet. Furthermore, it improves the ontological structure of WordNet by transforming the implicit relations between the polysemous senses at lexical level into explicit semantic relations. The manual treatment in two phases of the approach guarantees the quality of the approach result. Beside the manual evaluation of our approach, we plan to run an indirect evaluation to test the effects of our approach in terms of precision and recall. We applied our approach on all polysemous nouns in WordNet. In a subsequent phase, we are going to extend our algorithm to handle verbs, adjectives and adverbs.

The main contributions of this work are at two levels: at the conceptual level, we are providing a new foundation towards the problem of polysemy. At the implementation level, we aim to improve the quality of NLP and

knowledge-based applications, especially in the field of the semantic search.

REFERENCES

- [1] A. A. Freihat, F. Giunchiglia, and B. Dutta, "Approaching Regular Polysemy in WordNet," in proceedings of the 5th International Conference on Information, Process, and Knowledge Management (eKNOW), 2013, Nice, France, pp. 63-69.
- [2] G. A. Miller, "WordNet: a lexical database for English," Communications of the ACM 38 (11), November 1995, pp. 39 – 41.
- [3] J. Gonzalo, "Sense Proximity versus Sense Relations," Proc. of the Second Global WordNet Conference, Brno, Czech Republic, January 20-23, 2004, pp. 5-6.
- [4] P. P. Buitelaar, "CORELEX: Systematic Polysemy and Underspecification," PhD thesis, Brandeis University, Department of Computer Science, 1998.
- [5] R. Mihalcea and D. I. Moldovan, "EZ.WordNet: Principles for Automatic Generation of a Coarse Grained WordNet," FLAIRS Conference, 2001, pp. 454-458.
- [6] L. Barque and F. R. Chaumartin, "Regular Polysemy in WordNet", JLCL, vol. 24, no. 2, 2009, pp. 5-18.
- [7] B. Nerlich and D. D. Clarke, "Polysemy and flexibility: introduction and overview," B. Nerlich, Z. Todd, V. Herman and D. D. Clarke (Hg.), Polysemy: Flexible Patterns of meaning in Mind and Language, Berlin, New York: Mouton de Gruyter, 2003, pp. 3-29.
- [8] J. Gonzalo, I. Chugur, and F. Verdejo, "Sense clusters for Information Retrieval: Evidence from Semcor and the EuroWordNet InterLingual Index," ACL-2000 Workshop on Word Senses and Multi-linguality, Association for Computational Linguistics, pp. 10-18.
- [9] J. Pustejovsky, The Generative Lexicon, Cambridge: MIT Press, 1995.
- [10] R. Mihalcea, "Turning WordNet into an Information Retrieval Resource: Systematic Polysemy and Conversion to Hierarchical Codes," IJPRAI, vol. 17, no. 5, 2003, pp. 689-704.
- [11] W. Peters, "Detection and Characterization of Figurative Language Use in WordNet," PhD thesis, Natural Language Processing Group, Department of Computer Science, University of Sheffield, 2004.
- [12] R. Navigli, "Word sense disambiguation: a survey," ACM Comput. Surv., vol. 41, no. 2, 2009.
- [13] N. Tomuro, "Systematic Polysemy and Inter-Annotator Disagreement: emirecal Examinations,"

Proc. of the First International Workshop on Generative Approaches to Lexicon, 2001.

- [14] M. Palmer, H. T. Dang and C. Fellbaum, "Making fine-grained and coarse-grained sense distinctions, both manually and automatically," *Natural Language Engineering (NLE)*, vol. 13, no. 2, 2007, pp. 137-163.
- [15] N. Verdezoto and L. Vieu, "Towards semi-automatic methods for improving WordNet", *Proc. of the 9th International Conference on Computational Semantics*, Oxford, UK, 2011.
- [16] J. Apresjan, "Regular Polysemy," *Linguistics*, vol. 142, 1974, pp. 5-32.
- [17] W. Peters and I. Peters, "Lexicalized systematic polysemy in WordNet," *Language Resources and Evaluation*, 2000.
- [18] T. Veale, "A Non-Distributional Approach to Polysemy Detection in Wordnet," doi:10.1.1.146.5566.

A Human Surface Prediction Model Based on Linear Anthropometry

Ameersing Luximon, Huang Chao
Institute of Textiles and Clothing
The Hong Kong Polytechnic University
Hong Kong
E-mail: tcshyam@polyu.edu.hk,
huang.chao@connect.polyu.hk

Yan Luximon
School of design
The Hong Kong Polytechnic University
Hong Kong
E-mail: sdtina@polyu.edu.hk

Abstract—Body information is needed in product design, medical, archaeological, forensic and many other disciplines. Therefore, anthropometric studies and databases have been developed. Anthropometric measures are useful to some extent, but due to technological innovations, there is a shift toward surface anatomy. As a result, there is a need to shift from linear anthropometry tables to surface model databases. This study provides a general modelling technique, to convert linear anthropometry to complex surface model using recursive regression equations technique (RRET) and scaling technique. The technique makes use of similarities and differences between people. The similarities or standard shape are represented by averaging, while the differences are captured by using anthropometric measures. In order to build the surface model, some scanned data is needed for generating the standard shape. Using RRET techniques a few anthropometric measures are used to predict more anthropometric measures that are then used to scale the standard shape in order to generate a predicted 3D shape. Results indicate that the prediction model is accurate to few millimeters. This level of error is acceptable in different applications. This technique can be applied to generate 3D shape from anthropometry of external shape as well as internal organs. This model is essential to convert the existing large scale anthropometric databases into surface models. It can be applied to product design, sizing and grading, reconstructive surgery, forensic, anthropology and other fields.

Keywords—anthropometry, surface antropometry, digital human model, recursive regression equation.

I. INTRODUCTION

Shape modelling from linear anthropometry is a new field of study with important applications [1]. It takes advantages of anthropometry, is an old field of study dating back to Renaissance [1,2] and merging it with new technologies to create human shape models. Luximon and Chao [1] proposed a basic model for shape modelling from linear anthropometry. This paper expands the previous paper to including validation of the model on foot model.

Anthropometry emerged in the nineteenth century largely by German investigators in the physical anthropology discipline, while they needed to study the quantitative description of the human body reliably [3,4]. Anthropometry techniques can be applied to humans as well as plants and animal. According to the World Health Organization [5], Anthropometry is a method to assess size proportion and

composition of the human body. Some of anthropometric measures include weight, lengths (e.g., foot length), widths (e.g., Head width), heights (e.g., Stature), girths (e.g., Chest Girth), angles (e.g., hand flexion angle), and calculated indexes (e.g., Body mass Index (BMI)). Anthropometric studies are carried out internationally because it requires inexpensive, non-invasive simple tools such as rulers, tapes, callipers and goniometers. These tools can be easily transported to any location. World Health Organization [5], uses anthropometric measures to assess medical conditions of people by comparing with average values.

The basic anthropometric techniques developed during the nineteenth century are still used today [3]. The different anthropometric measures are represented in percentile values in anthropometric tables [2,6] and since the values are statistical values, they cannot be combined to create a single human body [7]. The anthropometric percentage values are generally used to compare different populations and to design for a given population. Anthropometric data has been widely used in fields ranging from engineering to arts. It has been widely used in product and workplace design [7,8] to determine sizing, grading, proper fit and comfortable design based on different body sizes and proportions. It has been used in forensic investigation [9,10] for better estimation and narrow down the forensic search. It has been used in growth and nutrition evaluation [5,11] worldwide to check malnutrition, proper growth and early detection of growth and nutrition problems. It has been used in medicine [12-14] and reconstructive surgery for screening problem, evaluations and corrections. It has been used in archaeology and cultural studies [15-17] to identification and classification. It has been used in many other studies including sports science and fitness evaluation [18-20], since the body proportion and composition is different for different sports.

Even now, anthropometric studies are carried out because of its non-invasiveness [5], inexpensiveness [4], simplicity [5], portability [21] and reliability [22]. Furthermore, as the anthropometric dimensions vary among different groups of population, anthropometric tables have been developed based on age, race, region, and occupation [5]. Anthropometric studies are very important; however, in many applications, anthropometric data alone is not sufficient. Surface geometry is required. Recently, due to

technological innovations, it is possible to acquire whole body or selected body parts surface data.

Surface model describe the size and shape as well as the 3D surface geometry of the human body [23]. It is possible to combine surface scan data and internal measurements [24], thus anthropometric techniques can be used to find the size, shape and proportion of the external as well as internal structures of the body. Thus in this modern world, data collected from Magnetic Resonance Imaging (MRI), Computed Tomography (CT) or Computed Aided Tomography (CAT), sound, optical (laser or structured light) or any scanning devices can be used to create surface model of the external as well as the internal structures of the body. Since surface model provides information on the complex surfaces of the human parts, in addition to the common linear anthropometry measures, it can be used for many applications such as planning and assessment of facial surgery, design and manufacture of implants and prostheses, facial reconstruction in forensic applications, archaeology, psychology, genetics, and comparative and evolutionary anatomy [25]. In addition, there are growing uses of synthesize 3D digital animated images of human models in science fiction movies and 3D digital dummies for equipment testing [26]. Although surface model seems to be very useful, it has several disadvantages. Surface scanning equipment is relatively expensive and is not widely available. It is relatively difficult to operate and require special skilled technicians to capture the dynamic and complex body shape. Furthermore, the data obtained from the scanning equipment requires additional processing and statistical data analyses are not trivial resulting in only few large-scale studies. Still, surface model is very useful for different applications and there is a need to simplify the method to acquire surface model, reduce the cost of equipments and develop surface model databases. In this study, a method to predict accurately the surface model by using simple, reliable and non-invasive linear anthropometry measurement techniques are proposed. As a result, the use of expensive surface scanning devices is minimized to model building and simple cost effective linear anthropometry measures can be used for surface model prediction.

II. RELATED WORK

The basic technique for 3D Surface (Surface model) prediction is to determine the similarities and then modify based on differences (Figure 1). For example, although we know there are variations in shape and sizes of face, we can easily know a human face from other body parts or other animals. There are similarities between faces. The similarities can be grouped based on age, gender, and culture. The differences are related to some dimensions which can be captured based on anthropometric studies. Surface model prediction involves spline curve and surface fitting [27], recursive regression equations [28,29] and scaling techniques [28]. In this study, a general prediction model with several variations is provided so that more body parts can be predicted using this simple method. Then, an example related to foot prediction is discussed. The main parts are data pre-processing and standard shape generation;

model development; surface model prediction; and prediction model validation.

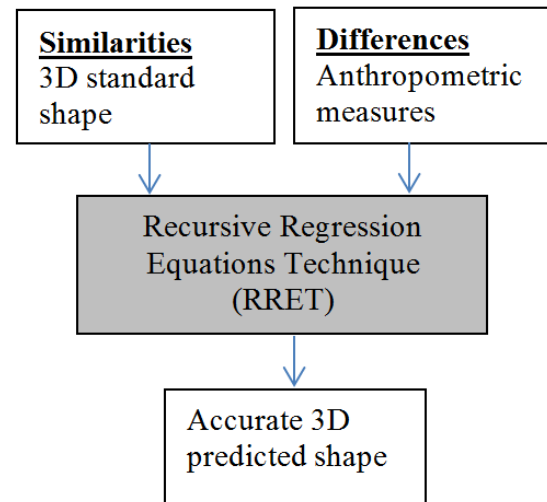


Figure 1. Method for surface prediction model.

The paper is organization in five major sections in order to develop a general surface model from linear anthropometry. These include data pre-processing; standard shape generation; model development; surface model prediction; and validation. During the model building a generic model method is used that can be applied to all body parts. For the validation, an accurate foot shape model prediction has been developed and error calculated. The conclusion and future work provide the importance of this method and its future application are further emphasized.

III. DATA PRE-PROCESSING

Most scanning technologies have error resulting in missing data or noisy data points. The flow chart for data preprocessing is shown in Figure 2. The steps involve surface data acquisition. Then the data is 'cleaned'. Human body shape has variations hence, there is a need for careful alignment of the data. The cleaned and alignment data provide a representation of the body shape, which may be affected by accuracy of scanning system, but in this study this data is considered the 'true' shape.

A. 3D scanning

For 3D scanning, any type of scanner to capture the external shape of the body or any specific part can be used. Figure 3 shows a whole body scanner at the Hong Kong polytechnic university. There are also specific scanners for head [29] and foot [28], since the whole body scanner does not provide accurate data for these extremities (Figure 4). Since a general method to build the prediction model has been proposed, some changes may be required to adopt for specific parts. It is assumed that N_s number of participants is used for the model development. In this formulation, left and right sides of the parts are not distinguished, but during the

formation of a specific part, the differences between left and right sides can be included as in Luximon and Goonetilleke [29]. For the i^{th} participant the scanned part has P_{si} number of points. The points are p_{ik} (where $i = 1, \dots, N_s$; $k = 1, \dots, P_{si}$). The coordinates of the point p_{ik} is (x_{ik}, y_{ik}, z_{ik}) . The point cloud from the scanner is unstructured and includes hundreds of thousands of data points.

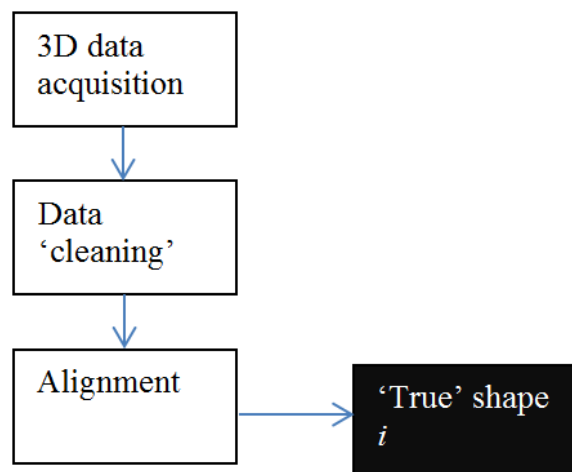


Figure 2. Flow chart for data preprocessing.



Figure 4. Laser scanned data.

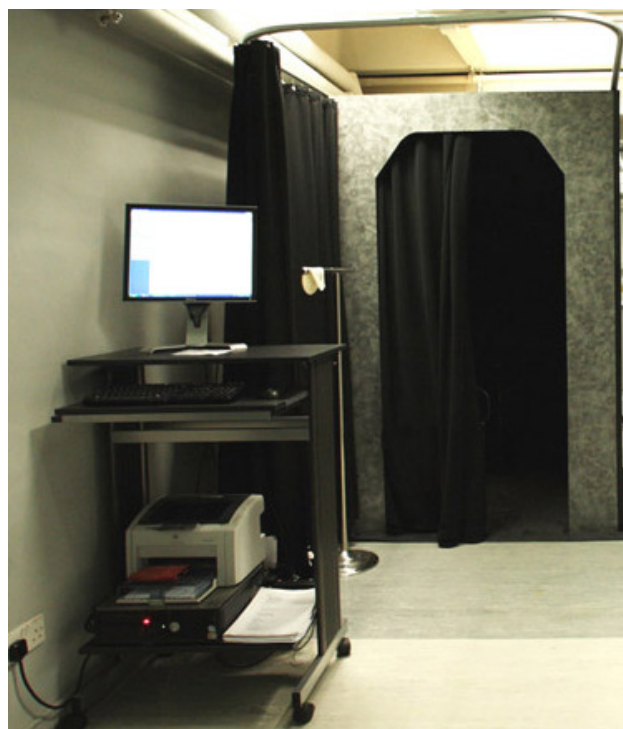


Figure 3. TC² whole body scanner (www.itc.polyu.edu.hk).

B. Data cleaning

Since scanning is usually disturbed by noises arising from various sources, the noise can be cleaned manually using software such as Rapidform (www.rapidform.com) or using algorithm methods such as Adaptive Moving Least Squares method [30]. Furthermore, there are cases of missing data that are filled using commercial software Rapidform2006 software. The points after data cleaning is represented by p_{ik} (where $i = 1, \dots, N_s$; $k = 1, \dots, P_i$). P_i is the number of points.

C. Alignment

Since all the scanned part might not be aligned in the same reference axis, all the parts have to be aligned on a consistent axis. The axis of alignment can be based on some anthropometric landmarks, commonly used axis or based on mathematical and statistical methods (such as principle component methods). For example, for the case of the human foot, heel centre line is commonly used [31]. For the arm, leg and body principal component can be used. For head data, eye and ear landmarks can be used for alignment [32]. After alignment, the coordinates of point $a p_{ik}$ is $(x_{ik,a}, y_{ik,a}, z_{ik,a})$ as shown in Figure 5. The part is aligned to have the axis with the highest variation along the Z-axis.

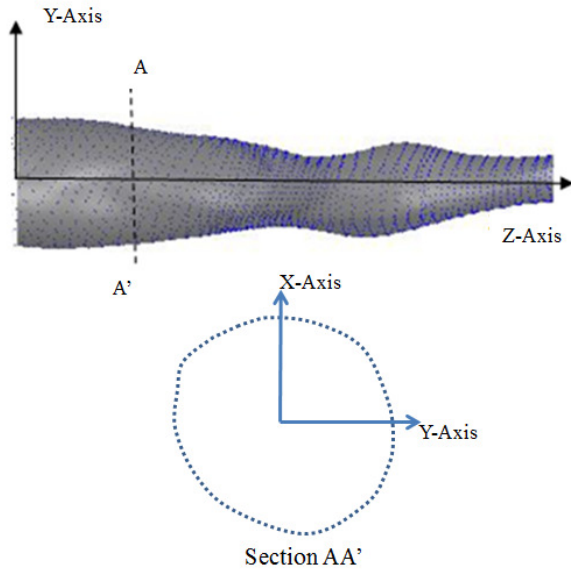


Figure 5. Aignment.

IV. STANDARD SHAPE GENERATION

In order to generate the standard shape, first the scanned data that has been cleaned and aligned ('true' shape) is sectioned and sampled (Figure 6). Sampling creates same number of points for all participants. The participants' data are then averaged to create a standard shape.

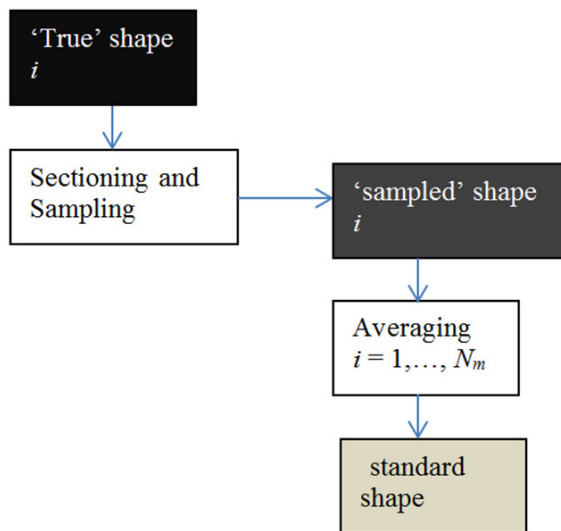


Figure 6. Flow chart to create standard shape.

A. Sectioning and sampling

During sampling fixed number of points is created for all participants. Different sampling methods can be used. When using cylindrical coordinate system, the part can be sectioned

along the Z-Axis, which represents the maximum variation (Figure 7).

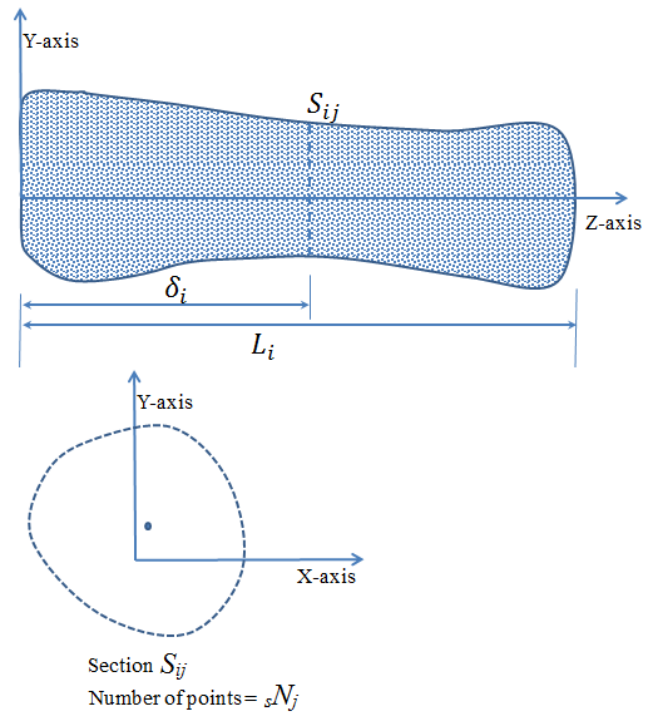


Figure 7. Sectioning and sampling in cylindrical coordinate system.

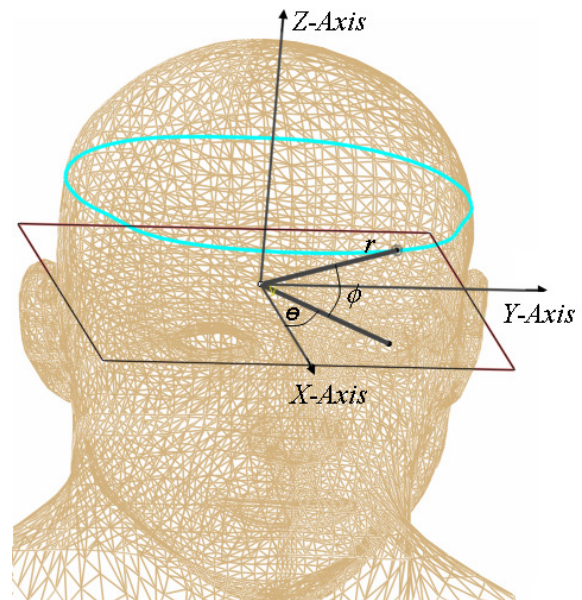


Figure 8. Sectioning and sampling in spherical coordinate system.

Cylindrical coordinate system can be used for most body parts. Once the part has been aligned, cross sections are extracted perpendicular to the Z-axis, called the 'main' axis. The length of the aligned part along the main axis is L_i (where $i = 1, \dots, N_s$). Cross-sections perpendicular to the main

axis at δ_j (where $j = 1, \dots, N_{sec}$) of L_i are extracted, where N_{sec} is the total number of cross-sections extracted (Figure 7). δ_j is monotonically increasing with j . The separation between the sampled cross sections needs not be uniform, but it has to be consistent between the different participants. The extracted sections are S_{ij} (where $i = 1, \dots, N_s$; $j = 1, \dots, N_{sec}$) and the z -value for the sections are given by Equation (1). Then, for each section, a fixed number of points are extracted using different sampling methods [30]. When using spherical coordinate system, the shape can be sectioned based on angle β (Figure 8). Spherical coordinate system is more appropriate for the head and face model [32].

Once the shape has been sectioned either using cylindrical coordinate system or spherical coordinate system, data points are extracted from the sections. Uniform polar sampling at β degrees intervals (Figure 9) and uniform sampling at δ mm along the edge (Figure 10) are common methods. The number of points for section S_{ij} is sN_j . For participant i , the number of points is same. The points after sampling are sP_{ijk} , where $i = 1, \dots, N_s$; $j = 1, \dots, N_{sec}$; $k = 1, \dots, sN_j$.

$$sZ_{ij} = \delta_j * L_i \quad (1)$$

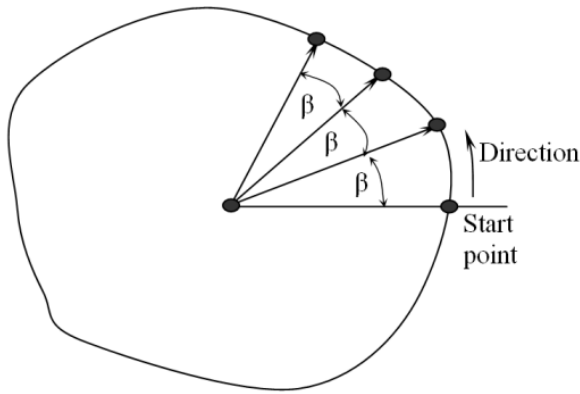


Figure 9. Polar sampling.

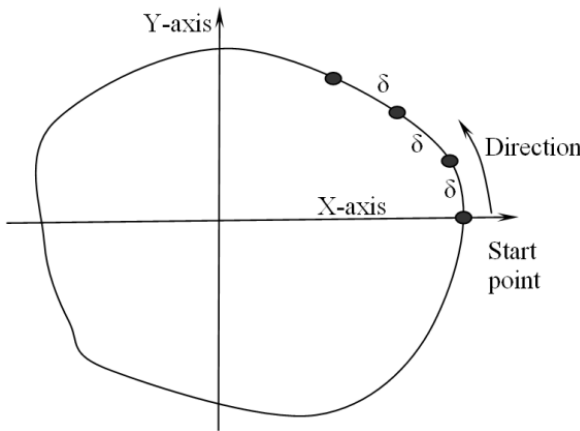


Figure 10. Uniform sampling.

B. Standard shape

Some of the part shape data can be used to generate the model, while other shape data can be used for model validation. Assuming that the standard model is generated using part shape data of N_m subjects where $N_m < N_s$. The coordinates of the point used to generate the standard part are $(sX_{ijk}, sY_{ijk}, sZ_{ijk})$ where $i = 1, \dots, N_m$; $j = 1, \dots, N_{sec}$; $k = 1, \dots, sN_j$. The standard shape is the representation of the given part for a given population. There can be several methods to generate the standard shape, based on different statistical methods such as geometric mean, arithmetic mean, mode, median etc. Equations (2), (3), and (4) show the x , y , and z coordinates of the standard shape when arithmetic mean is used. The standard foot shape has N_{sec} number of sections. The standard shape represents the shape of a given population and it can be stored in a database.

$$\bar{x}_{jk} = \frac{1}{N_m} \sum_{i=1}^{N_m} sX_{ijk} \quad (2)$$

$$\bar{y}_{jk} = \frac{1}{N_m} \sum_{i=1}^{N_m} sY_{ijk} \quad (3)$$

$$\bar{z}_{jk} = \frac{1}{N_m} \sum_{i=1}^{N_m} sZ_{ijk} \quad (4)$$

V. MODEL DEVELOPMENT

While the standard shape is being developed, parameters of each section can be extracted from the cross sections of the sampled shape. The flow chart for the RRET model development is shown in Figure 11. The number of parameters per section will determine the accuracy of the model. Parameters for a section can be length and widths. Then regression equations are developed to generate equation between parameter of one section to another section.

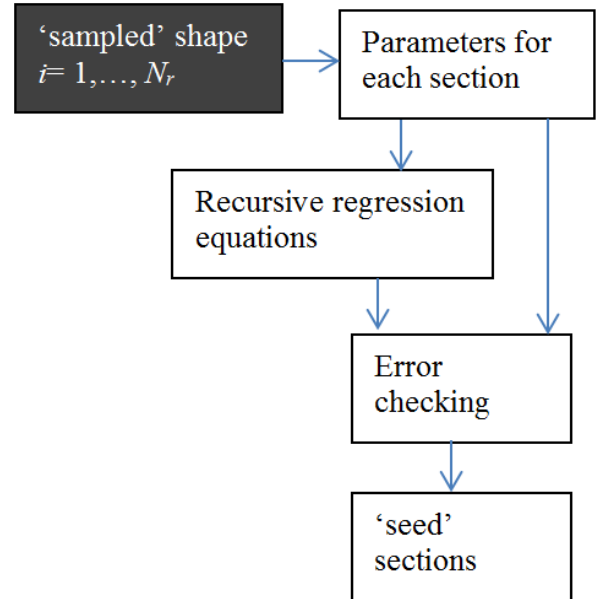


Figure 11. Flow chart for model development.

Since there are many equations between the parameters, if we know the value of one parameter we will be able to predict the value of other parameters. Hence there is a need to determine the best starting parameter or 'seed' parameter.

A. Parametization

Each cross section can be parameterized using several anthropometric variables. Figure 12 shows some of the parameters that can be used, such as maximum y deviation (H^+), minimum y deviation (H^-), maximum x deviation (W^+), minimum x deviation (W^-), height (H), width (W) and radius (R_θ) at θ degrees and circumference (C). The number of parameterization will determine the accuracy and complexity of the model. Furthermore, anthropometric studies are needed to determine the importance of the different parameters. Goonetilleke et al. [33] and Luximon and Goonetilleke [34] have used principle component and factor analysis to find the relative importance of different foot related parameters.

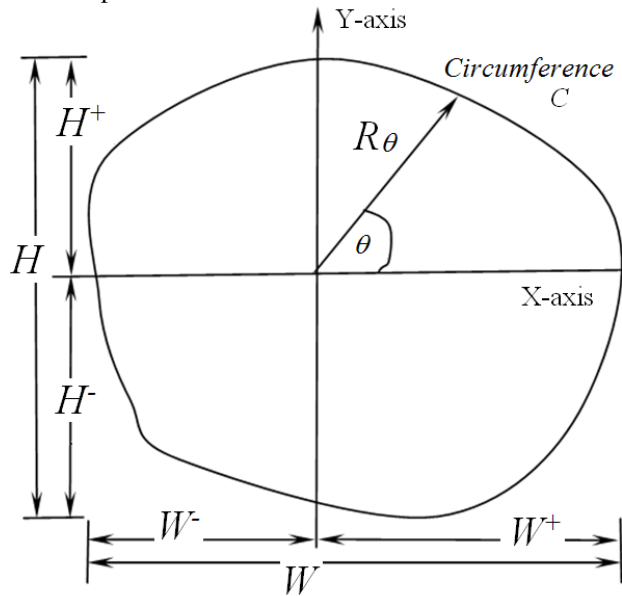


Figure 12. Anthropometric parametrization.

B. Recursive regression equation

The purpose of the recursive regression equation is to find the relationship of the anthropometric dimensions of all the sections of the part given the anthropometric dimension of one section. For example, one regression equation is build from anthropometric measure height (H) at section i and height at section j . The R^2 values are also recorded. If we have N_a anthropometric measures and N_{sec} sections, we can generate $N_a \times (N_{sec} - 1)$ equations if we consider consecutive sections. Using these regressions equations, knowledge of one set of values for N_a anthropometric measures ('seed section'), we will be able to predict all the $N_a \times N_{sec}$ anthropometric measure values.

C. 'Seed' section

There are some ways to find the best 'seed' section and build the regression equations. Luximon and Goonetilleke [29] developed linear regression equations between the anthropometric measures of adjacent sections. The best 'seed' section was found by using different 'seed' section to predict the anthropometric measures and choosing section that provided the highest correlation between the original set of anthropometric measures. For complex models $N_a \times (N_{sec} - 1) \times (N_{sec} - 2)$ equations may be needed. This problem can be solved using travelling salesman method [34]. The salesman need to travel between cities (1,2,...,8). The traveling distance between each city is given. The salesman needs to travel all the cities without repetition and take the least distance. In the case of the proposed method, the distance can be substituted by $1/R^2$.

VI. SURFACE MODEL PREDICTION

Using few anthropometric measures, the parameters of the seed section is predicted. The flow chart for shape prediction is shown in Figure 13. Using the parameter values of the seed section and the linear regression equations between the parameter of the sections, the parameter values of all the sections are predicted recursively. The predicted parameter values are used to scale the sections of the standard shape creating a predicted shape.

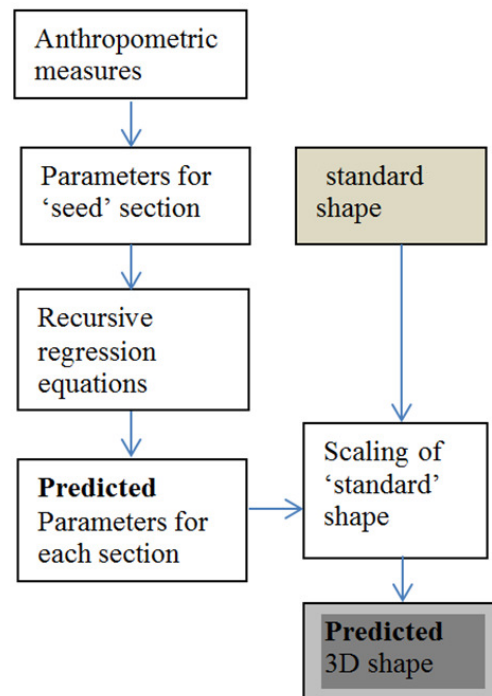


Figure 13. Shape prediction flow chart.

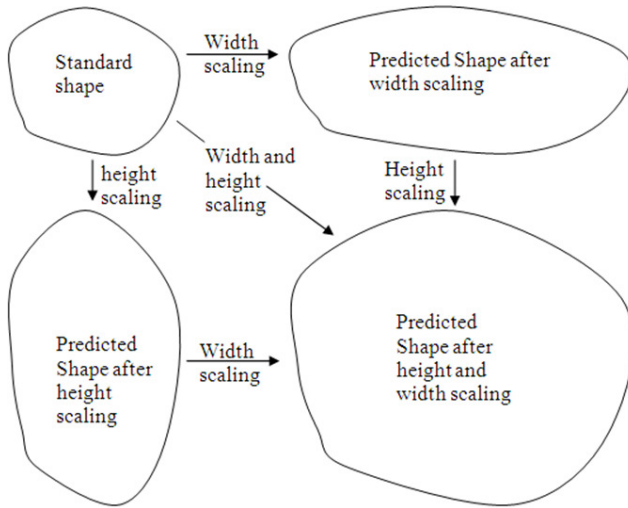


Figure 14. Scaling.

The model can be validated using 3D scanned data of a different set of N_v participants where $N_v < N_s$ and $N_v + N_m = N_s$. The model validation involves measurement or extraction of parameters of the 'seed' section, prediction of parameters of all the section based on the 'seed' section, scaling of the standard shape. Once the shape is predicted, the prediction error can be calculated when we compare it with the original data. Once we have the predicted parameters of the sections, the standard shape has to be scaled. There can be different scaling methods based on the different parameters. Luximon and Goonetilleke [29] have discussed proportional scaling. If the parameters are orthogonal (such as width and height) then the sections can be scaled independently (Figure 14). However, if the parameters are not orthogonal different scaling methods need to be developed. After scaling, the predicted shape has coordinates $(p_{xijk}, p_{yijk}, p_{zijk})$ where $i = 1, \dots, N_v; j = 1, \dots, N_{sec}; k = 1, \dots, sN_j$.

VII. PREDICTION MODEL VALIDATION

The flow chart for the validation of the predicted model is shown in Figure 15. The main component is the comparison of the 'true' shape with the predicted shape. For participant i the original shape after alignment has coordinates $(a_{xik}, a_{yik}, a_{zik})$, where $i = 1, \dots, N_v; k = 1, \dots, P_i$. The coordinates of the predicted foot is $(p_{xijk}, p_{yijk}, p_{zijk})$ where $i = 1, \dots, N_v; j = 1, \dots, N_{sec}; k = 1, \dots, sN_j$. The error is computed based on the shortest distance [Equation (5)] from the predicted foot to the real foot [35]. The error can have signed (+ or -) to indicate either the predicted point is inside or outside the original shape. Different statistics can easily be calculated to compare prediction accuracy. Error plots are also useful to show the error distribution at different regions [29].

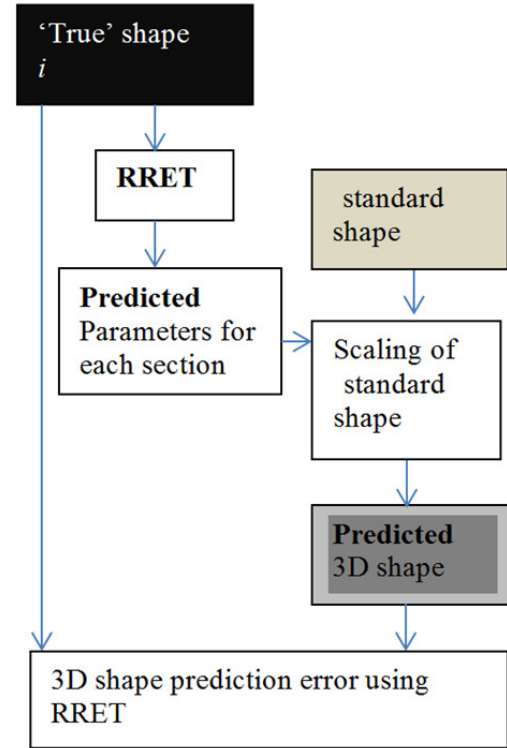


Figure 15. Error checking and validation.

$$e_{ijk} = \min \left\{ \sqrt{(p_{xijk} - a_{xil})^2 + (p_{yijk} - a_{yil})^2 + (p_{zijk} - a_{zil})^2} \right\}$$

where $i = 1, \dots, N_v; j = 1, \dots, N_{sec}; k = 1, \dots, sN_j; l = 1, \dots, P_i$. (5)

VIII. FOOT SHAPE MODELING

The accuracy of the RRET technique is illustrated by using foot modeling as an example. The data was collected using a foot scanner (Figure 16) shows the scanned foot data. P_i is about 100,000 points. The sampled foot is shown in Figure 17. The foot is sectioned at 1% interval creating 99 sections. The extracted sections for participant i are S_{ij} (where $j = 1, \dots, 99$). For each section points are sampled at 1 degree interval based on polar coordinate sampling. The points after sampling are p_{ijk} , where $j = 1, \dots, 99; k = 1, \dots, 360$.

Using data from 40 participants, a standard foot shape was created (Figure 18). The average age of the participants used for generating the standard shape was 22 years (standard deviation = 3.6). The average weight was 62.8 Kg (standard deviation = 8.5). The average stature was 171 cm (standard deviation = 5.5). The average foot length was 245mm (standard deviation = 11mm). The average foot width was 99 mm (standard deviation = 5mm).

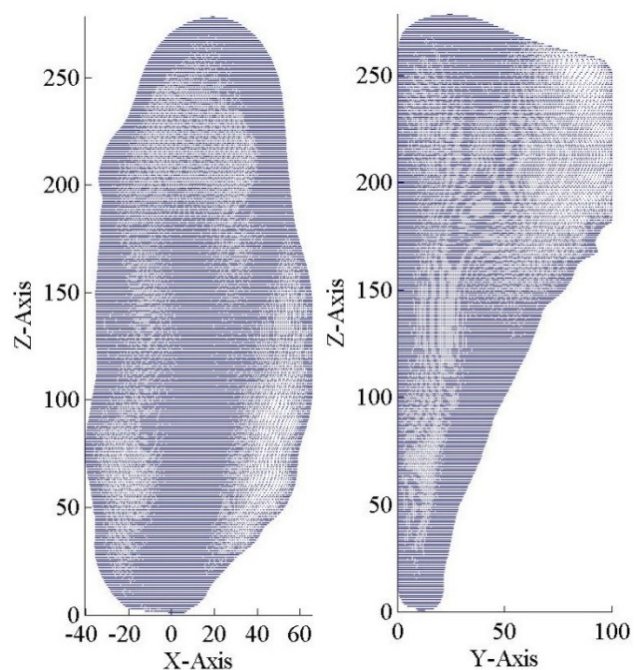


Figure 16. Foot laser scanned data (unit mm).

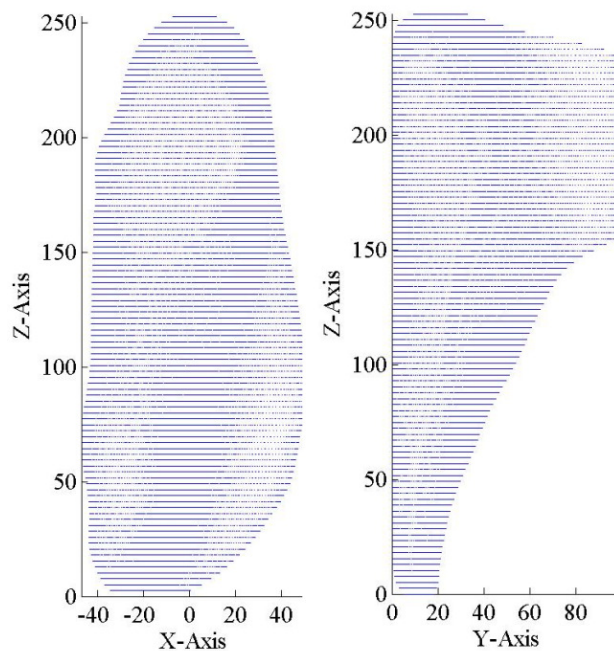


Figure 18. Standard foot shape (unit mm).

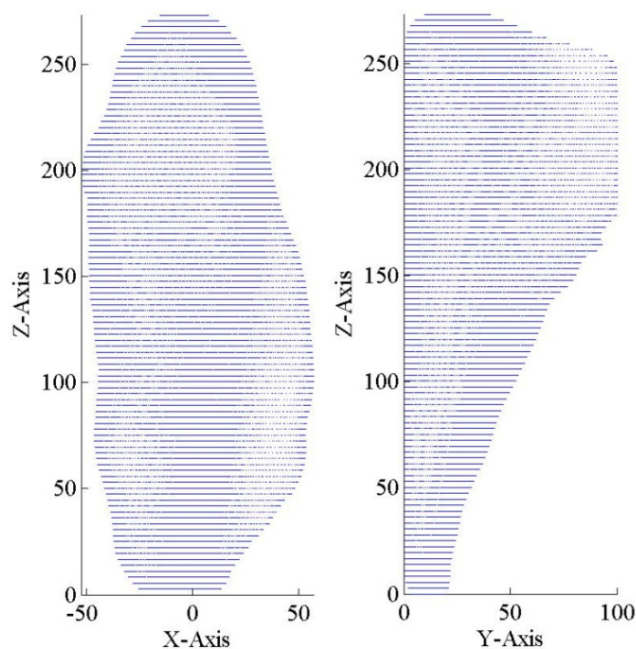


Figure 17. Sampled foot shape(unit mm).

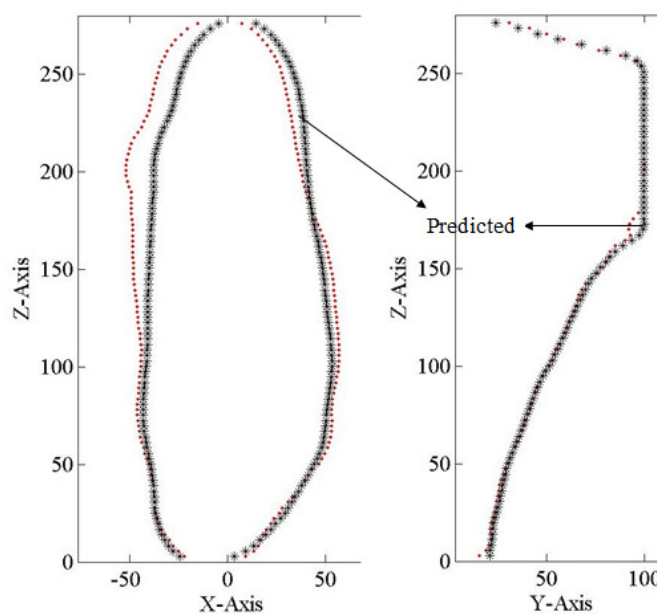


Figure 19. Parameter prediction (unit mm).

The prediction model was developed using foot length, maximum x deviation (W^+), minimum x deviation (W), and maximum y deviation (H^+). For left foot, the seed section for W^+ was at 7% foot length. The seed section for W was at 10% foot length. The seed section for H^+ was at 57% foot length. For right foot, the seed section for W^+ was at 9% foot length. The seed section for W was at 8% foot length. The

seed section for H^+ was at 57% foot length. Figure 19 shows the predicted and actual values for a participant. Figure 20 shows the 3D sampled and predicted foot shape. The mean prediction error is 2.93 mm and the standard deviation is 4.34 mm. Figure 21 shows the color-coded error on the surface of the foot.

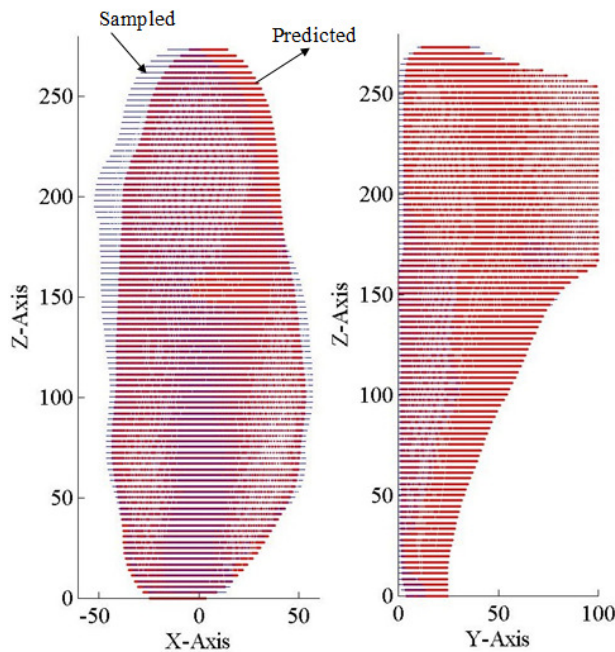


Figure 20. 3D prediction (units mm).

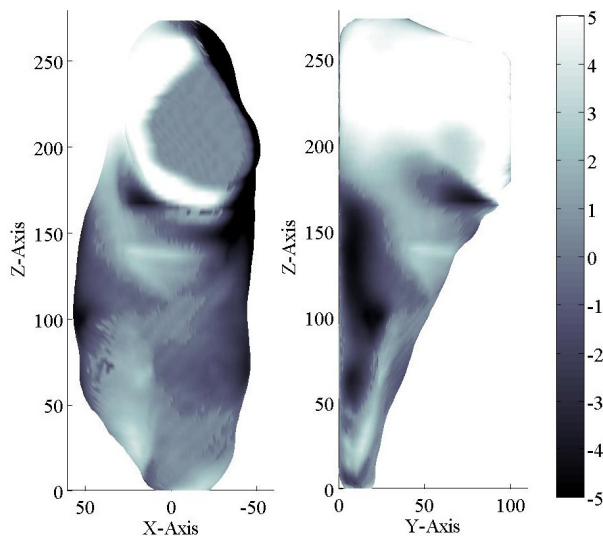


Figure 21. Prediction error plot (unit mm).

IX. CONCLUSION AND FUTURE WORK

Linear anthropometrics have existed for centuries and it is relatively very easy to measure linear anthropometric dimensions. Linear anthropometric is widely available based on age, race, region, and occupation and methods to capture linear anthropometry are non-invasive, inexpensive, simple, portable and reliable. For instance, if we want to buy custom-made shoes through the internet, it is much easier to provide a set of anthropometric measures (such as length, width, height, heel width).

Anthropometric measures have been widely used, but recently, there is a shift from linear anthropometric measures to surface anthropometric data in order to satisfy the ever-changing needs of the society. Anthropometry may be useful for sizing and selecting product, but in the design phase 3D shape information is required. For example, it is difficult to design shoes with only few measures. People are constantly looking for comfortable and 'proper' fitting wearable that not only match the linear anthropometric dimensions but also accommodate the complex surface of the body. In addition, more surface information is needed in medical, archaeological and forensic disciplines. As a result, the linear anthropometric table even though useful is not able to satisfy with the current demands. Thus, in order to have accurate information on body dimensions, surface model database has to be developed. Hence, many types of equipment (e.g., 3D scanners) to capture surface model have been developed.

In one side we have low cost traditional anthropometric measures and the other side we have accurate but expensive 3D scanners. Both of them are use full in some applications. The general method is to acquire anthropometric measures from surface scan data. This implies that existing huge database on anthropometric measures are not well utilized. In this study, a general model was proposed to generate surface model from linear anthropometry and standard shape. The standard shape can be stored in database based on age, sex, race, gender, etc. Simple recursive regression equations technique and scaling technique were used to build the prediction model. Model building involved data collection, alignment, cross sectioning, point sampling, averaging and regression equations development. Once the model has been built, given a few anthropometric measures, the standard shape can be scaled to generate a predicted 3D shape. Studies in foot modelling have shown that this method can predict the foot shape accurately using only 4 parameters including length, width, height and curvature. The accuracy of the predicted shape will generally be higher if more anthropometric measures are used. The model parameters can be adjusted to obtain the required accuracy depending on different applications. The application of this study is reconstructive surgery, forensic, anthropology, design, psychology, and other fields involving digital human models.

Further studies include making use of the most common anthropometric measures to create whole surface model; accurate model for specific parts; sensitivity analysis on the

use of number of anthropometric measures; and the use of the predicted data in product design.

ACKNOWLEDGMENT

We would like to thank the Hong Kong polytechnic university for providing support. This study was supported by RGC General Research Fund (B-Q26V).

REFERENCES

- [1] A. Luximon and H. Chao, "Shape Modeling: From Linear anthropometry to surface model," *Proc. ACHI 2013: The Sixth International Conference on Advances in Computer-Human Interactions*, Mar. 2013, pp. 420-425.
- [2] WHO Physical Status: The Use and Interpretation of Anthropometry, Technical Report No 854. Geneva: World Health Organization, 1995.
- [3] S. Pheasant, *Bodyspace: Anthropometry, ergonomics, and the design of work*. London: Taylor and Francis, 1996.
- [4] A. R. Tilley and Henry Dreyfuss Associates, *The measure of man and woman: human factors in design*. New York: John Wiley and sons, 2002.
- [5] F.E. Johnston, "Anthropometry," in *The Cambridge encyclopaedia of human growth and development*, S.J. Ulijaszek, F.E. Johnston and M.A. Preece, Eds. Cambridge: Cambridge University Press, 1998, pp. 26-27.
- [6] M. H. Al-Haboubi, "Statistics for a composite distribution in anthropometric studies," *Ergonomics*, vol. 40, pp. 189-198, 1997.
- [7] W. E. Woodson and D. W. Conover, *Human engineering guide for equipment designers* (2nd ed.). California: University of California Press Berkeley, 1964.
- [8] K. Gielo-Perczak, "The golden section as a harmonizing feature of human dimensions and workplace design," *Theoretical Issues in Ergonomics Science*, vol. 2, pp. 336-351, 2001.
- [9] A. Ozaslan, M. Y. Iscan, I. Ozaslan, H. Tugcu, and S. Koc, "Estimation of stature from body parts," *Forensic Science International*, vol 132, pp. 40-45, 2003.
- [10] M. Henneberg, E. Simpson, and C. Stephan, "Human face in biological anthropology: Craniometry, evolution and forensic identification," in *The human face: measurement and meaning*, M. Katsikitis, Ed. Boston : Kluwer Academic Publishers, 2003, pp. 29-48.
- [11] K. R. Fontaine, G. Gadbury, S. B. Heymsfield, J. Kral, J. B. Albu, and D. Allison, "Quantitative prediction of body diameter in severely obese individuals," *Ergonomics*, vol 45, pp. 49-60, 2002.
- [12] O. Giampietro, E. Virgone, L. Carneglia, E. Griesi, D. Calvi, and E. Matteucci, "Anthropometric indices of school children and familiar risk factors," *Preventive Medicine*, vol 35, pp. 492-498, 2002.
- [13] N.N. Prasad and D.V.R. Reddy, "Lip-Nose complex anthropometry," *International Journal of Cosmetic Surgery and Aesthetic Dermatology*, vol. 4, pp. 155-159, 2002.
- [14] R. Z. Stolzenberg-Solomon, P. Pietinen, P. R. Taylor, J. Virtamo, and D. Albanes, "A prospective study of medical conditions, anthropometry, physical activity, and pancreatic cancer in male smokers (Finland)," *Cancer Causes and Control*, vol. 13, pp. 417-426, 2002.
- [15] B. Bogin and R. Keep, "Eight thousand years of economic and political history in Latin America revealed by anthropometry," *Annals of Human Biology*, vol. 26, pp. 333-351, 1999.
- [16] R. Floud, "The dimensions of inequality: Height and weight variation in Britain, 1700-2000," *Contemporary British History*, vol. 16, pp. 13-26, 2002.
- [17] T.K. Oommen, "Race, religion, and caste: Anthropological and sociological perspectives," *International Journal of Comparative Sociology*, vol. 1, pp. 115-126, 2002.
- [18] T. Reilly, J. Bangsbo, and A. Franks, "Anthropometric and physiological predispositions for elite soccer," *Journal of Sports Sciences*, vol. 18, pp. 669-683, 2000.
- [19] P. Tothill and A. D. Stewart, "Estimation of thigh muscle and adipose tissue volume using magnetic resonance imaging and anthropometry," *Journal of Sports Sciences*, vol. 20, pp. 563-576, 2002.
- [20] M. Westerstaahl, M. Barnekow-Bergkvist, G. Hedberg, and E. JANSSON, "Secular trends in body dimensions and physical fitness among adolescents in Sweden from 1974 to 1995," *Scandinavian Journal of Medicine and Science in Sports*, vol. 13, pp. 128-137, 2003.
- [21] C.M. Worthman, "Recumbent anthropometry," in *The Cambridge encyclopaedia of human growth and development*, S.J. Ulijaszek, F.E. Johnston and M.A. Preece, eds. Cambridge: Cambridge University Press, 1998, pp. 29.
- [22] S. Ulijaszek, "Measurement error," in *The Cambridge encyclopaedia of human growth and development*, S.J. Ulijaszek, F.E. Johnston and M.A. Preece, Eds. Cambridge: Cambridge University Press, 1998, pp. 28.
- [23] P. R. M. Jones and M. Rioux, "Three-dimensional surface anthropometry: Applications to the human body," *Optics and Lasers in Engineering*, vol. 28, pp. 89-117, 1997.
- [24] V. A. Deason, "Anthropometry: The human dimension," *Optics and Lasers in Engineering*, vol 28, pp. 83-88, 1997.
- [25] A. D. Linney, J. Campos, and R. Richards, "Non-contact anthropometry using projected laser line distortion: Three dimensional graphic visualization and applications," *Optics and Lasers in Engineering*, vol. 28, pp. 137-155, 1997.
- [26] D. B. Chaffin, *Digital human modeling for vehicle and workplace design*. Warrendale, PA: Society of Automotive Engineers, 2001.
- [27] B. K. Choi, *Surface modeling for CAD/CAM: Advances in industrial engineering*, vol. 11. Amsterdam: Elsevier, 1991.
- [28] A. Luximon, *Foot shape evaluation for footwear fitting*. Hong Kong: Hong Kong University of Science and Technology, Hong Kong, 2001.
- [29] A. Luximon and R.S. Goonetilleke, 2003, "Foot shape modeling," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol 46, pp. 304-315, 2004.
- [30] T.K. Dey and J. Sun, *Adaptive MLS Surfaces for Reconstruction with Guarantees*, *Proc. Eurographics Symposium on Geometry Processing*, 2005.
- [31] R.S. Goonetilleke and A. Luximon, "Foot flare and foot axis," *Human Factors*, vol. 41, pp. 596-607, 1999.
- [32] Y. Luximon, R. Ball, and L. Justice, "The 3D Chinese Head and Face Modeling" *Computer-Aided Design*, vol. 44, pp. 40-47, 2012.
- [33] R. S. Goonetilleke, C.-F. Ho, and R. H. Y. So, "Foot anthropometry in Hong Kong," *Proc of the ASEAN 97 Conference*, 1997, pp. 81-88.
- [34] A. Luximon and R.S. Goonetilleke, *Dimensions for footwear fitting*, *Proc of the International Ergonomics Association 2003 conference*, 2003.
- [35] G. Gregory and P.P. Abraham *The traveling salesman problem and its variations*. Boston: Kluwer Academic Publishers, 2002.
- [36] A. Luximon, R.S. Goonetilleke, and K.L. Tsui, "Foot landmarking for footwear customization," *Ergonomics*, vol. 46, pp. 364-383, 2003.

Kinematic Description of Bimanual Performance in Unpredictable Virtual Environments

A Lifespan Study

Andrea H Mason, Drew N Rutherford, and
Andrew R Minkley

Department of Kinesiology
University of Wisconsin - Madison
Madison, USA
amason@education.wisc.edu, anrutherford@wisc.edu,
aminkley@wisc.edu

Patrick J Grabowski

Department of Physical Therapy
University of Wisconsin – La Crosse
La Crosse, USA
pgrabowski@uwlax.edu

Abstract— Immersive virtual environments show great promise for use in applications such as design and prototyping, data visualization, and rehabilitation of motor impairments. However, our understanding of how people of various ages process and use sensory information to complete tasks within these environments is limited. The purpose of the research described here was to characterize motor performance in virtual environments across the lifespan on simple, foundational skills. Our results indicated that children and older adults used different strategies when performing the task when compared to young adults. While older adults adjusted for the virtual environment by planning a slower movement, children compensated for the artificial environment by relying on feedback to a greater extent. Movement strategies for the youngest and oldest groups were also different in the virtual environment when compared to results from natural environment experiments. We conclude that children and older adults do not plan movements or make use of sensory information in a similar fashion to young and middle-aged adults when performing in a virtual environment. The design implications of these results are related to differences in needed sensory information between children, young and older adults, the transfer of training effects between virtual and real environments, and important differences between performance and learning applications.

Keywords- *virtual environment; aging; motor control; kinematic analysis; bimanual reach to grasp*

I. INTRODUCTION

With the expansion of the role of computers in schools, the workplace, and homes, the population of users who make regular use of computing technology has grown exponentially. Unfortunately, Human-Computer Interaction (HCI) research has not reflected this demographic reality. The purpose of the research reported here is to begin to fill this large gap in knowledge by determining how the age of the user influences motor performance in virtual environments. In [1], we asked children, young adults, middle age adults and older adult participants to perform simple and bimanual reach to grasp movements, and

movements where targets were visually displaced. We found that children and older adults used different strategies when performing reach to grasp tasks in a virtual environment when compared to young adults. Specifically, older adults adjusted for the virtual environment by planning a slower movement, while children compensated for the artificial environment by relying on feedback to a greater extent. In the current paper we extend on [1] by including a more detailed literature review, additional and detailed kinematic analyses and a more thorough discussion of the implication of these results.

The structure of this paper is as follows. In Section II we present a thorough review of the literature upon which this work is motivated. In Section III we describe the experimental method used in this paper to investigate the roles of age and vision on the performance of reach to grasp movements in virtual environments. In Section IV, the results of the statistical analyses are presented and finally, in Section V the results are discussed in the context of the current state of knowledge and potential applications/implications of this work.

II. REVIEW OF LITERATURE

To adequately frame the theory and methods used in the research study presented here, this review of literature first covers recent work on the influence of age on human computer interaction (Section A). Next, we review the current knowledge regarding the control of simple and bimanual movements in both natural and computer generated environments (Section B) as well as the motivation behind using a perturbation task in this study (Section C). Finally, in Section D we present the hypotheses for the current study.

A. HCI and Age

Results of the 2010 US Census show that 17.5% of the US population is between the ages of 5 and 18 and a further 40% of the population is above the age of 45 [2]. It has also been reported that Europe is experiencing an aging

population, with projections of 35% of the population being above the age of 65 by 2025 [3]. Still most HCI research is focused on younger people, often university or college students [4]. Rather than representing the true population of computer users, most experimental HCI research is biased heavily towards the cognitive and motor abilities of young adults.

In order to understand how age may influence performance on tasks requiring human-computer interaction, we must first take a step back and understand the influence of age on movement performance in general. The human body is a constantly changing entity throughout the lifespan and all systems of the human body, including the sensorimotor system, undergo changes. Movement programming functions are organized quite differently in children than in young adults [5], [6]. Further studies have shown that children process sensory information from visual and proprioceptive receptors differently than adults [7], [8]. Children tend to rely on visual feedback to a greater extent [9]. There is also a general indication that both the processing of afferent information, or incoming signals to the central nervous system (CNS), and the production of efferent information, or outgoing commands, steadily changes as a function of age in the developing human. Once beyond the “development” stage of the lifespan, into older adulthood, the volume of research on age-related changes greatly expands. Multiple authors demonstrate physical changes in brain tissues [10]-[12], changes in the activation of motor neurons in the brain [13], and a general loss of nerve tissue [11], [12]. These tissue changes then result in myriad functional declines within the CNS. There is a general deterioration of motor planning [14], [15] and anticipatory control [16], as well as slowing of central processing [17]-[20].

These transformations in the sensorimotor system have a resultant effect on motor performance in daily life. Children tend to show less accuracy, decreased smoothness of movement, and decreased speed when compared to young adults [21]. Many of these same manifestations become apparent as adults age. According to Schut [22], most physiologic processes begin to decline at a rate of 1% per year beginning at age 30. In general, aging adults demonstrate decreases in movement speed [14], [18], accuracy [17], strength [23], hand dexterity [24], and postural control [25], and increases in reaction time.

So, how do these lifespan changes in information processing within the CNS affect humans as they use computer interfaces? Where age-specific research has been conducted, the majority relates to the design of standard computer interface systems for various age groups. In particular, research has focused on ways to improve cognitive performance through specific training or tutorial methods (e.g., [26], [27]), or on the age-appropriate design of input devices (e.g., [28]-[31]). There is also a modest body of scientific literature which explores the areas of motor control in human computer interaction (HCI) as a

function of age [27], [29]. Most of this information centers on the input device, specifically mouse usage in children and older adults. It is reported that there are many age-related changes, and in general it is quite difficult for children and older individuals to use a mouse [27], [28]. Maintaining adequate pressure and the act of double-clicking seem to consistently be the most problematic. Difficulty with cursor control is named as a top complaint among older individuals [4], [26]. It has also been shown that performance within a standard computer interface is slower and results in a greater number of errors with increased age of the operator.

Much less is known about how age influences performance within immersive three-dimensional (3-D) virtual environments (VEs) [32]-[34]. Immersive VEs are becoming more prominent as the costs of the relevant tracking and display technologies decrease. VEs are commonly used in design and prototyping, data visualization, medical training, architecture, education, and entertainment. Further, recent research has focused on the utility of VEs for rehabilitation of motor impairments such as stroke in the elderly and attention deficit hyperactivity disorder (ADHD), developmental coordination disorder and cerebral palsy in the young [35], [36]. However, because there is a paucity of information on how healthy children and older adults interact in VEs, it is likely that the success of these systems will struggle. Specifically, it is nearly impossible to extrapolate design characteristics from healthy young adults to special-needs children and older adults. Results of the few studies conducted on performance across age-groups within virtual environments indicate relevant disparities in reactions to environmental immersion, usage of various input devices, size estimation ability, and navigational skills [32]-[34]. According to Allen et al. [32], “these results highlight the importance of considering age differences when designing for the population at large.”

The purpose of the research described here is to characterize motor performance in virtual environments across the lifespan. To do this we asked participants ranging in age from 7 to 90 years to perform a foundational skill (bimanual reach to grasp) within a table-top virtual environment. In the following sections, we describe the importance of the skill we chose to study.

B. Bimanual Reach to Grasp Skills

The performance of many everyday activities requires the completion of asymmetric but coordinated movements with our two hands. For example, touch typing, tying our shoelaces, and even reaching for a mug with one hand and a coffee pot with the other require the performance of two separate but coordinated movements. Many asymmetric bimanual tasks such as the ones described above can be performed quite effortlessly in natural environments. This seamless control is possible because we use feedforward sensory information (vision and proprioception) to pre-plan

our movements and feedback sensory information for on-line corrections during movement execution.

Recently, bimanual tasks have been targeted as important skills to (re)train in rehabilitation protocols employing natural environments and virtual reality [37]. In rehabilitation training after stroke, these types of tasks are important for functional recovery because they require the areas of the brain most commonly afflicted by stroke to work with areas usually left undamaged, thereby maximizing the potential for positive neuroplastic changes [38].

While the study of bimanual movements has received some attention in natural environments, very little is known regarding the performance of these types of movements in virtual environments [39]. Further, no studies have looked at how the control of bimanual skills changes as a result of age in VEs. In order to successfully implement rehabilitation and training protocols that make use of these types of tasks it is imperative that we first obtain a baseline understanding of how neurologically “normal” people across the lifespan perform bimanual skills in VEs and how they use sensory information for the performance of these skills.

In natural environments, results from bimanual movement studies have indicated that when the two limbs are used to accomplish both symmetric and asymmetric task goals, coupling between the limbs for certain parameters occurs in the temporal domain [40], [41]. In particular, movement onset, duration, and end times tend to be similar for the two hands when subjects aim toward or reach to grasp targets of different sizes or at different locations [40], [41]. However, timing differences between the hands have been shown, and results indicate that these differences are associated with insufficient visual feedback for movement control [42]. In the current study we investigated whether the same patterns of results are seen in virtual environments and whether these patterns change with age. We employed a target perturbation to specifically investigate how sensory (visual) information is used on-line by participants of various ages to modify their movements. These paradigms are discussed in more detail in the following section.

C. Unpredictable Environments: Perturbation Paradigms

An experimental paradigm that has been successfully used to investigate the role of on-line visual information for the performance of goal directed tasks uses target perturbation to study adjustments to ongoing movements. The use of this type of paradigm allows us to discern how long it takes the nervous system to adapt to an unexpected visual change as well as the efficiency of the adaptation.

In a target perturbation paradigm, the participant is unexpectedly presented with the requirement to alter their original movement plan either prior to or after movement onset. An example of a typical perturbation paradigm is as follows. A visual stimulus is presented to the participant prior to movement initiation and the participant generates a

movement plan appropriate to the acquisition of the target at this initial location. Shortly prior to or after movement onset the stimulus is suddenly replaced by a second stimulus presented at an alternative location. The participant is thus required to reorganize their movement to successfully grasp the target at its new position. Results of studies using perturbation paradigms in both natural [43] and virtual environments [39] have indicated increased movement times to displaced targets and double velocity peaks in kinematic recordings.

Studying the performance of bimanual perturbation tasks in a VE can provide us with important information about how participants make use of visual information during the execution of a skill. This is particularly important given that the use of sensory information changes across the lifespan [44], [45] and all the visual information presented to users of VEs must be synthetically created. By comparing results in the VE to studies performed in the “real” world we can determine whether performance is similar within these two environments.

D. Hypotheses

We asked participants ranging from 7 to 90 years of age to perform bimanual reach to grasp movements in a virtual environment. In the first set of trials, target objects remained at their initial position throughout the task, giving us a baseline performance for each participant. Based on previous literature on age differences and motor performance, we expected that younger children and older adults would perform the bimanual tasks more slowly than the young adults. Further, we expected that temporal synchronization between the two hands would be less strong in the youngest and oldest participants due to their reliance on visual feedback. These results would replicate the results of studies performed in natural environments. When considering performance in the perturbation conditions, we expected the youngest and oldest participants to show a decreased ability to respond to the visual displacement of the target when compared to the young adults. Specifically, it is known that young children and the elderly process sensory/visual information more slowly than young adults [21], [22]. Since responding to the perturbation relies on the speed of visual information processing, we hypothesized that children and older adults would respond more slowly and would show less coordinated movements in the perturbed conditions than the young adults.

III. METHOD

In the following section we detail the method used to determine how age influences performance in virtual environments. We begin by describing our participant pool and the experimental apparatus. Next we describe the tasks performed by each participant. Finally we describe our data analysis methods.

A. Participants

Fifty-one participants were divided into four age categories: Children (7-12 years, $n=13$), Young adults (18-30 years, $n=12$), Middle adults (40-50 years, $n=12$) and Older adults (60+ years, $n=12$). Due to problems with data collection final data analysis was conducted on 12 participants in the “Children” group and 11 participants in the “Older adult” group. Decades of motor control research have indicated that a sample size of 10-12 participants provides sufficient statistical power in this type of reach to grasp study. All participants were self-reported right-handers and had normal or corrected-to-normal vision. All participants provided informed consent before taking part in the experiment. The protocol was approved by the University of Wisconsin-Madison Social and Behavioral Science Institutional Review Board.

B. Experimental Apparatus

This experiment was conducted in the Wisconsin Virtual Environment (WiscVE) at the University of Wisconsin-Madison (Fig. 1). In this environment, subjects see three-dimensional graphical representations of target objects but interact with physical objects. Graphic images of two target cubes were displayed on a downward facing computer monitor. A half-silvered mirror was placed parallel to the computer screen, midway between the screen and the table surface. The graphic image of the cubes was reflected in the mirror and appeared to the participant to be located in the workspace on the table surface. Three light emitting diodes (LEDs) were positioned on the top surface of two wooden target cubes (38 mm). A VisualEyez 3000 motion capture system (Phoenix Technologies, Inc., Burnaby) tracked the three-dimensional position of the LEDs on the physical target cubes. This data was used to generate the



Figure 1. Experimental apparatus.

superimposed graphical representations of the cubes. The lag between motion of the LED and its graphical representation was indiscernible to participants. A shield was placed below the mirror to prevent subjects from seeing the real environment or their hands as they performed the reach-to-grasp task.

Participants wore CrystalEYES™ goggles to obtain a stereoscopic view of the graphic images being projected onto the mirror. Three LEDs were fixed to the goggles and were used to provide the subject with a head-coupled view of the virtual environment on the work surface. Thus, when the subject moved his/her head, the displayed scene was adjusted appropriately for the magnitude and direction of head movement. LEDs were also positioned on the subject's right and left thumbs, index fingers and wrists. Data from all LEDs was collected at a sampling rate of 120 Hz and was stored for data analysis purposes.

C. Design and Procedure

Each trial began with the illumination of two blue circular start positions (radius 5 mm) located 12.5 cm to the left and right of the participants' midline. The participants moved their hands from the periphery of the workspace to place their index fingers and thumbs over the start positions, which were haptically indicated by small metal hex nuts. When the participants' hands were correctly positioned, the start positions turned yellow. Once both of the participants' hands remained stationary at the start positions for 1 s, the two graphic target cubes appeared at a location 20 cm from the start position. The task was to reach forward with the right and left hands to grasp and lift the two target cubes. Grasps were made with a precision grasp (i.e., index finger and thumb only) and participants were asked to move at a comfortable pace once the target cubes appeared.

Participants experienced trials in four experimental conditions. As shown in Fig. 2, in the control condition both targets remained at their initial location throughout the trial (left target no jump/right target no jump; NN). As shown in Fig. 3, in the three perturbation conditions one or both targets were displaced 9 cm toward the participant at movement onset (defined as a displacement of 5 mm of the

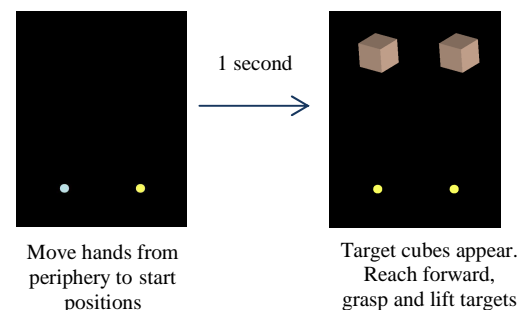


Figure 2. Time course a control trial (top-down view).

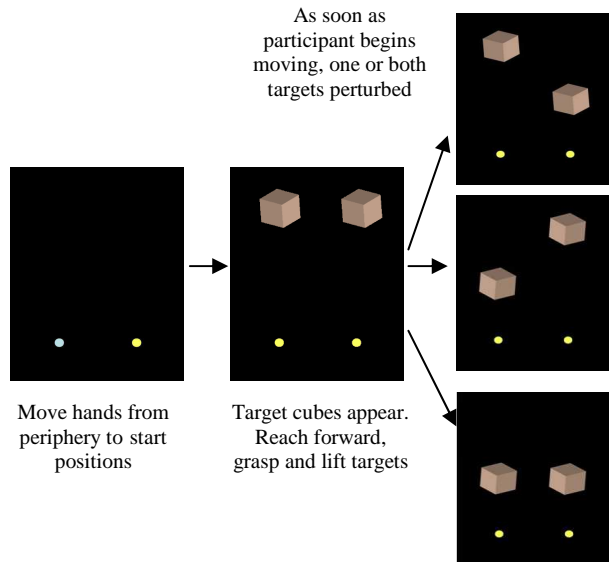


Figure 3. Time course of perturbation trials (top down view).

thumb LED). The perturbation conditions consisted of: 1) left target jump/right target no jump (JN), 2) left target no jump/right target jump (NJ), 3) left target jump, right target jump (JJ).

Participants performed a total of 100 trials. The first 10 trials were always control trials (NN). This allowed participants to become comfortable with the task and also gave us the opportunity to analyze a set of “control” trials where participants had no expectation of a perturbation. The remaining 60 control and 30 perturbation trials, 10 in each condition, were presented in a random order.

D. Data Analysis

Human motor control, biomechanics and neuroscience research has provided a comprehensive description of how humans reach to grasp and manipulate objects in natural environments under a variety of sensory and environmental conditions. By using the same measurement techniques as those employed to monitor human performance in natural environments, we can compare movement in virtual environments to decades of existing human performance literature. The comparisons allow us to develop comprehensive cognitive models of human performance under various sensory feedback conditions. Simple timing measures such as movement time provide a general description of upper limb movements. However, in motor control studies, more complex 3-D kinematic measures such as displacement profiles, movement velocity, deceleration time, and the formation of the grasp aperture (resultant distance between the index finger and thumb for a precision pinch grip) have also been used to characterize object acquisition movements. By observing regularities in the 3-D kinematic and kinetic information, inferences can be made

regarding how movements are planned and performed by the neurocontrol system.

Peak velocity can be used to measure the open-loop processes occurring during target acquisition tasks and is thought to reflect motor planning. In contrast, the time from peak velocity represented as a percentage of movement time can be used as a measure of closed-loop control, where a longer time spent decelerating toward the target is equated with a greater reliance on feedback. These measures combined with movement time allow us to completely describe a target acquisition task in terms of open and closed loop control.

For tasks that involve grasping objects, a measure of the opening and closing of the hand is also required. Aperture can be used to quantify grasp formation. In human performance literature larger apertures have been associated with more complex tasks that demand greater attentional resources [46]. It is believed that a larger aperture is used as a compensatory strategy to avoid missing or hitting the target. This detailed movement information essentially provides a window into the motor control system and allows the determination of what sensory feedback characteristics are important for movement planning and production.

We quantified the above kinematic measures of movement using position data from the block LED as well as LEDs on the wrists of both hands. Start of movement was defined as the point where resultant wrist velocity increased above a threshold of 5 mm/s and continued increasing to a peak. End of movement was defined as the point where vertical block lift velocity increased above 5 mm/s and continued increasing to a peak. Based on these two temporal measures we calculated Movement Time (MT) for both hands. The position data were differentiated and peak resultant velocity (PV) was extracted. Percent time from peak velocity (PTFPV) was defined as $(MT - \text{Time of peak velocity}) / MT * 100$. We also quantified temporal coupling of the two hands by determining whether the hands started and ended movement at similar times. To do this we calculated the Absolute Start Offset (ASO: Start Left Hand – Start Right Hand) and Absolute End Offset (AEO: End Left Hand – End Right Hand). To quantify the grasp, we extracted the peak aperture (PA) achieved by the index finger and thumb of each hand during the course of the movement.

Data were statistically analyzed in two ways. First, to quantify control performance in the first 10 trials, we conducted a 4 Group (Children, Young Adult, Middle Adult, Older Adult) X 2 Hand (left, right) repeated measures ANOVA on MT, PV, PTFPV and PA.

To quantify bimanual coupling during the control trials a 4 Group (Children, Young Adult, Middle Adult, Older Adult) repeated measures ANOVA was performed on ASO and AEO. To quantify performance during the perturbation trials we conducted separate 4 Group (Children, Young Adult, Middle Adult, Older Adult) X 4 Condition (JJ, JN, NJ, NN) repeated measures ANOVAs for each hand and

dependent measure. Post-Hoc analysis on significant main effects was done using the Fisher LSD method. When significant interactions occurred, these were further explored using simple main effects with Condition as the factor. An a priori alpha level was set at $p < 0.05$.

IV. RESULTS

The results of our statistical analyses are shown in the following sections. In Section A we present the results for the initial set of bimanual control trials that each participant performed at the beginning of the experimental session. In Section B we present the results for trials where the target could be displaced unexpectedly (i.e., perturbed).

A. Initial Performance: Control Trials

The control trials allow us to determine how bimanual performance changes as a function of age within virtual environments and whether patterns of performance in VEs replicate those seen in natural environments. Typical velocity profiles for children, young adults (middle adults resembled young adults) and older adults in the control condition are shown in Fig. 4A. Note that velocities are higher for the children and young adults than the older adults. Also note that movement times (as indicated by the end of the trace on the time axis) are longer for the children and older adults than the young adults. Finally, note that velocity profiles for the young and older adults appear smoother than those produced by the children. The decreased smoothness represented in the children's profiles reflects a greater reliance on sensory feedback and error correction during movement production. Results of the statistical analyses on the individual kinematic measures are presented below.

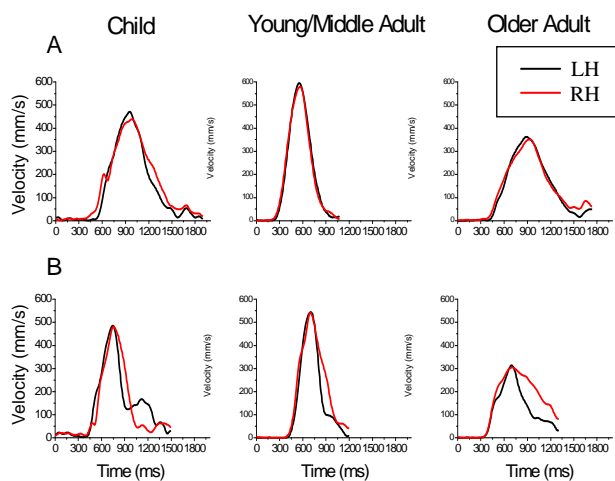


Figure 4. Typical velocity profiles for the children, young/middle adults and older adults in the A) NN condition, B) JN condition.

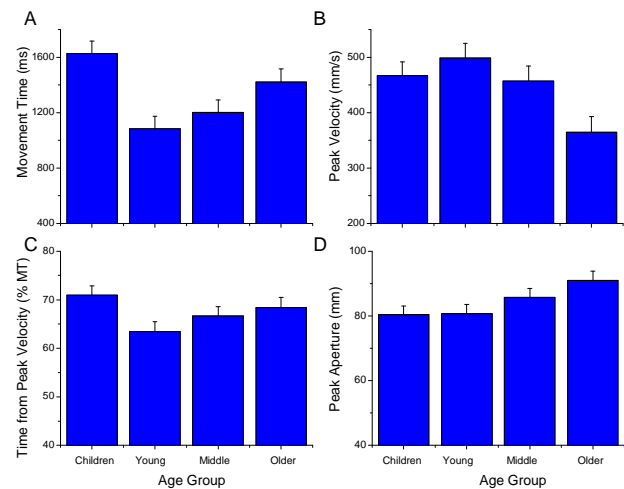


Figure 5. Main effect of Group on movement time, peak velocity, time from peak velocity as a percent of movement time and peak aperture in the control condition.

Main effects of Group were found for movement time ($F_{3,43} = 7.053$, $p=0.001$), peak velocity ($F_{3,43} = 4.335$, $p=0.01$), and peak aperture ($F_{3,43} = 3.2$, $p=0.033$). The main effect of Group for percent time from peak velocity was marginally significant ($F_{3,43} = 4.335$, $p=0.06$). Results indicated that the fastest movement times were found in the young and middle aged adults. Children were significantly slower than the young and middle aged adults, whereas older adults were only significantly slower than the young adults (Fig. 5A). Further decomposition of the movement into its velocity profile indicated that the longer movement time used by the older adults was the result of a significantly lower peak reaching velocity when compared to all other groups ($p<0.05$). In contrast, the children achieved a similar peak reaching velocity as the young and middle aged adults (Fig. 5B). For the children, the additional movement time when compared to the young adults came as the result of a longer time spent decelerating toward the target ($p=0.09$)(Fig. 5C). In contrast, the older adults spent a similar proportion of the movement decelerating toward the target as the middle-aged and young adults ($p>0.05$). These results suggest that although both the children and older adults perform the reach to grasp task more slowly than the young adults, the reason for this slowing is different for the two age groups. Finally, when considering grasp aperture, results indicated that the older adults produced a significantly larger hand opening when reaching for the targets than the young adults ($p<0.05$). In contrast, the aperture used by the children was similar to the young adults (Fig. 5D).

When looking at coupling between the left and right hands, main effects of Group were found for ASO ($F_{3,43} = 14.03$, $p<0.001$) and AEO ($F_{3,43} = 4.74$, $p=0.006$). The post-

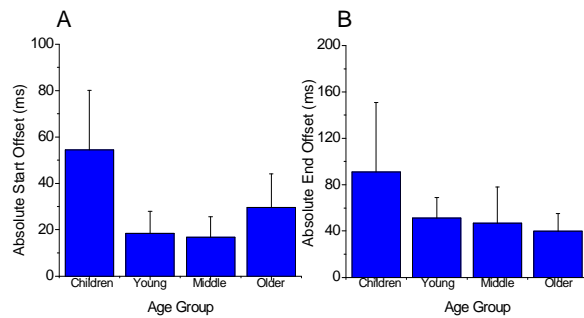


Figure 6. Main effect of Group on ASO and AEO.

hoc LSD indicated that children had significantly larger offsets at both the start (Fig. 6A) and end (Fig. 6B) of movement than any of the other age groups.

B. Perturbation Performance

The perturbation trials allowed us to investigate whether differences in the use of on-line visual feedback occur across age groups and for different perturbation conditions. Fig. 4B shows velocity profiles for the right and left hand in the JN condition for the children, young adults and older adults. First note that the young adults adjust smoothly to the perturbation and efficiently decouple the movements of the two hands to effectively grasp the perturbed target at its new location. In contrast, note that the children produce velocity profiles that are less smooth and efficient. These profiles provide evidence that the children have greater difficulty making use of online sensory information when reorganizing for the perturbation. For the older adults, note the much lower peak velocity. This suggests that older adults pre-plan a more conservative movement.

We analyzed the data separately for the right and left hands to simplify interpretation. An interaction between Condition and Group ($F_{9,129} = 2.934$, $p=0.003$) was found for MT of the right hand. Children had significantly longer MTs than all other groups in the NN, JN and JJ conditions (Fig. 7A). However, they did have similar MTs to the older adults in the NJ condition. The young and middle adults had similar MTs across all conditions but the older adults were significantly slower than the young adults in the NN and NJ conditions only. Further decomposition of the movement of the right hand into its velocity profile revealed a main effect of Condition ($F_{3,129} = 12.5$, $p<0.001$) and Group ($F_{3,43} = 2.75$, $p=0.055$) for peak velocity. Velocities were highest in the NN condition (474.5 ± 14 mm/s), lowest in the JJ condition (438.5 ± 14 mm/s) and moderate when only one target was perturbed (JN = 456.3 ± 15 mm/s; NJ = 453.0 ± 15.7 mm/s). The main effect of Group indicated that older adults had significantly lower peak velocities than the young adults (Fig. 7B).

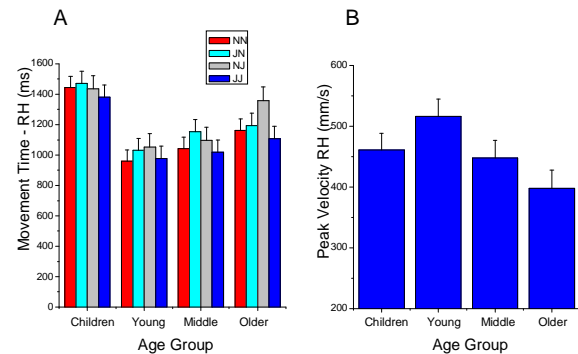


Figure 7. Group X Condition interaction for MT and main effect of Group for peak velocity of the right hand.

When considering how participants used visual feedback to decelerate toward the target, an interaction between Condition and Group ($F_{9,129} = 2.0$, $p=0.045$) was found for the right hand. As seen in Fig. 8, children had longer deceleration times than the three other age groups in all conditions except NJ.

Finally, for the grasp portion of the movement, main effects of Group ($F_{3,43} = 5.8$, $p=0.002$) and Condition ($F_{3,129} = 2.8$, $p<0.044$) were found for peak aperture for the right hand. The main effect of Group indicated that peak apertures were larger for the older adults (92.3 ± 3 mm) than the three other groups (children = 77.4 ± 2.5 mm; young adult = 80.3 ± 2.7 mm; middle adult = 81.2 ± 2.7 mm). The main effect of Condition indicated that peak apertures were larger when neither object was perturbed (83.7 ± 1.2 mm) when compared to the other three conditions (JJ = 82.0 ± 1.3 mm; JN = 82.5 ± 1.5 mm; NJ = 83.0 ± 1.4 mm).

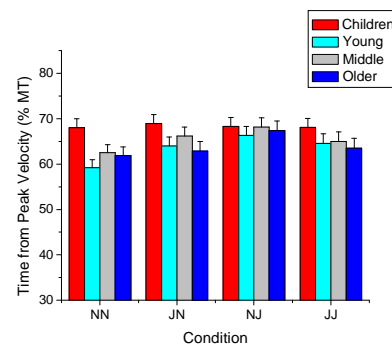


Figure 8. Interaction between Condition and Group for deceleration time of the right hand.

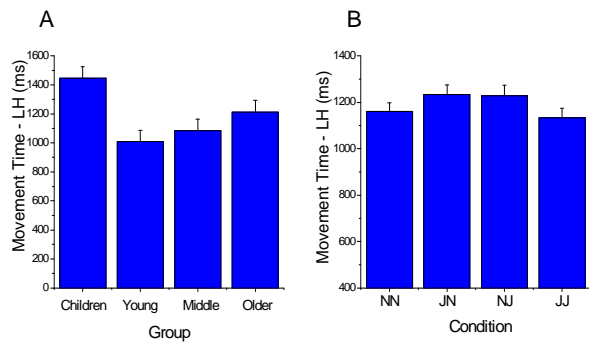


Figure 9. Main effects of Group and Condition on MT of the left hand.

For MT of the left hand, main effects of group ($F_{3,43} = 6.04$, $p=0.002$) and condition ($F_{3,129} = 10.6$, $p<0.001$) were found. The group main effect indicated that the children were significantly slower than the young and middle adults. No other significant differences were found (Fig. 9A). For the main effect of condition, results indicated that MTs for the left hand were significantly faster in the NN and JJ conditions than in the JN and NJ conditions (Fig. 9B).

Decomposition of movement time into kinematic features indicated a main effect of Condition ($F_{3,129} = 17.1$, $p<0.001$) and a marginally significant main effect of Group ($F_{3,43} = 2.4$, $p<0.082$) for peak velocity. The Group effect revealed lower peak velocities for the older adults (397.4 ± 30.5 mm/s) when compared to the young adults (508.7 ± 29 mm/s) ($p < 0.05$). Peak velocities for the children (440.5 ± 28 mm/s) and middle adults (454.2 ± 29 mm/s) were similar to all other groups. The main effect of Condition revealed higher peak velocities in the NN condition (468.3 ± 14 mm/s) than all other conditions (JJ = 432.7 ± 15 ; JN = 439.7 ± 146 ; NJ = 460.1 ± 15 mm/s). An interaction between Group X Condition ($F_{3,43} = 2.1$, $p<0.03$) was also found for deceleration time (see Fig. 10). As with the right hand, these results indicated that children had longer deceleration times than the young and older adults in all conditions.

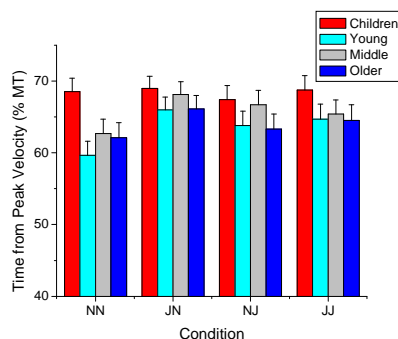


Figure 10. Interaction between Condition and Group for deceleration time of the left hand.

For the grasp portion of the movement main effects of Group ($F_{3,43} = 6.9$, $p=0.001$) and Condition ($F_{3,129} = 3.3$, $p<0.02$) were found for peak aperture of the left hand. The main effect of Group indicated that peak apertures were larger for the older adults (92.1 ± 3 mm) and middle age adults (86.2 ± 2.5 mm) than the young adults (79.2 ± 2 mm) and children (77.8 ± 2 mm). The main effect of Condition indicated that peak apertures were larger when neither object was perturbed (84.7 ± 1.2 mm) or when the right object was perturbed (84.4 ± 1.2) than the other two conditions (JJ = 83.0 ± 1.4 mm; JN = 83.2 ± 1.4 mm).

When looking at coupling between the two hands during perturbation trials, a main effect of group ($F_{3,43} = 15.9$, $p<0.001$) indicated that children had significantly larger offsets at movement initiation than any other age group (Fig. 11). For the end of movement, a Group X Condition interaction ($F_{9,129} = 2.232$, $p=0.024$) indicated that children had significantly larger offsets than all other groups in the NN condition (Fig. 12). The older adults had longer offsets than the young adults in the NJ condition. All groups had statistically similar offsets in the JN and JJ conditions.

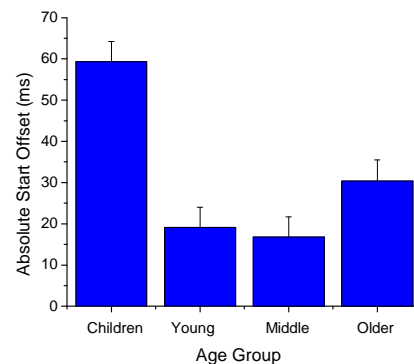


Figure 11. Main effect of Group on ASO of the left hand.

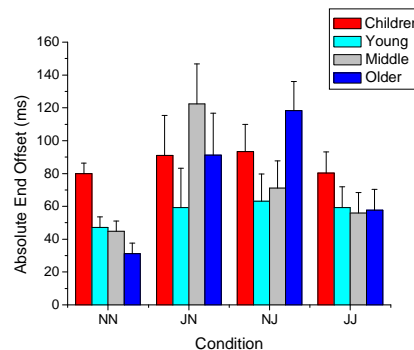


Figure 12. Main effect of Group on AEO.

V. CONCLUSION AND FUTURE WORK

A. Performance of Bimanual Movements in VEs across the lifespan: Control and Perturbation Conditions

Each participant began the experiment by performing a block of simple bimanual trials without perturbation. These trials allowed us to determine whether age-specific patterns of bimanual performance in VEs are similar to the patterns seen in the natural environment. When considering overall MT, research in natural environments has indicated that children and the elderly typically complete both simple and complex tasks more slowly than young adults [47], [48]. A similar pattern of results was found in the current study, indicating some similarities between VEs and natural environments. With respect to bimanual coupling in natural environments, prior studies have indicated that both young children and older adults exhibit greater offsets at movement initiation and movement completion than young adults [49], [50]. These results were replicated for the children; however, the older adults used similar movement offset patterns as the young and middle adults. This difference in movement coupling for the elderly subjects suggests that they use different control strategies in natural compared to virtual environments. Timing differences between the hands in bimanual tasks have been associated with the requirement to shift visual attention between the targets to obtain sufficient feedback [42]. In older adults, slowing of visual sensory processing due to aging should result in even greater timing differences between the hands [45]. The smaller offsets seen in the current study suggest that the elderly subjects may have been relying on a predominantly feedforward strategy to complete the task instead of the typical feedback-based strategy that is seen in natural environments. This conclusion is supported by the detailed kinematic measures reported in this study. In particular, peak velocity, which is typically reached early in the movement, can be used as a measure of movement planning. Older adults used a lower peak velocity in the control trials than subjects in all other groups. This suggests that the older adults were in fact using feed-forward planning to execute a cautious reach strategy in our virtual environment. The larger grasp apertures used by the older adults also support the notion of a cautious reaching strategy.

In a previous study investigating age differences on a simple reach-to-grasp task in a VE, we also found that older adults relied more heavily on a feedforward-based strategy [50]. Deceleration time results for the older adults also indicated that they spent a similar time using sensory information to home-in on the target as the younger adults. In contrast, results from reach-to-grasp studies in natural environments have indicated that elderly participants typically use longer deceleration times than their younger counterparts [51]. Again, this points to a difference in strategy in the virtual environment when compared to natural environments. We hypothesize that the impoverished

and unnatural feedback available in the virtual environment may have made this feedback less useful to the older adults. Therefore, the current findings add support to the notion that older adults may not rely on similar movement planning and execution strategies when performing tasks in VEs when compared to similar tasks in a natural environment.

Unlike the older adults, the children appear to use similar strategies in both virtual and natural environments for bimanual grasping. Specifically, it has been found that children tend to rely heavily on sensory feedback when grasping objects in natural environments [9]. In the current study, children produced peak reaching velocities that were similar to the young adults, yet their movement times were slower. These longer movement times were the result of an increased amount of time spent decelerating toward the target. Increased deceleration times can be used to infer a greater reliance on sensory information. Since sensory feedback is important for movement execution in children, our results suggest that providing an enriched sensory experience may improve their overall performance in VEs. Of interest in future work will be to determine whether children can achieve higher levels of performance in VEs if sensory information is enhanced/augmented when compared to natural environments. This could have significant implications with respect to motor skill learning for interaction tasks.

The perturbation conditions allowed us to investigate age differences in the visual control of movement in VEs. Overall, MT and offset results indicated similar movement performance between the ages of 18 and 50 years. These results suggest that design principles extracted from studies done on young adults may be applicable to middle-aged adults as well. In contrast, children and older adults exhibited distinct performance differences as a function of perturbation condition. While their performance was similar to the young and middle age groups for certain parameters and on certain conditions, the youngest and oldest age groups were slower and their movements were less coupled in other conditions. Further, the children continued to show an increased reliance on sensory feedback (i.e., longer deceleration times) whereas the older adults continued to rely on cautious movement planning (i.e., lower peak velocities). Overall, these results suggest that task conditions and age are critical factors when considering the design and functionality of VEs. Children and older adults do not plan movements or make use of sensory information in a similar fashion to young and middle-aged adults. Further, results are clearly task specific. This suggests that it is dangerous for designers to extrapolate performance in one task to other tasks. Instead, our results suggest that age-related performance must be investigated on a task by task basis for the generation of design principles.

B. *Implications for the Design of Training and Rehabilitation VEs*

Virtual environments have recently been touted as promising tools for training and rehabilitation [35]-[37]. A key consideration when designing a fully immersive virtual environment is that all sensory information provided to the user must be synthetically generated. As such, designers must make informed decisions about what sensory information to provide to the user and when that information should be provided. Several studies have been conducted to determine how to effectively provide sensory information to users for the performance of simple tasks in VEs [52]-[54]. Unfortunately, many of those studies have focused exclusively on the performance of young adults. As we begin to consider the multitude of applications for which VEs show promise, it is clear that users of all ages need to be considered (i.e., education and rehabilitation). The results of this study allow us to make concrete suggestions to designers of VEs. These relate to differences in needed sensory information between children, young and older adults, the transfer of training effects between virtual and real environments, and important differences between performance and learning applications.

Research has shown striking differences in the use of sensory feedback by children, young adults and the elderly as they perform motor tasks in natural environments [7],[8],[14]. Results of the current study replicate those findings and extend them to virtual environments. These findings suggest that designers of virtual environments may want to consider enhancing sensory feedback provided to the youngest and oldest users of virtual environments as a method of improving performance in those age groups. Grabowski & Mason [50] found that elderly participants performed simple reach to grasp tasks more effectively when luminance contrast was increased in the visual display. In contrast, young adult participants experienced a point of diminishing returns. Our current results and the results of this previous work suggest that sensory information tailored to a participant's age could lead to superior performance in virtual environments.

When considering education and rehabilitation applications, the capacity for virtual environments to enhance learning hinges on the user's ability to transfer gains made in the VE to improvements in performance in the real world. It has long been known in the human motor learning literature that successful transfer occurs when similarities in movement strategies between the practice and performance environment are greatest [55]. This phenomenon is called the encoding specificity principle [55]. In the current study we found that children, young, and middle-aged adults used similar bimanual strategies in the control condition to those reported for natural environments. This indicates that the sensory information available in our setup was sufficient to produce "normal" motor performance in the younger participant groups and could lead to positive transfer between the virtual and real

environments. In contrast the strategies used by the older adults in the VE were different than those reported in natural environments. These results suggest that the sensory characteristics present in our virtual environment did not sufficiently mimic natural environment conditions for our elderly participants. It is important to note that visual feedback in this study was impoverished and relatively crude (i.e., no hand representation, simple table surface and object representation, low luminance contrast levels). These results suggest that when designing environments for older adults, it may be necessary to design tasks and environmental feedback conditions that better mimic the visual feedback conditions available in the real world in order to elicit positive transfer between the two environments. In contrast, younger participants may see positive transfer with less realistic visual feedback conditions.

Finally, it is important to consider an apparent contradiction between the two previously mentioned implications. Specifically, we first suggested that designers may want to enhance the sensory feedback available in the virtual environment for younger and older participants. These enhancements could lead to sensory feedback that is more detailed and easily processed than what is available in natural environments (i.e., increased luminance contrast). Our next suggestion implies that sensory feedback may need to perfectly mimic what is available in natural environments in order to ensure positive transfer in learning applications for older users. This contradiction illustrates a third point that designers need to consider when determining how to provide sensory feedback; the task or application. It is clear that sensory feedback needs to be tailored, not only to the age of the user, but also to the application at hand. Specifically, applications that are performance based may benefit from sensory information that surpasses what is available in natural environments, whereas learning/transfer applications may need to better mimic the real world. A compromise between these two suggestions may be to use a "fading" technique where sensory feedback is initially enhanced to elicit improved performance but is faded during practice towards more realistic levels to enhance transfer of learned skills [55]. We are planning future studies to test this hypothesis.

ACKNOWLEDGEMENT

We would like to thank Brandon Bernardin, Stephanie Ehle, and Nick Yeutter for their assistance with data collection and analysis. This work was funded by the National Science Foundation, grant number 0916119.

REFERENCES

- [1] A.H. Mason, D.N., Rutherford, and P.J. Grabowski. "Bimanual performance in unpredictable virtual environments: A lifespan study." Proceedings of the Sixth International Conference on Advances in Computer-Human

- Interactions, ACHI 2013, IARIA Press, March 2013, pp. 263-268.
- [2] L.M. Howden and J.A. Meyer, "Age and Sex Composition 2010: Census Brief," Retrieved from URL <http://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf>, 11.25.2013.
 - [3] European Commision, "Eurostat Population Projections," Retrieved from URL http://epp.eurostat.ec.europa.eu/portal/page/portal/population/data/main_tables, 11.25.2013
 - [4] A. Dickinson, J. Arnott and S. Prior, "Methods for human-computer interaction research with older people," Behavioral Information Technology, Vol. 26, No. 4, July-August 2007, pp. 343-352.
 - [5] R.A. Muller, R.D. Rothermel, M.E. Behen, O. Muzik, T.J. Mangner, and H.T. Chugani, "Developmental changes of cortical and cerebellar motor control: A clinical positron emission tomography study with children and adults," Journal of Child Neurology, Vol 13, No. 11, November 1998, pp. 550-556.
 - [6] R.C. Chisholm and R. Karrer, "Movement-related brain potentials during hand squeezing in children and adults," International Journal of Neuroscience, Vol 19, No. 1-4, May 1983, pp. 243-258.
 - [7] M.L. Casselbrant, E.M. Mandel, P.J. Sparto, M.S. Redfern, and J.M. Furman, "Contribution of vision to balance in children four to eight years of age," Annals of Otolaryngology, Rhinology, and Laryngology, Vol. 16, No. 9, September 2007, pp. 653-657.
 - [8] L. Hay, C. Bard, C. Ferrel, I. Olivier, and M. Fleury, "Role of proprioceptive information in movement programming and control in 5 to 11-year old children," Human Movement Science, Vol 24, No. 2, April 2005, pp. 139-154.
 - [9] P.J. Sparto, M.S. Redfern, J.G. Jasko, M.L. Casselbrant, E.M. Mandel, and J.M. Furman, "The influence of dynamic visual cues for postural control in children aged 7-12 years," Experimental Brain Research, Vol 168, No. 4, January 2006, pp. 505-516.
 - [10] H.K. Kuo and L.A. Lipsitz, "Cerebral white matter changes and geriatric syndromes: Is there a link?" Journals of Gerontology Series A: Biological Sciences and Medical Sciences, Vol. 59, No. 8, August 2004, pp. 818-826.
 - [11] B.B. Andersen, H.J. Gundersen, and B. Pakkenberg, "Aging of the human cerebellum: A stereological study," Journal of Comparative Neurology, Vol. 466, No. 3, November 2003, pp. 356-365.
 - [12] N. Raz and K.M. Rodrigue, "Differential aging of the brain: Patterns, cognitive correlates and modifiers," Neuroscience & Biobehavioral Reviews, Vol 30, No. 6, June 2006, pp. 730-748.
 - [13] P. Rossini, M. Desiato, and M. Caramia, "Age-related changes of motor evoked potentials in healthy humans: Non-invasive evaluation of central and peripheral motor tracts excitability and conductivity," Brain Research, Vol 593, No. 1, October 1992, pp. 14-19.
 - [14] J.H. Yan JH, J.R. Thomas, and G.E. Stelmach, "Aging and rapid aiming arm movement control," Experimental Aging Research, Vol 24, No. 2, April-June 1998, pp. 155-168.
 - [15] A. Sterr and P. Dean, "Neural correlates of movement preparation in healthy ageing," European Journal of Neuroscience, Vol 27, No. 1, January 2008, pp. 254-260.
 - [16] J.H. Hwang, Y.T. Lee, D.S. Park, and T.K. Kwon, "Age affects the latency of the erector spinae response to sudden loading," Clinical Biomechanics, Vol 23, No. 1, January 2008, pp. 23-29.
 - [17] S. Chaput and L. Proteau, "Aging and motor control," Journals of Gerontology Series B: Psychological Sciences and Social Sciences, Vol 51, No. 6, November 1996, pp. 346-355.
 - [18] K.E. Light, "Information processing for motor performance in aging adults," Physical Therapy, Vol 70, No. 12, December 1990, pp. 820-826.
 - [19] N. Inui, "Simple reaction times and timing of serial reactions of middle-aged and old men," Perceptual Motor Skills, Vol 84, No. 1, February 1997, pp. 219-225.
 - [20] R.K. Shields, S. Madhavan, K.R. Cole, J.D. Brostad, J.L. Demeulenaere, C.D. Eggers, et al, "Proprioceptive coordination of movement sequences in humans," Clinical Neurophysiology, Vol 116, No. 1, January 2005, pp. 87-92.
 - [21] J.L. Contreras-Vidal, "Development of forward models for hand localization and movement control in 6- to 10-year-old children," Human Movement Science, Vol 25, No. 4-5, October 2006, pp. 634-645.
 - [22] L.J. Schut, "Motor system changes in the aging brain: What is normal and what is not," Geriatrics, Vol 53, Suppl 1., September 1998, pp. S16-S19.
 - [23] M.R. Roos, C.L. Rice, and A.A. Vandervoort, "Age-related changes in motor unit function," Muscle Nerve, Vol 20, No. 6, June 1997, pp. 679-690.
 - [24] R.D. Seidler, J.L. Alberts, and G.E. Stelmach, "Changes in multi-joint performance with age," Motor Control, Vol 6, No. 1, January 2002, pp. 19-31.
 - [25] D.H. Romero and G.E. Stelmach, "Changes in postural control with aging and parkinson's disease," IEEE Engineering in Medicine and Biology, Vol 22, No. 2, March-April 2003, pp. 27-31.
 - [26] K. Günther, P. Schäfer, B. J. Holzner, and G. W. Kemmler, "Long-term improvements in cognitive performance through computer-assisted cognitive training: A pilot study in a residential home for older people," Aging & Mental Health, Vol. 7, No. 3, May 2003, pp. 200-206.
 - [27] D. Hawthorn, "Interface design and engagement with older people," Behavioral Information Technology, Vol. 26, No. 4, July 2007, pp. 333-341.
 - [28] E. Strommen, "Children's use of mouse-based interfaces to control virtual travel," In Proceedings of the ACM Conference on Human Factors in Computing Systems: Celebrating Interdependence: CHI '94, April 1994, ACM Press, New York, NY, pp. 405- 410.
 - [29] M.W. Smith, J. Sharit, and S.J. Czaja, "Aging, motor control, and the performance of computer mouse tasks," Human Factors, Vol. 41, No. 3, September 1999, pp. 389-396.
 - [30] B. Laursen, B.R. Jensen, and A. Ratkevicius, "Performance and muscle activity during computer mouse tasks in young and elderly adults," European Journal of Applied Physiology, Vol. 84, No. 4, April 2001, pp. 329-336.
 - [31] A. Worden, N. Walker, K. Bharat, and S. Hudson, "Making computers easier for older adults to use: area cursors and sticky icons," Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '97), ACM Press, April 1997, pp. 266-271.
 - [32] R.C. Allen, M.J. Singer, D.P. McDonald, and J.E. Cotton, "Age differences in a virtual reality entertainment environment: A field study," Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society (IEA/HFES 2000), July 2000, pp. 542-545.
 - [33] F.A. McCreary, and R.C. Williges, "Effects of Age and Field-of-View on Spatial Learning in an Immersive Virtual Environment," Proceedings of the 42nd Annual Meeting of the Human Factors and Ergonomics Society (HFES 1998), October 1998, pp. 1491-1495.
 - [34] D.P. McDonald, D.A. Vincenzi, R.V. Muldoon, R.R. Tyler, A.S. Barlow, and J.A. Smither, "Performance differences between older and younger adults for a virtual environment

- locomotor task, In M.W. Scerbo and M. Mouloua (Eds.), *Automation technology and human performance: Current research and trends*, Lawrence Erlbaum Associates: Mahwah, New Jersey, 1999, pp. 262-269.
- [35] M.K. Holden, "Virtual environments for motor rehabilitation: review," *Cyberpsychology and Behavior*, Vol. 8, June 2005, pp. 212-219.
- [36] M. Wang and D. Reid, "Virtual reality in pediatric neurorehabilitation: attention deficit hyperactivity disorder, autism and cerebral palsy," *Neuroepidemiology*, Vol. 36, No. 1, November 2010, pp. 2-18.
- [37] S. Bermúdez i Badia and M. da Silva Cameirão, "The Neurorehabilitation Training Toolkit (NTT): a Novel Worldwide Accessible Motor Training Approach for at-Home Rehabilitation after Stroke," *Stroke Research and Treatment*, February 2012, Article ID 802157, 13 pages, doi:10.1155/2012/802157
- [38] J.H. Cauraugh, S.B. Kim, and A. Duley, "Coupled bilateral movements and active neuromuscular stimulation: intralimb transfer evidence during bimanual aiming," *Neuroscience Letters*, Vol. 382, July 2005, pp. 39-44.
- [39] A.H. Mason, "Coordination and control of bimanual prehension: Effects of perturbing object location," *Experimental Brain Research*, Vol. 188, No. 1, June 2008, pp. 125-139.
- [40] J.A.S. Kelso, D.L. Southard, and D. Goodman, "On the coordination of two-handed movements," *Journal of Experimental Psychology: Human Perception and Performance*, Vol. 5, May 1979, pp. 229-238.
- [41] G.M. Jackson, S.R. Jackson, and A. Kritikos, (1999). "Attention for action: Coordination of bimanual reach-to-grasp movements," *British Journal of Psychology*, Vol. 90, No. 2, May 1999, pp. 247-270.
- [42] G.P. Bingham, K. Hughes, and M. Mon-Williams, "The coordination patterns observed when the hands reach-to-grasp separate objects," *Experimental Brain Research*, Vol. 184, No. 3, January 2008, pp. 283-293.
- [43] J. Diedrichsen, R. Nambisan, S. Kennerley, and R.B. Ivry, "Independent on-line control of the two hands during bimanual reaching," *European Journal of Neuroscience*, Vol. 19, No. 6, April 2003, pp. 1643-1652.
- [44] P.J. Sparto, M.S. Redfern, J.G. Jasko, M.L. Casselbrant, E.M. Mandel, and J.M. Furman, "The influence of dynamic visual cues for postural control in children aged 7-12 years," *Experimental Brain Research*, Vol. 168, January 2006, pp. 505-516.
- [45] S. Chaput and L. Proteau, "Aging and motor control," *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, Vol. 51, November 1996, pp. 346-355.
- [46] A.M. Wing, A. Turton, A., and C. Fraser, "Grasp size and accuracy of approach in reaching," *Journal of Motor Behavior*, Vol. 18, No. 3, August 1986, pp. 245-260.
- [47] S. Zoia, E. Pezzetta, L. Blason, A. Scabar, M. Carrozzini, M. Bulgheroni, and U. Castiello, "A comparison of the reach-to-grasp movements between children and adults: a kinematic study," *Developmental Neuropsychology*, Vol. 30, No. 2, June 2006, pp. 719-738.
- [48] G.E. Stelmach, P.C. Amrhein, and N.L. Goggin, "Age differences in bimanual coordination," *Journal of Gerontology Series B: Psychological Sciences and Social Sciences*, Vol. 43, January 1988, pp. 18-23.
- [49] A.H. Mason, J.L. Bruyn, and J.C. Lazarus, "Bimanual coordination in children: Manipulation of object distance," *Experimental Brain Research*, Vol. 231, No. 2, November 2013, pp. 153-164.
- [50] P. Grabowski and A.H. Mason. *Vision for Motor Performance in Virtual Environments Across the Lifespan*, Virtual Reality and Environments, Dr. Cecília Sık Lányi (Ed.), ISBN: 978-953-51-0579-4, InTech, DOI: 10.5772/37341. April 2012. Available from: <http://www.intechopen.com/books/virtual-reality-and-environments/vision-for-motor-performance-in-virtual-environments-across-the-lifespan>
- [51] P.L. Weir, B.J. Mallat, J.L. Leavitt, E.A. Roy, and J.R. Macdonald. "Age-related differences in prehension: the influence of task goals," *Journal of Motor Behavior*, Vol. 30, No. 1, January 1998, pp. 79-89.
- [52] A.H. Mason, "An experimental study on the role of graphical information about hand movement when interacting with objects in virtual reality environments," *Interacting with Computers*, Vol. 19, No. 3, May 2007, pp. 370-381.
- [53] A.H. Mason and B.J. Bernardin, "The role of visual feedback when grasping and transferring objects in a virtual environment," *Proceedings of the 5th International Conference on Enactive Interfaces*, November 2008, pp. 111-116.
- [54] A.H. Mason and C.L. MacKenzie, "The effects of visual information about selfmovement on grip forces when receiving objects in an augmented environment," *Proceedings of the 10th international IEEE symposium on haptic interfaces for virtual environments and teleoperator systems*, March 2002, pp. 105-112.
- [55] R.A. Magill (2011) *Motor Learning and control: Concepts and Applications* (9th edition), New York, McGraw-Hill.

High End Computing Using Advanced Archaeology and Geoscience Objects

Claus-Peter Rückemann

Westfälische Wilhelms-Universität Münster (WWU),
Leibniz Universität Hannover,
North-German Supercomputing Alliance (HLRN), Germany
Email: ruckema@uni-muenster.de

Abstract—This paper presents the results from the creation of advanced long-term knowledge resources. Focus goals are multi-disciplinary knowledge documentation, discovery, and sustainability. Application scenarios with complex knowledge architectures require different computational workflows and resources. The paper discusses comprehensive case studies with advanced knowledge objects from archaeology and geosciences disciplines. It delivers results and experiences on creating intelligent and sustainable Integrated Information and Computing System components and systems developed for more than twenty years. The new universal knowledge resources and flexible collaboration framework allow multi-disciplinary documentation for any object as well as advanced scientific computing access and enable overall flexibility for interfacing High End Computing resources, which get increasingly important for integration, improving the quality of result matrices, dynamical tasks and highly efficient discovery workflows and processes. This way, structured objects and universal classification are a central means for long-term integration of information systems and supercomputing resources with any kind of workflow and discipline.

Keywords—*Information Systems; Knowledge Resources; Advanced Scientific Computing; Classification; Archaeology; Geosciences; UDC; Integrated Systems; High End Computing.*

I. INTRODUCTION

The more the demand increases for creating sustainable and flexible knowledge resources, the more we require long-term conceptional and computational methods for discovery in multi-disciplinary and heterogeneous resources [1]. Even widely accessible common collections of information and data, are missing reliability, validation, and long-term sustainability – problems, which are passed far into the future with long-term tasks. This is resulting from principle problems of implementation: There are no foundations of a suitable long-term strategy, documentation, tools, and resources. This demands static and dynamical components in all parts of an implementation.

Multi-disciplinary knowledge profits from long-term documentation. Creating extensive knowledge resources requires long-term means. Making use of extensive knowledge resources, especially over long periods of time requires the development of information and knowledge itself and an integration of media and features for transporting and working with information. Content has to be developed for long periods of time. This includes new research, including historical information, and extending multi-disciplinary references. The focus question for documentation, operating on information and computing is: How can complex systems be built, developed, and extended over the necessarily long periods of time?

With the application components, e.g., information system components like databases, mostly form monolithic and even proprietary blocks. Their life cycle is mostly much shorter than long-term content development. On the side of computing challenges it is possible to create solutions for very perishable present resources. Any more complex problem cannot be considered “solved” for future architectures and applications. Many information and knowledge resources cannot be used without the original context, e.g., computing resources any more. Up to now context of information science cannot be described by common means to a reasonable extend. This leads to another essential question: What information and knowledge on content and context can be preserved for medium- and long-term usage when the complexity of an overall system will be unpredictably high?

The long-term strategy created here is based on an implementation architecture, which includes long-term knowledge resources with the resources and development. In this paper we concentrate on the archaeological and geosciences topics being part of the knowledge resources. The foundations should enable the essential processing of archaeological, geoscientific, geophysical, geological, spatial and other data as well as a thorough documentation of all aspects of content and context and the exploitation of advanced scientific computing methods and resources for maximum flexibility.

Regarding the status of long-term knowledge we have to distinct between two important main parts, which this paper presents: The knowledge resources and the Integrated Information and Computing Systems (IICS), both already used in practice. The presentation of the long-term issues and the new features and results presented in this paper shows the potential of IICS being based on multi-disciplinary documentation for this purpose. Anyhow, it can only describe a tiny fraction of the multitude of possible features.

This paper is organised as follows. Section II describes the motivation and Section III summarises related work. Section IV introduces architecture and implementation for the IICS. Sections V and VI describe the long-term strategy and discuss the advanced knowledge object creation. Sections VII, VIII, IX, and X show high-end implementation case studies of advanced objects in context of a geoscience-archaeology IICS: A digital archaeology library, the computing features and components, the integration of external information, and basic mechanisms and workflows. Section XI presents an evaluation including processing, computing, and classification aspects. Section XII summarises the conclusions and future work.

II. MOTIVATION

For small volumes of data, small numbers of objects or primitive workflows most computing and storage requirements are neglectable. Using available data with complex processing workflows soon shows up with requirements increasing more than exponential. With a large amount of data and large numbers of objects this will become a huge challenge. The more, as commonly there are many different object types and even inside a type any of these objects will differ. One group of objects may consist of some thousand digital samples from a media database, another group may be physical objects carrying specific descriptions, and a third may be data sets and workflows for seismic processing and simulation. Besides that the workflows can be arbitrary complex, the requirements on the groups can widely differ. With one group the permanent storage may involve Terabytes of data and less computation. With another group the input data and parametrisation may be much smaller in size but the computational requirements can take up to several days per workflow step.

Currently, other approaches fail on integrative multi-disciplinary concepts as well as on the integration of structure, long-term preservation, and computational and information system facilities. An important reason is that those available approaches start from an implementation of features and not from the knowledge itself. There are no frameworks providing the necessary concepts and features for integration of data, workflows, knowledge and computing resources, and operation. For example, a lot of advanced geoscientific processing cannot be reproduced after a few years, even if the data and results are still existing. Means for extended long-term interpretation and analysis are missing.

It is a huge challenge that, besides data creation not being able to support sufficiently comprehensive documentation, the widely used technology, e.g., document formats, Uniform Resource Locators (URL), and Web Services are not persistent over longer periods of time, e.g., for static objects file formats do change, for applications and services the implementations and features will change. Therefore, information structures built from such technologies will become inaccessible. Long-term knowledge creation cannot rely on this. From the complex systems' point of view any of those building elements are not suitable for describing objects and creating long-term knowledge resources. Anyhow the original sources and building elements are needed for documentation of the original content and context. Therefore, knowledge creation has to separate the essentials of knowledge from technology, resources, and other tools while at the same time respecting their importance. Even worse, that workflows, algorithms, resources and their management cannot be guaranteed for long-term availability.

The topic is very complex and experiences with long-term knowledge creation are out of scale of the time interval of most researchers. Especially, it has been found to be less difficult for groups with a strong background of classical academic education to understand the problem itself, than it is for groups with a "technical-only" background to realise the multiple benefits of classification.

Examples of long-term creation of knowledge and the implementation of applications building on these resources are presented in the following sections, showing that such scenario

can hardly be managed in a comparable way with other available methods and concepts. The case studies from archaeology and geosciences disciplines are using the architecture, knowledge resources, computing interfaces, and features. The resources and interfaces have been continuously developed and improved for more than twenty years already. The following passages describe the requirements in the context of this new study and implementation. Further details on requirements have been described in several case studies cited.

A. Value of data

The value of data is a central driving force for creating sustainable knowledge resources, the more as data is increasingly important for long period of times. Long-term in cases of sustainable high-value data means many decades of availability and usability. Therefore, usability, security, and archiving are most important aspects of the value of data sets. Value is not the price a data set can be sold as there are many individual factors. The long-term studies, as the "Cost of Data Breach" study initiated by Symantec at the Ponemon Institute [2] summarise that the costs related to data loss are high and do increase every year [3]. Straight approaches for calculating individual risks and data loss, as with the Symantec Data Breach Calculator [4] illustrate the effects.

B. Sustainability and long-term issues

1) Recommendations of the German science council:

With information systems containing content that has to be created and cared for a sustainable long-term operation and provisioning, the main aspect is not the technology for a limited implementation. The focus is the specification and structure of the knowledge, which must be defined with the owner of the data. In many cases an exploitation for scientific discovery is imperative. The German science council (Wissenschaftsrat) recommends the exploitation of content and its public provision for content of special significance [5].

The science council provides recommendations for information infrastructures, which should be considered for wider implementation. The central aspects are exploitation, standardisation, and long-term archiving and archival storage. This is of prescind importance as far as the content and services are of public interest and can be defined core business of an institutional body. As from the recommendations, the planning phase for infrastructures should at least consider medium duration. Project funding of existing structures is not adequate. Regarding the context of the information, the operation must be considered a permanent task. Nevertheless, the German science council recommends an accompanying, infrastructure related research, which may be considered on a project base. This can be especially suitable for the component development, whereas data, content, and structures are non project matter. Aspects of funding, profiles, and digitising are of special concern within federal structures [6].

2) *Experiences from national and international studies:* Up to now, knowledge integration and its discovery aspects are widely discussed within national and international context. The existing concepts the data on knowledge and accompanying recommendations are too specialised and too strict [7], [8], [9], [10], [11], [12], [13], especially for the required long-term and multi-disciplinary use. As the concepts are artificial,

the integration with natural structures is most difficult. Even if developed over years, the basic concepts and structures rarely fit for many disciplines, services, and resources. Neither creation nor operation can profit from the existing practice and frameworks. Knowledge resources have to be considered universal, for any data as well as any workflow. Practically, there is no difference between explicit and tacit knowledge as promoted with various previous approaches. Auditing is not depending on a special kind of knowledge resources and from knowledge resources point of view it cannot be practically separated into asset structures, data auditing for projects, organisations or maintenance.

Knowledge resources should be defined for sustainability and long-term aspects. The most important aspect for decision making is experience. This has to be considered for creating all aspects of knowledge resources. Universal knowledge resources require a high level of multi-disciplinary comprehension. The workflows may not influence the structures of the knowledge resources. They should be created for selecting a focus. They should be able to incorporate and handle data of disciplines, content, algorithms, and documentation. They should allow integration with implementation and operation. They should allow for different required views.

III. RELATED WORK

There is no wider concept and implementation known comparable to the solution presented, described and implemented here. Nevertheless, there are concepts for components, implementations, and terminology. Previous work [14], [15], [16], [17], [18] has delivered important basic concepts and components, e.g., Integrated Information and Computing Systems, dynamical components, and taxonomy. Efficiently structuring, classifying, e.g., with UDC, and storing data are key issues with flexible and sustainable long-term knowledge resources. Taxonomy is the science and practice of classification. An important facetted classification is the Universal Decimal Classification (UDC) [19]. According to Wikipedia currently about 150000 institutions, mostly libraries, are using basic UDC classification worldwide [20], e.g., with documentation of their resources, library content, bibliographic purposes, for digital and realia objects. This is mostly restricted to publications and references but not of general knowledge and applications. Some aspects can be studied from the goals of knowledge discovery [21], which is becoming increasingly important. Other aspects are handled with search algorithms, which currently are still very primitive regarding knowledge creation and usage. Developing general categories and classifications is a long-term process [22]. Many generations of researchers and institutions have contributed to its development [23], [19]. A general usage for any kind of objects as it had been originally intended with the “universal” classification is still not practiced. A significant part of the long-term problem for putting a general application into practice is the complexity of multi-disciplinary knowledge and workflows.

IV. ARCHITECTURE AND IMPLEMENTATION

As far, it is not commonly possible to treasure content currently used for being preserved in order to create really long-term usable content. Even much more difficult that an implementable solution for any form of long-term context is

even in wide distance. In general, only a very small percentage of disciplines and researchers are familiar with knowledge classification and applications. The more, multi-disciplinary classification is currently only in the focus of third parties. Operational resources and features are considered to be short-term issues whereas information, knowledge, and respective resources and features must be considered of long-term significance. Case studies showed that long-term development requires a strong sustainability of content, context, and computation. This means, the most important part of these systems is the knowledge resources containing the content and context documentation to any extent necessary for describing the activities and isolating the perishable components for context documentation.

A solid classification cannot be done automatically. The more, it cannot be done automatically for use with IICS. This does not interfere with or limit in any way the amount of data handled by a workflow making use of classified knowledge resources. Anyhow, in fact that different views are possible, it is reasonable to have classification views from the origin, from main disciplines or from the developers in order to increase the quality of references. The architecture respects these conditions. The architecture integrates several components especially developed for long-term integration:

- Universal classification,
- Structure,
- Multi-disciplinary knowledge documentation,
- Resources support, e.g., computation and storage,
- Creating, describing, executing workflows, ...

All the implemented components have proved to fulfill the requirements for long-term vitality and extendability.

The following sections explain in detail how a successful implementation of an integrated system has been created and operated using knowledge resources and classification for information system usage.

A. Architecture for documentation and development

The architecture implemented for an economical long-term strategy is based on different development blocks. Figure 1 shows the three main columns: Applications resources, knowledge resources, and originary resources. The central block in the “Collaboration house” framework architecture [18], are the knowledge resources, scientific resources, databases, containers, and documentation (e.g., LX [14], databases, containers, list resources). These can be based on and refer to the originary resources and sources (photos, scientific data, literature). Application resources and components (Active Source, Active Map, local applications) are implementations for analysing, utilising, and processing data and making the information and knowledge accessible. These three blocks are supported by services interfaces. The interfaces interact with the physical resources, in the local workspace, in the compute and storage resources the knowledge resources are situated, and in the storage resources for the originary resources. All of these do allow for advanced scientific computing and data processing as well as the access of compute and storage resources via services interfaces. The resources’ needs depend on the application scenarios to be implemented for user groups.

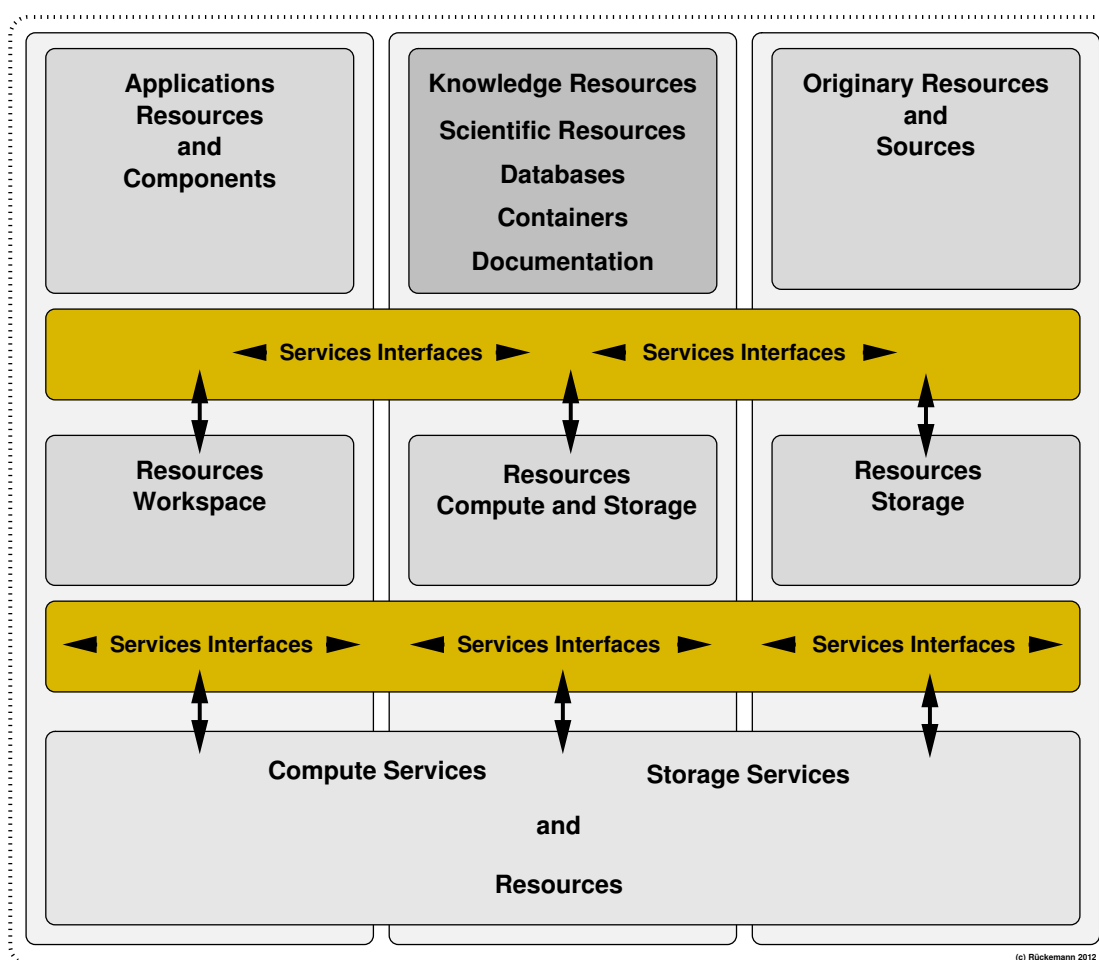


Figure 1. Architecture: Columns of practical dimensions. The knowledge resources are the central component within the long-term architecture. They are aligned by application resources and origiary resources. The knowledge resources are used as a universal component for compute and storage workflows.

B. Components: Applications, knowledge, and sources

The main information, data, geo-referencing, and algorithms for all presented components and examples are provided by the LX Foundation Scientific Resources [14]. This deploys the structure and classification of objects necessary for a reasonable implementation. Besides the LX structure the already established Universal Decimal Classification (UDC) [19] has been integrated for objects [15] as it provides a hierarchical and multi-lingual, faceted classification for any topic and allows implementing a faceted analysis with enumerative scheme features, as well as to create new classes by using relations and grouping. In multi-disciplinary object context, this empowers to use workflows combining keywords, enumerative concepts and full-text analysis with a faceted analysis.

Besides the academic, industrial, and business application scenarios in focus of the GEXI collaborations' case studies [16] it is an important factor to integrate the necessary documentation and computing facilities with systems like an Universal IICS (UIICS). An implementation of interfaces for using structure and classification with appropriate Archaeological IICS system components has been created for several simple (SAMPLE, COLLECTION, CONTEXT, DISCIPLINE) and slightly more complex workflows (CONNECT, REFERTO-TOPIC, REFERTO-SPATIAL, VIEW-TO, VIEW-FROM) [18].

For the topics and content discussed here, geoscientific and archaeological information and processing are the core content. Data in this context necessarily includes applications and algorithms. Besides the above implemented features, it is optional to support any visualisation tool, processing algorithms, cartographic and mapping features and many more tools from secondary sciences, e.g., spatial algorithms and components, UDC (1-0/-9). These features can be used with the objects in any way that will be necessary to describe data and automate workflows.

C. Classification, keywords, and interfaces

The interfaces allow to use the various resources. A central element is the classification and structure of the knowledge resources as it increases the flexibility of the long-term development. Table I compares some features of classification and keywords used for object description.

TABLE I. UDC CLASSIFICATION AND KEYWORDS COMPARISON.

| UDC | Keywords |
|----------------------|--|
| Internationalisation | Methodical support, partial internationalisation |
| Codes | Code table support |
| High level of detail | Medium level of detail |

Interfaces can be used in order to access and use objects. This includes filtering, combination, workflow and data processing and so on. In summation, this allows the integration of all data, objects, and resources available: scientific and discipline data, lexicographical and bibliographical data GPS data, geospatial information, processing algorithms, executable software, and many more, including realia objects.

The result means, we need to integrate multi-disciplinary information, allowing different views on the same context and allow even different paths for exploring knowledge.

V. LONG-TERM STRATEGY

The immanent attribute of preservation is time. Time is one of the factors limiting life and existence, e.g., of realia objects in archives, museums, and libraries as well as with implementations of mathematically based electronic machines and components.

Whereas looking from inside a traditional discipline, information seems to be complete and appears to increase slowly. On the other hand there is huge complementary information that cannot be described isolated by one discipline and tendency is increasing over the time and complexity.

Table II presents the result of a reasonable categorisation that has been found from practicing the knowledge resources creation and use for several decades. It shows a more detailed compilation of categorised features and components for an expected actuality time range. In this context, the goal for long-term means > 50 years, medium-term > 15 years, and short-term < 15 years.

TABLE II. TIME-RANGE GOALS WITH IICS COMPONENTS (SELECTION).

| <i>Long-term</i> | <i>Medium-term</i> | <i>Short-term</i> |
|----------------------------|----------------------|-----------------------|
| Knowledge | Applications | Context |
| Containers | Interfaces | Sources |
| LX Scientific Resources | DOI, URN, PURL | URL |
| UDC | Converters | Media |
| Keywords | Active Source | Converters |
| Virtualisation information | Storage resources | Computing resources |
| Algorithms | Distributed services | Compiler, Executables |
| Content | Virtualisation | MPI, OpenMP |
| Context information | Complex implement. | Batch systems |
| Relations & references | Application features | Web Services |
| Internationalisation | OS features | Communication |
| Processing & workflows | Library features | Middleware |

These components, described by a representative selection in Table II, can cover all aspects of knowledge creation, application, and system implementation. For example, with the implementation, the resources and containers are consisting of thousands of pages. For the presentation within the following sections a small excerpt of the objects and classification can be shown.

The long-term objects must be able to contain the essential knowledge, even as medium- and short-term objects cannot be preserved or made persistent as, e.g., DOI (Digital Object Identifier), URN (Uniform Resource Name), URL (Uniform Resource Locator), and PURL (Persistent Uniform Resource Locator) will vanish and context and sources may fade away

as well as OS (Operating System) features used. Therefore, we have to distinct between the real instance of a DOI and URL or a context situation and a descriptive reference of these objects. These descriptive references can contain as much information and knowledge as possible (for example DOI, URL, context description, sources).

VI. ADVANCED KNOWLEDGE OBJECT CREATION

A. Knowledge objects

The LX Scientific Resources [14], [18] are used as base knowledge resources for the following case studies, delivering structure, content, classification, and providing methodologically exploitable information. The elementary knowledge objects can have any required extent. Some objects are naturally small, e.g., translations or acronym expansions, others can have hundreds of pages with references and an arbitrary number of subobjects.

An object may contain any number of subobjects. As soon as it seems reasonable from the content and context a subobject can become an object or a number of subobjects can be grouped into a container. The object data excerpts show source, structure, references, and other features used. Media samples are referenced, but the extensive data is not explicitly given here.

Objects and information are naturally distributed, for example, spatially and logically. Ranking within any object matrix is subjective, especially when only the information from a few isolated and not comprehensive sources is considered. This holds true for any object referred to, for example, media objects, sources and publication. In the end, the isolated subset leads to a reduced quality within the selection. Therefore, some external resources have been included as examples.

B. Classification

The operated knowledge resources, based on the LX Foundation Scientific Resources [18], incorporate UDC classification for any discipline and purpose, e.g., for knowledge discovery and workflows.

A practical summarising excerpt subset of UDC codes used for computation with the available knowledge resources is given in Table III. In this context, the following examples explain multi-disciplinary documentation, views, and computational issues from several disciplines and topics, grouped by application:

- Digital archaeological library examples,
- Computing aspects,
- Knowledge resources and external information,
- Mechanisms and workflow case study on geo-objects.

TABLE III. ARCHAEOLOGY KNOWLEDGE RESOURCES CLASSIFICATION.

| UDC Code | Description |
|------------|--|
| UDC:902 | Archaeology |
| UDC:903.2 | Artefacts |
| UDC:904 | Cultural remains of historical times |
| UDC:738 | Ceramic arts. Pottery |
| UDC:738.8 | Various ceramic objects |
| UDC:629.5 | Watercraft engineering. Marine engineering. Boats. Ships ... |
| UDC:656 | Transport and postal services |
| UDC:741 | Drawing in general |
| UDC:691.2 | Natural stones. Other mineral materials |
| UDC:664.7 | Cereal technology. Flour and corn milling. Grain processing |
| UDC:664 | Food industry |
| UDC:641.5 | Preparation of foodstuffs and meals. Cookery |
| UDC:(32) | Ancient Egypt |
| UDC:(37) | Italia. Ancient Rome and Italy |
| UDC:(38) | Ancient Greece |
| UDC:(4) | Europe |
| UDC:(44) | France. French Republic. République Française |
| UDC:(450) | Italy. Republic of Italy. Repubblica Italiana |
| UDC:(460) | Spain. Kingdom of Spain, Reino de España |
| UDC:(495) | Greece. Hellenic Republic. Elliniki Dimokratia |
| UDC:(23) | Above sea level. Surface relief. Above ground ... |
| UDC:(24) | Below sea level. Underground. Subterranean |
| UDC:069.51 | Museum pieces |
| UDC:002 | Documentation. Books. Writings. Authorship |
| UDC:770 | Photography and similar processes |

The figures shown in the case study examples are computed from the content of knowledge resources and filtered with the Integrated Information and Computing Systems (IICS), using photo media samples (media samples and object entries © C.-P. Rückemann, 2011, 2012, 2013). It must be emphasised that the applications can provide any type of objects, high resolution media, and detailed information. An example how an environment and context can look like has been shown and discussed with special features of the knowledge resources and IICS (Figure 1 in [24]).

On the one hand, the object entries demonstrate the structure and references, e.g., to the attributes of the objects as classification, form, material, and location. On the other hand, the figures illustrate the result matrices for collections, context, and integration of multi-disciplinary information.

VII. DIGITAL ARCHAEOLOGICAL LIBRARY EXAMPLES

In combination with the above shown features, objects in digital archaeological libraries have been enriched with various information, e.g., on museum, library information, archives, network information, mapping services, locations and Points Of Interest (POI).

Due to the knowledge resources organisation, the objects can be used in references as well as in the cache for interactive components at any stage within the workflow process. Combining the structure and classification with the silken selection algorithms leads to very flexible, multi-disciplinary interfaces.

Each group of digital images shows a result matrix from a selection process. The following figures illustrate resulting objects from the digital library of the LX Scientific Resources with multi-disciplinary background, in these examples regarding material, function, and model or reconstruction purposes.

A. Knowledge objects, classification, and computation

The following example (Figure 2) shows two result matrices from the pottery and amphores context for the presentation of realia objects (row 1) and available documentation and reconstruction presentation of realia objects (row 2). Both rows of media samples in the computed result matrices are by option aligned left to right from West to East:

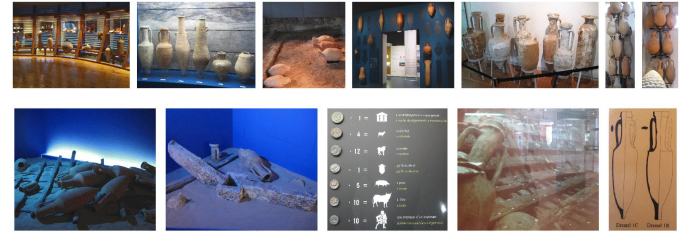


Figure 2. Result matrices – Pottery, UDC: 902, 738, 069.51, 002, ...
Row 1: Realia objects presentation, aligned West to East.
Row 2: Documentation and reconstruction presentation, aligned West to East.

Listing 1 shows the associated knowledge object entries for the realia objects. in Figure 2, row 1.

```

1 Valentia [Archaeology, Geophysics, ...]:
2   Object:      Amphores.
3   Object-Type:  Realia object.
4   Object-Location: Valentia Edetanorum,
5                 València, Spain.
6   Object-Relocation: Museum, Diputacio de
7                 Valencia, Centre Cultural la Beneficencia,
8                 València, Spain.
9   %%IML: media: ... img_7454.jpg
10  %%IML: UDC-Object
11  : [902+903.2+904]+738+738.8+(37)+(4)+(24)
12  %%IML: UDC-Relocation:069.51+(4)+(460)+(23)
13 Valentia [Archaeology, Geophysics, ...]:
14   ...
15 Valentia [Archaeology, Geophysics, ...]:
16   ...
17 Barcino [Archaeology, Geophysics, ...]:
18   ...
19   Object-Type:  Presentation, arrangement.
20   Object-Location: Museu d'Arqueologia de
21                 Catalunya, Barcelona, Spain.
22   %%IML: UDC-Object:002+770
23   %%IML: UDC-On-Content
24   : [902+903.2+904]+738+738.8+(37)+(4)+(24)
25   %%IML: UDC-Location:069.51+(4)+(460)+(23)
26 Cemenelum [Archaeology, Geophysics, ...]:
27   Object:      Amphores.
28   ...
29   Object-Location: Cemenelum, Nice-Cimiez,
30                 France.
31   Object-Relocation: Musée et Site Archéologique
32                 Cemenelum, Nice-Cimiez, Ville du Nice, France.
33   %%IML: UDC-Object
34   : [902+903.2+904]+738+738.8+(37)+(4)+(24)
35   %%IML: UDC-Relocation:069.51+(4)+(44)+(23)
36 Altinum [Archaeology, Geophysics, ...]:
37   Object:      Amphores.
38   ...
39   Object-Location: Altinum, Altino, Venice,
40                 Italy.
41   Object-Relocation: Museo Archeologico Nazionale
42                 di Altino, Venice, Italy.
43   %%IML: UDC-Object
44   : [902+903.2+904]+738+738.8+(37)+(4)+(24)
45   %%IML: UDC-Relocation:069.51+(4)+(450)+(23)
46 Altinum [Archaeology, Geophysics, ...]:
47   Object:      Amphores.
48   ...

```

Listing 1. Knowledge resources – Entries for realia objects.

The computation attributes are: Realia objects, presentation, Greek, Roman, pottery, amphores, ships, Mediterranean, location West to East.

For this example, the object entries are reduced to the significant information used for the computation. The sort order has been decided from the geo-coordinate latitudes of the location information references by the knowledge objects.

The first row (Figure 2 and Listing 1) show the resulting objects: Valentia Edetanorum (València), Barcelo (Barcelona), Cemenelum (Nice-Cimiez), Altinum (Venice), Altinum (Venice).

Second row: 3× Valentia Edetanorum (València), 2× Cemenelum (Nice-Cimiez). Computation attributes are: Documentation, reconstruction presentation, Greek, Roman, pottery, amphores, ships, Mediterranean, location West to East. Listing 2 shows associated knowledge object entries for the documentation entries in Figure 2, row 2.

```

1 Valentia [Archaeology, Geophysics, ...]:
2   Object:      Amphores in ship wreck.
3   Object-Type: Documentation, arrangement.
4   Object-Location: Valentia Edetanorum,
5     València, Spain.
6   Object-Relocation: Museum, Diputacio de
7     Valencia, Centre Cultural la Beneficencia,
8     València, Spain.
9   %%IML: media: ... img_7455.jpg
10  %%IML: UDC-Object:002+770
11  %%IML: UDC-On-Content
12  : [902+903.2+904]+738+738.8+(37)+(4)+(24)
13  %%IML: UDC-Location:069.51+(4)+(460)+(23)
14 Valentia [Archaeology, Geophysics, ...]:
15   Object:      Amphores and anchor and ship
16     wreck.
17   ...
18 Valentia [Archaeology, Geophysics, ...]:
19   Object:      Amphores, values, coins.
20   ...
21 Cemenelum [Archaeology, Geophysics, ...]:
22   Object:      Amphores in ship wreck.
23   Object-Type: Documentation, photo.
24   Object-Location: Cemenelum, Nice-Cimiez,
25     France.
26   Object-Relocation: Musée et Site Archéologique
27     Cemenelum, Nice-Cimiez, Ville du Nice, France.
28   %%IML: media: ... img_0017.jpg
29   %%IML: UDC-Object:002+770
30   %%IML: UDC-On-Content
31   : [902+903.2+904]+738+738.8+(37)+(4)+(24)
32   %%IML: UDC-Location:069.51+(4)+(44)+(23)
33 Cemenelum [Archaeology, Geophysics, ...]:
34   Object:      Amphores type.
35   Object-Type: Documentation, sketch.
36   ...

```

Listing 2. Knowledge resources – Entries for documentation objects.

Special external media material from digital libraries and private collections [25] can be referenced from within the knowledge resources, too.

B. Classification and painted pottery

Figure 3 presents the computed result matrix for painted pottery. The example shows a result matrix on painted vases and amphores. The knowledge resources provide information that within ceramic products the amphores is a subgroup of vases, both being pottery.



Figure 3. Result matrix – Painted pottery. UDC:902,738,741,...
Row 1: Vases, painted, various types, Greek.
Row 2: Vases, painted, amphores type, Greek and Etruscan.

The results in the first row show painted vases whereas the results in the second row show painted amphores only.

C. Classification and transport pottery

Figure 4 shows the computed result matrix on pottery used for transport of products. The knowledge resources provide additional object information, e.g., on the cultural background.



Figure 4. Result matrix – Transport pottery. UDC:902,738,656,...
Row 1: Amphores, transport, Roman.
Row 2: Amphores, transport, Visigoth.

All objects show transport amphores. The results in the first row show Roman transport amphores whereas the results in the second row show Visigoth transport amphores. For these groups, the Listings 3 (Figure 3, row 2, objects 2 and 3) and 4 (Figure 4, row 1, objects 1, 2, and 3), show excerpts of several amphores object entries, containing additional media data.

```

1 Object: Amphora, painted.
2 Object-Type: Realia object.
3 Object-Location: Unknown.
4 Object-Relocation: Museu d'Arqueologia de Catalunya,
  Barcelona, Spain.
5 %%IML: media: ... img_5673.jpg
6 %%IML: media: ... img_5673.jpg
7 %%IML: UDC-Object:[902+903.2+904]+738+738.8+741+(37)+(4)
8 %%IML: UDC-Relocation:069.51+(4)+(460)+(23)
9 %%IML: labellanguage: Catalan
10 %%IML: labelcomment: written label, documentation on
  photo media
11 %%IML: label: {MUSEUM-Description: Àmfora grega de
  figures negres}
12 %%IML: label: {MUSEUM-Date: Cultura grega (s. VI
  aC)}
13 %%IML: label: {MUSEUM-Material: Ceràmica}
14 %%IML: label: {MUSEUM-Origin: Procedeix d'un taller
  d'Atenes}
15 %%IML: label: {MUSEUM-Inventory: Nùm inv. 11311}
16
17 Object: Amphora, painted - decorated. ...
18 %%IML: label: {MUSEUM-Description: Àmfora etrusca}
19 %%IML: label: {MUSEUM-Date: Cultura etrusca (final
  s. VII aC)}
20 %%IML: label: {MUSEUM-Material: Ceràmica de bucchero
  lueido}

```

Listing 3. Knowledge resources – Entries for painted amphores objects.

```

1 Object: Amphora, transport.
2 Object-Location: Unknown.
3 Object-Relocation: Museu d'Arqueologia de Catalunya,
  Barcelona, Spain.
4 %%IML: media: ... img_5831.jpg
5 %%IML: media: ... img_5831.jpg
6 %%IML: UDC-Object:[902+903.2+904]+738+738.8+656+(37)+(4)
7 %%IML: UDC-Relocation:069.51+(4)+(460)+(23)
8 %%IML: labellanguage: Catalan
9 %%IML: label: {MUSEUM-Description: Àmfora per a vi}
10 %%IML: label: {MUSEUM-Date: Cultura romana (
  primera meitat s. I aC - I dC)}
11 %%IML: label: {MUSEUM-Material: Ceràmica}
12 %%IML: label: {MUSEUM-Origin: Procedència
  desconeguda}
13 %%IML: label: {MUSEUM-Inventory: Nùm inv. 27701}
14
15 Object: Amphora, transport. ...
16 Object-Location: Southern Italy.
17 Object-Relocation: Museu d'Arqueologia de Catalunya,
  Barcelona, Spain. ...
18 %%IML: label: {MUSEUM-Description: Àmfora per a vi}
19 %%IML: label: {MUSEUM-Date: Cultura romana (inici
  s. II aC - inici s. I aC)}
20 %%IML: label: {MUSEUM-Origin: Sud d'Itàlia}
21
22 Object: Amphora, transport. ...
23 Object-Location: Andalusia, Spain.
24 Object-Relocation: Museu d'Arqueologia de Catalunya,
  Barcelona, Spain. ...
25 %%IML: label: {MUSEUM-Description: Àmfora per a salaò}
26 %%IML: label: {MUSEUM-Date: Cultura romana (finals
  s. I aC - I dC)}
27 %%IML: label: {MUSEUM-Origin: Andalusia}

```

Listing 4. Knowledge resources – Entries for transport amphores objects.

D. Discover, Complete, Explore

1) *Knowledge attribute based discovery*: The knowledge resources allow to create references by specifying criteria and associations. With the criteria Leibniz; pottery; Italy; economy and the association pottery : amphores : storage : transport the following reference (Listing 5) can be computed.

```

1 POI-person-place-date-comment: Gottfried Wilhelm Leibniz
  :: Venedig (Venezia, Venecia, Venice), Italien (Italy,
  Italia) :: \isodate{1689}{03}{ } :: Rome Travel, Italy,
  Travel to Rome

```

```

2 POI-person-place-date-comment: Gottfried Wilhelm Leibniz
  :: Venedig (Venezia, Venecia, Venice), Italien (Italy,
  Italia) :: \isodate{1690}{02}{ } :: Rome Travel, Italy,
  Travel from Rome
3 ...
4 POI-person-place-date-comment: Vespasian; Cesar :: Roma (
  Rom, Rome), Italia (Italy, Italien) :: \isodate
  {0069}{ } --\isodate{0079}{ } A.C. :: reign
5 POI-person-place-date-comment: Vespasian; Cesar :: Roma (
  Rom, Rome), Italia (Italy, Italien) :: : Amphoras.
6 POI-person-place-date-comment: Vespasian; Cesar :: Roma (
  Rom, Rome), Italia (Italy, Italien) :: : Taxes.
7 POI-person-place-date-comment: Vespasian; Cesar :: Roma (
  Rom, Rome), Italia (Italy, Italien) :: : Clothiers.
8 POI-person-place-date-comment: Vespasian; Cesar :: Roma (
  Rom, Rome), Italia (Italy, Italien) :: : (lat.) ``
  pecunia non olet``. A saying that may have been created
  in this context.
9 POI-person-place-date-source: Sueton, Vespas. 23 :: Roma
  (Rom, Rome), Italia (Italy, Italien) :: : Vespasian to
  his son Titus.

```

Listing 5. Pottery and associations (LX Resources).

The reference computed from the resources contains context associated with the criteria and association.

2) *UDC based completion*: The potential to dynamically create combined classification and context information does provide huge benefits. The example 'amphores in the net' will show how to find classified information. There is no general method nor a tool for systematically discovering classified material in the Internet. Even the amount of material available is not known as the awareness about classification and as an result the demand is not very developed. When looking for information a search even not supporting classification can give some hints. Looking for examples of UDC amphores classification we have currently (2013-08-11) retrieved a few slightly complex classifications via Google (Listing 6).

```

1 904:738(497.5-37 Rovinj)
2 904:738.8(497.5 Dalmacija)"-04/-01"
3 904:738.8(497.5 Vinkovci)"652"

```

Listing 6. Amphores UDC classification (retrieved via Google).

All three showing UDC classifications regarding amphores context in Croatia, mediterranean area, using an explicit specification of the respective location. The classification usage is very well comparable with those commonly used for amphores objects within the objects of the knowledge resources. This increases the numbers of suitable positive results within the results matrices, especially for the documentation information on references and sources.

3) *Knowledge based exploration*: Listing 7 shows an excerpt of several amphores object entries. These object entries do contain various information. In context with the request on a certain region, these objects build a group by their classification and context and the information that additional media data for these objects is not available and target of further investigation.

```

1 Akandia [Archaeology, Geophysics, ...]:
2 Greek city, Rhodos Island, Dodekanese, Greece.
3
4 Object: Ship wreck.
5 Object-Location: 550\UD{m} NE of Akandia Harbor
6
7 %%SRC: ...
8 %%IML: UDC-Object:[902+903.2+904]+629.5+(38)
  +(4)+(24)

```

```

9      Subobject: Amphoras.
10     Subobject-Find-Depth: 36\UD{m}. ...
11     Subobject-Description: Rhodian type.
12     Subobject-Date: 1st century B.C. -- early 2nd
13     century A.C.
14     %%IML: UDC-Object
15     : [902+903.2+904]+738+738.8+(38)+(4)+(24)
16     %%IML: UDC-Relocation:069.51+(4)+(495)+(23)
17 Lindos [Archaeology, Geophysics, ...]: ...
18     Object: Ship wreck.
19     Object-Location: 500\UD{m} SE of Hagios Pavlos
20     Harbor.
21     %%IML: UDC-Object:[902+903.2+904]+629.5+(38)
22     +(4)+(24) ...
23 Rhodos [Archaeology, Geophysics, ...]: ...
24     Object: Ship wreck.
25     Object-Location: Rhodes Merchant Harbor. ...
26     %%IML: UDC-Object:[902+903.2+904]+629.5+(38)
27     +(4)+(24) ...
28
29     Subobject: Pottery.
30     Object-Relocation: 20 objects in Archaeological
31     Museum Rhodos City ...
32     %%IML: script: {INSCRIPT: APICTAPXOC} {
33     TRANSCRIPT: Aristarchos}
34     %%SRC: Nikos Th. Nikolitsis: Archäologische
35     Unterwasser-Expedition bei Rhodos ...

```

Listing 7. Knowledge resources – Amphores exploration context.

E. On-water transport and utilisation

The knowledge resources provide information that amphores are tightly associated with on-water transport. Especially, the associations are:

- Amphora sites :: finding historical transport routes.
- Amphores :: geographic origin of the cargo.
- Amphores :: valuable means for dating ship wrecks.

Selecting archaeological objects, watercraft engineering, marine engineering, boats, ships, boat building, ship building, and being models having origin from ancient Egypt and the Mediterranean (UDC:902+629.5+(32), (37), (38)) results in a subset from the ship model collection. The following examples (Figures 5 and 6) show two subsets.



Figure 5. Result matrix – Ship + war.

Figure 5 presents a result matrix of media samples for ancient war ships whereas Figure 6 presents a result matrix for ancient civil transport ships.



Figure 6. Result matrix – Ship + civil transport.

Going into this result matrix, the first sample shows a Greek ship used for commercial trading, including amphores.

The classification refers to this ship model into the amphores context as it also carries amphores in this model. The second sample shows a Roman commercial transport ship, also associated with amphores transport. The third and fourth sample are a Llagut type and an Egyptian ship.

F. Objects and geological information

Figure 7 excerpts a result on mills, stone, and crop. Adding volcanic to the target matrix delivers a more specific object (Figure 8), a rotary mill made from volcanic stone, found at an excavation at Badalona, the historical Baetulo.



Figure 7. Result matrix – Mills + stone + crop.



Figure 8. Result matrix – Mills + stone + crop + volcanic (one object).

The object entry contains a reference to appropriate media objects showing the realia object. The original label does not contain the information on the specific material. The computed object entry also holds the reference to an appropriate material sample of the volcanic stone. It identifies the stone as Basalt and refers to further relevant geological objects. The computed object also links to comparable material and sources where those materials can be found and in this case it delivers reasons why this material might have been used for this product. Listing 8 shows an excerpt of the first media object being a subobject entry of the more comprehensive Baetulo object.

```

1 Baetulo [Archaeology, Geophysics, ...]:
2 Object: Rotary mill.
3 Object-Type: Realia object.
4 Object-Location: Baetulo, Spain.
5 Object-Relocation: Museu d'Arqueologia de
6 Catalunya, Barcelona, Spain.
7 %%IML: media: ... img_5833.jpg
8 %%IML: UDC-Object
9 : [902+903.2+904]+664.7+691.2+664+641.5+(37)
10 +(4)+(23)
11 %%IML: UDC-Relocation:069.51+(4)+(460)+(23)
12 %%IML: label: {MUSEUM-Description: Moli
13 rotatori}
14 %%IML: label: {MUSEUM-Date: Cultura
15 romana (s. I-II dC)}
16 %%IML: label: {MUSEUM-Material: Pedra
17 volcànica}
18 %%IML: label: {MUSEUM-Origin: Badalona}
19 %%IML: label: {MUSEUM-Inventory: Nùm inv.
20 22038}
21 %%IML: objectcomment: {MATERIAL-class:
22 Igneous rock, volcanic}

```



```

15 %%IML: objectcomment: {MATERIAL-stone:
    Basalt} ...
16 %%IML: objectcomment: {MATERIAL-usage:
    Basaltic mill stones sharpen
    themselves with ongoing usage.}
17 %%IML: objectcomment: {MATERIAL-comparable:
    Basalt lava stones from Mendig,
    Eifel, Germany, have been exported in Roman
    times for producing mill stones.}
18 %%IML: objectcomment: {MATERIAL-comparable-
    attributes: Basalt lava stone from Mendig,
    Eifel, Germany, are of grey colour and rich or
    pores.}
19 %%IML: objectreference: s. Basalt ...

```

Listing 8. Knowledge resources – Rotary mill subobject.

The object refers to other objects Basalt and Mendig, in the text as well as with explicit references. These are integrated into the volcanology and geology context, which can deliver more detailed references and information.

VIII. COMPUTING: FEATURES AND COMPONENTS

The knowledge resources can be flexibly accessed and used for computational purposes. They have been used with processing components on High End Computing resources, e.g., dynamical and interactive support on the one hand and diffraction support and seismic stacking on the other hand [26], [18] and may be used for comparable concepts, e.g., the Smart Stacking [27].

A. Selected features

The following passages present some features implemented for processing of objects, e.g., translations and transcription support, support of exceptions or correction, historical writing support, and ranking. Many compute intensive methods of extending the object pool for specific discovery workflows use the modification within the available resources. These can include correction or modification of typographical appearances. These methods can be used for content as well as for keywords or classification, in centralised or distributed resources. For computational issues with data and computing centred workflows dynamical process communication as well as batch and envelope support is shown with the later examples.

1) *Computing translations and transcriptions:* The translation resources are one option to use translation and associated transcription data (Listing 9, excerpt). The workflow components can support computing references and links into multi-lingual context by exploiting these resources.

```

1 Catalan: {amphora}
2 English: {amphore, amphorae / amphoras (pl.)}
3 French: {amphora}
4 German: {Amphore}
5 Greek: {amphora, amphoreas}{\alpha\mu\varphi i o\rho\
epsilon\alpha\varsigma}
6 Italian: {anfora, anfore}
7 Latin: {amphora}
8 Spanish: {amfora}
9
10 Catalan: {basalto}
11 Danish: {basalt, mørk vulkanske bjergart, vulkanske
    klipper}
12 English: {basalt}
13 Finnish: {basaltti}
14 French: {basalte}
15 German: {Basalt}

```

```

16 Greek: {basaltes, eidos petromatos}{\betaeta\alpha\
sigma\alpha\lambdaambda\tau\eta\varsigma$, \epsilon\i\deltaelta
o\varsigma$ \pi\epsilon\tau\tau\eta\omega\mu\alpha\tau\epsilon\varsigma}
17 Icelandic: {basalt}
18 Italian: {basalto}
19 Norwegian: {basalt}
20 Russian: {bazalt}
21 Cyrillic: {bazalt}{\setcyr{bazal\soft t}}
22 Spanish: {basalto, basanita}

```

Listing 9. Object translation and transcription data (LX Resources).

The entries amphores and basalt are examples from knowledge resources objects. They show a subset of languages and translation and transcription data. The data sets are integrated within the appropriate knowledge objects and referenced from an arbitrary number of associated objects. As shown in the example, the transcriptions can be defined generic and portable. The data itself can be handled by various algorithms. For example, the transcription data can be iterated automatically with special algorithms in order to generate alternatives for adding references as well as using the information for typesetting or character replacements. Listing 10 shows an excerpt of retrieved examples (lxaztrdb) from the general purpose translation and transcription database.

```

1 Catalan: Ceràmica
2 Spanish: Cerámica
3 English: ceramics
4 German: Keramik
5 French: céramique
6
7 Catalan: antiga salsa romana
8 Spanish: salsa romana antigua
9 English: ancient Roman sauce
10 German: antike römische Soße
11
12 Catalan: Moli rotatori
13 Spanish: Molino rotatorio
14 English: Rotary mill
15 German: Drehmühle
16
17 Catalan: volcànica
18 Spanish: volcánico
19 English: volcanic
20 German: vulkanisch
21 French: volcanique

```

Listing 10. Translation and transcription database (LX Resources).

The translation and transcription database allows to have an arbitrary algorithm and structure for an implementation within a workflow. It can also associate terms with the translations, like associating “Garum” with “ancient Roman sauce” or creating black lists for translations.

2) *Computing keyword translation terms:* Listing 11 shows generated (lxazkw_de2en.sh) keyword translations.

```

1 s/\[(.*)Archäologie\(.*)\]/\[\1Archaeology\2\]:/g
2 s/\[(.*)Geologie\(.*)\]/\[\1Geology\2\]:/g
3 s/\[(.*)Geophysik\(.*)\]/\[\1Geophysics\2\]:/g
4 s/\[(.*)Mineralogie\(.*)\]/\[\1Mineralogy\2\]:/g
5 s/\[(.*)Vulkanologie\(.*)\]/\[\1Volcanology\2\]:/g
6 s/\[(.*)Fernerkundung\(.*)\]/\[\1Remote Sensing\2\]:/
g
7 s/\[(.*)Seefahrt\(.*)\]/\[\1Seafaring\2\]:/g
8 s/\[(.*)Etymologie\(.*)\]/\[\1Etymology\2\]:/g
9 s/\[(.*)Einheit\(.*)\]/\[\1Unit\2\]:/g

```

Listing 11. Keyword translation terms (LX Resources, excerpt).

Keywords have a prominent role with objects and can be used independently from other parts and attributes. Their

translation and definition can therefore be separated from other translations.

3) *Computing translation groups*: Comparable to the keyword handling the resources allow for definition of translation groups for strings, which might even not be translations. Listing 12 shows an excerpt of translation groups.

```
1 Archaeology; Archeology; Archäologie; Archaeologia;
  Archeologia;
2 ...
3 Volcanism; Vulcanism; Vulkanismus;
4 ...
5 Vulcanologia; Volcanology; Vulkanologie; Vulcanology;
```

Listing 12. Translation groups (LX Resources, excerpt).

These groups can be used for defining various character strings being used alternatively for a term.

4) *Computing ligatures*: Listing 13 shows an excerpt of the ligature handling algorithms (`lx_ligature`) generated from the knowledge resources components.

```
1 if (/([Aa]uf)(falt)/) { ...
2 if (/([Aa]uf)(füll)/) { ...
3 if (/([Rr]elief)(f[oö]rm)/) { ...
4 if (/([Ss]umpf)(fieber)/) { ...
```

Listing 13. Ligature handling algorithms (LX Resources, excerpt).

For providing a means for any purpose, the algorithms can be used in both directions, separating and ligating. For example, the ligating direction can be helpful for even finding links that would not be recognised with a regular expression search on documents coded with different ligations.

5) *Computing on historical writing objects*: Listing 14 is an excerpt (`lx_histcorr.sh`) of a historical writing algorithm used for the Leibniz cases presented here.

```
1 auff :: auf
2 Cammer :: Kammer
3 darff :: darf
4 delinquenten :: Delinquenten
5 Galeren :: Galeeren
6 gemischt :: gemischt
7 Golphe :: Golf (von Venedig)
8 Gondolier :: Gondoliere
9 mußen :: müssen
10 navis :: Schiff
11 niederlaßet :: niederläßt
12 salviren :: salvieren; retten
13 schiffe :: Schiffe
14 sizend :: sitzend
15 trincken :: trinken
16 ufer :: Ufer
17 verfertiget :: verfertigt
18 waßer :: Wasser
19 wirfft :: wirft
```

Listing 14. Historical writings algorithm, Leibniz texts (LX Resources).

The example lists historical notation and spelling from transliterations of original handwritten documents from Gottfried Wilhelm Leibniz (1646–1716).

6) *Computing typographical corrections*: Listing 15 shows examples (`lxazkw_tycorr.sh`) from the correction algorithms generated via the available typographical database.

```
1 iab archäology archaeology
2 iab Archäology Archaeology
3 ...
4 iab laoratories laboratories
```

```
5 iab Laoratories Laboratories
6 ...
7 iab vulkanology volcanology
8 iab Vulkanology Volcanology
```

Listing 15. Typographical correction algorithms (LX Resources, excerpt).

The `tycorr` routines based on this database can be used for extending workflows at any step, from input, content, context, intermediate data up to finals results. With non-batch workflows the same database is used for editing corrections and publishing support. Any object, attribute or component can be handled that way. On the one hand, even objects or keyword labels as generated for special languages only can be integrated into a regular discovery. On the other hand, original and historical writing, different transcriptions or “authentic” errors can be included in the discovery process and used within workflows. With the existing components many thousands of variants, transcription sets, and automatic corrections have been created and build an extended representation of the knowledge resources content. Applying these within the workflows can improve the quality of the discovery, exactly for the purpose of the application scenario.

7) *Ranking*: Algorithms applied to isolated input data stored in proprietary database cannot result in a generally valid data ranking. This ranking is only appropriate for the specific limited context of the collected data. Any comparisons of the ranking between separate groups of data is not reasonable.

The basic environment is made up by the algorithm and the input data. Input data consists of the knowledge resources and external data. Everything that may be used for discovery can be integrated with the knowledge resources. Besides data sets from natural sciences modelling or simulation, and documentation also references, descriptive data, citations, publications with content and bibliographic information can be exploited.

- Alphanumeric ranking,
- Ranking based on context,
- Ranking based on classification,
- Ranking based on number of items, like regular expressions or references,
- Silken selection, like phonetic algorithms support.

B. Components

Using the following concepts, we can implement for mostly any system:

- Application communication via IPC.
- Application triggering on events.
- Storage object requests based on envelopes.
- Compute requests based on envelopes.

For demonstration and studies flexible and open Active Source Information System components have been used for maximum transparency. This allows OO-support (object, element) on application level as well as multi-system support. Listing 16 shows a simple example for application communication with framework-internal and external applications (Inter-Process Communication, IPC).

```
1 catch { send {rasmol #1} "$what" }
```

Listing 16. Application communication (IPC).

This is self-descriptive Tcl syntax. In this case, the IPC `send` is starting a molecular graphics visualisation tool and catching messages for further analysis by the components.

Listing 17 shows an example of how the communication triggering can be linked to application components.

```
1 text 450.0 535.0 -tags {itemtext relictrotatex} -fill
  yellow -text "Rotate_x" -justify center
2 ...
3 $w bind relictrotatex <Button-1> {sendAllRasMol {rotate x
  10}}
4 $w bind relictballsandsticks <Button-1> {sendAllRasMol {
  spacefill 100}}
5 $w bind relictwhitebg <Button-1> {sendAllRasMol {set
  background white}}
6 $w bind relictzoom100 <Button-1> {sendAllRasMol {zoom
  100}}
```

Listing 17. Application component triggering.

Tcl language objects like `text` carry tag names (`relictrotatex`) and dynamical events like `Button` events are dynamically assigned and a user defined subroutine `sendAllRasMol` is executed, triggering parallel visualisation. Storage object requests for distributed resources can be done via OEN. Listing 18 shows a small example of a generic OEN file.

```
1 <ObjectEnvelope><!-- ObjectEnvelope (OEN)-->
2 <Object>
3 <Filename>GIS_Case_Study_20090804.jpg</Filename>
4 <Md5sum>...</Md5sum>
5 <Shalsum>...</Shalsum>
6 <DateCreated>2010-08-01:221114</DateCreated>
7 <DateModified>2010-08-01:222029</DateModified>
8 <ID>...</ID><CertificateID>...</CertificateID>
9 <Signature>...</Signature>
10 <Content><ContentData>...</ContentData></Content>
11 </Object>
12 </ObjectEnvelope>
```

Listing 18. Storage object request (OEN).

OEN are containing element structures for handling and embedding data and information, like `Filename` and `Content`. An end-user public client application may be implemented via a browser plugin, based on appropriate services. With OEN instructions embedded in envelopes, for example as XML-based element structure representation, content can be handled as content-stream or as content-reference. Algorithms can respect any meta-data for objects and handle different object and file formats while staying transparent and portable. Using the content features the original documents can stay unmodified. The way this will have to be implemented for different use cases depends on the situation, and in many cases on the size and number of data objects. However, the hierarchical structured meta data is uniform and easily parsable. Further, it supports signed object elements (`Signature`), validation and verification via Public Key Infrastructure (PKI) and is usable with sources and binaries like Active Source. Compute requests for distributed resources are handled via CEN interfaces [28]. Listing 19 shows a generic CEN file with embedded compute instructions.

```
1 <ComputeEnvelope><!-- (CEN) --><Instruction>
2 <Filename>Processing_Batch_GIS612.pbs</Filename>
3 <Sha512sum>...</Sha512sum>
4 <DateCreated>2013-09-15:210917</DateCreated>
5 <Signature>...</Signature>
6 <Content><DataReference>https://doi...
7 <Script><Pbs>
8 <Shell>#!/bin/bash</Shell>
9 <JobName>#PBS -N myjob</JobName>
10 <Oe>#PBS -j oe</Oe>
11 <Walltime>#PBS -l walltime=00:20:00</Walltime>
12 <NodesPpn>#PBS -l nodes=8:ppn=4</NodesPpn>
13 <Feature>#PBS -l feature=ice</Feature>
14 <Partition>#PBS -l partition=hannover</Partition>
15 <Accesspolicy>#PBS -l naccesspolicy=singlejob ...
16 <Module>module load mpt</Module>
17 <Cd>cd $PBS_O_WORKDIR</Cd>
18 <Np>np=$(cat $PBS_NODEFILE | wc -l)</Np>
19 <Exec>mpiexec_mpt -np $np ./dyna.out 2>&1</Exec>
20 </Pbs></Script></Instruction>
21 </ComputeEnvelope>
```

Listing 19. Compute request: Compute envelope (CEN).

Content can be handled as content-stream or as content-reference (`Content`, `ContentReference`). Compute instruction sets are self-descriptive and can be pre-configured to the local compute environment. In this case, standard PBS batch instructions like `walltime` and `nodes` are used. The way this will have to be implemented for different use cases depends on the situation, and in many cases on the size and number of data objects. An important benefit of content-reference with high performant distributed or multicore resources is that references can be processed in parallel on these architectures. The number of physical resources and the transfer capacities inside the network are limiting factors.

IX. INTEGRATION OF EXTERNAL INFORMATION

Besides the computational aspects, it had to be analysed if and how external structures of traditional information sources can contribute to the knowledge resources. The following example describes how an intelligent use of Integrated Information and Computing Systems knowledge resources can support on-site knowledge discovery as well as external distributed discovery. Regarding the examples in the last section, information from an external resource has been analysed and compared for finding complementary references for objects and result matrices of the knowledge database and for detecting targets for future enhancements of the object and media database. Using the LX Scientific Resources and data from external sources, in this case the publicly available Leibniz information (concept glossaries [29], manuscript collections and catalogues [30], [31], and critical editions [32], emended and commented), several new references can be created for the resources. The critical editions contains the texts on base of the original drafts. As there is no flexible interface or standard format available, the referenced files have to be fetched, processed, and evaluated in order to retrieve the required information.

A. Leibniz and archaeology case

The knowledge resources have been used for discovering paths into external resources. Listing 20 shows an excerpt of the object paths used in context with Gottfried Wilhelm Leibniz (1646–1716).

```

1 Ceramics : Keramik : Tonerde, Material, Porzellan
2 Amphoras : Amphoren : Transport, Lastentransport
3 Venice : Venedig : Schiff, Schiffsbau
4 Ship : Schiff, navis, Ägypten

```

Listing 20. Object paths for discovery (LX Resources).

The general knowledge resources translation component implicitly supports a significant number of the gathered terms (Listing 21, excerpt), translating from German into English.

```

1 Ägypten :: Egypt
2 Keramik :: Ceramic
3 Porzellan :: Porcelain
4 Schiff :: Ship
5 Ton :: Clay
6 Tonerde :: Alumina
7 Transport :: Transport

```

Listing 21. Relevant terms from the translation component (LX Resources).

An excerpt of complementary information regarding references from the external resources is shown in Listing 22.

```

1 Transportwesen - chinesische Methoden
2 Transportwesen - technische Lösungen, Lastentransport
3 Transportwesen - technische Lösungen, allgemein

```

Listing 22. Knowledge object references on “transport” in Leibniz context.

These refer to the appropriate descriptions and sources, e.g., [33]. Complementary associated information regarding ceramics, ships, and “Venice” is shown in Listings 23, 24, and 25.

```

1 Schiffsankerfunde
2 Schiffsentführung
3 Schiffsfunde - Niedersachsen, Schweden
4 Schiffsreste - Funde

```

Listing 23. Knowledge object references on “ship” in Leibniz context.

```

1 Frankreich - Flotte - Schiffe, Golf von Venedig
2 Griechen - in Venedig

```

Listing 24. Knowledge object references on “Venice” in Leibniz context.

```

1 Ton, Tonerde
2 Tonerde

```

Listing 25. Knowledge object references on “clay” in Leibniz context.

```

1 Ägypten - Altertümer
2 Ägypten - Erdbeben
3 Ägypten - Kornkammer
4 Ägypten - Weltwunder
5 Ägypten - Wissenschaft
6 Ägypten - Transport von Schiffen

```

Listing 26. Knowledge object references on “Egypt” in Leibniz context.

These refer to the appropriate descriptions and sources, e.g., [34], [35], [36], [37] and [38], [39], [40], [41] as well as [42]. It is significant that some important passages are not containing the terms in either language or transcription but some context of these. In the case of [36] neither “Venedig”, “Venice” or another term is mentioned in the texts but the term “Golphe”, which in this case refers to the “Gulf of Venice”. Therefore, it is most important to have a context description and to process this description for a successful discovery.

B. Leibniz and geo resources case

The knowledge resources not only enclose textual references. They can also deliver references and context for non-textual material, different terminology, and sources:

Within the knowledge resources, the context of the “Leibniz” object also delivers references to non-textual material Gottfried Wilhelm Leibniz had access to, e.g., the Witsen Map [43]. The knowledge resources can provide transfer of context and terminology, e.g., for different disciplines, epochs, languages, and cultures. For example, it refers the term “seismology” to the term “terrae motus” used in that time [44], [45], as has been shown in [24]. It further links the term to the “Vesuvius” and earthquake related context and from this delivers source, e.g., for the correspondence Leibniz had regarding geoscientific phenomena [46], [47].

C. Integration of the Leibniz resources

The external resources themselves have revealed some deficits, especially the do not use a unique terminology. In addition, the structure of these resources is not unique. There is no general harmonisation of the implementations. The implementations show that different institutions and projects with different intentions have been working on the topic. Only a low level of interoperability is available. The preparation of the data in its current organisation and formatting is not well suited for further electronical use in information systems. The numbering of sources and references is not intended to be used automatically. The provided implementation of transcriptions, translations, and references to original sources, scans, and bibliographic data is not consistent. In addition, it cannot be interfaced in order to enable a full discovery of the content.

X. MECHANISMS AND WORKFLOWS CASE STUDY

A filter chain can be used to compute resulting object sets. Based on the available system the following steps can be separated, in an example from geosciences and archaeology:

- Select topic from knowledge base (volcanology).
- Select region from results (Europe, Caribbean).
- Select volcano from results (Vesuvius, La Soufrière).
- Select object entries (geosciences and archaeology).
- Select media objects and references.
- Select application resources and interfaces.

A. Workflows and algorithms

The knowledge resources block is the central resource in the long-term strategy. The knowledge resources can contain any kind of content. Application components can be migrated into the knowledge resources for documentation purposes and re-use. The services can access archived and historical data as well as live data and feed it into the workflows. Services interfaces allow to build complex workflows using arbitrary algorithms. The knowledge resources can be accessed from applications to extract suitable information and trigger the use of compute and storage resources. Objects can be selected by any algorithm, e.g., combinatory, search, and filter algorithms. Examples for the universal use of documentation and algorithms as used with the disciplines presented here are media data like digital images, video, hypertexts, Portable


```

1 create_archaeology_planet_view_topic.sh \
2 volcano_guadeloupe_soufriere_viewto.jpg \
3 volcano_guadeloupe_soufriere_viewfrom.jpg \
4 volcano_saba_mts scenery_viewto.jpg \
5 volcano_saba_mts scenery_viewfrom.jpg

```

Listing 30. Generated core data for Topicview processing.

As the knowledge resources' objects carry references to any kind of detailed processable data, distribution maps and satellite views can be computed and passed on.

D. Resulting object entries on geosciences and archaeology

The following object entries are excerpts from the calculated cross-links table (Table IV). The excerpts contain some, structure, UDC classification, keywords, references, and satellite image reference. The references for the geopositioning are created via classification and can be used for any purpose. Listing 31 shows an excerpt of an LX Resources object entry [14], "Vesuvius" volcano.

```

1 Vesuvius [Volcanology, Geology, Archaeology]:
2 (lat.) Mons Vesuvius.
3 (ital.) Vesuvio.
4 (deutsch.) Vesuv.
5 Volcano, Gulf of Naples, Italy.
6 Complex volcano (compound volcano).
7 Stratovolcano, large cone (Gran Cono).
8 Volcano Type: Somma volcano,
9 VNUM: 0101-02=,
10 Summit Elevation: 1281\UD{m}.
11 The volcanic activity in the region is observed
12 by the Oservatorio
13 Vesuviano. The Vesuvius area has been declared a
14 national park on
15 \isodate{1995}{06}{05}. The most known antique
16 settlements at the
17 Vesuvius are Pompeji and Herculaneum.
18 Syn.: Vesaevus, Vesevus, Vesbius, Vesvius
19 s. volcano, super volcano, compound volcano
20 s. also Pompeji, Herculaneum, seismology
compare La Soufrière, Mt. Scenery, Soufriere
%%IML: UDC
: [911.2+55]: [57+930.85]: [902] "63" (4+23+24)
=12=14
%%IML: GoogleMapsLocation: http://maps.google.de
/maps?hl=de&gl=de&vpsrc=0&ie=UTF8&ll
=40.821961,14.428868&spn=0.018804,0.028238&t=h&
z=15

```

Listing 31. Knowledge resources – object entry "Vesuvius" volcano.

The example contains a reference and VNUM for the Vesuvius volcano, various secondary objects, UDC classification, satellite image reference (Satelliteview in Table IV). It refers to "Soufriere", "La Soufrière", and "Mt. Scenery". Listing 32 shows an excerpt of the "Soufriere" object entry.

```

1 Soufriere [Volcanology, Geology]:
2 A common name for a volcanic feature
3 resulting from the
4 french term for \periref{tgt:PeriSulfur}{
5 Sulfur}.
6 The name soufriere is used for a volcanic
7 crater
8 or other area in combination with solfataric
9 activity.
10 The name is mostly used in French speaking
11 regions,
12 especially in the West Indies.
Very well known are, for example:
La Soufrière volcano, Guadeloupe, F.W.I.
Soufriere Hills, F.W.I.
Soufriere St. Vincent, F.W.I.
Syn.: Soufrière

```

```

13 s. also La Soufrière, F.W.I., volcano,
14 seismology
%%IML: UDC
: [911.2+55]: [57+930.85]: [902] "63" (7+23)
=84/=88

```

Listing 32. Knowledge resources – secondary object entry "Soufriere".

This secondary object entry, "Soufriere", also refers to the La Soufrière volcano (Listing 33), which itself refers to various data and objects, e.g., satellite image references.

```

1 La Soufrière [Volcanology, Geology]:
2 La Soufrière volcano, Guadeloupe, F.W.I.
3 Volcano Type: Stratovolcano,
4 Country: France,
5 Subregion Name: West Indies, Caribbean,
6 VNUM: 1600-06=,
7 Summit Elevation: 1467\UD{m}.
8 Syn.: Soufriere
9 s. volcano
10 s. also Soufriere, F.W.I., lava,
11 lava sand, OVSG
%%IML: UDC: [911.2+55]: [57+930.85]: [902] "63"
(7+23+24)=84/=88
%%IML: GoogleMapsLocation: http://maps.
12 google.com/?ie=UTF8&ll
=16.043153,-61.663374&spn
=0.003088,0.003262&t=k&z=18&vpsrc=6&lci=
weather

```

Listing 33. Knowledge resources – secondary object "La Soufrière".

The secondary object entry "Mt. Scenery" (Listing 34) also contains classifications and media, and further data references for the Mt. Scenery volcano on Saba. Extracted examples are volcano type, VNUM, region, status, elevation, and UDC classification views as well as the geo-references, which with this request are used to automatically compute views and distribution maps for classified objects in the result matrix.

The classification groups themselves show references to associated objects. The data and media object in a processed reference chain can be used for further analysis, creating special features. That way, using UDC classifications, e.g., places from a region or context that can be associated with volcanology and associated with archaeological sites can be selected and media objects can be processed and realia referred.

```

1 Mt. Scenery [Volcanology, Geology]:
2 Volcano, Saba, Netherlands Antilles, D.W.I.,
3 The Netherlands
4 Volcano Type: Stratovolcano,
5 Country: Netherlands,
6 Subregion Name: West Indies, Caribbean,
7 VNUM: 1600-01=,
8 Volcano Status: Historical,
9 Last Known Eruption: in or before \isodate
10 {1640}{}{}},
Summit Elevation: 887\UD{m}.
%%IML: UDC
: [55+56+911.2]: [902+903+904]: [57+930.85]
"63" (7+23+24)=84/=88
%%IML: GoogleMapsLocation: http://maps.
11 google.com/maps?f=q&source=s_q&hl=en&
geocode=&q=mt+scenery,+saba+netherlands+
12 antilles,+google+maps&aq=&sll
=17.633225,-63.236961&sspn
=0.048997,0.052185&vpsrc=0&t=h&ie=UTF8&hq=
mt+scenery,+saba+netherlands+antilles,&
hnear=&z=14&lci=weather
s. also Saba, D.W.I., volcano, seismology

```

Listing 34. Knowledge resources – object entry "Mt. Scenery".

Dynamical components can even benefit from precalculation and precomputation of objects. This includes precalculated classification and weights ("PreUDC"). The following section presents examples calculated from the above classified objects and the figures are showing the results of selected attributes, including the classification and geo-references used basic visualisation.

E. Resulting features selection for cross-links processing

The following (Figure 9) is an excerpt of the secondary objects computed above for the Caribbean region volcanoes and a selection with "UDC : (23) , (24)".



Figure 9. Topicview – volcanoes, La Soufrière (left), Mt. Scenery (right), VIEW-TO (green), VIEW-FROM (blue).

Figure 10 illustrates the computed objects (Topicview), e.g., here volcanic samples after processing, all showing the variety of material from the top of the La Soufrière volcano.



Figure 10. Topicview – related volcanic samples (La Soufrière, 2011).

Any of these objects being part of the resulting matrix for a request, e.g., photos for object entries as well as media data for physically available samples, have been found via references and UDC from the knowledge base. The realia references for the objects refer to a collection where the samples are stored. Further analysis for the samples is available via the knowledge resources. Figure 11 shows the geo-locations [14] for computed geoscientific and archaeological object samples on a configurable object map.

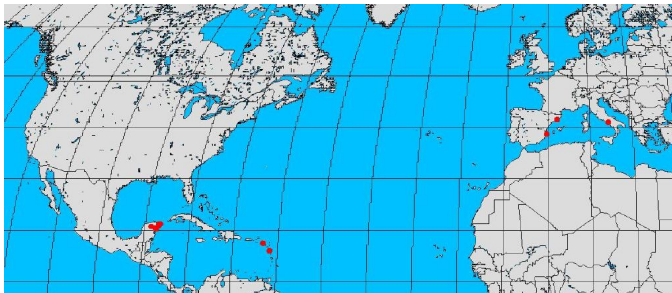


Figure 11. Objectmap – computed map for related objects (red, excerpt).

A sample distribution of volcanic features is depicted in Figure 12. It shows a comparison of volcanic data in a projection identical to the computed Objectmap (Figure 11). Although the knowledge matrix of this example is most complex (Table IV), the workflow for producing a view can be specified very easy like for spatial presentation. The map generated with the workflow as described with the case study presents the related objects from the context available in the geophysical research database. It can visualise various aspects of the classified objects. In this case of volcanoes and geological samples a reasonable view is the spatial distribution of the referenced selection.

Anyhow it must be emphasised that the number of possible views is not limited, neither from the knowledge base nor from the implementation. Spatial and cartographic methods provide only a very restricted tool set for supporting sciences for their complex tasks. For example, more complex examples from the same context could use more advanced presentation methods than available from spatial procedures. As it is obvious from this, the implementation of the knowledge resources architecture can be used for any purpose.

With the suggested workflow, the objects from the knowledge resources can be processed by any means like phonetic search, e.g., via classical or modified Soundex algorithms. This includes the flexible development of a non-limited number of extensions for dynamical search and analysis. It, too, provides a multiplicity of granularity regarding objects and classification.

Any features and data shown, based on the knowledge resources [14] and further sources, even if much less structured like online encyclopedia material, are resulting from the request and workflow, e.g., selection of classification, topics, object, secondary data, area, map projection, applications and so on [48], [49], [50].

One possible example of an algorithm for the interface workflow, with one request iteration is:

- Knowledge base request,
- Keyword filtering,
- Object processing,
- UDC filtering,
- Object element processing,
- Object container retrieval,
- Media retrieval,
- Media and
- Container processing,
- Building resulting media,
- Visualisation,
- Provisioning results.

This can be used to create multi-disciplinary text and media results, e.g., dynamical distribution maps, from requests, using calculation, processing, and computation of objects. A workflow can enter the knowledge matrix from different directions, e.g., from topic to related topic or from overview to detailed view as well as vice versa.

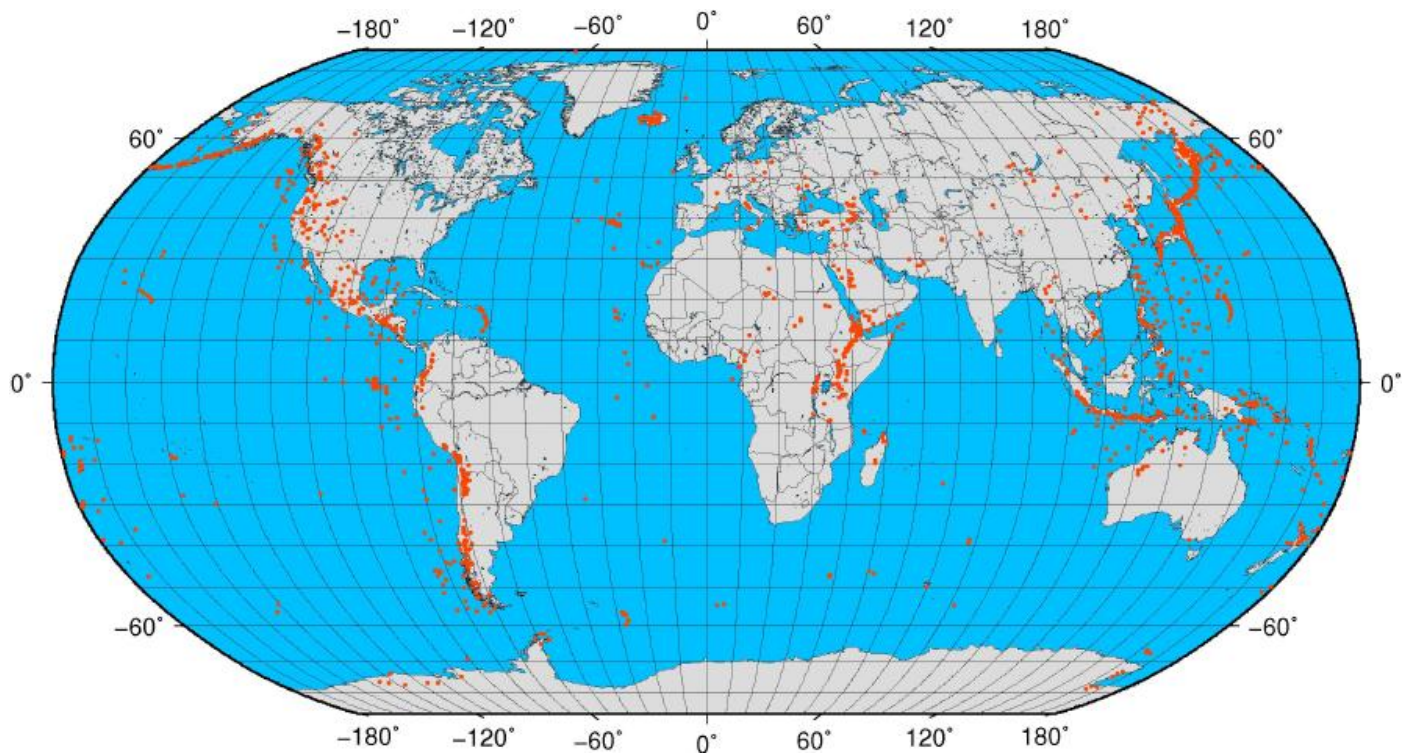


Figure 12. Volcanomap – worldmap of referenced volcanoes (orange). The computed spatial distribution map shows the selection of those object entries classified as volcanic features within the knowledge resources.

XI. EVALUATION

The architecture and resources have been able to support any development required so far. Even as it is not possible in any paper to describe all disciplines and features, the case study shows the important features of the solution regarding the involved disciplines! Especially, the solution including the appropriate architecture and implementation does have all the required features, presented and discussed in the case studies allowing:

- A universal classification of all possible multi-disciplinary objects and views,
- To integrate and use any known type of document and structure, e.g., media and realia,
- Multi-disciplinary knowledge documentation, e.g., with archaeology and geosciences,
- To address any type of resources, e.g., computation and storage, including a feasible concept for collaboration and operation,
- To create, describe, and execute workflows, e.g., for discovery and development.

The means are not limited in any way to the presented examples. The facilities are only limited by the inventive spirit of the implementing groups.

The integration of structure and classification allows to use the benefits of algorithms like filtering for any possible use of processing and computing. Structuring the content and context documentation allows a flexible balance for redundancy of data and compute requirements for various application scenarios, even with identical data.

The faceted classification and multi-disciplinary data have proved to provide significant benefits for knowledge re-use and discovery. This includes various ways of describing aspects correctly. From one view a glass of water is half full. From another view the same glass of water is half empty. The two groups representing the classical views might argue that the other view is unintelligible. Both are generally not good as they only represent views. An alternative view will be describing the status giving a filling percentage. In addition, this reduces the limitation of unprecise references.

Most content, tasks, and developments handled with information and computing systems are not suitable for any long-term use, the more there are requirements for functional long-term documentation. The use of the universal knowledge resources and collaboration framework has shown to be very flexible and extendable with implementations and technologies over several decades.

It has been found that standard search and pattern recognition on information is by far not sufficient to gain reasonable results for long-term knowledge herding and evaluation processes. In contrast, the implementation shows excellent results with opening multi-discipline data for IICS, advanced computing, and processing. Statistically, filtering 1 GB of unstructured data delivers less quality than using 10 MB structured classified knowledge base data. The Quality of Data (QoD) must be drastically improved in order to get better results. This can help to reduce compute times, storage volume, and besides overall costs it can help to decrease energy consumption in the end. Using UDC in this context, the availability of a full UDC catalog, and an implementation allowing classification views,

combined classification, and ranking priority has proven to drastically increase the QoD. With multi-disciplinary networks, there is even need for a tolerance of individual classification.

In common environments it is only feasible to do one implementation for a specific application, as has been done with these components. Anyhow it has been possible to implement the applications on various architectures providing different resources. Workflows support the use of remote resources (Table V). In case of a 1000 knowledge-objects reference chain, with 1–10 elements per object, performance will increase much with low latencies.

TABLE V. WORKFLOW PROCESSES (REMOTE, ETHERNET, 1000 NODES).

| <i>Remote Workflow Process</i> | <i>Elements</i> | <i>Response Time</i> |
|--------------------------------|-----------------|----------------------|
| Knowledge base request | 1000 | 5 s |
| Processing (object, media) | 10 | 7 s |
| Building result | 10 | 5 s |
| Visualisation | 2 | 25 s |

When using one of the described very basic application scenarios on a certain resources architecture the efficiency mostly depends on the decision for the depth of the cross-links to be considered and on the processing requirements for the media data for the originary resources. With current sizes for digital photos and a low depth of five to ten for the cross-links a medium sized application can easily use about one-hundred parallel processes. On a common compute resource without a queue configured for the jobs the response time will be less than a minute.

That way, implementing components with IICS on many compute nodes can profit from using various technologies as suited for different purposes, using task and thread parallelism to the extend needed to handle a problem remotely:

- *High level:* Integrated Systems, collaboration frameworks, dynamical application components, Partitioned Global Address Space (PGAS) models.
- *Virtualisation level:* Parallel Virtual Machine (PVM).
- *Low level:* Message Passing Interface (MPI), OpenMP, and comparable.

Increasing requirements on resources are mostly focussed on compute, communication, and storage. The requirements can increase by various reasons, from classical High End Computing, currently Peta-Scale or towards Exa-Scale, or Advanced Scientific Computing components up to multi-user performance within an application or resource. The concept can handle

- many computation tasks associated with a process,
- many processes associated with a workflow,
- many workflows associated with an application instance,
- many application instances associated with an application scenario.

The integrated and dynamical concept allows a scalable use of components, implementing components based on the appropriate level. Besides this, the components, which require

an appropriate physical configuration of the resources, for example, in order to be usable with interactive dynamical applications are considered are considered challenging at the lower level.

XII. CONCLUSION AND FUTURE WORK

It has been demonstrated how complex systems for multi-disciplinary documentation and computing can be built, developed, and extended based on creating long-term-knowledge resources supported by a universal classification and implementing IICS systems. This paper presented the successful implementation of a new universal framework for an integrated system, integrating knowledge resources and implementation components for long-term knowledge creation and use, including the facilities for High End Computing and processing resources.

The geoscientific and archaeological knowledge resources have been, structured, extended, and developed for several decades now, having been successfully used with various technology over time. Huge benefits creating new instances of objects and components result from enabling a long-term stepwise development for all parts of the knowledge and application space and a free extendability of the knowledge base. The previous work, which this implementation is built on has been discussed.

The architecture allows any kind of documentation and algorithm for content, context, information and resources usage. The services and resources usage is very economic and only limited by the limiting implementation factors, e.g., capacities and policies. This solution goes far beyond data and text mining or image analysis and pattern recognition. As shown, classification, as well as spatial data should be integrated with the objects. In no case is it suitable regarding the long-term goals of knowledge creation to “fix” knowledge objects with an application or implementation, neither simple or complex, nor closed or open licensed.

The comparison showed that the possibility of combining methods (UDC, keyword, full-text analysis) does lead to unique benefits. Comparable precision, reliability, performance, and scalability is not available from any isolated method. For any advanced knowledge resources and improved QoD, a flexible classification is indispensable. Bringing the integration of universal classification and IICS into wider acceptance can provide a time-capsule against the transience of knowledge and open new synergetic long-term possibilities.

Complex systems can be created and extended over the necessarily long periods of time, using IICS and UDC. Advanced scientific computing is supported by interfaces, accessing compute and storage resources. With the available architecture implementations of these compute and storage resources are meant to be short- to medium-term tools for supporting the long-term knowledge resources. Further it has been shown what information and knowledge on content and context can be preserved for medium- and long-term usage even for large complexity of an overall system.

With collaborative implementations a catalog of universal criteria is needed for feature development and exploitation of knowledge resources. This catalog has to consider long-term multi-disciplinary and international aspects. The overall

operative system components must be self-learning. Anyhow, central components will always afford an editorially managed operation.

The case study on knowledge resources and external information showed a number benefits. Especially the workflows and results can very much profit from an integration. Regarding the external Leibniz resources, an impressive amount of valuable data has been collected by various institutions within the last decades. Currently, only a small percentage is being developed for exploitation and further integration. Based on the vast experiences gathered within the last generations, clearly structured resources, interfaces, and tools are required for future knowledge discovery. For sustainability issues it will be very desirable to see a suitable long-term data structure being created in the future, using a unique terminology and supporting an improved knowledge discovery and reuse.

Workflows utilising future storage and computational resources can profit from autonomous intelligent units, e.g., multi-agent system components [51].

The basic architecture has been presented using a long-term knowledge base (LX), documentation, and classification of objects, the “Collaboration house” framework, flexible algorithms, workflows and dynamical and Active Source components for creating future IICS. Besides that, there is a strong demand for future education and teaching in all disciplines of academia and research in order to mediate and disseminate the basics of knowledge creation and classification.

ACKNOWLEDGEMENTS

We are grateful to all national and international academic, industry, and business partners in the GEXI cooperations for the innovative constructive work and the Science and High Performance Supercomputing Centre (SHPSC) for long-term support of collaborative research and the LX Project for providing suitable resources.

Many thanks to the scientific colleagues at the Leibniz Universität Hannover, the Institute for Legal Informatics (IRI), and the Westfälische Wilhelms-Universität (WWU), sharing experiences on ZIV, HLRN, Grid, and Cloud resources and for participating in fruitful case studies as well as the participants of the EULISP Programme for prolific scientific discussion over the last years.

We are grateful to the UDC Consortium for continuously providing, extending, and improving the excellent universal decimal classification for public use.

We are grateful to the Gottfried Wilhelm Leibniz Bibliothek (GWLb), Hannover, Germany, for the collection and public provisioning of most complete information on Gottfried Wilhelm Leibniz and related work. Our thanks go to the Akademie der Wissenschaften zu Göttingen and the Akademie der Wissenschaften zu Berlin for the successful implementation of information system components enabling an advanced provisioning and integration of information.

Thanks for excellent inspiration, support, and photo scenery go to the Saba Conservation Foundation, Saba Marine Park and National Heritage Foundation St. Maarten (D.W.I.), National Park Guadeloupe and Museum St. Martin (F.W.I.),

Museu d’Arqueologia de Catalunya, Barcelona, Spain, Museu d’Història de la Ciutat de Barcelona, Spain, Diputació de València, Centre Cultural la Beneficència, València, Spain, Museo Archeologico Nazionale di Altino, Venice, Italy, Musée et Site Archéologique Cemenelum, Nice-Cimiez, Ville du Nice, France, and Museu Egipci Barcelona, Spain, as well as Canon for the photo equipment. We do thank the international colleagues from geosciences, informatics, and archaeology in the present collaborations and the peer reviewers for constructive feedback and proof-reading this paper.

REFERENCES

- [1] C.-P. Rückemann, “Advanced Scientific Computing and Multi-Disciplinary Documentation for Geosciences and Archaeology Information,” in *Proceedings of The International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2013)*, February 24 – March 1, 2013, Nice, Cote d’Azur, French Riviera, France. XPS Press, 2013, pp. 81–88, Rückemann, C.-P. (ed.), ISSN: 2308-393X, ISBN: 978-1-61208-251-6, URL: http://www.thinkmind.org/download.php?articleid=geoprocessing_2013_4_10_30035 [accessed: 2013-05-26], URL: <http://www.iaria.org/conferences2013/ProgramGEOProcessing13.html> (Program) [accessed: 2013-05-26].
- [2] Ponemon Institute, “Data-Security,” 2013, Ponemon Institute, URL: <http://www.ponemon.org/data-security> [accessed: 2013-07-10].
- [3] Ponemon Institute, “Cost of Data Breach 2011,” 2011, Ponemon Institute / Symantec, URL: <http://www.ponemon.org/library/archives/2012/03> [accessed: 2013-07-10].
- [4] Symantec, “Symantec Data Breach Calculator,” 2013, Symantec, URL: <https://databreachcalculator.com/> [accessed: 2013-07-10].
- [5] Wissenschaftsrat, “Übergreifende Empfehlungen zu Informationsinfrastrukturen,” *Wissenschaftsrat, Deutschland, Drs. 10466-11, Berlin*, 28.01.2011, 2011, URL: <http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf> [accessed: 2013-08-09].
- [6] Wissenschaftsrat, “Stellungnahme zum Akademienprogramm,” *Wissenschaftsrat, Deutschland, Drs. 9035-09, Saarbrücken*, 28.05.2009, 2009, URL: <http://www.wissenschaftsrat.de/download/archiv/9035-09.pdf> [accessed: 2013-08-09].
- [7] di Maio, P., “What Are Knowledge Resources?” 2012, URL: <https://sites.google.com/site/kaframework/what-are-knowledge-assets> [accessed: 2013-08-10].
- [8] C. W. Holsapple and K. D. Joshi, “Knowledge Management Support of Decision Making, Organizational knowledge resources,” *Decision Support Systems*, vol. 31, no. 1, pp. 39–54, May 2001, ISSN: 0167-9236, Elsevier, URL: [http://dx.doi.org/10.1016/S0167-9236\(00\)00118-4](http://dx.doi.org/10.1016/S0167-9236(00)00118-4) [accessed: 2013-08-10].
- [9] di Maio, P., “Why Audit Knowledge Assets?” 2012, URL: <https://sites.google.com/site/kaframework/why-audit-knowledge-assets> [accessed: 2013-08-10].
- [10] di Maio, P., “A Global Vision: Integrating Community Networks Knowledge,” *Community Wireless Symposium, European Community, EC Infoday, Barcelona, 5th October 2012*, 2012, URL: http://people.ac.upc.edu/leandro/misc/CAPS_Paola.pdf [accessed: 2013-08-10].
- [11] I. Nonaka and H. Takeuchi, *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, Oxford, 1995, ISBN: 0195092694.
- [12] S. Lavington, N. Dewhurst, E. Wilkins, and A. Freitas, “Interfacing knowledge discovery algorithms to large database management systems,” *Journal on Information and Software Technology, special issue on Knowledge Discovery and Data Mining*, vol. 41, no. 9, pp. 605–617, Jun. 1999.
- [13] M. Liu, X. Wang, Z. Wang, and Y. Huang, “A Knowledge Discovery Algorithm Based on Genetic Algorithm,” vol. 1. Proceedings of the 3rd World Congress on Intelligent Control and Automation, June 28 – July 2, 2000, Hefei, P.R China, 2000, pp. 549–552, ISBN: 0-7803-5995-X.
- [14] “LX-Project,” 2013, URL: <http://www.user.uni-hannover.de/cpr/x/rprojs/en/#LX> (Information) [accessed: 2013-07-27].

- [15] C.-P. Rückemann, "Integrating Information Systems and Scientific Computing," *International Journal on Advances in Systems and Measurements*, vol. 5, no. 3&4, pp. 113–127, 2012, ISSN: 1942-261x, LCCN: 2008212470 (Library of Congress), URL: http://www.thinkmind.org/index.php?view=article&articleid=sysmea_v5_n34_2012_3/ [accessed: 2013-05-26] (ThinkMind(TM) Digital Library), URL: http://www.ariajournals.org/systems_and_measurements/sysmea_v5_n34_2012_paged.pdf [accessed: 2013-06-09].
- [16] "Geo Exploration and Information (GEXI)," 1996, 1999, 2010, 2013, URL: <http://www.user.uni-hannover.de/cpr/x/rprojs/en/index.html#GEXI> (Information) [acc.: 2013-07-27].
- [17] C.-P. Rückemann, "Beitrag zur Realisierung portabler Komponenten für Geoinformationssysteme. Ein Konzept zur ereignisgesteuerten und dynamischen Visualisierung und Aufbereitung geowissenschaftlicher Daten," Dissertation, WWU, Münster, Deutschland, 2001, 161 (xxii+139) S., OPAC, OCLC: 50979238, URL: <http://www.math.uni-muenster.de/cs/u/ruckema/x/dis/download/dis3acro.pdf> [accessed: 2012-01-15], URL: <http://www.user.uni-hannover.de/cpr/x/publ/2001/dissertation/wwwmath.uni-muenster.de/cs/u/ruckema/x/dis/download/dis3acro.pdf> [accessed: 2013-07-28].
- [18] C.-P. Rückemann, "Enabling Dynamical Use of Integrated Systems and Scientific Supercomputing Resources for Archaeological Information Systems," in *Proceedings of the International Conference on Advanced Communications and Computation (INFOCOMP 2012), October 21–26, 2012, Venice, Italy*. XPS, Xpert Publishing Services, 2012, pp. 36–41, Rückemann, C.-P. and Dini, P. and Hommel, W. and Pankowska, M. and Schubert, L. (eds.), ISBN: 978-1-61208-226-4, URL: http://www.thinkmind.org/download.php?articleid=infocomp_2012_3_10_10012 [accessed: 2013-06-09].
- [19] "Universal Decimal Classification Consortium (UDCC)," 2013, URL: <http://www.udcc.org> [accessed: 2013-07-27].
- [20] "Universal Decimal Classification (UDC)," 2013, Wikipedia, URL: http://en.wikipedia.org/wiki/Universal_Decimal_Classification [accessed: 2013-07-27].
- [21] B. A. Worley, "Knowledge Discovery from Data," *International Panel on Future High End Systems: Chances and Challenges for Intelligent Applications and Infrastructures, October 22, 2012, The International Conference on Advanced Communications and Computation (INFOCOMP 2012), October 21–26, 2012, Venice, Italy*, 2012, URL: <http://www.aria.org/conferences2012/ProgramINFOCOMP12.html> [accessed: 2013-07-28].
- [22] E. de Grolier, "A study of general categories applicable to classification and coding in documentation," 1962, UNESCO, United Nations, UNESDOC, URL: <http://unesdoc.unesco.org/images/0002/000250/025055eo.pdf> [accessed: 2013-08-18].
- [23] D. C. Weeks, M. Benton, and M. L. Thomas, "Universal Decimal Classification: A Selected Bibliography of UDC Literature," *Prepared for: U.S. Air Force Office of Scientific Research, Washington, D.C.*, Jan. 1971, URL: <http://www.dtic.mil/dtic/tr/fulltext/u2/727076.pdf> [accessed: 2013-08-18].
- [24] C.-P. Rückemann, "Archaeological and Geoscientific Objects used with Integrated Systems and Scientific Supercomputing Resources," *International Journal on Advances in Systems and Measurements*, vol. 6, no. 1&2, pp. 200–213, 2013, pages 200–213, ISSN: 1942-261x, LCCN: 2008212470 (Library of Congress), PPN: 756934176, URL: http://www.thinkmind.org/download.php?articleid=sysmea_v6_n12_2013_15 [accessed: 2013-08-17] (ThinkMind(TM) Digital Library), Leibniz-Bibliography, Germany: URL: <http://www.leibniz-bibliographie.de> [accessed: 2013-08-17], URL: <http://lccn.loc.gov/2008212470> [accessed: 2013-08-17] (LCCN Permalink).
- [25] M. Aydemir, "World of Amphora," 2013, URL: http://worldofamphora.com/site/index_en.html [accessed: 2013-08-10].
- [26] C.-P. Rückemann, "High End Computing for Diffraction Amplitudes," in *Symposium on Advanced Computation and Information in Natural and Applied Sciences, Proceedings of the 11th International Conference of Numerical Analysis and Applied Mathematics (ICNAAM), September 21–27, 2013, Rhodes, Greece, Proceedings of the American Institute of Physics (AIP), AIP Conference Proceedings 1558, Volume 1558, Two-part Book*. AIP Press, American Institute of Physics, Melville, New York, USA, 2013, pp. 305–308, ISBN: 978-0-7354-1184-5, ISSN: 0094-243X, DOI: 10.1063/1.4825483, URL: <http://proceedings.aip.org> [accessed: 2013-12-01], URL: <http://link.aip.org/link/?APCPCS/1558/305/1> [accessed: 2013-12-01].
- [27] M. A. Rashed, "Smart stacking: A new CMP stacking technique for seismic data," *The Leading Edge*, vol. 27, pp. 462–467, Apr. 2008, ISSN: 1070-485X, DOI: 10.1190/1.2907176, URL: <http://t.le.geoscienceworld.org/content/27/4/462.abstract> [accessed: 2012-12-16].
- [28] C.-P. Rückemann and B. Gersbeck-Schierholz, "Object Security and Verification for Integrated Information and Computing Systems," in *Proceedings of the Fifth International Conference on Digital Society (ICDS 2011), Proceedings of the International Conference on Technical and Legal Aspects of the e-Society (CYBERLAWS 2011), February 23–28, 2011, Gosier, Guadeloupe, France / DigitalWorld 2011*. XPS, 2011, pp. 1–6, ISBN: 978-1-61208-003-1, URL: http://www.thinkmind.org/download.php?articleid=cyberlaws_2011_1_10_70008 [accessed: 2013-07-28].
- [29] Berlin-Brandenburgische Akademie der Wissenschaften, "Leibniz Edition, Reihe VIII, Naturwissenschaftliche, medizinische und technische Schriften," 2013, Glossary, Concepts, BBAW, Berlin, URL: <http://leibnizviii.bbaw.de/glossary/concepts/> [accessed: 2013-05-26] (concepts glossary), URL: http://leibnizviii.bbaw.de/Leibniz_Reihe_8/Aus+Otto+von+Guericke,+Experimenta+nova/LH035,14,02_091v/index.html [accessed: 2013-05-26] (transcription), URL: http://leibnizviii.bbaw.de/pdf/Aus+Otto+von+Guericke,+Experimenta+nova/LH035.14,02_091v/LH035,14!02_091+va.png [accessed: 2013-05-26] (scan).
- [30] Gottfried Wilhelm Leibniz Bibliothek (GWLb), Niedersächsische Landesbibliothek, "GWLb Handschriften," 2013, hannover, URL: <http://www.leibnizcentral.de/CiXbase/gwlbhss/> [accessed: 2013-05-26].
- [31] "LeibnizCentral," 2013, URL: <http://www.leibnizcentral.com/> [accessed: 2013-02-10].
- [32] "LeibnizEdition," 2013, Akademie-Ausgabe, (critical edition), URL: <http://www.leibniz-edition.de/> [accessed: 2013-09-08].
- [33] Berlin-Brandenburgische Akademie der Wissenschaften, "Leibniz Edition, Reihe IV, 4, Politische Schriften," 2013, BBAW, Berlin, Akademie-Ausgabe, (critical edition), URL: http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/leibniz_potsdam/bilder/IV7text.pdf#page=224 [accessed: 2013-09-08].
- [34] Berlin-Brandenburgische Akademie der Wissenschaften, "Leibniz Edition, Reihe IV, 1, Allgemeiner, politischer und historischer Briefwechsel," 2013, BBAW, Berlin, Akademie-Ausgabe, (critical edition), URL: http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/leibniz_potsdam/bilder/IV7text.pdf#page=226 [accessed: 2013-09-08].
- [35] Berlin-Brandenburgische Akademie der Wissenschaften, "Leibniz Edition, Reihe IV, 7, Politische Schriften," 2013, BBAW, Berlin, Akademie-Ausgabe, (critical edition), URL: http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/leibniz_potsdam/bilder/IV7text.pdf#page=224 [accessed: 2013-09-08], URL: http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/leibniz_potsdam/bilder/IV7text.pdf#page=226 [accessed: 2013-09-08].
- [36] Niedersächsische Landesbibliothek, "Leibniz Edition, Reihe I, 21B, Allgemeiner, politischer und historischer Briefwechsel," 2013, NLB, Hannover, Leibniz-Archiv, Akademie-Ausgabe, (critical edition), URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/I21B.pdf#page=86> [accessed: 2013-09-08].
- [37] Niedersächsische Landesbibliothek, "Leibniz Edition, Reihe I, 17A, Allgemeiner, politischer und historischer Briefwechsel," 2013, NLB, Hannover, Leibniz-Archiv, Akademie-Ausgabe, (critical edition), URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/I17A.pdf#page=11> [accessed: 2013-09-08].
- [38] Niedersächsische Landesbibliothek, "Leibniz Edition, Reihe III, 5A, Mathematischer, naturwissenschaftlicher und technischer Briefwechsel," 2013, NLB, Hannover, Leibniz-Archiv, Akademie-Ausgabe, (critical edition), URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III5A.pdf#page=408> [accessed: 2013-09-08].
- [39] Niedersächsische Landesbibliothek, "Leibniz Edition, Reihe III, 5B, Mathematischer, naturwissenschaftlicher und technischer Briefwechsel," 2013, NLB, Hannover, Leibniz-Archiv, Akademie-Ausgabe, (critical edition), URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III5B.pdf#page=212> [accessed: 2013-09-08], URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III5B.pdf#page=232> [accessed: 2013-09-08], URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III5B.pdf#page=234> [accessed: 2013-09-08].

- 08], URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III5B.pdf#page=292> [accessed: 2013-09-08].
- [40] Niedersächsische Landesbibliothek, "Leibniz Edition, Reihe III, 6A, Mathematischer, naturwissenschaftlicher und technischer Briefwechsel," 2013, NLB, Hannover, Leibniz-Archiv, Akademie-Ausgabe, (critical edition), URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III6A.pdf#page=102> [accessed: 2013-09-08], URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III6A.pdf#page=117> [accessed: 2013-09-08].
- [41] Niedersächsische Landesbibliothek, "Leibniz Edition, Reihe III, 6B, Mathematischer, naturwissenschaftlicher und technischer Briefwechsel," 2013, NLB, Hannover, Leibniz-Archiv, Akademie-Ausgabe, (critical edition), URL: <http://www.nlb-hannover.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III6B.pdf#page=180> [accessed: 2013-09-08].
- [42] Berlin-Brandenburgische Akademie der Wissenschaften, "Leibniz Edition, Reihe IV, 1, Politische Schriften," 2013, BBAW, Berlin, Akademie-Ausgabe, (critical edition), URL: http://www.bbaw.de/bbaw/Forschung/Forschungsprojekte/leibniz_potsdam/bilder/IV1text.pdf#page=385 [accessed: 2013-09-08].
- [43] N. Witsen, "Tartaria, sive magni Chami Imperium ex credendis amplissimi viri," 1705, Amsterdam, URL: http://upload.wikimedia.org/wikipedia/commons/5/5f/Witsen_-_Tartaria.jpg [accessed: 2013-08-18] (Wikipedia), URL: http://de.wikipedia.org/wiki/Nicolaas_Witsen [accessed: 2013-08-18] (Wikipedia).
- [44] M. Fogel, "Brieffragmente (Letter fragments) about 16xx, Historici Pragmatici universal, Terrae motus, Physica," manuscript ID: 00016293, Source: Gottfried Wilhelm Leibniz Bibliothek (GWLb), Niedersächsische Landesbibliothek, GWLB Handschriften, Hannover, URL: <http://www.leibnizcentral.de/CiXbase/gwlbhss/> [accessed: 2013-05-26].
- [45] M. Fogel, "Brieffragmente (Letter fragments) about 16xx, Terrae Motus in Nova Francia," manuscript ID: 00016278, Source: Gottfried Wilhelm Leibniz Bibliothek (GWLb), Niedersächsische Landesbibliothek, GWLB Handschriften, Hannover, URL: <http://www.leibnizcentral.de/CiXbase/gwlbhss/> [accessed: 2013-05-26].
- [46] E. W. von Tschirnhaus, "Brief (Letter), Ehrenfried Walther von Tschirnhaus an Leibniz 17.IV.1677," pp. 59–73, 1987, Gottfried Wilhelm Leibniz, Sämtliche Schriften und Briefe, Mathematischer, naturwissenschaftlicher und technischer Briefwechsel dritte Reihe, zweiter Band, 1667 – 1679, Leibniz-Archiv der Niedersächsischen Landesbibliothek Hannover, Akademie-Verlag Berlin, 1987, herausgegeben unter Aufsicht der Akademie der Wissenschaften in Göttingen; Akademie der Wissenschaften der DDR.
- [47] G. F. von Franckenau, "Brief (Letter), Georg Franck von Franckenau an Leibniz 18. (28.) September 1697, Schloss Frederiksborg, 18. (28.) September 1697," pp. 568–569, Gottfried Wilhelm Leibniz Bibliothek (GWLb), Leibniz-Archiv der Niedersächsischen Landesbibliothek Hannover, URL: <http://www.gwlb.de/Leibniz/Leibnizarchiv/Veroeffentlichungen/III7B.pdf> [accessed: 2013-05-26].
- [48] B. Steinberger, "Plumes in a convecting mantle: Models and observations for individual hotspots," *JGR*, vol. 105, pp. 11 127–11 152, 2000.
- [49] "GMT - Generic Mapping Tools," 2013, URL: <http://imima.soest.hawaii.edu/gmt> [accessed: 2013-07-27].
- [50] GDAL Development Team, *GDAL - Geospatial Data Abstraction Library*, Open Source Geospatial Foundation, 2013, URL: <http://www.gdal.org> [accessed: 2013-07-27].
- [51] U. Inden, D. T. Meridou, M.-E. C. Papadopoulou, A.-C. G. Anadiotis, and C.-P. Rückemann, "Complex Landscapes of Risk in Operations Systems Aspects of Processing and Modelling," in *Proceedings of The Third International Conference on Advanced Communications and Computation (INFOCOMP 2013), November 17–22, 2013, Lisbon, Portugal*. XPS Press, 2013, pp. 99–104, ISSN: 2308-3484, ISBN: 978-1-61208-310-0.

Using Semantic Web Technologies to Follow the Evolution of Entities in Time and Space

Benjamin Harbelot*, Helbert Arenas[†], and Christophe Cruz[‡]
 Laboratoire Le2i, UMR-6302 CNRS, Département Informatique
 University of Burgundy
 Dijon, France

*benjamin.harbelot@checksem.fr, [†]helbert.arenas@checksem.fr, [‡]christophe.cruz@u-bourgogne.fr

Abstract—In this paper we present the “continuum model”. Our work follows a “perdurantism” approach and is designed to handle dynamic phenomena extending the 4D-fluent with the use of semantic web technologies. In our approach we represent dynamic entities as constituted by timeslices each with semantic, geometric, temporal and identity components. Our model is able to link the diverse representations of an entity and allows the inference of qualitative information from quantitative one. The inference results are later added to the ontology in order to improve knowledge about the phenomenon. The model has been implemented using OWL and SWRL. Our preliminary results are promising and we plan to further develop the model in the near future to increase the suitable data sources.

Keywords—spatio-temporal; semantics; GIS; perdurantism.

I. INTRODUCTION

For the design of a spatio-temporal knowledge system, it is necessary to consider the three components of an entity representation: 1) Spatial: consisting in the geometry, 2) Temporal: which defines the interval of existence of the geometries and finally 3) Semantic: which defines a meaning for the entity beyond the purely geographic one [1] [2]. Most of the current GIS tools focus on analysis and presentation of geographic data. However, nowadays due to the increasing availability of spatial/temporal data, it is necessary to have tools with inference capabilities, capable of assisting researchers in analysing large datasets. This new kind of tools should be able to identify patterns and perform reasoning with datasets corresponding to dynamic phenomena.

Modeling a real dynamic phenomenon can be seen as tracking the transition of phenomenon composing entities from one state to another. This transition is called: filiation relationship. Along time, entities with spatial components, can maintain different spatial and semantic relations with other entities. A natural way to model dynamic phenomena is to represent the evolution as a graph, in which entities and their states are represented as vertices and relations between entities as edges. A phenomenon would then generate a complex graph composed by different types of relations such as: temporal, semantic, spatial or filiation.

An alternative to classic GIS tools are Semantic Web technologies. Using these technologies it is possible to develop data models called ontologies specifically designed for reasoning and inference with software mechanisms. Ontologies allow for any given domain, the representation of relevant high level concepts as well as their properties and the relationships between concepts and entities. In this research we use Semantic Web technologies to develop the “continuum model”, an ontology that allows us to represent diverse dynamic entities and analyse their relationships along time. Traditionally ontologies are static in the sense that the information represented in them does not change in time or space. In this paper we introduce the continuum model, an ontology that extends the 4D-fluent. Our ontology provides the mechanisms required to keep track of spatial and semantic evolution of entities along time.

In Section II, we discuss related work in the field of spatio-temporal knowledge representation. In Section III, we introduce the continuum model. In Section IV, we present the model specification using description logics. In Section V, we show some examples of GeoSPARQL used to implement the model. In Section VI, we describe how the model operates using an urban growth example and later we indicate our conclusions and future work.

II. RELATED WORK

The development of a spatial-temporal knowledge system involves two aspects, first the representation of the knowledge and second, the necessary mechanisms to perform analysis and querying.

A. Representing temporal data

The two main philosophical theories concerning the representation of object persistence over time are: *endurantism* and *perdurantism*. The first one, *endurantism*, considers objects as three dimensional entities that exist wholly at any given point of their life. On the other hand, *perdurantism*, also known as the four dimensional view, considers that entities have temporal parts, “timeslices” [3]. From a perdurantism point of view the temporal dimension of an entity is composed by all its timeslices. Therefore, it

represents the different properties of an entity over time as *fluent*. A *fluent* is a property valid only during certain intervals or moments in time. From a designer point of view, the *perdurantism* approach offers advantages over the *endurantism* one, allowing richer representations of real world phenomena [4].

The implementation of a *perdurantism* approach within an ontology, requires the conversion of static properties into dynamic ones. The two primary Semantic Web languages are OWL and RDF, unfortunately both of them provide limited support for temporal dynamics [5]. The OWL-Time ontology describes the temporal content of web pages and temporal properties of web services. Moreover, this ontology provides good support for expressing topological relationships between times or time intervals, as well as times or dates [6]. However, OWL allows only binary relations between individuals. In order to overcome this limitation several methodologies have been proposed for the representation of dynamic objects and their properties. Among the most well known are: temporal RDF, versioning, reification, N-ary relationships and the 4D-fluent approach.

Temporal RDF [7] proposes an extension of the standard RDF for naming properties with the corresponding time interval. This allows an explicit management of time in RDF. However, temporal RDF uses only RDF triples; therefore, it does not have all the expressiveness of OWL for instance, it is not possible to employ qualitative relations. Reification is a technique used to represent n-ary relations, extending languages such as OWL that allow only binary relations [8]. In [5], the authors developed a lightweight model using Reification. The model is designed to be deployed on top of existing OWL ontologies extending their temporal capabilities. The model also implements a set of SWRL (Semantic Web Rule Language) operators to query the ontology. Reification allows the use of a triple as object or subject of a property. But this method has also its limitations, for instance the transformation from a static property into a dynamic one increases substantially the complexity of the ontology, reducing the querying and inference capabilities. Additionally reification is prone to redundant objects which reduces its effectiveness. Versioning is described as the ability to handle changes in ontologies by creating and managing multiple variants of them [9]. However, the major drawback of Versioning, is the redundancy generated by the slightest change of an attribute. In addition, any information requests must be performed on multiple versions of the ontology affecting its performance.

An alternative to the previously mentioned approaches is the 4D-fluent, which is an approach based on the *perdurantism* philosophical theory. It considers that the existence of an entity can be expressed with multiple representations, each corresponding to a defined time interval. In the literature, 4D-fluent is the most well known method to handle dynamic properties in an ontology. It has a simple structure

allowing to easily transform a static ontology into a dynamic one [10]. Unfortunately, the 4D fluent approach has also some limitations; although it allows the recording of frequent timeslices, it can not handle explicit semantics. This fact causes two problems: 1) It is difficult to maintain a close relationship between geometry and semantics; and 2) It increases the complexity for querying the temporal dynamics and understanding the modelled knowledge. Furthermore, this approach does not define qualitative relations to describe the type of change that has occurred or to describe the temporal relationships between objects. Then we can not know which entities have undergone a change and what entities might be the result of that change. Regardless of its limitations the 4D-fluent approach offers a solid starting point for the representation of temporal information in OWL. An interesting previous work using this approach is [11]. Here the authors developed SOWL, which uses 4D-fluent to extend the ontology OWL-time making it able to handle qualitative relations between intervals, such as “before” or “after” even with intervals with vague ending points.

B. Querying the ontology

In [12], the authors introduce a model in which spatial-temporal information contained in a database and a spatial-temporal inference system work together. However, no information is given on the Semantic Web technologies, only the Java language is quoted as a component of the inference engine; therefore, the universality and effectiveness of the inference system can be questioned. Another work is [13] in which the authors propose a reasoning system that combines the topological calculus capabilities of a GIS and the inference capabilities of the semantic web field. However, the notion of time is not incorporated in this model.

The capability of switching from quantitative to qualitative data is only possible with a reasoning system. In the case of SOWL this is possible thanks to the implementation of SWRL built-in. In SOWL, the built-ins allow the system to infer topological, directional and metric relations between entities. Qualitative information can be inferred from quantitative one and can be used as an alternative in case of missing quantitative data. In order to query the ontology the developers of SOWL implemented a language similar in syntax to SQL. This language performs simple spatial-temporal querying for both static and dynamic data [11]. However, the work does not support identity relationships. It does not provide mechanisms to follow the changes an entity might experience by analysing its different representations.

Due to the nature of spatial-temporal datasets, we need a system able to handle large datasets. Traditionally, SPARQL has been the most common language to query an ontology. SPARQL is a W3C recommendation that operates at the level of RDF graphs. There are extensions, such as st-SPARQL and geoSPARQL that have been developed in order to allow SPARQL to operate on spatial entities [14].

These extensions define datatypes, functions and operations allowing spatial analysis, however, there is limited temporal support. St-SPARQL is based on an extension of RDF called st-RDF that integrates contact geometries and incorporates time in RDF. GeoSPARQL offers similar capabilities, however, it has the advantage of being an OGC supported standard.

Our previous work [1] introduced the *Continuum* model using Java and SWRL rules to implement it. The rules were executed via a graphical interface using the Jena API to connect to the ontology and JDBC to access a database. The application automatically detects the presence of spatial built-ins in SWRL rules and performs the necessary calculations in the database. The system can automatically rewrite SWRL rules containing spatial built-ins. On one hand, this prevents repeating calculations that have already been done. On the other hand, it also generates SWRL rules without spatial built-ins but rather based on qualitative relationships expressed through properties defined in the ontology. However, we note three limitations to this model: 1) The treatment of a query containing a spatial built-ins can be very long depending on the number of geometries involved in spatial analysis, 2) The execution of SWRL rules containing spatial built-ins currently depends on our application and cannot be executed from other sources, for instance the traditional plugin SQWRL Tab Query of the Protégé tool, 3) There are limitations in the size of the datasets that can be managed by the application.

Although SWRL is a potent inference tool, it is not fully supported in current available triplestores. A triplestore is a software mechanism able to store large datasets with semantic annotations, providing query and retrieval capabilities. Some of the available triplestores support GeoSPARQL allowing users to perform complex spatial queries [15] [16].

In order to overcome this limitations identified in [1], we decided to modify the system architecture and implement a new version using a triplestore with spatial capabilities for spatial calculations and data storage. After evaluating the available options we opted for Parliament. In this paper we present a further development of the model first presented in [1], using in this case GeoSPARQL/SPARQL. In the next section we will describe how we implemented this approach in the continuum model.

III. THE CONTINUUM MODEL

The spatial evolution of an object involves movement or a change of shape [17]. In the case of a movement, it is easy to identify and locate the entity before and after the event. However, when an entity suffers a succession of changes a key question arises: how much can it change before its identity is modified? And if there is a semantic change, then how do we know that this is the same entity at different times?

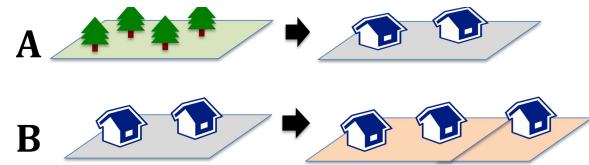


Figure 1. Evolution examples: A) Two different semantic objects for the same geometry. B) Two related geometries for the same semantic object.

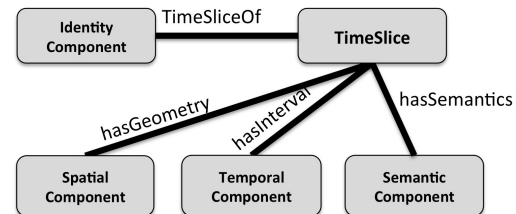


Figure 2. The four components of a timeslice within the continuum model.

The 4D-fluent approach does not allow an entity to change its nature, only allows the change of the value of some of its properties. However, the semantics associated with a geometry may change. For example, a land parcel may change from being forest into being urban. In this example the geometry has not changed, however, a semantic change has occurred (see Figure 1A). It is equally possible that the semantics might not change while the geometry evolves. For instance, a given urban land parcel might expand by purchasing neighbouring parcels (see Figure 1B).

In order to represent a dynamic entity in the continuum model we create a set of timeslices, each corresponding to a representation of the entity during a determined period of time. Each timeslice is constituted by four components as depicted in Figure 2: 1) Semantic: To describe the knowledge associated with the entity, 2) Spatial: It is the graphical representation, 3) Temporal: It represents the interval or time instant that describe the temporal existence, and 4) The identity component, that allows us to group timeslices belonging to the same entity.

The goal of the continuum model is to follow the evolution of entities through time. To achieve this goal the model records the changes that entities might go through in their semantic or spatial components along time. For this purpose the model creates a new representation every time a change occurs (spatial, semantic or identity). There is a parent-child relationship between the resulting timeslices. A resulting child timeslice retains all the unchanged characteristics from the original parent timeslice. Figure 3 depicts how we can represent the evolution of an object in which only the semantic part has changed. Figure 4 depicts the evolution of an object in which the spatial component varies, while the rest of the components remain constant. Each change adds to the genealogy of the spatio-temporal components. The parent-child relation is recorded in the system, allowing

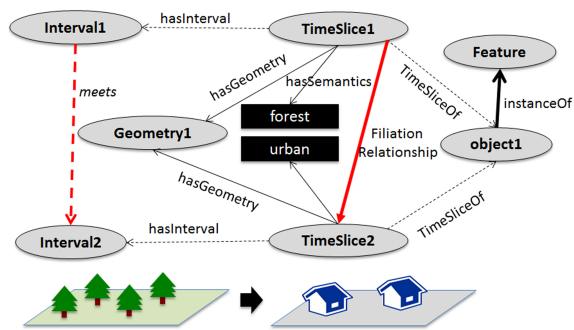


Figure 3. Evolution example: A semantic change with the same geometry.

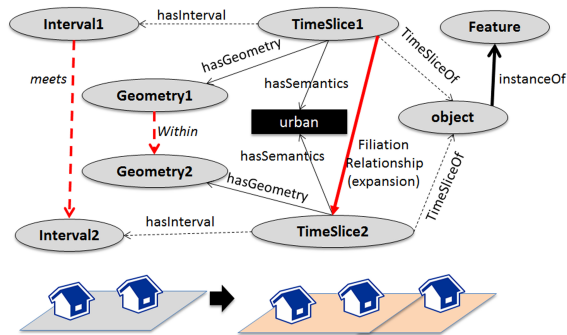


Figure 4. Evolution example: A spatial change with the same semantic

the analysis and querying of the information. The model enforces a coherency between the time intervals of timeslices contained in the system. By using this representation we are able to establish relationships between the components of two different timeslices.

Figure 5 depicts an example of objects genealogy. In this example we have the objects o_1, o_2, \dots, o_6 . Each of the object evolve along time. A set of timeslices compose the temporal representation of each object, thus $o_1 : [ts_1, ts_2]$, $o_2 : [ts_3, ts_5]$, $o_3 : [ts_4, ts_6]$, $o_4 : [ts_8, ts_9, ts_{10}]$, $o_5 : [ts_{11}, ts_{12}, ts_{13}, ts_{14}, ts_{15}]$ and $o_6 : [ts_{16}, ts_{17}, ts_{18}]$. The system enforces temporal coherency, children objects can not occur before the parents.

The continuum groups related timeslices, which have a valid time interval of existence. The model links individual timeslices to their context. For instance, a timeslice can have a child that corresponds to a new object, then the identity component might be different between parent and child. Our system allows the definition of qualitative relations between timeslices, even when the timeslices belong to different objects. Figure 5 depicts the evolution of objects and how the continuum model is used to study them.

In our model we have implemented qualitative temporal relations based on binary and mutually exclusive relations as proposed by Allen [18] (see Figure 6). The addition of Allen relations increase the expressive power of the system by adding qualitative information in addition to the quantitative

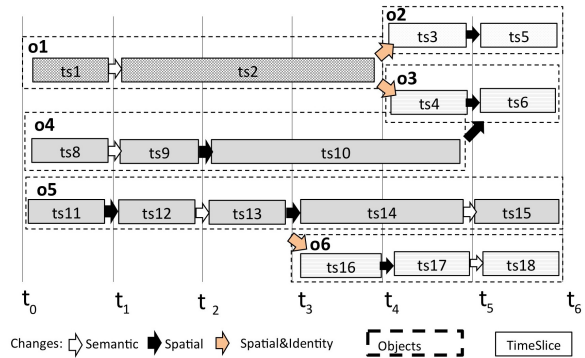


Figure 5. Using the continuum model to represent the evolution of an entity.

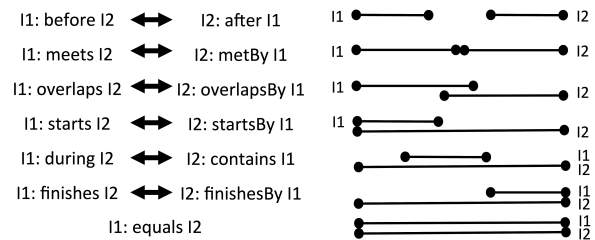


Figure 6. Allen temporal relations.

one. By using defined Allen relations between intervals we can obtain qualitative information even from intervals with vague endpoints in a similar fashion to [10]. For example, Figure 7 depicts intervals "I1", "I2" and "I3". While we know the start and ending points of "I1", we do not know the ending point of "I2", and we do not know the starting point of "I3". However, we know that "I1" meets "I2" and that "I2" contains "I3". Then we can infer that because "I2" contains "I3", then "I3" must be after "I1", even if the information about start and ending points is incomplete. Lack of knowledge caused by semi closed intervals is largely filled by the integration of Allen relations to the model (see Figure 7).

In GIS, objects or regions are represented by points, lines, polygons or other more complex geometries based on these primitives. All these geometries are defined using the coordinates of points which are quantitative information. There are mainly three types of relationships between geometries: directional, metric, and topological relationships. The topological analysis between two objects is done using the models: Dimensionally Extended Nine-Intersection Model (DE-9IM) or Region Connection Calculus (RCC8)

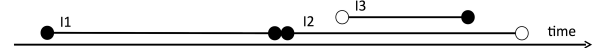


Figure 7. Using Allen temporal relations to infer new knowledge.

Table I
TOPOLOGICAL PREDICATES AND THEIR CORRESPONDING MEANINGS.

| Topological | Predicate Meaning |
|-------------|---|
| Equals | The Geometries are topologically equal. |
| Disjoint | The Geometries have no point in common. |
| Intersects | The Geometries have at least one point in common (the inverse of Disjoint). |
| Touches | The Geometries have at least one boundary point in common, but no interior points. |
| Crosses | The Geometries share some but not all interior points, and the dimension of the intersection is less than that of at least one of the Geometries. |
| Overlaps | The Geometries share some but not all points in common, and the intersection has the same dimension as the Geometries themselves. |
| Within | Geometry A lies in the interior of Geometry B |
| Contains | Geometry B lies in the interior of Geometry A (the inverse of Within) |

[19]. In both cases, we obtain an equivalent set of topological relationships for specific regions. To calculate the spatial relationships between two geometries the DE-9IM model takes into account the inside, the outside, and the contour of the geometries leading to the analysis of nine intersections as described in [19]. There are eight possible spatial relationships of the resulting analysis-9IM (see Table I).

IV. MODEL SPECIFICATION

The relationships based on quantitative information can be translated later into qualitative data [17]. By analysing the relationships between temporal, spatial and identity components of timeslices it is possible to deduce qualitative topological relationships between them. The results of the analysis can be used to specify more semantically complex constructions. In this section we use Tarski-style formalisms to specify the components of our model.

A. Temporal components

To represent time intervals we follow the semantics suggested by Artale and Franconi (1998). We can think of the temporal domain as a linear structure \mathcal{T} composed by a set of temporal points \mathcal{P} . The components of \mathcal{P} follow a strict order $<$, which forces all points between two temporal points t_1 and t_2 to be ordered. By selecting a pair $[t_1, t_2]$ we can limit a closed interval of ordered points. The set of interval structures in \mathcal{T} is represented by \mathcal{T}_{\leq}^* [20].

Temporal Points (\mathcal{P}):

$$\mathcal{P}^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \quad (1)$$

Time Intervals (\mathcal{T}_{\leq}^*):

$$[to, tf] \doteq \{x \in \mathcal{P} | to \leq x \leq tf, to \neq tf\} in \mathcal{T} \quad (2)$$

where to and tf are the initial and ending points of the interval respectively.

To define the relations identified by Allen [18] (see Figure 6) we first define two intervals $i1$ and $i2$: $\mathcal{T}_{\leq}^*(i1)$, $\mathcal{T}_{\leq}^*(i2)$,

being i_{to} the starting point and i_{tf} the ending point of the intervals:

$$Before(i1, i2) \rightarrow (i1_{tf} < i2_{to}) \quad (3)$$

$$Meets(i1, i2) \rightarrow (i1_{tf} = i2_{to}) \quad (4)$$

$$Overlaps(i1, i2) \rightarrow (i1_{tf} > i2_{to}) \wedge (i1_{tf} < i2_{tf}) \quad (5)$$

$$Starts(i1, i2) \rightarrow (i1_{to} = i2_{to}) \wedge (i1_{tf} < i2_{tf}) \quad (6)$$

$$During(i1, i2) \rightarrow (i1_{to} > i2_{to}) \wedge (i1_{tf} < i2_{tf}) \quad (7)$$

$$Finishes(i1, i2) \rightarrow (i1_{to} > i2_{to}) \wedge (i1_{tf} = i2_{tf}) \quad (8)$$

$$Equals(i1, i2) \rightarrow (i1_{to} = i2_{to}) \wedge (i1_{tf} = i2_{tf}) \quad (9)$$

B. Spatial Components

The spatial representation of an object is given by the coordinates representing its geometry and characteristics associated to it (Spatial reference system, accuracy, precision, format, etc). It is represented by \mathcal{G} . The spatial topological relations between geometries are defined by the Extended Nine-Intersection model (DE-9IM) (see Table I [19]).

Additionally, we can implement the operation *Union* valid for geometries:

$$[x_g, y_g, z_g] \in \mathcal{G} | Equals(z_g, Union(x_g, y_g)) \quad (10)$$

In this case, the combination of geometries x_g and y_g will result in a new geometry z_g .

C. Semantic Component

The semantic component of the objects describes the nature of the entities and can be composed by one or more alphanumeric properties.

D. Timeslices

An object representation in time is composed by a set of timeslices. Each timeslice \mathcal{TS} in the model has four components: 1) A time interval \mathcal{T}_{\leq}^* 2) A geometry \mathcal{G} , 3) An identity \mathcal{O} and 4) A semantic component representing all other potential alphanumeric properties associated to a timeslice. We represent all these properties as $\overline{\mathcal{TS}}$, as suggested in [21]. $\overline{\mathcal{TS}}$ represents all the qualities that distinguish the class timeslice from other classes.

$$\mathcal{TS} \equiv \forall hasGeometry. \mathcal{G} \sqcap \forall hasTime. \mathcal{T}_{\leq}^* \sqcap \overline{\mathcal{TS}} \sqcap \forall isTimeSliceOf. \mathcal{O} \quad (11)$$

E. Filiation relationships between timeslices

In the continuum model the existence of an object is defined by a set of timeslices representing the state of the object during a defined period of time. In the model, a new timeslice is generated when a original timeslice suffers a change in any of its components. The relation between the original and the new timeslice follows a *parent - child*, filiation pattern. For this relation to exist the interval of the *parent* timeslice must *meets* the interval of the *child*

timeslice (see Figure 6). In order to exist the filiation *parent-child* relationship between timeslices at least one of the components (geometry, semantics or identity) must remain constant. The filiation relationship is specified as:

$$\begin{aligned} & \forall hasFiliation.TS \\ & \{ts_1 \in TS^I | \forall ts_2. (ts_1, ts_2) \in hasFiliation^I \\ & \rightarrow ts_2 \in TS^I \wedge \exists \leq_2 ((ts_{1g} \neq ts_{2g}) \vee (ts_{1s} \neq ts_{2s}) \vee \\ & (ts_{1o} \neq ts_{2o})) \wedge (meets(ts_{1i}, ts_{2i}))\} \end{aligned} \quad (12)$$

where: $\{ts_1, ts_2\} \in TS$, $\{ts_{1g}, ts_{2g}\} \in \mathcal{G}$, $\{ts_{1s}, ts_{2s}\} \in TS$ and $\{ts_{1i}, ts_{2i}\} \in \mathcal{I}$

The filiation relationship can be further specialized by setting or not constraints in the identity (\mathcal{O}) component, then we have two possible filiation relationships: *hasContinuation* and *hasDerivation* [22] [23].

1) *Continuation relationship*: In this case, the identity component of parent and child timeslices is the same.

$$\begin{aligned} & \forall hasContinuation.TS \\ & \{ts_1 \in TS^I | \forall ts_2. (ts_1, ts_2) \in hasContinuation^I \\ & \rightarrow ts_2 \in TS^I \wedge ((ts_{1g} \neq ts_{2g}) \vee (ts_{1s} \neq ts_{2s})) \wedge \\ & (ts_{1o} = ts_{2o}) \wedge (meets(ts_{1i}, ts_{2i}))\} \end{aligned} \quad (13)$$

where: $\{ts_1, ts_2\} \in TS$, $\{ts_{1g}, ts_{2g}\} \in \mathcal{G}$, $\{ts_{1s}, ts_{2s}\} \in TS$ and $\{ts_{1i}, ts_{2i}\} \in \mathcal{I}$

2) *Derivation relationship*: In this case, there is a difference between the identity component of parent and child, while there is at least one component (geometry or semantic) that remains constant.

$$\begin{aligned} & \forall hasDerivation.TS \\ & \{ts_1 \in TS^I | \forall ts_2. (ts_1, ts_2) \in hasDerivation^I \\ & \rightarrow ts_2 \in TS^I \wedge \exists ((ts_{1g} = ts_{2g}) \vee (ts_{1s} = ts_{2s})) \wedge \\ & (ts_{1o} \neq ts_{2o}) \wedge (meets(ts_{1i}, ts_{2i}))\} \end{aligned} \quad (14)$$

where: $\{ts_1, ts_2\} \in TS$, $\{ts_{1g}, ts_{2g}\} \in \mathcal{G}$, $\{ts_{1s}, ts_{2s}\} \in TS$ and $\{ts_{1i}, ts_{2i}\} \in \mathcal{I}$

F. Topological filiation relationships

By identifying the topological relationships between the geometric component of the timeslices we can define specific filiation relationships in which the spatial components evolve (see Figure 8).

1) *Expansion*: In this case, the geometric component of the child timeslice contains the geometry of the parent timeslice. There is no change in the identity component of the timeslice, both parent and child timeslices belong to the same object (see Figure 8).

$$\begin{aligned} & \forall hasExpansion.TS \\ & \{ts_1 \in TS^I | \forall ts_2. (ts_1, ts_2) \in hasExpansion^I \\ & \rightarrow ts_2 \in TS^I \wedge (ts_{1g} \neq ts_{2g}) \wedge (ts_{1o} = ts_{2o}) \wedge \\ & meets(ts_{1i}, ts_{2i}) \wedge hasWithin((ts_{1g}, ts_{2g}))\} \end{aligned} \quad (15)$$

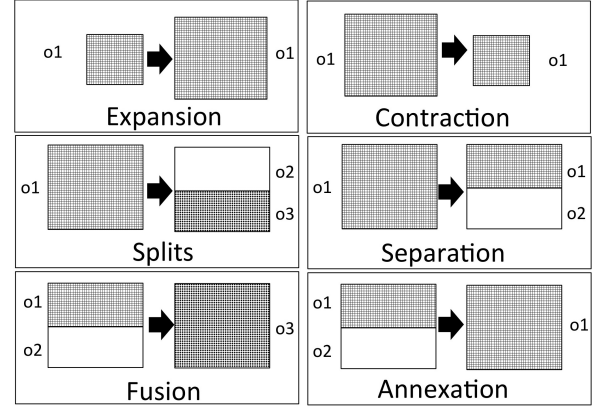


Figure 8. Topological filiation relationships.

where: $\{ts_1, ts_2\} \in TS$, $\{ts_{1g}, ts_{2g}\} \in \mathcal{G}$ and $\{ts_{1i}, ts_{2i}\} \in \mathcal{I}$

2) *Contraction*: This process is the opposite to *expansion* (see Figure 8).

$$\begin{aligned} & \forall hasContraction.TS \\ & \{ts_1 \in TS^I | \forall ts_2. (ts_1, ts_2) \in hasContraction^I \\ & \rightarrow ts_2 \in TS^I \wedge (ts_{1g} \neq ts_{2g}) \wedge (ts_{1o} = ts_{2o}) \wedge \\ & meets(ts_{1i}, ts_{2i}) \wedge hasContains((ts_{1g}, ts_{2g}))\} \end{aligned} \quad (16)$$

where: $\{ts_1, ts_2\} \in TS$, $\{ts_{1g}, ts_{2g}\} \in \mathcal{G}$ and $\{ts_{1i}, ts_{2i}\} \in \mathcal{I}$

3) *Splits*: In this relationship, the object identified as the parent timeslice identity (\mathcal{O}) ceases to exist. The geometry of the parent timeslice is then the origin of two new geometries corresponding to timeslices whose identity is new. The union of the geometries of the resulting children timeslices is equal to the geometry of the parent timeslice (see Figure 8).

$$\begin{aligned} & \forall hasSplits.TS \\ & \{ts_1 \in TS^I | \forall (ts_2, ts_3). (ts_1, (ts_2, ts_3)) \in hasSplits^I \\ & \rightarrow (ts_2, ts_3) \in TS^I \wedge \\ & (ts_{1g} \neq ts_{2g}) \wedge (ts_{1g} \neq ts_{3g}) \wedge (ts_{2g} \neq ts_{3g}) \wedge \\ & (ts_{1o} \neq ts_{2o}) \wedge (ts_{1o} \neq ts_{3o}) \wedge \\ & meets(ts_{1i}, ts_{2i}) \wedge meets(ts_{1i}, ts_{3i}) \wedge \\ & equals(ts_{1g}, Union(ts_{2g}, ts_{3g}))\} \end{aligned} \quad (17)$$

where: $\{ts_1, ts_2, ts_3\} \in TS$, $\{ts_{1g}, ts_{2g}, ts_{3g}\} \in \mathcal{G}$ and $\{ts_{1i}, ts_{2i}, ts_{3i}\} \in \mathcal{I}$

4) *Separation*: In this case, the parent entity continues existing, however, its geometry originates a new geometry corresponding to a new entity. A *hasSeparation* relationship is similar to a *hasSplits* relationship with the difference that in *hasSeparation* at least one of the children timeslices must have the same entity as the parent timeslice (see Figure 8).

$$\begin{aligned}
& \forall hasSeparation. \mathcal{TS} \\
& \{ts_1 \in \mathcal{TS}^I \mid \forall (ts_2, ts_3). (ts_1, (ts_2, ts_3) \in hasSeparation^I \\
& \rightarrow (ts_2, ts_3) \in \mathcal{TS}^I \wedge \\
& (ts_{1g} \neq ts_{2g}) \wedge (ts_{1g} \neq ts_{3g}) \wedge (ts_{2g} \neq ts_{3g}) \wedge \\
& \exists_{=1}((ts_{1o} = ts_{2o}) \vee (ts_{1o} = ts_{3o})) \wedge \\
& meets(ts_{1i}, ts_{2i}) \wedge meets(ts_{1i}, ts_{3i}) \wedge \\
& equals(ts_{1g}, Union(ts_{2g}, ts_{3g}))\}
\end{aligned}
\tag{18}$$

where: $\{ts_1, ts_2, ts_3\} \in \mathcal{TS}$, $\{ts_{1g}, ts_{2g}, ts_{3g}\} \in \mathcal{G}$ and $\{ts_{1i}, ts_{2i}, ts_{3i}\} \in \mathcal{I}$

5) *Fusion*: In this relationship, the two parent entities merged and cease to exist to give rise to a new geometry corresponding to a new entity. Inverse to a *hasSplits* relationship. The resulting geometry is equal to the union of the former geometries.

$$\begin{aligned}
& \forall hasFusion. \mathcal{TS} \\
& \{ts_1 \in \mathcal{TS}^I \mid \forall (ts_2, ts_3). (ts_1, (ts_2, ts_3) \in hasFusion^I \\
& \rightarrow (ts_2, ts_3) \in \mathcal{TS}^I \wedge \\
& (ts_{1g} \neq ts_{2g}) \wedge (ts_{1g} \neq ts_{3g}) \wedge (ts_{2g} \neq ts_{3g}) \wedge \\
& (ts_{1o} \neq ts_{2o}) \wedge (ts_{1o} \neq ts_{3o}) \wedge (ts_{2o} \neq ts_{3o}) \wedge \\
& meets(ts_{1i}, ts_{3i}) \wedge meets(ts_{2i}, ts_{3i}) \wedge \\
& equals(Union(ts_{1g}, ts_{2g}), ts_{3g})\}
\end{aligned}
\tag{19}$$

where: $\{ts_1, ts_2, ts_3\} \in \mathcal{TS}$, $\{ts_{1g}, ts_{2g}, ts_{3g}\} \in \mathcal{G}$ and $\{ts_{1i}, ts_{2i}, ts_{3i}\} \in \mathcal{I}$

6) *Annexation*: In this case, the two parent entities merge but the resulting entity keeps the identity of one of its parents.

$$\begin{aligned}
& \forall hasAnnexation. \mathcal{TS} \\
& \{ts_1 \in \mathcal{TS}^I \mid \forall (ts_2, ts_3). (ts_1, (ts_2, ts_3) \in hasAnnexation^I \\
& \rightarrow (ts_2, ts_3) \in \mathcal{TS}^I \wedge \\
& (ts_{1g} \neq ts_{2g}) \wedge (ts_{1g} \neq ts_{3g}) \wedge (ts_{2g} \neq ts_{3g}) \wedge \\
& ((ts_{1o} = ts_{3o}) \vee (ts_{2o} = ts_{3o})) \wedge (ts_{1o} \neq ts_{2o}) \wedge \\
& meets(ts_{1i}, ts_{3i}) \wedge meets(ts_{2i}, ts_{3i}) \wedge \\
& equals(Union(ts_{1g}, ts_{2g}), ts_{3g})\}
\end{aligned}
\tag{20}$$

where: $\{ts_1, ts_2, ts_3\} \in \mathcal{TS}$, $\{ts_{1g}, ts_{2g}, ts_{3g}\} \in \mathcal{G}$ and $\{ts_{1i}, ts_{2i}, ts_{3i}\} \in \mathcal{I}$

V. IMPLEMENTATION

This is an evolving work, continuously we are adding new capabilities to the continuum model. In our latest implementation we have deployed our ontology in a Parliament triplestore. In order to populate our ontology we have developed customized tools able to read information stored in shapefiles, GML, WFS and postgresSQL/postGIS data repositories and upload it into our triplestore. The harvesting tools have been developed using Java with Jena and Geotools libraries.

In this section we show how we can identify some of the filiation relationships between timeslices using GeoSPARQL.

Continuation:

```

SELECT
?ts1 ?ts2
WHERE{
?ts1 a abc:TimeSlice.
?ts2 a abc:TimeSlice.
?o1 a abc:Object.
?ts1 abc:isTimeSliceOf ?o1.
?ts2 abc:isTimeSliceOf ?o1.
?ts1 abc:hasInterval ?i1.
?ts2 abc:hasInterval ?i2.
?ts1 geo:hasGeometry ?geo1.
?ts2 geo:hasGeometry ?geo2.
?geo1 geo:asWKT ?geo1wkt.
?geo2 geo:asWKT ?geo2wkt.
?ts1 abc:hasSemantic ?s1.
?ts2 abc:hasSemantic ?s2.
FILTER (
((!geof:sfEquals(?geo1wkt,?geo2wkt)) &&
(?s1=?s2))) ||
((geof:sfEquals(?geo1wkt,?geo2wkt)) &&
(?s1!=?s2))) &&
(temporal:meets(i1,i2)) )
}

```

Derivation:

```

SELECT
?ts1 ?ts2
WHERE{
?ts1 a abc:TimeSlice.
?ts2 a abc:TimeSlice.
?o1 a abc:Object.
?o2 a abc:Object.
?ts1 abc:isTimeSliceOf ?o1.
?ts2 abc:isTimeSliceOf ?o2.
?ts1 abc:hasInterval ?i1.
?ts2 abc:hasInterval ?i2.
?ts1 geo:hasGeometry ?geo1.
?ts2 geo:hasGeometry ?geo2.
?geo1 geo:asWKT ?geo1wkt.
?geo2 geo:asWKT ?geo2wkt.
?ts1 abc:hasSemantic ?s1.
?ts2 abc:hasSemantic ?s2.
FILTER (
((!geof:sfEquals(?geo1wkt,?geo2wkt)) &&
(?s1=?s2))) ||
((geof:sfEquals(?geo1wkt,?geo2wkt)) &&
(?s1!=?s2))) &&
(temporal:meets(i1,i2)) &&
(?o1!=?o2) )
}

```

```
}

```

Topological filiation relationships: Splits:

```
SELECT
?ts1 ?ts2 ?ts3
WHERE{
?ts1 a abc:TimeSlice.
?ts2 a abc:TimeSlice.
?ts3 a abc:TimeSlice.
?o1 a abc:Object.
?o2 a abc:Object.
?o3 a abc:Object.
?ts1 abc:isTimeSliceOf ?o1.
?ts2 abc:isTimeSliceOf ?o2.
?ts3 abc:isTimeSliceOf ?o3.
?ts1 abc:hasInterval ?i1.
?ts2 abc:hasInterval ?i2.
?ts3 abc:hasInterval ?i3.
?ts1 geo:hasGeometry ?geo1.
?ts2 geo:hasGeometry ?geo2.
?ts3 geo:hasGeometry ?geo3.
?geo1 geo:asWKT ?geo1wkt.
?geo2 geo:asWKT ?geo2wkt.
?geo3 geo:asWKT ?geo3wkt.
FILTER (
(?s1!=?s2) &&
(?o1!=?o3) &&
(?o2!=?o3) &&
(temporal:meets(i1,i2)) &&
(temporal:meets(i1,i3)) &&
(!geof:sfEquals
(?geo1wkt,geof:union(?geo2wkt,?geo3wkt)))
}
```

VI. EXAMPLE CONTINUUM

The continuum model is flexible enough to be adapted in multiple fields. In this example, we use it to represent the urban evolution of the city of New Orleans. This city is the largest in Louisiana. It is located between the Mississippi river and the lake Pontchartrain. The oldest part of the city is placed on the banks of the Mississippi river, on the natural levees of the river. Since its beginning it was the main settlement in the area, however, New Orleans was not the only one. In the vicinity, other cities such as Jefferson, Lafayette or Greenville were established. These cities were absorbed into New Orleans in the XIXth. century. From 1718 to 1900 the urban growth was only into areas that were by nature *high and dry*. However, after 1900, technical developments allowed municipal authorities to drain the swamps located between the river and the lake Pontchartrain, creating artificially dry land for urban development. The final result is a city that occupies an area that resembles a bowl, with large neighborhoods lying on the bottom, in areas with low elevation, some of them even below sea

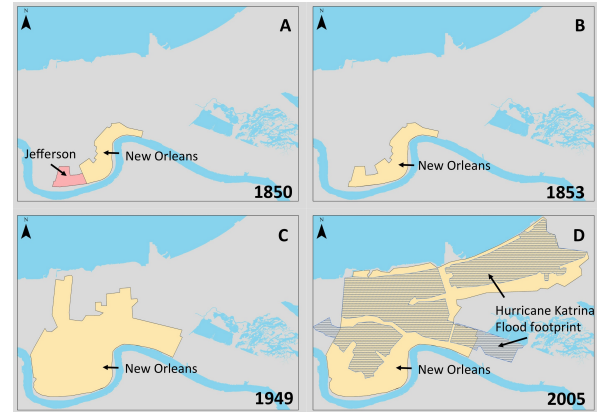


Figure 9. City of New Orleans along time.

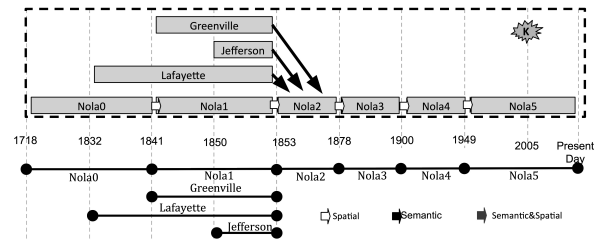


Figure 10. Time frame or urban evolution

level, which make them vulnerable to floods. If we add to this, the subsiding soil phenomenon occurring in the area, and the erosion of the coast line, we end up with a city in a particularly vulnerable location [24]. Figure 9 depicts different stages of the urban evolution of New Orleans. Figure 9D depicts the extension of the city around 2005, when it was flooded by Hurricane Katrina.

In order to use the continuum model to represent the urban evolution, first we define the class *Human Settlement* (*HS*) which represents cities that evolve through time. The temporal existence of each of the entities belonging to this class is represented by a set of timeslices which have the four components: 1) Semantic: representing properties associated with the entity, valid for the specific time interval, 2) Spatial: It is the graphical representation, in this case, the footprint of the human settlement, 3) Temporal: It represents the valid interval of existence for this timeslice, and 4) The identity component, that allow us to group timeslices belonging to the same human settlement.

By using the continuum model we are able to identify processes such as *conurbation*. The conurbation process involves two cities merging. Using the model we can represent the process as:

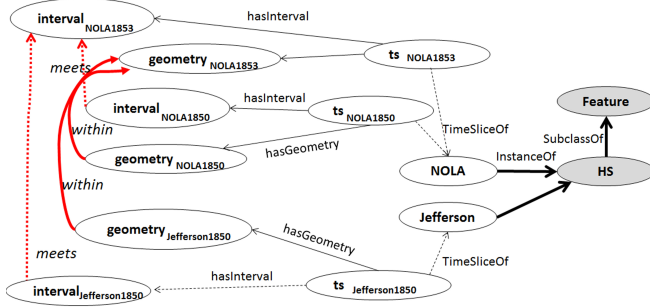


Figure 11. Representation of a conurbation process using the continuum model

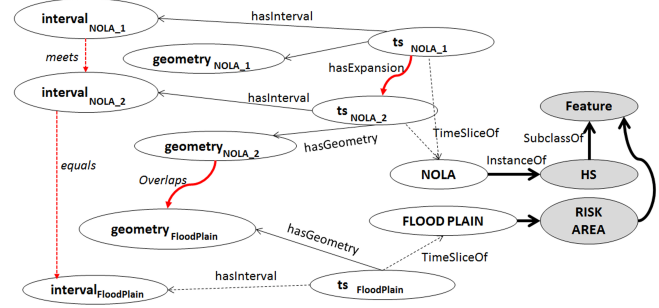


Figure 12. Representation urban growth in risk areas using the continuum model

$\forall hasConurbation.TS$

$$\{ts_1 \in \mathcal{TS}^I \mid \forall (ts_2, ts_3). (ts_1, (ts_2, ts_3) \in hasConurbation^I \rightarrow (ts_2, ts_3) \in \mathcal{TS}^I \wedge (ts_{1g} \neq ts_{2g}) \wedge (ts_{1g} \neq ts_{3g}) \wedge (ts_{2g} \neq ts_{3g}) \wedge ((ts_{1o} = ts_{3o}) \vee (ts_{2o} = ts_{3o})) \wedge (ts_{1o} \neq ts_{2o}) \wedge meets(ts_{1i}, ts_{3i}) \wedge meets(ts_{2i}, ts_{3i}) \wedge equals(Union(ts_{1g}, ts_{2g}), ts_{3g}) \wedge ([ts_{1o}, ts_{2o}, ts_{3o}] \in \mathcal{HS})\}$$

where: $\{ts_1, ts_2, ts_3\} \in \mathcal{TS}$, $\{ts_{1g}, ts_{2g}, ts_{3g}\} \in \mathcal{G}$ and $\{ts_{1i}, ts_{2i}, ts_{3i}\} \in \mathcal{I}$. This can be expressed in a more compact form as:

$$hasConurbation((ts_1, ts_2), ts_3) \equiv hasAnnexion((ts_1, ts_2), ts_3) \mid ([ts_{1o}, ts_{2o}, ts_{3o}] \in \mathcal{HS})$$

Figure 11 depicts how the model is used in the conurbation New Orleans example. *NOLA* (New Orleans) and *Jefferson* are instances of the class *human settlements HS*. $ts_{NOLA1850}$ and $ts_{NOLA1853}$ are timeslices of the entity *NOLA*, while $ts_{Jefferson1850}$ is a timeslice of the entity *Jefferson*. In the graphic we can see the relationships that can be established between the geometries and intervals of the different timeslices. Based on the analysis of the spatial-temporal relationships of the components of the timeslices we can infer qualitative information such as the identification of *Conurbation* processes.

Using the continuum model it is also possible to combine timeslices of objects of different nature. For instance, we can model the process *urban growth in risk area*. In order to represent the risk area we will use the footprint of the flood caused by Hurricane Katrina in 2005 (see Figure 9D). We create a new class *risk areas* as \mathcal{RA} . Then we can identify the process *growth in risk area* as:

$$UrbanGrowthInRiskArea(hs_1, hs_2, ra_1) \equiv hasExpansion(hs_1, hs_2) \wedge \neg Overlaps(hs_1, ra_1) \wedge Overlaps(hs_2, ra_1)$$

where: $\{hs_1, hs_2\} \in \mathcal{HS}$ and $ra_1 \in \mathcal{RA}$

Figure 12 depicts the relationships that are necessary to analyse to determine the urban growth process in risk areas.

VII. CONCLUSION

In Figures 11 and 12 we represent the relationships between geometries and intervals used in our analysis. By using these relations we can detect complex transitions between timeslices. Understanding data semantics is at the core of our work providing an easier way to manage data and reduce queries complexity. When using reasoning capabilities specific to the Semantic Web, the system may increase the knowledge stored in the ontology.

The continuum model is based on the 4D-fluent representation and develops the continuum concept in the context of a spatial-temporal GIS in order to preserve best understandable semantics for the objects represented. The continuum model handles time and space independently for each object allowing the inclusion or not of time and space in queries of spatial, temporal or spatial-temporal nature. Currently, the system is capable of tracking the evolution of objects along time. This model introduces a novel approach for the handling of properties and attributes for each object. The semantic management of the properties and attributes for each object will be part of further research in order to develop a complete system for the semantics of spatial-temporal information.

Our model offers explicit semantic and flexibility for semantics interoperability between information systems and data sharing. Currently, we are doing research in the field of *smart queries*, a term coined by [25]. The term refers to the combination of heterogeneous datasources in order to solve complex problems. Using the continuum model we will be able to integrate vector data sources available on the web [26]. We plan to apply these capabilities to complex modelling scenarios such as Land Use/Land Cover change.

REFERENCES

- [1] B. Harbelot, H. Arenas, and C. Cruz, "The spatio-temporal semantics from a perdurantism perspective," in *Proceedings of the Fifth International Conference on Advanced Geographic*

Information Systems, Applications, and Services GEOProcessing, Nice, France, February-March 2013, pp. 114–119.

- [2] M. Yuan, “Use of a three-domain representation to enhance GIS support for complex spatial-temporal queries,” *Transactions in GIS*, vol. 3, pp. 137–159, March 1999.
- [3] C. Welty and R. Fikes, “A reusable ontology for fluents in OWL,” in *Proceedings of 2006 conference on Formal Ontology in Information Systems (FOIS 2006)*, 2006, pp. 226–236.
- [4] M. Al-Debei, M. Mourhaf Al Asswad, S. de Cesar, and M. Lycett, “Conceptual modelling and the quality of ontologies: Endurantism vs. perdurantism,” *International Journal of Database Management Systems*, vol. 4, no. 3, June 2012.
- [5] M. O’Connor and A. Das, “A method for representation and querying temporal information in OWL,” in *Proceedings of Biomedical Engineering Systems and Technologies BIOSTEC 2010*, 2010, pp. 97–110.
- [6] J. Hobbs and F. Pan, “Time ontology in OWL,” (Online) <http://www.w3.org/TR/owl-time/>, (Accessed on November 2012).
- [7] C. Gutierrez, A. Hurtado, and A. Vaisman, “Introducing time into RDF,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, pp. 207–218, February 2007.
- [8] P. Hayes, “RDF semantics, W3C Recommendation, 10 february 2004,” (Online) <http://www.w3.org/TR/rdf-mt/>, 2004, (Accessed on November 2012).
- [9] M. Klein and D. Fensel, “Ontology versioning on the Semantic Web,” in *Proceedings of the First International Semantic Web Working Symposium SWWS’01*, Stanford, July 2001, pp. 75–91.
- [10] S. Batsakis and E. Petrakis, “SOWL: Spatio-temporal representation reasoning and querying over the semantic web,” in *Proceedings of the 6th. International Conference on Semantic Systems I-SEMANTICS 2010*, Graz, Austria, September 2010, pp. 15:1–15:9.
- [11] —, “SOWL: a framework for handling spatio-temporal information in OWL2.0,” *Rule Based Reasoning, Programming, and Applications Lecture Notes in Computer Science*, vol. 6826, pp. 242–249, 2011.
- [12] K. Ryu and Y. Ahn, “Application of moving objects and spatiotemporal reasoning,” 2001, a TIMECENTER Technical Report.
- [13] A. Karmacharya, C. Cruz, F. Boochs, and F. Marzani, “Integration of spatial processing and knowledge processing through the semantic web stack,” in *GeoSpatial Semantics*, ser. Lecture Notes in Computer Science, C. Claramunt, S. Levashkin, and M. Bertolotto, Eds. Springer Berlin Heidelberg, 2011, vol. 6631, pp. 200–216. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-20630-6_13
- [14] M. Koubarakis and K. Kyzirakos, “Modeling and querying metadata in the semantic sensor web: The model stRDF and the query language stSPARQL,” in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, L. Aroyo, G. Antoniou, E. Hyvnen, A. Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, Eds. Springer Berlin Heidelberg, 2010, vol. 6088, pp. 425–439. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-13486-9_29
- [15] I. Emmons, *Parliament User Guide*, Raytheon BBN Technologies, 2012.
- [16] R. Battle and D. Kolas, “Enabling the geospatial semantic web with parliament and GeoSPARQL,” *Semantic Web*, 2012.
- [17] N. Brisaboa, I. Mirbel, and B. Pernici, “Constraints in spatio-temporal databases: A proposal for classification,” in *Proceedings of the 3th. International Workshop on Evaluation of Modeling Methods in System Analysis and Design*, Pisa, 1998.
- [18] J. Allen, “Maintaining knowledge about temporal intervals,” *Communications of the ACM*, vol. 26, no. 11, pp. 832–843, Nov. 1983. [Online]. Available: <http://doi.acm.org/10.1145/182.358434>
- [19] C. Strobl, *Encyclopedia of GIS Springer*. Springer, 2008, ch. Dimensionality Extended Nine-Intersection Model (DE-9IM), pp. 240–245.
- [20] A. Artale and E. Franconi, “A temporal description logic for reasoning about actions and plans,” *Journal of Artificial Intelligence Research*, vol. 9, pp. 463–506, 1998.
- [21] F. Baader and W. Nutt, “The description logic handbook,” F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds. New York, NY, USA: Cambridge University Press, 2003, ch. Basic description logics, pp. 43–95. [Online]. Available: <http://dl.acm.org/citation.cfm?id=885746.885749>
- [22] G. D. Mondo, M. Rodriguez, C. Claramunt, L. Bravo, and R. Thibaud, “Modeling consistency of spatio-temporal graphs,” *Data & Knowledge Engineering*, vol. 84, no. 0, pp. 59 – 80, 2013.
- [23] G. D. Mondo, J. G. Stell, C. Claramunt, and R. Thibaud, “A graph model for spatio-temporal evolution,” vol. 16, no. 11, pp. 1452–1477, jun 2010.
- [24] R. Campanella, *Geographies of New Orleans: Urban fabrics before the storm*. Center for Louisiana Studies, 2006.
- [25] J. Goodwin, “What have ontologies ever done for us - potential applications at a national mapping agency,” in *OWL: Experiences and Directions (OWLED)*, 2005.
- [26] H. Arenas, B. Harbelot, and C. Cruz, “A Semantic Web Approach for Geodata Discovery,” in *Proceedings of 7th. SecoGIS Workshop*, Hong Kong, PRC, November 2013.

Designing for 3D User Experience in Tablet Context

Design and Early Phase User Evaluation of Four 3D GUIs

Minna Pakanen¹⁾, Leena Arhippainen²⁾, and Seamus Hickey¹⁾

¹⁾Department of Information Processing Science, ²⁾Center for Internet Excellence

¹⁾P.O. Box 3000, ²⁾P.O. Box 1001

FI-90014 University of Oulu, Finland

minna.pakanen@oulu.fi, leena.arhippainen@cie.fi, seamus.n.hickey@gmail.com

Abstract—This article focuses on a possibility to have a personal three dimensional graphical user interface inside a virtual environment on a tablet device. We describe the visual design process and early phase user experience evaluation of four 3D GUIs in a virtual environment. A user evaluation was conducted by using a structured pair evaluation procedure, where we adapted the concept walkthrough method with non-functional visually high quality prototypes. In addition, we conducted a self-expression task, where participants were able to draw their idea of a 3D GUI on a touch screen tablet device. This evaluation provided us a lot of user feedback for the design, which we utilized in the final iterated designs. In addition, we point out many design issues relating to the visual design of the personal GUI in virtual environments in a touch screen context. Our user evaluation indicated that participants would like to have their personal 3D GUI in a virtual environment. However, the visual design of the 3D GUI should create a secure and private feeling for them. Also, participants did not want the GUI to occlude excessively with the background. The visual indication is needed also when a user transfers items from a personal GUI to the virtual environment and for showing the user's active position between the GUI and virtual environment. We took all of these and other aspects into account when we designed the final iterated designs, which are also introduced in this article.

Keywords- visual design; user experience; 3D GUI; touch screen tablet device; HCI; self-expression method.

I. INTRODUCTION

Three dimensional (3D) user interfaces (UIs) have been studied from the late 1970s. Prior research has focused on PCs with several input devices [10] and larger touch screen devices [20][42][23], but there is not a large amount of research in the area of 3D UIs on mobile touch screen tablet sized devices [38]. There is a need for user experience based information because of the increasing amount of 3D applications, such as 3D games [33], developed for touch screen tablet devices, e.g., Apple iPad. Also, there have been interest in bringing 3D collaborative virtual environments (CVEs) such as Second Life (SL) [28] to the tablet devices. The first attempt is Lumiya, a 3D viewer for SL for Android tablets [30] which also has a limited view to the 3D virtual environment. In 3D CVEs there are a lot of 3D objects and avatars present in a 3D space. The challenge from a user experience point of view is that in current 3D CVEs, it is not

possible to carry out other activities, such as reading personal emails, browsing files or playing games, in parallel. To do activities like that, a user needs to switch to another application, which may weaken the 3D environment experience.

In this article, we explore a different approach to that problem by focusing on a possibility to have a personal 3D graphical user interface (GUI) inside a VE. By a personal GUI, we mean a private user interface (UI) showed only to the user, not visible publicly, in contrast to embedded elements in VEs which are visible to all users.

This article investigates users' subjective experiences of a personal 3D GUI in a collaborative VE in the early design phase and offers user feedback on visual design of 3D GUIs. Article extends our previous work from the ACHI 2013 conference paper by Pakanen et al. [1]. In this article, we discuss more designing for user experience, extend the design phase description and present new findings gathered by using the Self-Expression Template method [8]. Finally, we present new iterated concepts.

First, in Section III, we present the visual design of four 3D GUI metaphors and preparation of user experience (UX) evaluation material. In Section IV, we describe the UX study conducted with 40 participants by using non-functional visually high quality prototypes and the Self-Expression Template method. In Section V, we report the findings and factors that participants pointed out while evaluating our designs. In Section VI, we present our four iterated 3D GUI designs, based on our findings in the UX evaluation. Then, in Section VII, we discuss topics that designers should consider when designing 3D GUIs for CVEs on touch screen tablet devices. Finally, in Section VIII, conclusion and future work are presented.

II. RELATED RESEARCH

The research with 3D UIs and VEs have been extensive and is studied over many decades with PCs using several input devices. Touch screen technology has extended the research to new device areas, such as on larger touch displays on tables [20][42] and on the wall [23]. Despite of this, there is only little research done with tablet devices [38]. Bowman et al. [10] define a 3D UI as a UI that involves 3D interaction, which means human-computer interaction (HCI) where a user performs tasks directly in a 3D context. Based on this definition, a 3D interaction can be defined so that it comprises navigation, object manipulation, application

control [10][21][45] and visual design [14]. The focus of prior research has been on several topics. As 3D UI allows a larger set of items to be displayed at the same time in the UI space than 2D UIs, many earlier studies have focused on 3D file browsing and displaying hierarchical information [25][37][15]. Also different kinds of 3D menus [17] and metaphors have been investigated a lot over the years [2][17]. According to Gotchev et al. [19], the most popular 3D metaphors for mobile 3D media are: tree, mirror, elevator, book, art gallery, card and the hinged metaphor. As tablet devices have been used for reading books and magazines, a bookshelf metaphor [13] has become quite popular for displaying content, for example, in the Apple iPad [3]. Also 3D carousel metaphors have been under a large interest, both in industry and academy [23][36][48]. Different kinds of 3D and 2½D desktops have been designed and studied as well [2][27][41].

CVEs are social in their nature, but if there are personal items in a CVE, then their privacy should be clearly visualized to the user. Culnan [16] defines privacy as: *"The ability of individuals to control the terms under which their personal information is acquired and used"*. Privacy is a large research topic, but in this article, our emphasis is only on visual indication of the privacy in CVEs and VEs. The prior research has focused mainly on e-commerce applications for selling either real world products or virtual products for avatars [39] or for information exchange between avatars [26]. Butz et al. [12] introduced two visual indication practices (vampire mirror and privacy lamps) for indicating which items are shared and which are private.

Even though there has been an extensive amount of research in 3D UIs within the areas of navigation, application control and manipulation, the impact of visualization as a part of the 3D user experience is the least explored in the research. Also the research of personal 3D GUI elements, such as menu items, applications and files in a collaborative virtual environment is still lacking from the visual design, user experience and mobile tablet device points of view.

III. DESIGNING FOR USER EXPERIENCE

ISO 9241-110:2010 [22] defines user experience as: *"person's perceptions and responses resulting from the use and/or anticipated use of a product, system or service"*. Aesthetic aspects have become one of the largest parts of the UX with the modern tablet touch screen UIs, as the screen quality has improved within last years. Norman [32] claims that aesthetic design of objects can have a larger influence on user preferences than usability of the product. Also De Angeli et al. [18] found in their study with web pages that users preferred a more attractive webpage interface even though it was not as usable as the not so attractive version of it. Arhippainen's [4] 9th user experience heuristic says: *"Go for a perfect visual design"*. She explains that visual design can both make the UI aesthetically pleasurable and improve usability of the UI by making it more understandable, consistent and guiding [4]. Therefore, the visual design should be carefully understood and done by a visual designer. User experience cannot be designed, because it is in people, but it is possible to 'Design for experiencing' [40].

Also, material design for the evaluation should be carefully done by the designer and an effort should be made to make sure that examples are understandable and approachable to the evaluation participants. As Löwgren and Stolterman [31] stated, a designer has to be able to make his/her ideas "visible" to the evaluation participants so that subjects can "see", analyze and evaluate them. If participants cannot understand a new idea or a vision, it does not matter how good it is [31].

To 'design for experiencing', we used the industrial design process [44]. First, we explored approximately 40 existing 3D UIs and concept designs. Then, based on this benchmarking, literature and lessons learned from our earlier studies [6][38] with 3D UIs, we identified three design goals.

1) *Design a 3D GUI in a collaborative virtual environment.* The idea was to be able to use a 3D UI while spending time in a virtual space. 2) *Support the use of multiple applications within the 3D virtual environment.* In current touch screen tablet UIs, it is not possible to handle multiple applications within the same view. The aim was to make it possible to handle multiple parallel applications in a 3D UI within a VE. 3) *Design for 3D interaction on a touch screen.* The idea was to make it possible to select objects from the back rear of a carousel UI.

In the following sections, we describe the design phase of our 3D GUI metaphor designs. We had two design phases in our design process. The first phase included a preparation of the visual theme boards, one group design session, one individual design phase and expert evaluation of the concepts. The second design phase included an individual design phase, evaluation and expert evaluation for selecting the final ideas to the 3D modeling phase. Then, we also developed a Self-Expression Template method [8] for the UX evaluation. The aim of the template method was to get participants to express their ideas of 3D UI topic other way than just commenting on our designs. And finally, we prepared all material for the UX evaluation.

A. The First Design Phase

We started the first design phase with the preparation of five different styled *Visual theme boards* to help us create visuals for the concepts. Visual theme boards are almost as similar as *Mood boards* [29] and they are both used within the industrial design discipline as idea generation tools. For the preparation of visual theme boards, different kinds of inspirational images from the Internet were browsed through. When browsing through hundreds of images, we found interesting looking images of forms that could help us to create new visual styles for the 3D GUI. Each image group was named to represent the forms in the images. The given names were: Futuristic, Minimalistic, Natural/ Organic, Steampunk and Cartoon (Fig. 1). Futuristic images had smooth and streamlined forms with a little edge (Fig. 1, A). Minimalistic images had plain and simple forms and an idea of 'less is more' (Fig. 1, B). In the Natural/ Organic group, images had forms from nature, such as honey comb (Fig. 1, C). Steampunk style had gimmick, utopian and nostalgic forms (Fig. 1, D). Cartoon had sketch like, funny,

unexpected and exaggerated forms (Fig. 1, E). After selecting images to the groups, we built collages of the images on A4 sized boards to represent the titled visual style.

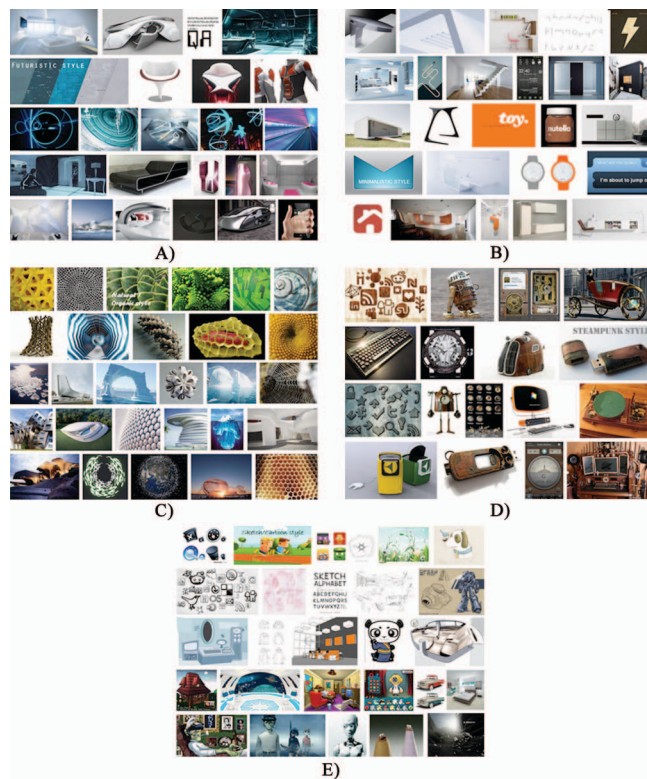


Figure 1. Visual theme boards: A) Futuristic, B) Minimalistic, C) Natural/Organic, D) Steampunk and E) Cartoon.



Figure 2. 1st expert evaluation.

Then, we started the designing. We started with a brainstorming session utilizing the visual theme boards with two industrial/interaction designers and a UX researcher. During this session, we wrote down different kinds of ideas, advice and needs for a 3D GUI. Next, we had a one-week individual sketching phase. Within that one week, we produced over 100 sketches of 3D GUI metaphors.

Then, we had an expert evaluation of the concepts with eight project members. Each one marked five of the most promising ideas with sticky-notes and wrote on it a short explanation (Fig. 2). The most promising ideas were

categorized to the five groups and named them as: 'Carry with you', 'Organic', 'History', 'Real life', 'Transparent' and 'Living' (Fig. 3). The names were given based on the content of the sketches. Sketches in the 'Carry with you' group looked like they could be carried in hand or in a pocket (Fig. 3, A). The 'Organic' category had natural forms and organisms, which grow from the ground when, for example, receiving an email (Fig. 3, B). The 'History' group had navigation signs and things faded out when the UI depth grew (Fig. 3, C). In the 'Real life' group, sketches had ideas of room based metaphors and things from real life, such as a letterbox (Fig. 3, D). In the 'Transparent' group, the sketches offered transparent and translucent solutions for 3D UIs (Fig. 3, E). The main idea was to be able to see through UI elements and being able to have a lot of items showed in parallel. In the 'Living' group, there was a lot of transitions and the form of the UI evolved as a reaction for user touch (Fig. 3, F).

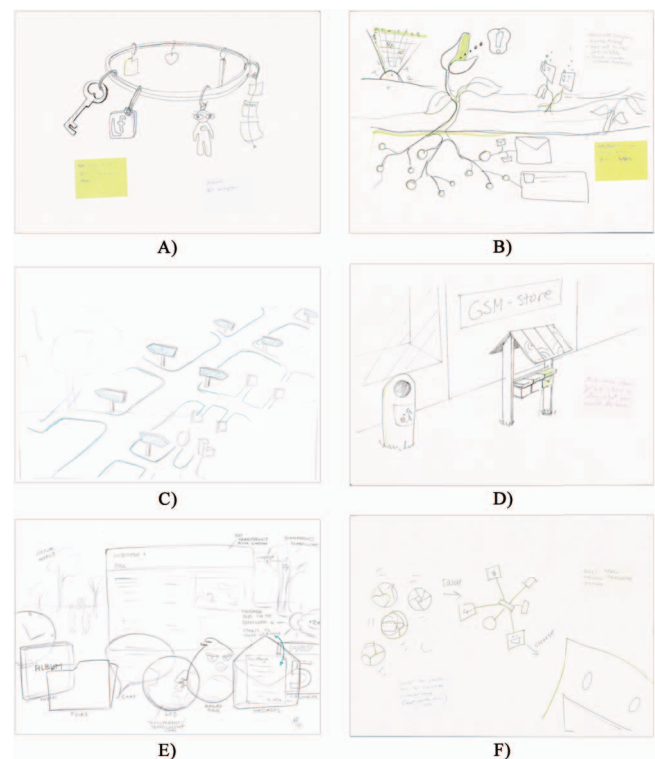


Figure 3. Example sketches of an each sketch group: A) Carry with you, B) Organic, C) History, D) Real life, E) Transparent, and F) Living.

B. The Second Design Phase

The second design phase was started with an individual sketching period. We developed the selected concept groups of the first design phase further and in more detail. In this phase, we also paid attention to user interaction steps, e.g., how the UI behaves if a user selects or taps something in it. Then, the sketches (approximately 50) were evaluated by eight UI and UX professionals. The evaluators were asked to give their vote or votes to the most interesting concepts according to their intuition. Intuition is often used in the design field for selecting the best concept. After evaluators

made their selections, they also discussed together to find out the best concepts for the user evaluation. Finally, the four 3D GUI metaphors: *Room*, *Shelves*, *Pie*, *Keyring* (Fig. 4, A-D, on the left) were selected for the 3D modeling and user evaluation phase. *Room* (Fig. 4, A) and *Shelves* (Fig. 4, B) metaphors have a similar visual style and both of them had a binder metaphor for files but a different amount of icons and depth of space. *Pie* (Fig. 4, C) and *Keyring* (Fig. 4, D) are both examples of the carousel metaphor, but with different visual style, hierarchy level structure and amount of icons. *Room* and *Shelves* concepts were examples of the 'Real life' UI metaphor group. *Pie* and *Keyring* were examples of 'Carry with you' UI metaphors.

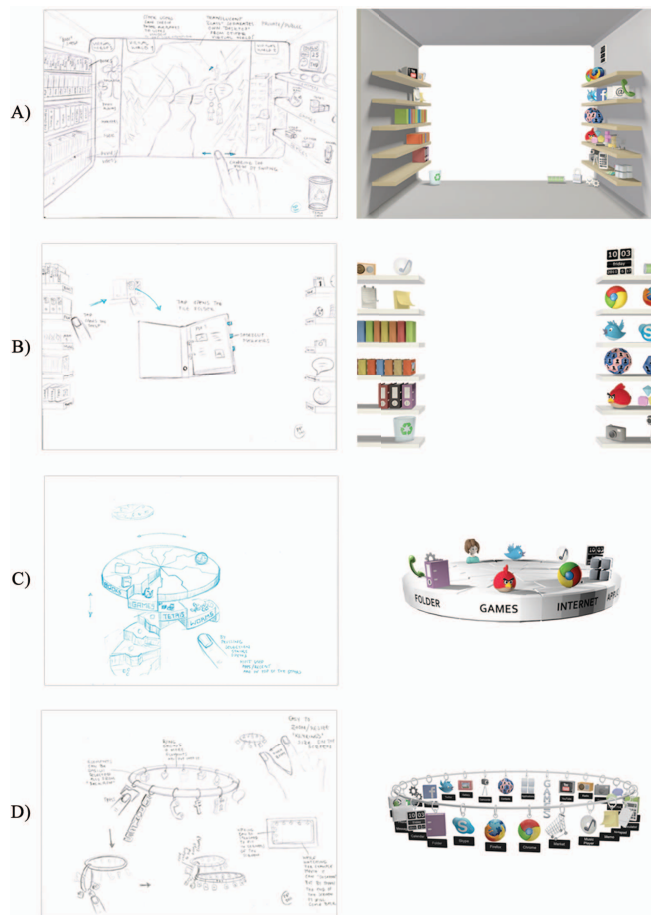


Figure 4. The four selected GUI concepts: A) Room, B) Shelves, C) Pie, and D) Keyring. Sketches on the left and 3D models on the right.

C. File Searching and Sharing Use Cases

As our sketches included hierarchical structures in the GUIs, therefore, we designed also a step-by-step use case (file searching and sharing) for each concept presented in Table I. The idea of the use case was to search for a PDF file (named as PDF 2), copy and share it to a pre-named contact.

D. Modeling of 3D GUI Metaphors and 3D Icons

The 3D models of the selected concepts were created by using the Blender program. First, we designed and modeled the GUI elements and 3D icons for the 3D GUIs. We selected applications that can be used in the tablet context (e.g., mail, phone, messaging, notebook, radio, maps, contacts, books, browsers, gallery, folder, trashcan, calendar, camera, games, music player and social media services). We had a set of 33 icons to be used in our GUI metaphor designs. The amount of icons in every design varied, because we wanted to have a different evaluation setup for each concept in order to evaluate the UI hierarchy structures and the amount of objects displayed in the UI metaphor and on the screen at once. There were 31 icons in the *Room* concept's first view, but in the *Shelves*, there were only seventeen, which are either fully or partially shown icons. The *Keyring* concept included 28 icons and *Pie* ten icons in the first menu hierarchy level (Fig. 4, A-D, on the right). Finally, we made compositions for 'file searching and sharing' use cases by moving and duplicating modeled UI menu elements.

E. Preparation of the Prototypes

We decided to evaluate our four designs as non-functional visually high quality prototypes at an early design phase as possible to get user feedback for the next iteration of our concepts with a fast, easy and cost-effective way. Because we were interested in finding out the user experiences of the visual aspects, it was important to make as high quality looking evaluation examples as possible [31]. Based on our design goals, we wanted to evaluate how users perceive the 3D GUIs in a virtual environment (Table I). Therefore, we selected one 3D model of a collaborative looking outdoor music VE from our earlier research work [6] and rendered out one image of it from Blender. Then, we rendered each image of the metaphors with the step-by-step use case and placed them on a VE background in Photoshop. We then added an authentic-sized 10 inch tablet frame around the images. Finally, we added images of hands, which were representing the touch gestures on top of the use case images (Table I) and saved the image series as PDFs.

F. Self-Expression Template Method

Involving users to the design of the user interfaces has become quite popular [49][24]. Prior research has shown that different self-reporting methods are good tools for gathering users' experiences, ideas and wishes for product development [4][6][43][24][5]. Therefore, we created a Self-Expression Template method to our UX evaluation. As we are studying tablet devices, we needed to make the template look realistic enough for the participants. Therefore, we placed an image of a real 10 inch tablet frame in the center of the A4 sized paper (21 x 29,5 centimeters) [8]. Self-Expression Templates were printed on heavy weight paper (200 grams). In the self-expression task, we gave to the participants color pencils for drawing their ideas on the template.

TABLE I. A FILE SEARCHING AND SHARING USE CASE IN EACH 3D GUI METAPHORS.

| Steps | Room | Shelves | Pie | Keyring |
|--------------------------------------|---|--|---|--|
| File searching | <p>User: Zooms in with a pinch zoom gesture.</p>  <p>User: Taps the PDF binder icon.</p>  <p>System: Opens the binder in the center of the screen.</p> <p>User: Taps the 'PDF 2' index marker</p>  <p>System: Turns the page and the intended PDF is in sight.</p> | <p>User: Tap the PDF binder icon on the shelf on the left side of the screen.</p>  <p>System: Activates and moves a shelf (that the binder is located on) near the center area of the screen and opens the binder in the center of the screen.</p>  <p>User: Taps the 'PDF 2' index marker</p>  <p>System: Turns the page and the intended PDF is in sight.</p> | <p>User: Taps the binder icon which was located on a one piece of the Pie.</p>  <p>System: The tapped piece of Pie drops one step down and the system opens three sub-pieces of the Pie on the same horizontal level. Three icons are located on top of the pieces; W (Word), PP (PowerPoint) and PDF (2nd hierarchy level).</p> <p>User: Taps the PDF icon.</p>  <p>System: Sub-pieces opens under the Pie GUI in the format of a hierarchical helical stairs (3rd hierarchy level).</p> | <p>User: Taps the binder icon which hangs from the Keyring.</p>  <p>System: Vertically orientated sub-ring with three icons; W, PP and PDF appears to hang from the original ring.</p> <p>User: Tap the PDF icon.</p>  <p>System: Another sub-ring opens horizontally to the icon's place.</p> <p>User: Zooms in (pinch zoom gesture).</p>  |
| File copying | <p>User: Long press the PDF icon</p> <p>System: The copied file icon appears on top of the PDF file.</p>  | <p>Copying is made similarly as in Room GUI.</p>  | <p>Copying is made similarly as in the Room GUI.</p>  | <p>Copying is made similarly as in the Room GUI.</p>  |
| File sharing by dragging | <p>User: Drags the copied file to the other side of the Room to the contact object (ball), and finally to the chosen contact.</p>  <p>System: Camera follows the file dragging and zooms in to the contact ball.</p>  | <p>User: Drags the copied file on another shelf on the other side of the screen with two contact objects (balls) on.</p>  <p>System: Moves the shelf with contact objects to the center area of the screen and closes folder.</p> <p>User: Drags the copied file to the contact ball, and finally to the chosen contact.</p> | <p>User: Drags the copied file on a contact piece in the Pie.</p>  <p>System: Opens sub-pieces in hierarchical helical stair format, where all the contacts are located on the steps of the 'stairs'.</p> <p>User: Drags the copied file to the chosen contact.</p> <p>System: Camera follows the file dragging.</p>  | <p>User: Drags the copied file to the contact object (ball) at the rear of the first hierarchy level ring.</p>  <p>System: Camera follows the file dragging and zooms in to the contact object.</p> <p>User: Drags the copied file to the chosen contact.</p> |
| Feedback indication to a user | <p>System: Shows a tiny version of the icon beside the contact, which disappears when it is sent.</p>  | <p>The system indication for sending is done the same way as in the Room GUI.</p>  | <p>System: Shows a tiny version of the icon beside the contact on the step of the stair, which disappears when the file is sent.</p>  | <p>The system indication for sending is done the same way as in Room GUIs.</p>  |

IV. USER EXPERIENCE EVALUATION

As we were interested in participants' subjective experiences, we conducted the study by using semi-structured pair evaluation settings, where we studied our four 3D GUI concepts as a part of a mixed methods evaluation procedure. Table II presents the contents of the whole evaluation procedure, which lasted from 90 to 120 minutes. This article focuses on findings gathered from the tasks number 2 & 3: four 3D GUI concepts with use cases and task number 7: self-expression task. Findings from the other tasks are presented in other publications. Subjects' preferences for 3D icons (task 1) can be found from [34]. How the participants perceived the depth of 3D space (task 5) is presented in [7]. Four 3D GUI concepts were studied by adapting the design walkthrough method in a controlled setting, which lasted from 25 to 59 minutes. In the end of evaluation session, the participants performed a Self-Expression Template drawing task, which lasted from 10-24 minutes. We used different methods to gather user feedback and experiences: video recording, semi-structured interviewing, and observing with user comments written down. First, in the beginning of the evaluation, subjects filled up a short background questionnaire, which had questions about the subjects' gender, age, prior touch screen and 3D experience.

The actual design walkthrough was conducted as follows for each 3D GUI concepts: Showing the 3D GUIs on a 3D VE with the 'file searching and sharing' use case on an authentic-sized tablet frame as a PDF from a laptop where the moderator changed the image and led the discussion. She asked participants to comment freely about what they are thinking and also asked additional questions every now and then. After the concept design walkthrough and other tasks, we had a self-expression task. Participants were given the Self-Expression Templates and color pencils and they were asked to draw or write a 3D UI for a touch screen tablet device. After participants finished their drawings, they were asked to explain their drawings.

TABLE II. THE UX EVALUATION PROCEDURE AND USER TASKS.

| No. | Task |
|-----|--|
| 1. | 2D/3D icon comparison tasks |
| 2. | Four 3D GUI concept evaluations |
| 3. | Four 3D GUI use case evaluation tasks |
| 4. | Contact and Square UI evaluations |
| 5. | 3D UI space and depth level selection tasks |
| 6. | Other 3D UI concept evaluations |
| 7. | Self-expression task |

A. Participants

In our user evaluation, we had 40 persons of which 63% were male. For recruiting participants, we used an online test user environment [35] and also sent email invitations to friends and colleagues to be distributed. The criterion for selecting participants was that each of them should have at least two months' experience [11] with touch screen devices (mobile phones or tablets). Almost all of the participants (93%) had prior touch screen experience with smart phones and 85% of them had tried or used tablet devices. The subjects' age varied from 23 to 52 years, with a mean of 35.

V. FINDINGS

All the material was qualitative, which we analyzed by applying the affinity diagram method [9]. We wrote down participants' comments on sticky-notes. Then, we made two analysis rounds for notes and grouped them based on their content. A summary of the analysis is presented in Table III. In the following subsections, we present the participants' perceived aspects and comments on the 3D GUIs in 3D VE and their wishes and needs for 3D GUI in a self-expression task.

TABLE III. A SUMMARY OF HOW PARTICIPANTS PERCEIVED FOUR 3D GUI METAPHORS ("+" ARE POSITIVE AND "-" ARE NEGATIVE ASPECTS).

| UX Factor | Four 3D GUI Metaphors | | | |
|--|---|--|--|--|
| | <i>Room</i> | <i>Shelves</i> | <i>Pie</i> | <i>Keyring</i> |
| Perceived visual appearance | + homely + things are ordered (garage/storage) + can see all the icons at once - unclear (icons are occluded/ too full) - childish and funny/ toy store | + clear - shelves are floating in the air (odd) - icons cut in half (ugly) - not possible to see all icons at once - floating in the air (odd) | + new / exciting / attractive + can see the most important icons at once - bulky / too thick/ chunky - masculine / engineering type / official - floating in the air (odd) | + new / different / interesting/ fun + can see all the icons at once - full / unclear (icons are overlapping) - feminine/ kitsch bracelet / swinging - floating in the air (odd) |
| Perceived 3Dness | + 3D space (Room) + enough depth + icons occluded | - not enough depth = 2D GUI - just 3D icons do not make 3D GUI - no occlusion | + 3D shape (round) + icons occluded + looks rotatable (interaction) | + 3D shape (round) + looks rotatable (interaction) - icons occluded |
| Perceived consumption of space from VE | + distinct from the background VE - consumes too much space from VE | + light/airy + does not consume too much space from VE | - consumes too much space from VE | + light/ airy + does not consume too much space from VE |
| Perceived privacy and safety | + clear visual separation from VE (walls) - can other users of VE see the content a shared item | - possible to share something to the VE by accident (no walls) - not clear visual separation from VE - can other users of VE see the content of own GUI or a shared item | - not clear visual separation from VE - can other users of VE see the content of own UI and a shared item | - not clear visual separation from VE - can other users of VE see the content of own UI and a shared item |
| Perceived ease of use | + looks simple/ easy to use - require more steps than 2D UI - too long dragging - needs camera & zooming controls | + no brainer to use + no camera controls required + shorter dragging | + brainless to use - carousals are difficult - menu hierarchy difficult and messy - too many steps (file search & sharing) | - difficult (can accidentally select a wrong icon) - menu hierarchy messy and weird - too many steps (file search & sharing) |
| Perceived utility by customization | + easy to categorize the content + easy to customize the GUI space | + easy to categorize the content | + could work as a launcher | + could work as a launcher + easy to categorize the content |

A. Perceived Visual Appearance

The *Room* metaphor (Fig. 4, A) was considered as a 'homely' GUI where one's own applications are in order. The *Room* metaphor was also called as 'garage' or 'storage', but it was also regarded as childish and funny like 'a toy store'. 18% of the participants thought that the *Shelves* concept (Fig. 4, B) was better, clearer, more approachable and pleasurable than the *Room* GUI. The *Pie* GUI metaphor (Fig. 4, C), in its turn, was perceived as interesting, new, exciting and visually attractive. On the other hand, the *Pie* was regarded as an official, masculine and engineering type of object and was called as 'a disk' or 'hard drive'. The visual style of the *Pie*'s plate was perceived to be bulky, chunky and too thick and it was called 'a concrete plate', 'tray', 'puzzle', 'Battle Star Galactic' or 'puck'. It was even suggested that the plate could be translucent. The visual style of the *Keyring* (Fig. 4, D) was considered to be new, different, interesting and fun. On the other hand, one participant commented that it is: "*a moment's wow*". Compared to the *Pie*, the *Keyring* was regarded as a feminine object and it was called as a kitsch bracelet. It was also referred to movement, for example, to 'a shower curtain rack', 'coat hanger rack', 'mobile', and 'janitor's key ring'. One person even said: "*I don't like if it's swinging*".

The participants liked the fact that they can easily get an overview of the GUI with one glance, with *Room*, *Pie* and *Keyring* GUIs. 15% of the participants did not like that all of the icons are not showing in the *Shelves* GUI. Also, it was perceived as odd and ugly that some of the icons on the shelves were cut in half. In contrast, 30% of the subjects liked the tighter view that the *Pie* concept offered even though there was even less content in sight. *Pie* and *Keyring* were perceived to look like launchers for applications. Participants thought that in the *Room* (18%) and *Keyring* (25%) GUIs, there were too many occluding application icons. It was perceived to be unclear and error prone while making selections.

The participants thought that all GUIs except the *Room*, looked weird and distressing with the virtual environment background, because they seemed to be floating in the air, for example, the *Pie* GUI was perceived as a UFO. Also, one participant commented the meaning of the *Pie* metaphor because of its location in the 3D environment: "*It looks like a tray when it is located near a bar*".

B. Perceived 3Dness

When the participants evaluated the 3Dness of the concepts, one factor was the depth of the space. Compared to other concepts, in the *Shelves* GUI, there was not enough depth to make it look like it was 3D and it was considered to be only a 2D GUI with 3D icons. As one participant commented on it: "*3D icons do not change the UI into 3D*". Another factor was the perceived interaction. The *Pie* and *Keyring* had the round shape which made them look rotatable; therefore, they were perceived as 3D. Also the icon occlusion was considered to be an important factor for creating a 3D feeling; thus, the *Shelves* concept was not

considered to be a 3D GUI. From the users' perspectives, 3Dness is made from occlusion, the shape of the UI and the depth of the space.

C. Perceived Consumption of Space from VE

The occlusion of the virtual environment by the GUI was evaluated by the participants. The *Room* and *Pie* were perceived to occlude too much from the VE. The *Room* GUI was showing the center area, but it was considered more like a little peak view to the VE. With the *Pie* GUI, the situation was quite the opposite; the plate of the *Pie* blocked the center area. In comparison, the *Shelves* and *Keyring* GUIs were considered to be lighter and airy on VE.

D. Perceived Privacy and Safety

The participants felt more secure with the *Room* concept, because there were walls separating the private area from the public background area. To create a secure feeling, there should be some kind of separation from the background environment. However, with the *Shelves*, *Pie* and *Keyring* GUIs, the participants had concerns for their privacy. For example, one participant commented on the *Pie* GUI: "*If I am in a public virtual space, can other people see my UI?*" The *Shelves* and *Keyring* GUIs were perceived as visually unclear and confusing, because behind the icons and UI elements, there were not any visual elements to separate it from the VE. With the *Shelves* GUI, the participants wished for a back plate or curtain behind the shelves.

There should also be a clear visual indication for showing the user's active position between the personal 3D GUI and collaborative virtual environment. This could be done with color or dimming effect on the non-active UI area. The participants thought that a possibility to interact between spaces and share content directly to a friend in the VE was good. On the other hand, they were concerned about a possibility to share something into the VE by accident. This could be prevented by giving a user a visual indication with a highlight color when something is moved from their personal GUI to the collaborative VE. There were also concerns such as, 'can someone else see a shared file and to whom it is shared to'. The shared content should be invisible to other users and it should look like it is protected for the user who is sharing it and who is receiving it. For example, as one participant suggested: "*Shared file could be protected with a folder?*"

E. Perceived Easy of Use

Even though we did not have a functional prototype, participants commented a lot how they perceived the usability aspects of each GUI metaphors with the 'file searching and sharing' use case. The participants thought that the *Shelves* GUI (Fig. 4, B) was better than other GUIs from the usability point of view. It was perceived to have shorter dragging, simpler hierarchy, fewer steps and camera movements, such as view rotation and zooming in too near to the UI elements. Also, one person said that in the *Shelves* GUI the interaction can be done more "*brainlessly*".

Even though a tablet has a gestural interface, the participants did not like long dragging, because it was perceived as difficult and prone to errors, such as an item is dragged to a wrong place. 15% of the participants suggested that instead of long dragging, a copy could be moved to a 'pocket' or virtual USB-memory stick and kept there until sharing. 15% of the participants suggested that the GUI could be intelligent, for example, the target object (in this case a contact), could automatically open beside of the binder while copying.

With the *Pie* and *Keyring* metaphor concepts, the 2nd and 3rd hierarchy levels were found to be distressing, because there were too many items illustrated at the same time and it looked too messy. When the 3rd hierarchy level opened in the *Keyring* GUI, 30% of the given comments were negative. As one participant commented: "*More and more jingling*". Also, the orientations of the sub-rings was found to be irritating, against the laws of physics and foolish. Therefore, it was suggested that rings could open horizontally either under the original ring, replacing it, or earlier opened rings could move deeper into the space when the new ring opens. With the *Pie* GUI, the 3rd hierarchy level opening as a form of helical stairs was unexpected by the participants and over 50% of the given comments were negative. For example, one participant described it: "*It exploded, went broken*". It was perceived as difficult, hard, complex and distressing. 13% of the participants commented that it looks like endless stairs. The helical stairs structure was perceived to prioritize the content. For example, with contacts, it creates a feeling that some of the contacts are more important than others. With PDF files, the structure was not that irritating, but the amount of items was considered to be critical for the controllability of the GUI. The participants suggested that instead of the 2nd and 3rd hierarchy levels in the *Pie* and *Keyring* GUIs, there could be a similar binder metaphor as in the *Room* and *Shelves*. Other suggestions for the *Pie* included: a drawer opening from it or another *Pie* could open under the first one. With the *Keyring*, there could be a binder metaphor or file cabinet instead of the 2nd and 3rd hierarchy levels.

F. Perceived Utility by Customization

The participants found customization interesting and useful within all of the GUIs. The smallest and simplest thing with the UI customization is to let the users adjust the amount of icons, change their places in the UI and categorize the GUI contents. For example, with the *Keyring*, the participants wanted to categorize icons in groups or pile them in stacks. With the *Room* and *Shelves* GUIs, the participants would have liked to organize their icons by placing work and leisure items on different sides of the room or on different shelves. Also, with the *Room* GUI, 30% of the participants were eager to change the overall visual style of the GUI, for example, with wallpapers.

G. Drawings on Self-Expression Template

Most of the participants (93%) drew on the Self-Expression Template, the rest of them wrote their needs for the 3D GUI (Fig. 5). The participants who did not draw their

ideas explained that they did not have skills to draw or did not know how to draw their ideas.

The participants had different ideas of 3D GUI on a touch screen tablet. 45% of them presented 3Dness in the UI by placing objects in either a room space (Fig. 6, A) or a realistic looking 3D world. 28% of the participants drew a carousel type of structures into their UI, so for them, the possibility to rotate makes a 3D GUI. 15% of the participants piled objects in the depth. So for them, a 3D GUI means adding the depth (z-axis) to the UI. The rest of the drawings were either 2D UIs with a 3D virtual environment background (Fig. 6, B) or 2D UIs with one 3D item in it. Also, the content of the drawings varied; 68% of them were based on our 3D GUI designs. The participants often combined our ideas, such as they drew a carousel with shelves. Also, there were other 3D objects combined with carousel GUIs, for example, in one drawing, there was a ball menu in the second hierarchical level of a *Keyring* GUI (Fig. 6, D). Most of the drawings (57%) included different applications and layout needs for the UI. In 38% of the drawings, there were rotatable UI elements presented, for example, in one drawing the whole GUI was rotatable (Fig. 6, C). Also in 15% of the drawings, there were some kinds of hierarchical structures showed, for example, Fig. 6, C and D.



Figure 5. A participant drawing on the template.

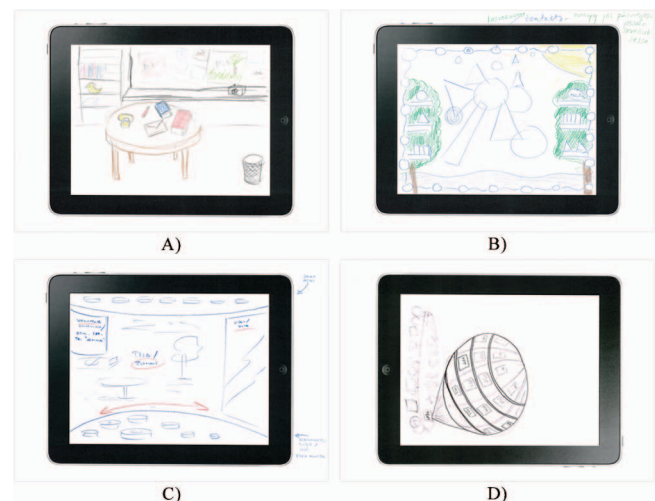


Figure 6. Participants' drawings A-D on the Self-Expression Templates.

VI. FINAL UX BASED DESIGN

Based on the participants' experiences on our designs and their own drawings in the self-expression task, we made design iteration for our earlier 3D GUIs. As the aim of UX studies is to help in selecting the best design solution and make sure that the development is on the right track [47], we made changes that we thought to be best for each GUI. One big change needed based on the participants' comments on the shown four GUIs was the need for a flat hierarchy structure. We needed to solve how we could display a lot of sub items in the *Pie* and *Keyring* concepts without hierarchical structures. In the following sub-sections, we go through all four 3D GUIs and how we iterated their design.

A. The Final Room GUI

The *Room* metaphor was iterated a little to make it more approachable and easier to use. Therefore, the floor plan was changed from a square to a circle to make it easier to see the content on the shelves and ease touch interaction with the content (Fig. 7, on the left). Also, we increased the view area to the virtual environment. Likewise, we decided to ease the interaction with the copied file by providing possible target objects beside the folder when the copy has been made (Fig. 7, on the right). Also, when files are browsed in the center of the screen, the view to the VE is dimmed, because we wanted to increase security and prevent accidental sharing to the VE. The dimming is automatically removed when a user stops browsing items in the center of the screen. If a user needs to share items from his/her private GUI to the virtual environment during the dimming effect is used, he/she just needs to drag an item on the dimmed area and then the visual indication is showed, a flash of light in the dimmed area, and then the item moves through the dimming effect to the virtual environment. We added a possibility to rotate the view with a swipe gesture with three fingers. We also made it possible to hide the GUI totally from the view by swiping with four fingers towards the screen corners and by doing vice versa to make it come back in sight.



Figure 7. The finalized design of *Room* GUI. On the left start view and on the right target objects (trashcan and contact balls) provided when a copy has been made.

B. The Final Shelves GUI

The *Shelves* GUI was not perceived as a 3D GUI in the user evaluation, thus we needed to make it to look more 3D. Therefore, we decided to add a little depth in it. Therefore, we rotated shelves according to z-axis. We also add the translucent white back plate to the shelves to make it distinct more on the background and make it also more secure to use,

e.g., users would not have to be worried about that they accidentally share their private items to the public virtual environment (Fig. 8, on the left). We did not want to change the possibility to pull just one shelf at sight, because we believe that works best in touch screen context. Also, then the GUI will not occlude too much with the view to the virtual environment. Also, with *Shelves* GUI, we decided to ease the interaction with the copied file by providing possible target objects beside the folder when the copy has been made (Fig. 8, on the right). Also, when files are browsed in the center of the screen, the view to the VE is dimmed, because we wanted to increase security and prevent accidental sharing to the VE. Also in *Shelves* GUI it is possible to share items to the VE even if the dimming effect is used. This was designed in the same way as in the *Room* GUI. As in the *Room* GUI, we made it possible to hide the *Shelves* GUI totally from the view by swiping with four fingers towards the screen corners and by doing vice versa to make it come back in sight.



Figure 8. The finalized design of *Shelves* GUI. On the left start view and on the right target objects (trashcan and contact balls) provided when a copy has been made.

C. The Final Pie GUI

The *Pie* GUI needed quite a lot of changes. First, we removed all the hierarchical helical stairs structures. In replacement for those, we decided to use drawers. The way how content is presented in the drawer depends on the content. For example, for files, there can be a file cabinet drawer (Fig. 9, on the right), and for games, there can be a normal drawer where the games are laying. To limit the occlusion of the VE by the GUI, we made the *Pie*'s plate less thick. We also made it possible to move the GUI's place and scale the size of it (Fig. 9, on the left). Also, to avoid accidental sharing of items in the rear of the GUI and to make GUI feel more private, we decided to add a dimming effect to the background when a user is interacting with the personal GUI (Fig. 9, on the right). The dimming effect on the *Pie* GUI is used when a user is interacting with the VE (Fig. 9, on the left). Moving between the GUI and the VE is designed to be easy by long pressing the dimmed UI to make it active. Sharing items into the VE was designed to be possible even if it is dimmed. If a user needs to share something from his/her private GUI to the virtual environment, he/she needs to drag an item on the dimmed area and then the visual indication is shown, a flash of light in the dimmed area, and then the item moves through the dimming effect to the virtual environment.



Figure 9. The finalized design of *Pie* GUI. On the left a view when user's focus is on the VE. On the right a view when user's focus is in the *Pie* and files.

D. The Final Keyring GUI

For the *Keyring* GUI, we removed all the hierarchical structures. Instead of presenting subrings, we decided to use different ways for presenting hierarchical information. For example, we used a similar folder for the files as in the *Room* and *Shelves* GUIs (Fig. 10, on the right). We decided to use a similar dimming effect as in the *Pie* also in the *Keyring*, as it was difficult to make it feel a private GUI without heavy additional structures (Fig. 10). Also, sharing through the dimming effect was designed in the same way as in the *Pie* GUI. To prevent the *Keyring* blockin the view to the VE, we made it possible to move the GUI's place and scale the size of it.

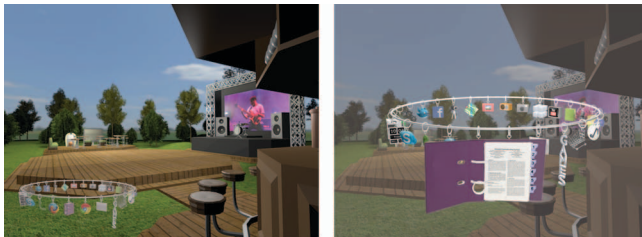


Figure 10. The finalized design of the *Keyring* GUI. On the left a view when user's focus is on the VE. On the right a view when user's focus is in the *Keyring* and files.

VII. DISCUSSION

In this study, we focused on designing 3D GUIs from the user experience point of view. We made a large design process which was focused on designing of personal 3D GUI in a virtual environment. As the outcome of the process, we had four 3D GUIs for the user evaluation. In the concept evaluation, we had also several other tasks and as a final task we had the self-expression template task. Our study provided a lot of user experience based information for the next iterative design phase of the four concepts, but in addition, it helped us to save a large amount of implementing costs and time.

The study indicated that the visual design of the GUI has an impact on the user's experience with the 3D GUIs on public VEs from privacy perspectives. Therefore, when designing 3D GUIs which are used in parallel with collaborative VEs, it is important to create a secure and private look for the 3D GUI. The *Room* GUI was perceived to be the most secure, because it had walls separating it from the background VE. Participants wished for visual elements,

such as walls or curtains, which will distinguish the GUI and its elements from the background environment. However, these elements should not excessively occlude the virtual environment; therefore, they could be also translucent. For the final design, we used a translucent dimming effect for the *Pie* and *Keyring* GUIs and a translucent back plate solution for the *Shelves*. By these solutions, in our opinion these three GUIs are now looking more secure and private.

The study indicated that there should be a clear visual indication for showing the user's active position between the personal 3D GUI and a collaborative virtual environment. Also, when something is transferred from the personal GUI to the public virtual environment, there should be a visual indication shown to the user. In the final iterated designs, we used a dimming of the background or UI for showing the active position and a highlight color for indicating object moving from a private GUI area to a public VE. By using the dimming effect, we believe we were also able to reduce the influence of the background space to the visual design of the 3D GUI. This was because subjects thought that all GUIs except the *Room* looked weird and distressing with the VE background, because they seem to be floating in the air and were unclear looking. With the dimming effect, they do not look anymore as a part of the 3D scene; therefore, the visual weirdness is not that big issue anymore.

From the participants' perspectives, 3Dness in a 3D GUI is made from icon occlusion, the shape of the UI and the depth of the UI space. All other GUIs except *Shelves* were perceived as 3D. Therefore, we needed to add depth to the *Shelves* GUI to make it look more 3D.

The hierarchy structure does not have to continue similarly through the hierarchy levels; it is more preferable to use flat hierarchy on touch screen devices. Therefore, we replaced deep hierarchical structures from the *Pie* and *Keyring* to the more flat solutions, for example in the *Pie* we used different kinds of drawers.

The study elicited also that users need to have a possibility to organize icons, UI elements and decorate the GUI space as they wish. Therefore, customization is an important aspect of the user experience of the 3D GUI. The shape of the GUI and amount of the icons depend on the user's personal preferences. One user wishes to see all of the applications at one glance and another one would just want to see only the most important applications in their GUI. Therefore, there should be different kinds of GUI designs available to the users.

Even though the procedure of the whole evaluation was very large and included several examples, the topic was interesting for the participants and the rhythm of the evaluation procedure was balanced with the tasks. Therefore, participants were not exhausted after the session. Instead, they were surprised how fast the time went and how much fun they had. Because we studied 3D UIs from several perspectives, it was very important to use a mixed methods procedure with different types of concept examples. Likewise, it was critical that the participants used the drawing template in the final task because they were able to use all showed examples as a source of inspiration for the drawing. Also, this case elicited the fact that a pair

evaluation setting is a good setup for experience elicitation and sharing, because the participant is expressing his/her experiences, wishes and ideas to the other participant (e.g., a friend or a colleague) and not only to the researcher.

Even though this evaluation had limitations from the interaction point of view, it nevertheless provided useful information to us for the next iteration of the four concepts. Although it was not possible to evaluate touch screen interaction with a non-operational prototype, it was possible to show the designed interaction ways and discuss about them with the participants. If we would have not done this early phase user evaluation, we would not have gained this valuable information, nor would we been able to update our designs to the next level. As prior research have shown, it is important to evaluate user experience in different phases of the development [47][4][46]. Also, the meaning of the visually high quality evaluation materials cannot be undervalued. As prior research has shown, for the successful evaluation, the designer need to be able to present his/her designs in the way that participants can understand them and are able to discuss and give a feedback about them [31]. Our study confirms this finding, but it also indicated that the use of visually high quality evaluation material can reveal big problems with visual, interaction and usability aspects, before any implementing and development hours are spend.

VIII. CONCLUSION AND FUTURE WORK

In this article, we focused on a possibility to have a personal 3D GUI inside a VE on a tablet device. We started with the visual design process of 3D GUIs with 3D icons and 'file searching and sharing' use cases. Then, we had an early design phase UX evaluation of four 3D GUIs in a virtual environment background with 40 participants. In the evaluation, we used non-functional visually high quality prototypes and a Self-Expression Template to involve participants in the design of 3D GUIs. This evaluation provided a lot of user feedback for the design, which we utilized in the final iterated designs. In addition, we pointed out many design issues relating to the visual design of the personal GUI in VEs in a touch screen tablet context.

We found that the participants liked the possibility of having their personal 3D GUI in a virtual environment. However, they wished that the visual design of the 3D GUI should create a secure and private feeling for them. In this case, the *Room* concept was perceived as the most secure, because it had walls separating it from the background space. For creating as a secure feeling also to other GUIs, participants wished for visual distinction of the GUI and its elements from the background VE. However, participants wished that the GUI should not occlude too much with the background VE. Therefore, we included in the final iterated designs either a translucent back plate or with a dimming effect of the background when a user is interacting with the GUI and vice versa. Also, when items are moved from the personal GUI to the virtual environment, participants wished for some kind of visual indication. Similarly, visual indication should be used for showing the user's active position between the GUI and VE. Therefore, we included in the final iterated design a flash of the dimmed area when a

user moves an item through it. In a touch screen tablet context, the participants found deep hierarchical structures distressing and difficult; therefore, they were replaced with more flat hierarchical solutions in the iterated designs.

In future studies, it would be interesting to evaluate the final GUI designs as fully functional prototypes to find out how participants perceive them and test user interaction as well. Especially we need more information about the animations of the final GUIs, e.g., how items will behave when a user is interacting with them. In addition, the dimming effect and indication while moving items between UI spaces need to be studied more.

IX. ACKNOWLEDGMENTS

This work was carried out in the Intel and Nokia Joint Innovation Center at the Center for Internet Excellence unit in the University of Oulu. We would like to thank our funders Intel, Nokia and Tekes. We want to also express our gratitude to Meiju Sunnari for taking part in the 1st and 2nd design phases and Julianna Hemmoraanta for 3D modeling of virtual environment, icons and GUIs for the user evaluation. Also warm thanks to our evaluation participants.

REFERENCES

- [1] M. Pakanen, L. Arhippainen, and S. Hickey, "Studying four 3D GUI metaphors in virtual environment in tablet context - Visual design and early phase user experience evaluation," Proc. Advances in Computer-Human Interactions (ACHI' 13), ThinkMind Press, March 2013, pp. 41–46.
- [2] A. Agarawala and R. Balakrishnan, "Keepin' it real: pushing the desktop metaphor with physics, Piles and the Pen," Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI'06), ACM Press, 2006, pp. 1283–1292, doi:10.1145/1124772.1124965.
- [3] Apple iPad bookshelf: <http://tinyurl.com/95hdvcz>. 10.9.2013.
- [4] L. Arhippainen, Studying user experience: issues and problems of mobile services – Case ADAMOS: User experience (im)possible to catch? Oulu, Finland, Oulu University Press, 2009.
- [5] L. Arhippainen, M. Pakanen, S. Hickey, and P. Mattila, "User experiences of 3D virtual learning environment," Proc. Academic MindTrek Conference (MindTrek'11), ACM Press, 2011, pp. 222–227, doi:10.1145/2181037.2181075.
- [6] L. Arhippainen, M. Pakanen, and S. Hickey, "Designing 3D virtual music club spaces by utilizing mixed UX methods: from sketches to Self-Expression Method," Proc. Academic MindTrek Conference (MindTrek'12), ACM Press, 2012, pp. 178–184, doi:10.1145/2393132.2393167.
- [7] L. Arhippainen, M. Pakanen and S. Hickey, "Studying depth in a 3D user interface by a paper prototype as a part of the mixed methods evaluation procedure. Early phase user experience study," Proc. Advances in Computer-Human Interactions (ACHI' 13), ThinkMind Press, March 2013, pp. 35–40.
- [8] L. Arhippainen and M. Pakanen, "Utilizing Self-Expression Template Method in user interface design - Three Design Cases," Proc. Academic MindTrek Conference (MindTrek'13), ACM Press, 2013, pp. 80–86, doi:10.1145/2523429.2523477.
- [9] H. Beyer and K. Holtzblatt, Contextual design: defining customer-centered systems. San Francisco, CA, USA, Morgan Kaufmann Publishers Inc, 1998.
- [10] D. A. Bowman, J. Chen, C.A. Wingrave, J. Lucas, A. Ray, N.F. Polys, Q. Li, Y. Haciahetoglu, J. Kim, S. Kim, R.

- Boehringer, and T. Ni, "New directions in 3D user interfaces," *The International Journal of Virtual Reality*, vol 5, no 2 2006, pp. 3–14.
- [11] R. Budi and N. Nielsen, "Usability of iPad apps and websites," Fremont, CA: Norman Nielsen Group. http://www.nngroup.com/reports/mobile/ipad/ipad-usability_2nd-edition.pdf (2011). 13.9.2013.
- [12] A. Butz, C. Beshers, and S. Feiner, "Of vampire mirrors and privacy lamps. Privacy management in multi-user augmented environments," *Proc. Symp. User interface software and technology (UIST '98)*, ACM Press, 1998, pp. 171–172, doi:10.1145/288392.288598.
- [13] S. Card, G. G. Robertson, and W. York, "The WebBook and the Web Forager: an information workspace for the World-Wide Web," *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI '96)*, ACM Press, 1996, pp. 111–117, doi:10.1145/238386.238446.
- [14] Z. Cipiloglu, A. Bulbul, and T. Capin, "A framework for enhancing depth perception in computer graphics," *Proc. Symp. Applied Perception in Graphics and Visualization (APGV'10)*, ACM Press, 2010, pp. 141–148, doi:10.1145/1836248.1836276.
- [15] A. Cockburn and B. McKenzie, "3D or not 3D? Evaluating the effect of the third dimension in a document management system," *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*, ACM Press (2001), 434–441, doi:10.1145/365024.365309.
- [16] M. J. Culnan, "Protecting privacy online: Is self regulation working?," *Journal of Public Policy and Marketing* vol 19, no 1, Spring 2000, pp. 20–26, <http://www.jstor.org/stable/30000484>.
- [17] R. Dachsel and A. Hubner, "Three-dimensional menus: a survey and taxonomy," *Computers & Graphics*, vol 31, 2007, pp. 53–65, <http://dx.doi.org/10.1016/j.cag.2006.09.006>.
- [18] A. De Angeli, A. Sutcliffe, and J. Hartmann, "Interaction, usability and aesthetics: what influences users' preferences?," *Proc. Designing Interactive systems (DIS'06)*, ACM Press, 2006, pp. 271–280, doi:10.1145/1142405.1142446.
- [19] A. Gotchev, G. B. Akar, T. Capin, D. Strohmeier, and A. Boev, "Three-dimensional media for mobile devices," *Proc. IEEE* vol 99, no 4, March 2011, pp. 708–741, doi:10.1109/JPROC.2010.2103290.
- [20] M. Hancock, T. ten Cate, and S. Carpendale, "Sticky Tools: full 6DOF force-based interaction for multi-touch tables," *Proc. Interactive Tabletops and Surfaces (ITS '09)*, ACM Press, 2009, pp. 133–140, doi:10.1145/1731903.1731930.
- [21] C. A. Hand, "Survey of 3D interaction techniques," *Computer Graphics Forum*, vol 16, no 5, December 1997, pp. 269–281, doi:10.1111/1467-8659.00194.
- [22] ISO 9241-210:2010, Ergonomics of human system interaction - Part 210: Human-centred design for interactive systems. International Standardization Organization. Switzerland.
- [23] G. Jacucci, A. Morrison, G. Richard, J. Kleimola, P. Peltonen, L. Parisi, and T. Laitinen, "Worlds of information: designing for engagement at a public multi-touch display," *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*, ACM Press, 2010, pp. 2267–2276, doi:10.1145/1753326.1753669.
- [24] M. Kynsilehto and T. Olsson, "Checkpoints, hotspots and standalones - placing smart services over time and place," *Proc. Nordic Conference on Human-Computer Interaction (NordiCHI'12)*, ACM Press, 2012, pp. 209–218, doi:10.1145/2399016.2399049.
- [25] A. Leal, C. A. Wingrave, and J. J. LaViola, "Initial explorations into the user experience of 3D file browsing," *Proc. British HCI Group Annual Conference on People and Computers (BCS-HCI'09)*, ACM Press, 2009, pp. 339–344.
- [26] C. Y. Lee and M. Warren, "Security issues within virtual worlds such as Second Life," *Proc. Australian Information Security Management Conference, Research online*, 2007, pp. 142–151.
- [27] J. Light and J.D. Miller, "Miramar: a 3D workplace," *Proc. Professional Communication Conference (IPCC'02)*, IEEE Press, 2002, pp. 271–282, doi:10.1109/IPCC.2002.1049110.
- [28] Linden Lab. Second Life. <http://secondlife.com/>. 13.9.2013.
- [29] A. Lucero, "Framing, aligning, paradoxing, abstracting, and directing: how design mood boards work," *Proc. Designing Interactive Systems Conference DIS'12*. ACM Press, 2012, pp. 438–447, doi:10.1145/2317956.2318021.
- [30] Lumiya viewer for Second Life: <http://www.lumiyaviewer.com/>. 13.9.2013.
- [31] J. Löwgren and E. Stolterman, *Thoughtful interaction design. A design perspective on information technology*. Massachusetts, USA, The MIT Press, 2007.
- [32] D. A. Norman. *Emotional Design: Why we Love (or Hate) Everyday Things*. New York, USA, Basic Books, 2004.
- [33] Order and Chaos: <http://orderchaosonline.com/>. 13.9.2013.
- [34] M. Pakanen, L. Arhippainen, and S. Hickey, "Design and evaluation of icons for 3D GUI on tablets," *Proc. Academic MindTrek Conference (MindTrek'12)*, ACM Press, 2012, pp. 203–206, doi:10.1145/2393132.2393171.
- [35] Patio test user forum: <http://www.patiolla.fi/en>. 13.9.2013.
- [36] D. Patterson, "3D SPACE: using depth and movement for selection tasks," *Proc. 3D web technology (Web3D'07)*, ACM Press, 2007, pp. 147–155, doi:10.1145/1229390.1229416.
- [37] G. Robertson, J. Mackinlay, and S. K. Card, "Cone Trees: animated 3D visualizations of hierarchical information," *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI '91)*, ACM Press, 1991, pp. 189–194, doi:10.1145/108844.108883.
- [38] K. Salo, L. Arhippainen, and S. Hickey, "Design guidelines for hybrid 2D/3D user interfaces on tablet devices – a user experience evaluation," *Proc. Advances in Computer-Human Interactions (ACHI'12)*, ThinkMind Press, 2012, pp.180–185.
- [39] R.-A. Shang, Y.-C. Chen, and S.-C. Huang, "A private versus a public space: Anonymity and buying decorative symbolic goods for avatars in virtual world," *Computers in Human Behavior*, vol 28, issue 6, November 2012, pp. 2227–2235, <http://dx.doi.org/10.1016/j.chb.2012.06.030>.
- [40] E. B.-N. Sanders, "Virtuosos of the experience domain," <http://www.maketools.com/papers-2.html>. 13.9.2013.
- [41] L. Staples, "Representation in virtual space: visual convention in the graphical user interface," *Proc. INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (INTERCHI '93)*, ACM Press, 1993, pp. 348–354, doi:10.1145/169059.169279.
- [42] O. Ståhl, A. Wallberg, J. Söderberg, J. Humble, L. Fahlén, A. Bullock, and J. Lundberg, "Information exploration using the Pond," *Proc. Collaborative virtual environments (CVE '02)*, ACM Press, 2002, pp. 72–79, doi:10.1145/571878.571890.
- [43] M. Tähti and L. Arhippainen, "A Proposal of collecting emotions and experiences," *Proc. HCI 2004*, vol. 2. 6-10.9. 2004.
- [44] K. Ulrich and S. Eppinger, *Product Design and Development*. New York, USA, McGraw-Hill, 2008.
- [45] A. van Dam, A. S. Forsberg, D. H. Laidlaw, J. J. Jr. LaViola., and R. M. Simpson, "Immersive VR for scientific visualization: a progress report," *Computer Graphics and Applications*, vol 20, no 6, November/December 2000, pp. 26–52, doi:10.1109/38.888006.
- [46] A. Vermeeren, E. Law, V. Roto, M. Obrist, J. Hoonhout, and K. Väänänen-Vainio-Mattila, "User experience evaluation

- methods: current state and development needs,” Proc. Nordic Conference on Human-Computer Interaction (NordiCHI’10), ACM Press, 2010, pp. 521–530, doi:10.1145/1868914.1868973.
- [47] K. Väänänen-Vainio-Mattila and M. Wäljas, “Developing an expert evaluation method for user eXperience of cross-platform web services,” Proc. MindTrek Conference (MindTrek’09). ACM Press, 2009, pp. 162–169, doi:10.1145/1621841.1621871.
- [48] S. Wang, M. Poturalski, and D. Vronay, “Designing a generalized 3D carousel view,” Proc. Extended Abstracts on Human Factors in Computing Systems (CHI EA’05), ACM Press, 2005, pp. 2017–2020, doi:10.1145/1056808.1057081.
- [49] B. Zaman, Y. Poels, N. Sulmon, J.-H. Annema, M. Verstraete, F. Cornillie, D. De Grooff, and P. Desmet, “Concepts and mechanics for educational mini-games. A human-centred conceptual design approach involving adolescent learners and domain experts,” *International Journal On Advances in Intelligent Systems*, vol 5, no 3 & 4, 2012, pp. 567– 576.

Constructing Autonomous Systems: Major Development Phases

Nikola Šerbedžija
Fraunhofer FOKUS,
Berlin, Germany,

Email: nikola.serbedzija@fokus.fraunhofer.de

Annabelle Klarl, Philip Mayer
Ludwig-Maximilians-Universität München
Munich, Germany

Email: {klarl|mayer}@pst.ifi.lmu.de

Abstract—Developing autonomous systems requires adaptable and context aware techniques. The approach described here decomposes a complex system into service components – functionally simple building blocks enriched with local knowledge attributes. The internal components’ knowledge is used to dynamically construct ensembles of service components. Thus, ensembles capture collective behavior by grouping service components in many-to-many manner, according to their communication and operational/functional requirements. To achieve such high level of dynamic behavior a complete development life cycle for ensemble based systems has been defined and supported by rigorous analyses and modeling methods, linguistic constructs and software tools. We focus here on the analysis, modeling, programming and deployment phases of the autonomous systems development life cycle. A strong pragmatic orientation of the approach is illustrated by two different application scenarios. The main result of this work is an integrated view on developing autonomous systems in diverse application domains.

Keywords—autonomous systems, component-based systems, context-aware systems

I. INTRODUCTION

Developing massively distributed systems has always been a grand challenge in software engineering [1], [2], [3], [4]. Incremental technology advances have continuously been followed by more and more requirements as distributed applications grew mature. Nowadays, one expects a massive number of nodes with highly autonomic behavior still having harmonized global utilization of the overall system. Our everyday life is dependent on new technology which poses extra requirements to already complex systems: we need reliable systems whose properties can be guaranteed; we expect systems to adapt to changing demands over a long operational time and to optimize their energy consumption [5], [6].

One engineering response to these challenges is to structure software intensive systems in ensembles featuring autonomous and self-aware behavior [7], [8]. The major objective of the approach is to provide formalisms, linguistic constructs and programming tools featuring autonomous and adaptive behavior based on awareness. Furthermore, making technical systems aware of their energy consumption contributes significantly to ecological requirements, namely to save energy and increase overall system utilization.

The focus here is to integrate the functional, operational and energy awareness into the systems providing autonomous functioning with reduced energy consumption. The rationale, expressing power and practical value of the approach are

illustrated on the e-mobility and cloud computing application domains. The two complex domains appear to be fairly different. However, taking a closer look at the requirements of the two scenarios it becomes noticeable that the problem domains share numerous generic system properties, especially when seen from the optimized control perspective.

The paper presents integrative work focusing on the development lifecycle of complex distributed control systems. It binds together methods and techniques to model and construct systems with service components and ensembles. The rationale and development lifecycle (Sections II) of the approach is presented through close requirements analysis (Section III), ensemble modeling (Section IV) and programming/deployment (Sections V and VI) of two concrete application scenarios. A strong pragmatic orientation as well as the general nature of the approach is shown on two different case studies. Finally, the approach is summarized giving further directions for the work to come in Section VII.

II. DEVELOPMENT LIFECYCLE

The engineering of autonomous systems includes all of the challenges of non-autonomous complex systems plus the added problem of achieving self-* properties allowing for autonomy. An autonomous system needs to be self-aware and self-adaptive. That means it has to maintain the knowledge of its own functional and operational requirements and it should be capable of performing appropriate changes without human intervention. To implement such behavior, a number of feedback loops within the system are needed, to deal with changes both in the controlled environment and the system per se.

The method to develop autonomous systems thus needs to focus more on the runtime side than traditional engineering approaches, as both outside changes and system adaptive responses happen in operation (i.e., live). Within this approach, the iterative development processes that are standard in industry today have been extended to include two main loops; one focuses on the design time and one on the runtime of the system. Both loops are connected to allow feedback. The resulting Ensemble Development Life Cycle (EDLC) is shown in Figure 1. The first loop is the design loop which begins with the analysis of requirements, continues on to the modeling and programming phase, and finally to the verification and validation of the system. This loop runs iteratively until the result is satisfying. A connecting arrow which corresponds to the deployment of the system leads to the second loop which

represents the runtime of the system. Usually, an ensemble-based system stays within this loop once deployed, using a feedback loop to achieve adaptation. This loop consists of a monitoring phase (both of the environment and itself), awareness built on the monitored information, and finally self-adaptation which leads back to monitoring.

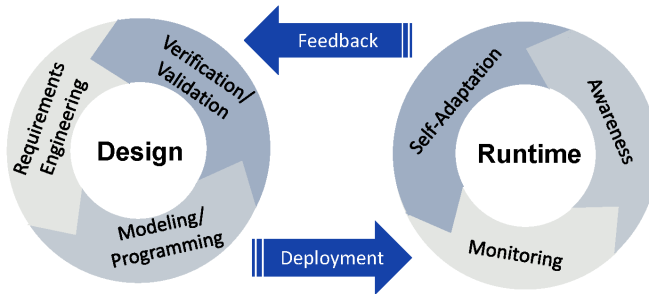


Figure 1. The Ensemble Development Life Cycle

The feedback arrow back to the design phase has two functions. First, feedback of normal system situations can be fed back to design to tweak the system or as input for a next version. Second, if a critical system encounters a non-adaptable situation, the feedback can be given immediately to human operators with the ability to reconfigure.

For each of the phases and transitions within the development life cycle a number of tools and methodologies have been developed, allowing for a formal and rigorous development of complex distributed autonomous systems [9].

Within this paper, we focus on two practical applications which are illustrated on the requirements and modeling phases, followed by programming and deployment. The other phases (validation and verification, awareness, monitoring, and self-adaptation) and the feedback transition are beyond the scope of this paper.

III. REQUIREMENTS

To explore the system requirements, two complex application domains are closely examined: e-mobility control and cloud computing.

E-mobility is a vision of future transportation by means of an electric vehicles network allowing people to fulfill their individual mobility needs in an environmental friendly manner (decreasing pollution, saving energy, sharing vehicles, etc).

Cloud computing is an approach that delivers computing resources to users in a service-based manner, over the Internet, thus reinforcing sharing and reducing energy consumption).

At a first glance electric vehicular transportation and distributed computing on demand have nothing really in common!

A. Common Characteristics

In a closer examination the two systems, though very different, have a number of common characteristics.

1) *Massive Distribution and Individual Interest:* E-mobility deals with managing a huge number of e-vehicles that transport people from one place to another taking into account numerous restrictions that the electrical transportation means imposes.

Each cloud computing user has also his/her individual application demands and interest to efficiently execute it on the cloud. The goal of cloud computing is to satisfy all these competing demands.

Both applications are characterized by a huge number of single entities with individual goals.

2) *Sharing and Collectiveness:* In order to cover longer distances, an e-vehicle driver must interrupt the journey to either exchange or re-charge the battery. Energy consumption has been the major obstacle in a wider use of electric vehicles. An alternative strategy is to share e-vehicles in a way that optimizes the overall mobility of people and the spending of energy. In other words: when my battery is empty – you will take me further if we go in the same direction and vice versa.

The processing statistics show that most of the time computers are idle – waiting for input to do some calculations. Computers belong amongst the fastest yet most wasteful devices man has ever made. And they dissipate energy too. Cloud computing overcomes that problem by sharing computer resources making them better utilized. In another words, if my computer is free – it can process your data and vice versa; or even better, let us have light devices and leave a heavy work for the cloud.

At a closer look “sharing and collectiveness” are common characteristics of both application domains!

3) *Awareness and Knowledge:* E-mobility can support coordination only if e-vehicles know their own restrictions (battery state), destinations of users, re-charging possibilities, parking availabilities, the state of other e-vehicles nearby. With such knowledge, collective behavior may take place, respecting individual goals, energy consumption and environmental requirements. Cloud computing deals with the dynamic scheduling of available computing resources within a wider distributed system. Maximal utilization can only be achieved if the cloud is “aware” of the users’ processing needs and the states of the deployed cloud resources. Only with such knowledge a cloud can make a good utilization of computers while serving individual users’ needs.

At a closer look “awareness” of own potentials, restrictions and goals as well as those of the others is a common characteristic. Both domains require self-aware, self-expressive and self-adaptive behavior based on a knowledge about those “self*” properties.

4) *Dynamic and Distributed Energy Optimization:* E-mobility is based on a distributed network that manages numerous independent and separate entities such as e-vehicles, parking slots, re-charge stations, and drivers. Through a collective and awareness-rich control strategy the system may dynamically re-organize and optimize the use of energy while satisfying users’ transportation needs.

Cloud computing actually behaves as a classical distributed operating system with the goal of maximizing operation and

throughput and minimize energy consumption, performing tasks of multiple users.

At a closer look “dynamic and distributed optimization” is an inherent characteristic of the control environment for both application domains.

B. Common Approach

This set of common features serve as a basis for analysis and modeling of such systems leading to a generic framework for developing and deploying complex autonomous systems (Table I). Respecting these characteristics in constructing such systems will help meeting the common requirements and provide major behavioral principles: adaptation, self-awareness, knowledge, and emergence. These principles are actually very close and inter-related: namely knowledge is needed for awareness which is further needed for adaptation which further leads to emergent behavior.

Table I. COMMON CHARACTERISTICS

| Common feature | Cloud computing | E-Mobility |
|------------------------------------|---|---|
| Single entity | Computing resource | Car, passenger, parking lot, charging station |
| Individual goal | Efficient execution | Individual route plan |
| Ensemble | Application, CPU pool | Free vehicles, free parking lots, etc. |
| Global goal | Resource availability, optimal throughput | Travel, journey, low energy |
| Self-awareness | Available resources, computational requirements, etc. | Awareness of own state and restrictions |
| Autonomous and collective behavior | Decentralized decision making, global optimization | Reaching all destinations in time, minimizing costs |
| Optimization | Availability, computation task, execution | Reaching destination in time, vehicle/infrastructure usage |
| Adaptation | According to available resources | According to traffic, individual goals, infrastructure, resource availability |
| Robustness | Failing resources | Range limitation, charging battery, traffic resources |

In this approach the adaptation is modeled as progress in a multi-dimensional space where each axis represents one aspect of system awareness (knowledge about its own functional, operational, or other states). Adaptation actually happens when the system state moves from one to another position within the space according to the pre- and post-condition on each of its awareness dimensions. This adaptation model is called SOTA (State Of The Affairs) [10].

The trajectory of an entity in the SOTA space is illustrated in Figure 2. Defining certain states in the SOTA space as desired goal states helps to understand and model goal-directed behavior of entities. We determine three kinds of (sets of) states: pre-conditions G^{pre} , utilities U , and post-conditions G^{post} . Starting from the desired state, G^{post} is the goal state that the entity should reach. This goal is only activated when the entity moves to the state G^{pre} so that it now has to try to move to G^{post} . On the way through the state space, the entity may have to adhere to specific constraints which are called utilities U .

The SOTA adaptation model is used to extract major application requirements and offers appropriate adaptation patterns

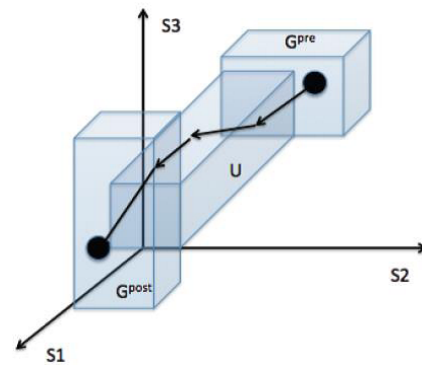


Figure 2. SOTA Adaptation Model

that effectively compose the system into numerous adaptation loops and guarantees the required behavior at run-time (more details on SOTA approach are described elsewhere, see [8] and [10]).

IV. ENSEMBLE MODELING

Control systems for both the cloud computing and e-mobility domains share the idea of groups of entities collaborating towards specific goals. Those groups are formed dynamically while each group exhibits a collective and goal-directed behavior on the basis of complex interactions between members of the group. Well-known techniques in component-based engineering [11] are not enough to capture the particular characteristics of those highly dynamic and collaborative systems. Component-based modeling merely determines the architectural and dynamic properties of the underlying system while ensemble modeling focuses more on the cooperative features on top of the component-based models.

The HELENA approach [12] provides the formal foundations for rigorous ensemble modeling. Each group of entities collaborating towards a goal is abstractly modeled as an ensemble. The specific functionalities and interaction abilities in the ensemble are captured in roles played by the components, and connectors between those roles. The first distinguishing feature of the HELENA approach is the ability of components to dynamically adopt and give up roles. This feature supports adapting to changing conditions since appropriate components can contribute to the ensembles on demand. It also increases robustness since defective members can easily be replaced by new components taking over responsibility for an abandoned role. Lastly, it also helps to efficiently use resources since roles can be given up as soon as they are no longer needed. The second distinguishing feature is that components can adopt multiple roles in parallel so that they may play different roles concurrently in the same or different ensembles. Thus, the components of a single component-based system may take part in multiple collaborations playing task-specific roles in each group – or the other way round, that is multiple ensembles perform their tasks building on top of the same resources of the underlying system.

Figure 3 shows a snapshot of a component-based system on which two ensembles are imposed as described in the HELENA approach. The bottom level shows the pool of all components available in the system. They provide the core

capabilities which are commonly available for all tasks. The components can be of different types, for example in the e-mobility scenario the basic components are persons, cars, and parking lots. However, they can also be of the same type as in the cloud computing scenario. The two upper levels of Figure 3 show different ensembles. Each requires specific roles to collaborate (here, we just model two roles R and R'). Each role provides additional capabilities which are required while performing the role-specific task. When a component adopts a certain role – depicted by an arrow in Figure 3 – it also adopts the role-specific capabilities. For example, a person may adopt the role of the driver of a car and thus gain the ability to decide on the velocity of the car, while in the role of a passenger she can only hop on and off.

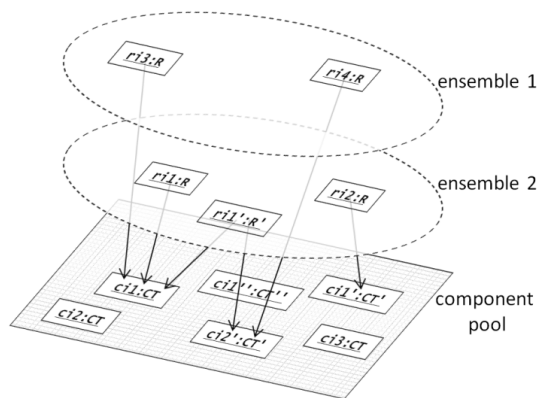


Figure 3. Ensembles in the HELENA Approach

At the same time, one component may participate in several ensembles playing the same or different roles. In each role, the component is equipped with particular properties and capabilities in addition to its core capabilities. Thus, these properties are the key features in the HELENA approach to enable task-oriented awareness. For example, adopting the role of the driver provides a person with the permission to retrieve the battery status of the vehicle. With this knowledge about the battery she can adapt her route accordingly.

Ensemble modeling on top of a component-based model is especially useful as a basis for subsequent development phases. Modeling with roles allows concentrating on the capabilities needed for a specific task. This increases coherence which leads to cleaner ensemble architectures. Furthermore, it decreases complexity of the models, thus providing a well-defined foundation for verification and validation as well as for detailed component-based designs implementing the required ensemble architecture.

Ongoing research in HELENA currently focuses on the derivation of component-based designs from ensemble architectures, description of ensemble behavior based on interacting roles, and checking goal satisfaction. For implementing ensemble architectures we investigate a systematic transition from HELENA models to abstract programming languages like SCCL [9] (cf. Section V-A).

The HELENA approach helps us to develop application scenarios based on top of a common basic component model. In both of the discussed application domains (e-mobility and

cloud computing), basic components provide the core capabilities such that we can build appropriate ensemble architectures for each scenario exploiting and enhancing the underlying model.

A. Modeling E-Mobility with Ensembles

In the e-mobility domain, persons, cars, and parking lots team up in ensembles to manage a fleet of cars serving travelers' needs. Scenarios range from journey planning and execution to management of the car park. For example, in the "journey scenario", we envisage an ensemble structure enabling a group of people to travel to (maybe different) destinations (cf. Figure 4).

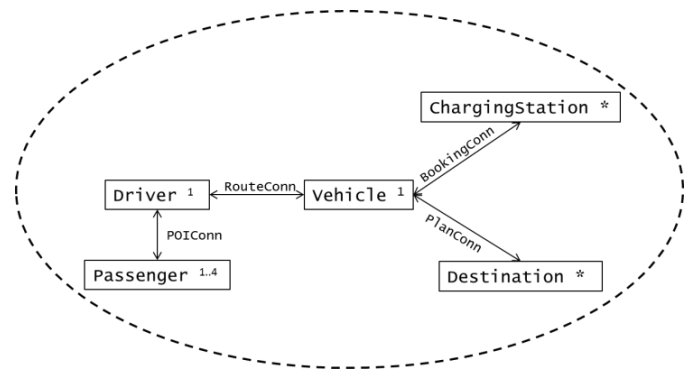


Figure 4. Ensemble Structure for the "Journey Scenario"

To implement the scenario, five roles of different multiplicities need to collaborate. The most important participants are the vehicle and the driver of the vehicle. They communicate with each other via the connector *RouteConn* to exchange, for example, destination requests and route points. It is crucial to recognize that the ensemble structure only mentions the particular roles that are needed for the collaboration and not the underlying components. The roles capture the role-specific capabilities; for example the need for a driver's license (*driver*) or passenger seats (*vehicle*). Particular components assume those roles, e.g., a person adopts the role of the driver, but in the case of self-driving cars a computer could also steer the vehicle.

Additionally, up to four passengers may join the collaboration who want to travel to some destinations and therefore communicate with the driver (via the connector *POIConn*) to announce their target locations. Usually, one of the passengers will also be the driver of the vehicle. This is where we benefit from the clear separation of roles and components. In the HELENA approach, one person is able to take different roles at the same time: on the one hand, she is a mere passenger just announcing her destination; on the other hand, she takes on responsibility for steering the vehicle. This dualism increases complexity when we try to model both responsibilities simultaneously in one component. Separating them into two roles as proposed in the HELENA approach facilitates the ensemble model by far. The collaboration is completed by an arbitrary number of destinations the passengers want to reach and an arbitrary number of charging stations needed to load the battery of the vehicle. Both roles can be taken by parking lots and are filled by appropriate parking lots on demand.

Note that other ensembles may be considered in parallel to the “journey ensemble”. For example, another ensemble may take care of relocating a car back to its home base after a one-way journey. This calls for an ensemble structure composed of different collaborating roles in comparison to the “journey ensemble”. However, even several instances of the same ensemble structure can run concurrently if we think of multiple journey groups each traveling to different destinations.

B. Modeling Cloud Computing with Ensembles

In a similar manner, computing entities may team up to provide a seamless platform of resources to users. We envision a cloud computing platform that allows executing applications in the cloud on a PaaS level. Managing the cloud system necessitates various collaborative tasks for distributed deployment and fail-safe execution which we want the system to perform in self-organizing teams. In Figure 5, the basic structure of an ensemble for the deployment of an application is given.

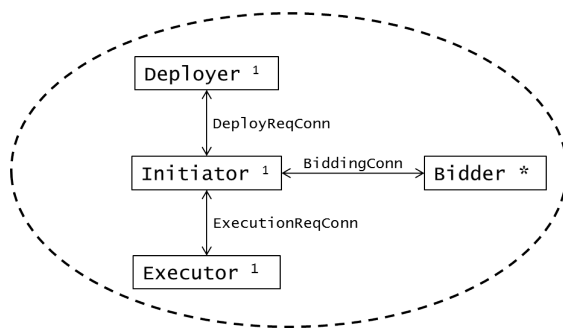


Figure 5. Ensemble Structure for the “Deployment Scenario”

In this scenario, a user wants to deploy an application in the cloud. As the user is outside the network, she needs to address her request to a deployer node inside the network which, from now on, serves as the origin of the request for the collaboration. The application comes with a set of requirements for the executing node such that the cloud has to search for an appropriate node. This search is managed by the initiator. The initiator has three responsibilities: it announces the application with its requirements for execution in the network, calls for bids from possible execution nodes, waits for bids and finally selects one node to serve as executor for the current application. The possible execution nodes – and therefore the bidders for execution – are a highly dynamic and application-specific set of nodes which cannot be determined from the available resources beforehand. This is where the notion of a role facilitates the description of collaboration. The role of a bidder abstractly defines that any node adopting this role has to meet the requirements of the application to be executed. By adopting the role, the node assumes the behavior of the bidder role such that it is only equipped with the appropriate bidding capabilities on demand. This applies to the role of the executor as well. The ensemble structure simply defines that an executor is needed for this task; the role of the executor is then filled dynamically at runtime, the chosen node adopting the appropriate behavior for execution of the application.

The above example of ensembles which take care of executing an application, shows that multiple ensembles of

the same structure can be run in parallel, sharing resources through different collaborations. For each application to be run in the cloud, a new collaboration needs to be established. However, nodes can join different ensembles at the same time. For example, a node can be initiator for one application while being the executor for another application. It can also adopt the same role twice in different ensembles, e.g., executing two different applications in two different ensembles. The same node can even take responsibility for two roles in the same ensemble like being initiator and executor at the same time.

V. E-MOBILITY DEPLOYMENT

Finding a way from high-level modeling to development and deployment of software intensive systems is a complex endeavor. Reasoning and validation often require high-level abstractions, while implementation calls for detailed programming and low-level deployments. To bridge this gap a number of intermediate tools are being developed that assist in the engineering process [8].

A. SCEL Language Programming Abstractions

The challenge for developers of complex distributed systems is to find proper linguistic abstractions to cope with individual vs. collective requirements of system elements and their need to respond to dynamic changes in an autonomous manner. A set of semantic constructs has been proposed [9], [13] that represent behaviors, knowledge and composition supporting programming of awareness-rich systems. It provides linguistic abstractions for describing the behavior of a single component as well as the formation of ensembles.

The basic ingredient of SCEL – Software Component Ensemble Language – is the notion of an autonomic component $\mathcal{I}[\mathcal{K}, \Pi, P]$ that consists of:

- An interface \mathcal{I} providing structural and behavioral information about the autonomic component in the form of attributes visible to other components.
- A knowledge repository \mathcal{K} managing information about the component interface, requirements, major state attributes etc. Managing such knowledge allows for self-aware behavior and dynamic interlinking with other system components.
- A set of policies Π which controls internal and external interaction.
- A set of processes P defining component functionality specific to the application and internal management of knowledge, policies, and communication.

The structure and organization of the SCEL notation is illustrated in Figure 6.

The code in Table II shows a fraction of the SCEL syntax (with notation for S - systems, C - components, P - processes, a - actions and c - targets); a fully detailed presentation of SCEL syntax and semantics can be found in [9], [13], [14].

SCEL aggregates both semantics and syntax power to express autonomic behavior. At one side, being abstract and rigorous SCEL allows for formal reasoning about system behavior; at another, it needs further programming tools to

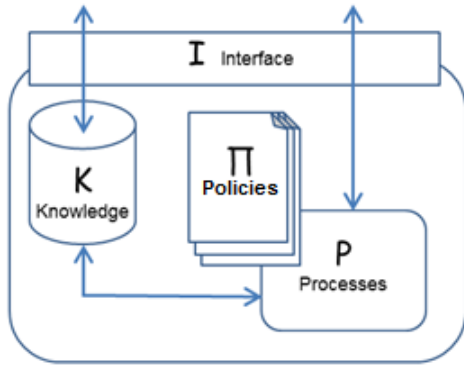


Figure 6. SCEL Elements

Table II. SCEL SYNTAX

| | |
|-------------|---|
| SYSTEMS: | $S ::= C \mid S_1 \parallel S_2 \mid (\nu n)S$ |
| COMPONENTS: | $C ::= \mathcal{I}[\mathcal{K}, \Pi, P]$ |
| PROCESSES: | $P ::= \text{nil} \mid a.P \mid P_1 + P_2 \mid P_1[P_2] \mid$ $X \mid A(\bar{p})$ |
| ACTIONS: | $a ::= \text{get}(T)@c \mid \text{qry}(T)@c \mid \text{put}(t)@c \mid$ $\text{fresh}(n) \mid \text{new}(\mathcal{I}, \mathcal{K}, \Pi, P)$ |
| TARGETS: | $c ::= n \mid x \mid \text{self} \mid P \mid p$ |

support system development and deployment. Formal reasoning, modeling and validation are covered in referenced articles about SCEL. Here, the focus is more on the pragmatic orientation on the given application scenario.

B. Java Framework for SCEL Programming and Model Checking

To execute SCEL programs, the jRESP framework has been developed. jRESP is a Java runtime environment providing means to develop autonomic and adaptive systems without any centralized control programmed in SCEL [15]. By relying on the jRESP API, a programmer can embed the SCEL paradigm in Java applications.

A prototypic statistical model-checker running on top of jRESP simulation environment has been implemented. Following this approach, a randomized algorithm is used to verify whether the implementation of a system satisfies a specific property with a certain degree of confidence. The statistical model-checker is parameterized with respect to a given tolerance t and error probability p . The used algorithm guarantees that the difference between the computed values and the exact ones is greater than t with a probability lower than p .

The model-checker included in jRESP can be used to verify reachability properties. These properties allow one to evaluate the probability to reach, within a given deadline, a configuration where a given predicate on collected data is satisfied [15].

C. Programming E-Mobility

Programming in jRESP is self-explaining and elegant. Considering the e-mobility application, we can easily program a vehicle supporting a user in her daily travel obligations. The vehicle is controlled by four modules:

```
1 VEHICLE =
2 ContactParkingLots[Planner[Book[MonitorPlanExecution]]]
```

Each of the modules has its own responsibility. The module ContactParkingLots retrieves all appointments of the passenger and searches for parking lots close to the points of interest. Afterwards, the module Planner plans a route how to reach all appointments in time using the available parking lots. Booking of the appropriate parking lots is done by the module Book. At last, the module MonitorPlanExecution monitors the current progress on the route and displays information about the reservation of the next parking lot on the route. The jRESP code (enriched with comments explaining the processes) is given in the following boxes.

```
1 ContactParkingLots =
2 //read the size of the calendar
3 //(i.e., the list of appointments)
4 qry("calendarSize", ?n)@self .
5 //scan the calendar
6 for(i := 0 ; i < n ; i ++){
7 //read an appointment of the calendar
8 qry("calendar", i, ?poi, ?poiPos, ?when, ?howLong)
9 @self .
10 //contact the parking lots near to the POI
11 //(this illustrates attribute-based communication
12 //typical in SCEL)
13 put("searchPlot", self, poi)
14 @{ I.type="Plot" & walkingDistance(poiPos, I.pos)} .
15 //ensemble predicate
16 }
17 //signal completion of the phase of data request
18 // from the parking lots
19 put("dataRequestSent")@self
```

```
1 Book =
2 //wait for the completion of the planning phase
3 get("planningCompleted")@self .
4 //read the size of plan list
5 //(i.e., the Plots to be booked)
6 get("planListSize", ?n)@self .
7 //scan the plan
8 for(i := 0 ; i < n ; i ++){
9 //read an entry of the plan list
10 get("plan", i, ?pLot, ?when, ?howLong)@self .
11 //send the booking request to the Plot
12 put("book", self, when, howLong)@pLot .
13 //wait for the reply of plot
14 // (we assume that booking requests always succeed)
15 get("bookingOutcome", true)@self .
16 //store the reservation in the list of reservations
17 put("reservation", i, pLot, when, howLong)@self .
18 }
19 //close the list of reservations
20 put("reservationListSize", n)@self .
21 //signal completion of the booking phase
22 put("bookingCompleted")@self
```



```

1 Planner =
2 //wait for the completion of the phase of
3 //requirement of data to the parking lots
4 get("dataRequestSent")@self .
5 //we intentionally leave this process unspecified
6 //input: collection of tuples of the form
7 // (poi,pLotId,pLotInfo) received from the PLOTS
8 //output: list of chosen planned PLOTS,
9 // i.e., (planListSize,n)
10 // (plan,0,pLotId0,when0,howLong0)...(plan,n-1,...)
11 //signal completion of the planning phase
12 put("planningCompleted")@self

```

```

1 MonitorPlanExecution =
2 //wait for the completion of the booking phase
3 get("bookingCompleted")@self .
4 //read the size of the reservation list
5 // (i.e., the PLOTS to be visited)
6 get("reservationListSize", ?n)@self .
7 //scan the reservation list
8 for(i := 0 ; i < n ; i ++){
9 //read a reservation
10 get("reservation", i, ?pLot, ?when, ?howLong)@self .
11 //display the info about the next reservation
12 //to the user
13 put("reservation", ?pLot, ?when, ?howLong)@screen .
14 //wait for the arrival at the parking lot
15 // (signaled by the user)
16 get("arrivedAt", pLot)@self .
17 }
18 //signal completion of the plan execution phase
19 put ("planExecuted")@self

```

The jRESP processes illustrate the expressive power of the language to cope with huge systems with complex interactions. A distinguishing feature of SCEL which is directly implemented in jRESP is implicit ensemble building by attributed-based communication. For example, while searching for parking lots close to a particular point of interest in the module ContactParkingLots, an ensemble of appropriate parking lots is implicitly formed by using the following predicate:

```

1 I.type="PLot" & walkingDistance(poiPos,I.pos)

```

This directly addresses the search request to the appropriate parking lots.

VI. SCIENCE CLOUD DEPLOYMENT

Cloud computing refers to provisioning resources such as full machines, storage space, processing power, or even applications to consumers "on the net": Consumers can use these resources without having to install hard- or software themselves and can dynamically add and remove new resources. Common use cases include renting virtual machines, external disk space, or ready-made applications for traditional office tasks. Cloud solutions are software products which offer this ability. They may be installed by dedicated cloud companies which only offer the cloud end results to users; however, a company working in a non-IT branch (for example, manufacturing) can also install a cloud solution in-house, thus creating a private cloud for its own employees. The same applies to universities and research institutions.

A. A Voluntary, Peer-to-Peer Platform as a Service

In the science cloud case study [16], [17], the focus is on implementing a cloud in a fully distributed, peer-to-peer, voluntary computing fashion. The cloud is intended for use by the scientific community; each scientist – or university – can contribute to the cloud with computing power and storage space, but can also retract their resources if they are required elsewhere which corresponds to the voluntary aspect of the cloud. Furthermore, there is no centralized control in the cloud; rather, individual nodes communicate in a peer-to-peer fashion to organize themselves.

The cloud itself offers services on a PaaS level, that is, it provides a platform for executing applications. Each application may have its own requirements (or service level agreement) which the cloud must do its best to satisfy while in general keeping all applications running and conserving energy. These aspects make such a distributed cloud-based systems complex and hard to design, build, test, and verify.

To take part in the science cloud, each partner must run an instance of the Science Cloud Platform (SCP). Such an instance, running on a physical or virtual machine, is considered to be a service component in the previous described sense.

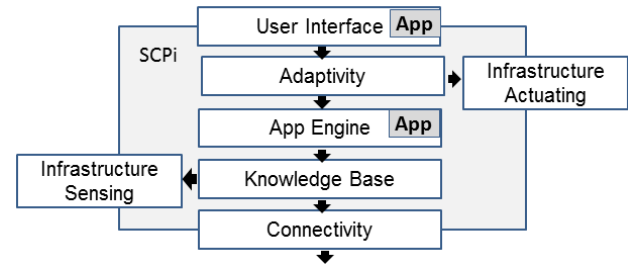


Figure 7. Science Cloud Platform Architecture

Figure 7 shows the logical components which make up a Science Cloud Platform instance (SCP). We explore three of them in more detail here: Connectivity, Knowledge, and Adaptivity.

a) *Connectivity*: Each SCPi has a connectivity component which enables it to talk to other SCPs over the network, and deals with the overlay structure the cloud imposes on the lower-level network layers. The protocol followed by these communications must enable SCPis to find one another and establish links, for example by manually entering a network address or by a discovery mechanism. Furthermore, SCPis must be able to query others for knowledge and at the same time distribute their own knowledge. Finally, the protocol must support exchange of data and applications.

There are different options for implementing such an overlay which in this case is built on top of the TCP/IP (Internet) network. As pointed out above, the science cloud takes a peer-to-peer approach to communication, and thus re-uses classical algorithms for peer routing (for example, DHT-based protocols like Pastry [18] are useful in this context).

b) *Knowledge*: Each SCPi has knowledge consisting of (1) its own properties (set by developers), (2) its infrastructure

(CPU load, available memory), and (3) other SCPis (acquired through the network). Since there is no global coordinator, each SCPi must build its own view and act upon the available knowledge. The SCPi may acquire knowledge about its infrastructure using an infrastructure sensing plug-in which provides information about static values, such as processor speed, available memory, available disk space, number of cores etc. and dynamic values, such as currently used memory, disk space, or CPU load.

SCPi properties are important when specifying conditions as Service Level Agreements (SLAs) [8], [16] for the applications. For example, when looking for a new SCPi to execute an application, low latency between the SCPis might be interesting. Other requirements may be harder: For example, an application may simply not fit on a SCPi because of the lack of space whereas another may require a certain amount of memory.

c) Adaptivity: As already outlined in the Ensemble Development Life Cycle (EDLC) in the second section of this paper, monitoring, awareness, and self-adaptation are key to managing autonomous systems. For this reason, each SCPi contains its own adaptivity component which uses the information available in the knowledge base.

Adaptation in the science cloud means several things. Applications will be deployed on the cloud and, depending on their SLA, must be executed on nodes which are able to fulfill these. If nodes become overloaded, or leave the system (which they may do at any time), applications need to be moved to different machines and restarted. This requires planning ahead for such situations, i.e., keeping redundant copies of both the applications' executable code and its data. The science cloud may, as indicated in Figure 7, work with an additional IaaS solution below it which allows the cloud to start new virtual machines and migrate to these machines if necessary. In the other direction, using such an IaaS allows shutting down machines if idle, thus conserving energy.

The adaptivity logic is exchangeable, application-independent, and has a direct relation to the SLAs of applications. The adaptivity logic itself can be written in a standard programming language or custom domain-specific languages or rules. It may take into consideration elements such as the reputation of nodes (previously established through their uptimes or capabilities), past performance, peak times experienced, and so on. Besides the connectivity, knowledge, and adaptivity components, each SCPi contains components for sensing the environment (for example, load, attached storage devices, etc.) and for acting on it (in the case an IaaS solution is available). Furthermore, an application engine executes applications locally; both the application interfaces and the SCPi meta-interface are available through a user interface component.

The science cloud is formed by connecting multiple SCPis together over a network. Within this cloud, we consider a subset of SCPis with certain properties as an ensemble which we call a Science Cloud Platform Ensemble (SCPe). The set of properties may be based on attributes of the SCPis and/or the SLAs of applications. In other words, an ensemble consists of SCPis which work together to run one application in a fail-safe manner and under consideration of the SLA of that application,

which may require a certain number of SCPis, certain latency between the parts, or have restrictions on processing power or on memory. An example of a science cloud with five SCPis grouped in two (overlapping) ensembles is shown in Figure 8.

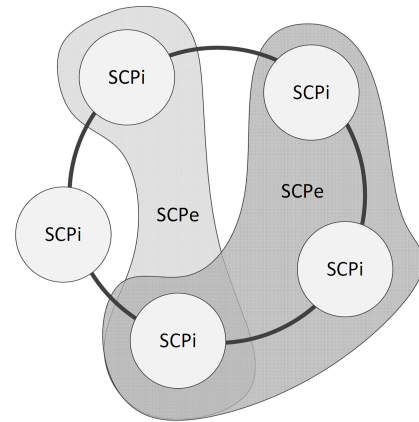


Figure 8. SCP Ensemble

At runtime, an ensemble may gain new SCPis or lose them depending on the behavior of the SCPis themselves; the load generated by applications, and the physical state of the underlying node (which may join and leave the network).

B. Programming the Science Cloud

Currently, a prototype of a science cloud platform is being developed and tested in a physical network connecting two universities [8]. The experimental platform features ad-hoc and voluntary behavior supporting dynamic re-configuration of physical layers and application migration on an upper level.

High-level SCEL modeling and model checking provide formal means for property proofs while a prototype implementation offers pragmatic means to test deployment and effectiveness of autonomous and self-aware behavior. The prototype we are currently investigating is based as much as possible on existing projects and scientific results. In particular, we are re-using the Pastry peer-to-peer substrate [18] and accompanying protocols for implementing the peer-to-peer and voluntary computing part, and in addition an interpretation of the ContractNET [19] protocol for the upper layer of application execution.

Our prototype is split into three layers which correspond roughly to handling the network addressing logic, data management, and application execution.

The first, i.e., the network layer, is based on the Pastry protocol. This protocol uses a hash-based addressing scheme similar to that of a Distributed Hash Table (DHT): Each node is assigned a random hash within a certain (wrapped) range; thus a network position agnostic overlay ring is formed. Routing works by sending messages to a certain hash target; the node whose hash is closest to the target hash receives the message. While routing is possible along the ring, Pastry also introduces shortcuts for reaching the target in $O(\log n)$ routing steps. It is important to note here that for each conceivable hash, exactly one node is the closest node which is an interesting property

exploited in upper layers. Pastry includes mechanisms for self-healing in case nodes drop out of the overlay network, and automatically integrates newly arriving nodes. Since Pastry only supports unicast routing, the SCRIBE protocol [20] is added on top which provides multicasting support using a form of publish/subscribe mechanism. In the science cloud platform, each node is implemented as a Pastry node.

The second layer, i.e., the data layer, is implemented by the PAST protocol [21] which implements the remaining features for realizing an actual DHT. Thus, it is possible to store data packages given their hash; again, the data is stored at the nearest node. However, PAST also supports redundancy since it allows storing not only one but k copies of a data package clustered around the nearest node. Thus, if the nearest node fails, another automatically takes its place.

The data layer is used in the science cloud platform for storing the applications to be executed (as byte code) as well as the data they keep during runtime. Both must be stored in a redundant fashion.

The application layer deals with the actual execution of applications. Here, we employ an implementation of the ContractNET protocol which is based on a bidding-like process. After deployment of an application by a user, an initiator node is chosen (based on hash nearness of the application code data package) which is from now on responsible for the execution of the application. Note that if this node drops out of the network, another node takes its place automatically according to changing hash nearness. The initiator will now request bids from other nodes through a SCRIBE-based communication channel, sending the application name and requirements to enable other nodes to evaluate whether they are capable of executing the app. This process is shown in Figure 9.

Having received all bids, the initiator decides on an executor node and sends further instructions to this node. The initiator then switches to a monitoring mode: If the executor fails, the initiator starts a new bidding process. The SCP implementation is open-source and can be downloaded from the ASCENS web site [8].

VII. CONCLUSION

This paper presents a unified approach to model, validate and deploy complex distributed systems with massive number of nodes that respect both individual and global goals. The non-centralized character of the approach allows for autonomic and self-aware behavior which is achieved by introduction of knowledge elements and enrichment of compositional and communication primitives with awareness of both system requirements and individual state of the computing entities.

The essence of the approach is to de-compose a complex system into a number of generic components, and then again compose the system into ensembles of service components. The inherent complexity of such ensembles is a huge challenge for developers. Thus, the whole system is decomposed into well-understood building blocks, reducing the innumerable interactions between low-level components to a manageable number of interactions between these building blocks. The result is a so-called hierarchical ensemble, built from service components, simpler ensembles and knowledge units

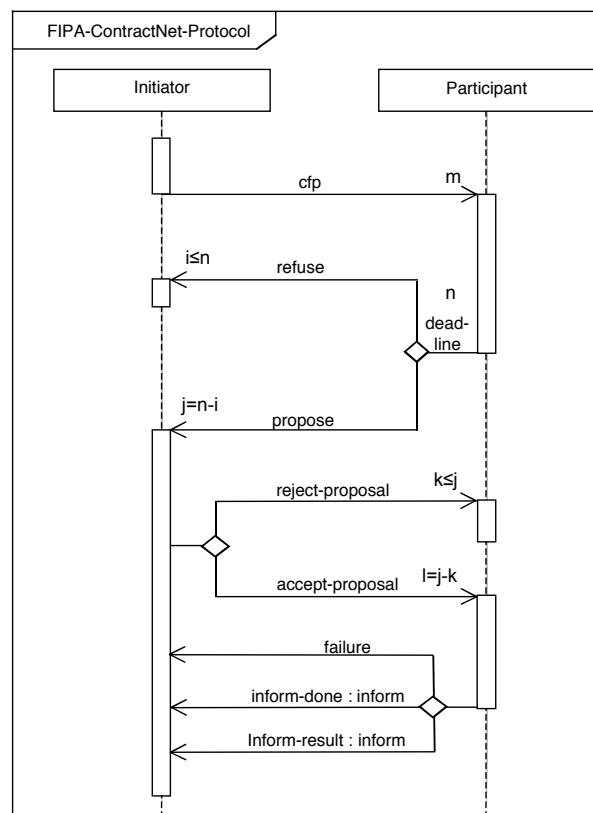


Figure 9. FIPA Contract NET Protocol (from [19])

connected via a highly dynamic infrastructure. Ensembles exhibit four main characteristics: adaptation, self-awareness, knowledge and emergence, yielding a sound technology for engineering autonomous systems [6], [8]. A number of linguistic constructs and validation and programming tools are under development and are being tested in different application scenarios.

This paper presents an integrated view (from high level modeling to application deployment) of a complex approach which has been described by a number of referenced papers, each focusing on different aspects of the work: Modeling ensembles using Helena [12] and SCEL [9], [13], system validation [15], adaptation aspects [10], knowledge management and deployments [10], [16] and engineering aspects [6], [8]. Further contribution of this paper is in optimized and energy-aware control based on autonomous behavior. Optimized distributed control with improved throughput and utilization of the cloud and e-mobility frameworks contribute significantly to the overall strategy to reduce energy consumption. Using the sharing principle instead of exclusive use of the transportation and computing means represents a significant challenge (requiring significant changes in our perception of vehicles and computers) in the application domains under consideration. This principle will undoubtedly play an important role in extending the application area.

ACKNOWLEDGMENT

Most of the work presented here has been performed in the context of the ASCENS project (project number FP7-

257414) [8], funded by the European Commission within the 7th Framework Programme, pervasive adaptation initiative. Special thanks go to the developers of SCEL and jRESP language (Rocco De Nicola from IMT Lucca and his group).

REFERENCES

- [1] N. Serbedzija, "Autonomous Systems: from Requirements to Modeling and Implementation," in *International Conference on Autonomic and Autonomous Systems (ICAS)*, 2013, pp. 1–6.
- [2] The InterLink Project. Accessed: 2013-18-11. [Online]. Available: <http://interlink.ics.forth.gr>
- [3] I. Sommerville, D. Cliff, R. Calinescu, J. Keen, T. Kelly, M. Z. Kwiatkowska, J. A. McDermid, and R. F. Paige, "Large-scale complex it systems," *Commun. ACM*, vol. 55, no. 7, pp. 71–77, 2012.
- [4] M. Hölzl, A. Rauschmayer, and M. Wirsing, "Engineering of software-intensive systems: State of the art and research challenges," in *Software-Intensive Systems and New Computing Paradigms*, ser. Lecture Notes in Computer Science, M. Wirsing, J.-P. Banâtre, M. M. Hölzl, and A. Rauschmayer, Eds. Springer, 2008, vol. 5380, pp. 1–44.
- [5] L. Xu, G. Tan, X. Zhang, and J. Zhou, "Energy Aware Cloud Application Management in Private Cloud Data Center," in *Proc. Cloud and Service Computing (CSC)*, 2011, pp. 274–279.
- [6] C. Seo, "Energy-Awareness in Distributed Java-Based Software Systems," in *Automated Software Engineering (ASE)*. IEEE Computer Society, 2006, pp. 343–348.
- [7] M. M. Hölzl and M. Wirsing, "Towards a System Model for Ensembles," in *Formal Modeling: Actors, Open Systems, Biological Systems*, ser. Lecture Notes in Computer Science, G. Agha, O. Danvy, and J. Meseguer, Eds., vol. 7000. Springer, 2011, pp. 241–261.
- [8] The ASCENS Project. Accessed: 2013-11-18. [Online]. Available: <http://www.ascens-ist.eu>
- [9] R. De Nicola, M. Loret, R. Pugliese, and F. Tiezzi, "SCEL: a Language for Autonomic Computing," IMT Institute for Advanced Studies Lucca, Italy, Tech. Rep., 2013.
- [10] D. B. Abeywickrama, F. Zambonelli, and N. Hoch, "Towards simulating architectural patterns for self-aware and self-adaptive systems," in *Self-Adaptive and Self-Organizing Systems (SASO) Workshops*. IEEE Computer Society, 2012, pp. 133–138.
- [11] C. Szyperski, *Component Software: Beyond Object-Oriented Programming*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2002.
- [12] R. Hennicker and A. Klarl, "Foundations for Ensemble Modeling - The Helena Approach," in *Specification, Algebra, and Software: A Festschrift Symposium in Honor of Kokichi Futatsugi (SAS 2014)*. LNCS, in press.
- [13] R. De Nicola, G. L. Ferrari, M. Loret, and R. Pugliese, "A language-based approach to autonomic computing," in *Formal Methods for Components and Objects (FMCO)*, ser. Lecture Notes in Computer Science, B. Beckert, F. Damiani, F. S. de Boer, and M. M. Bonsangue, Eds., vol. 7542. Springer, 2011, pp. 25–48.
- [14] M. P. Ashley-Rollman, S. C. Goldstein, P. Lee, T. C. Mowry, and P. Pillai, "Meld: A declarative approach to programming ensembles," in *Intelligent Robots and Systems (IROS)*. IEEE, 2007, pp. 2794–2800.
- [15] M. Loret, "jRESP: a Runtime Environment for SCEL programs," IMT, Institute for Advanced Studies Lucca, Italy, Tech. Rep., 2013. [Online]. Available: <http://rap.dsi.unifi.it/scel/>
- [16] P. Mayer, C. Kroiss, and J. Velasco, "The Science Cloud Case Study: Overview and Scenarios," Ludwig-Maximilians-Universität München, Munich, Germany, Tech. Rep. TR20120500, 2012.
- [17] P. Mayer and J. Velasco, "The Science Cloud Case Study: Overview and Scenarios," Ludwig-Maximilians-Universität München, Munich, Germany, Tech. Rep. TR20130300, 2013.
- [18] A. I. T. Rowstron and P. Druschel, "Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems," in *Distributed Systems Platforms (Middleware)*, ser. Lecture Notes in Computer Science, R. Guerraoui, Ed., vol. 2218. Springer, 2001, pp. 329–350.
- [19] FIPA Contract Net Interaction Protocol Specification. Accessed: 2013-11-18. [Online]. Available: <http://www.fipa.org/specs/fipa00029/SC00029H.html>
- [20] M. Castro, P. Druschel, A.-M. Kermarrec, and A. I. T. Rowstron, "Scribe: a large-scale and decentralized application-level multicast infrastructure," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 8, pp. 1489–1499, 2002.
- [21] A. Rowstron and P. Druschel, "Storage management and caching in past, a large-scale, persistent peer-to-peer storage utility," *SIGOPS Oper. Syst. Rev.*, vol. 35, no. 5, pp. 188–201, Oct. 2001.

Interactive Rigid-Body Dynamics and Deformable Surface Simulations with Co-Located Maglev Haptic and 3D Graphic Display

Peter Berkelman
Department of Mechanical Engineering
University of Hawaii at Manoa
Honolulu Hawaii, USA
Email: peterb@hawaii.edu

Sebastian Bozlee
Mathematics Department
University of Portland
Portland Oregon, USA
Email: sebbyj@gmail.com

Muneaki Miyasaka
Department of Mechanical Engineering
University of Washington
Seattle Washington, USA
Email: muneaki@uw.edu

Abstract—We have developed a system which can combine realtime dynamic simulations, 3D display, and magnetic levitation to provide high-fidelity co-located haptic and graphic interaction. Haptic interaction is generated by a planar horizontal array of cylindrical coils which act in combination to produce arbitrary forces and torques in any direction on magnets fixed to an instrument handle held by the user, according to the position and orientation sensed by a motion tracking sensor and the dynamics of a realtime physical simulation. Co-located graphics are provided by a thin flat screen placed directly above the coil array so that the 3D display of virtual objects shares the same volume as the motion range of the handheld instrument. Shuttered glasses and a head tracking system are used to preserve the alignment of the displayed environment and the interaction handle according to the user's head position. Basic interactive environments have been developed to demonstrate the system feasibility and operation, including rigid bodies with solid contacts, suspended mass-spring-damper assemblies, and deformable surfaces. Interactive physical simulation of these environments requires real-time collision detection between geometric models; numerical, discrete-time numerical integration to calculate the physics of networks of mass, spring, and damper elements; and calculation and actuation of interactive forces to the user in haptic rendering. Incorporating these functions into a single executable requires multiple program threads with various update rates, ideally performed using a multicore processor PC. Details and discussion of various simulations are given with experimental results.

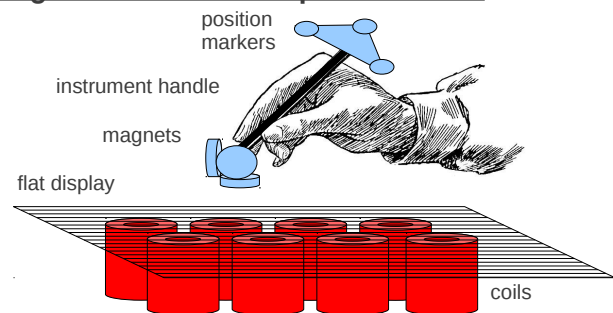
Keywords- haptics, interaction, simulation

I. INTRODUCTION

The ideal of virtual reality and haptic interfaces is to physically interact with simulated objects with the highest possible fidelity in both the graphical display and the kinesthetic forces and torques sensed by the user during interaction. Computer-generated graphics can produce highly realistic, dynamic 3D imagery in real time, but haptic interfaces are generally based on single point contact feedback, tactile cues, and linkage devices which have various limitations in their force and motion ranges, frequency response bandwidth, and resolution.

Our system combines a graphical display with a large range of motion magnetic levitation device, as shown in

Magnetic Levitation Haptic Interface:



3D Display of Virtual Environment to User:

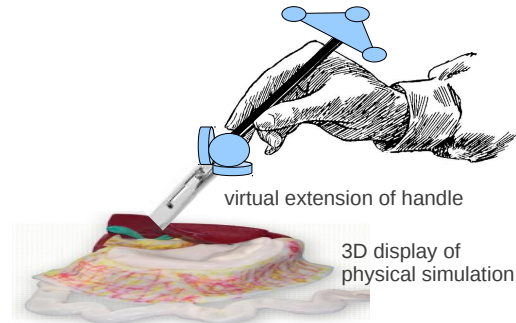


Figure 1: Co-located maglev haptic and 3D graphic display

Fig. 1. The graphical display is placed directly above a horizontal array of cylindrical coils and underneath the instrument handle held by the user, so that electromagnetic forces and torques can be generated on magnets embedded in the handle as the instrument is moved by the user into contact with the displayed simulated environment.

The magnets embedded in the instrument handle, the coil array with its current amplifiers, and the motion tracking sensor with its infrared LED markers, function together as a magnetic levitation system. Several variations of magnet configurations have been developed for stable levitation with the planar coil array, each providing a different combination

of parameters such as mass and size; force, torque and impedance ranges; and vertical translation and tilt (roll and pitch) rotation ranges. For the interactive simulations described here, two-magnet and four-magnet configurations were used, where the two-magnet handle provides greater feedback force and torque capabilities and the four-magnet handle is less massive, smaller, and provides somewhat greater vertical and rotational motion ranges. Both magnet configurations provide motion ranges of at least 100x100 mm horizontally, approximately 50 mm vertically, with unlimited yaw and tilt up to at least 35 degrees.

A secondary, slower and less precise motion tracking system tracks the position of the user's head so that the 3D views are generated correctly according to the position of each of the user's eyes. A pair of shuttered glasses, synchronized to the update rate of the graphics on the monitor, is worn by the user so that each eye sees a different image as the shutters alternate. In practice, this head tracking system allows the user to observe the handheld instrument and the 3D displayed environment together from the side and from above, in a natural ergonomic position for hand-eye coordination during dextrous manipulation of a handheld instrument or tool. Examples of relevant dextrous tool manipulation tasks include any writing, carving, or cutting tasks, operation of wrenches or screwdrivers, and medical needle manipulation for suturing, injections, and biopsy.

The real-time haptic interaction and graphical display are generated from a dynamic simulation which must perform collision detection, finite element deformation, and haptic rendering sufficiently quickly to support graphical updates at 30-60 Hz and haptic interaction and magnetic levitation at 800-1000 Hz. These tasks are sufficiently computationally intensive to be the limiting factor regarding the resolution and realism of the simulated environment.

This paper is an extended version of the previously published conference paper [1]. A survey of similar research in co-located haptics and graphics, magnetic levitation, and interactive physical simulation areas is given in Section II. The implementation details are given in Section III for the magnetic levitation subsystem and Section IV for the co-located 3D display subsystem. The physical simulation software and haptic rendering details are given in Section V. Force and motion experimental results for selected interactive simulations are given in Section VI. Continuing work is described in Section VII followed by a conclusion section.

II. RELATED WORK

The realization of our interactive system depends on the performance and integration of technology in the areas of maglev haptics, graphics, and physical simulation. Relevant prior work in each of these areas is surveyed below.

A. Co-Located Haptics and Graphics

3D graphics and haptic force and/or torque feedback can be generated at the same location by simply placing the 3D display behind the haptic interaction device, however, this method has two drawbacks. First, the body of the haptic interface device partially occludes the display, and second, there may be a significant difference produced between the perceived location of the displayed imagery and the surface of the screen, so that the convergence and focal distance of the user's eyes do not match, which is unnatural and may cause discomfort to the user.

ReachIn, *ImmersiveTouch*, and *SenseGraphics* systems [2] use a partially silvered mirror between the head and hand of the user, so that the display can be moved out of the way and the focal and convergence distances of the user's eyes can be matched. The haptic device and the user's hand do not occlude the 3D graphics behind them, but rather the real and virtual environments are superimposed and semitransparent due to the half-silvered mirror, which may be a distraction to the user.

The "what you see is what you feel" system [3] uses a thin flat display with a camera behind it. The video image of the user's hand is then extracted from the camera view using a green screen chroma-key technique, and rendered in the virtual environment. Holographic display [4] using a diffraction grid reflector and multiple projectors is another method which has been used for haptic and graphic co-location.

Other systems which have combined co-located haptics and graphics for user interaction have included mechanisms built into the display monitor [5], a cable-driven pen above the monitor [6], or linear induction motors with graphics projected from overhead [7]. The haptic feedback provided by these systems is limited, however, to only planar forces and torques, or predetermined locations, whereas the haptic interaction in our system provides full six degree-of-freedom rigid-body force and torque feedback over large ranges of translation and rotation.

Comparative studies have shown [8] [9] evidence of improved perception and performance from co-located haptic interaction.

B. Haptic Magnetic Levitation

Hollis and Salcudean first developed Lorentz force magnetic levitation devices [10] and applied them to haptic interaction and force-feedback teleoperation. Lorentz force magnetic levitation haptic interaction development continued with other more specialized device designs [11] [12] and larger range devices developed by Berkelman [13] [14].

Lorentz magnetic levitation is based on the Lorentz force F , which is directly proportional to both electric current I and magnetic field flux density B , integrated along the current path l , expressed as $F = \int B \times I dl$. Fixed magnet assemblies and a set of six coil windings on the levitated platform must be arranged so that forces and torques can

be produced in any direction as required for stable 6 DOF position feedback motion control and stable levitation. The advantage of the Lorentz actuation method compared to electromagnetic attraction and repulsion forces is that the force to current and flux density relationships are linear, and there is no direct dependence on position, so that the coil current to force and torque vector transformation is nearly constant over the motion range of the levitated body.

The range of motion in translation for Lorentz levitation is limited by the size of the gaps between the magnet faces in which the magnetic fields are concentrated, and the range in rotation is limited by the active areas of each coil in which the coil windings pass through the magnetic fields. To maximize the range of motion in both translation and rotation, it is best to arrange large-area flat-wound coils onto a thin hemispherical shell, with a user interaction handle mounted at its center.

Compared to linkage-based haptic devices such as the Sensable Phantom [15], the Novint Falcon [16], and the Force Dimension Delta [17], Lorentz levitation haptic interface devices can provide much greater forces and torques greater than 10 N and 1.0 N-m, and control stiffnesses of 10 N/mm are achievable without difficulty. Closed-loop position control bandwidths are greater than 100 Hz in all directions in both translation and orientation. Lorentz levitation motion ranges are much more limited, however, as the Carnegie Mellon and Butterfly Haptics devices have ranges of approximately 25 mm and 30 degrees of rotation, and a University of Hawaii prototype with a modified magnet and coil configuration has a range of 50 mm and 60 degrees of rotation [18]. Overheating of the actuation coils is not a problem, as the levitated hemispherical shell is quite thin with a large surface area and acts as an effective heat dissipator.

The general design and function of the planar coil array magnetic levitation system used here is described in [19]. This system uses a fixed planar array of cylindrical coils to levitate a platform of one or more cylindrical magnets. The yaw of the levitated platform is unlimited and its horizontal motion range is determined by the size of the planar array. Vertical levitation distances of up to 75 mm and tilt angles of 45 degrees are achievable, depending on the mass of the levitated platform and the dimensions of the magnets used.

Similar tabletop-scale large range magnetic levitation systems have been developed for suspension of models in wind tunnels [20] and for micromanipulation using pole pieces to shape magnetic fields [21].

C. Realtime Physical Simulation Libraries and Haptic Rendering Programming Interfaces

Realistic software simulations of dynamic physical environments have been developed by Baraff both for rigid [22] and deformable [23] objects, including efficient collision and reaction force detection and surface friction. Freely available

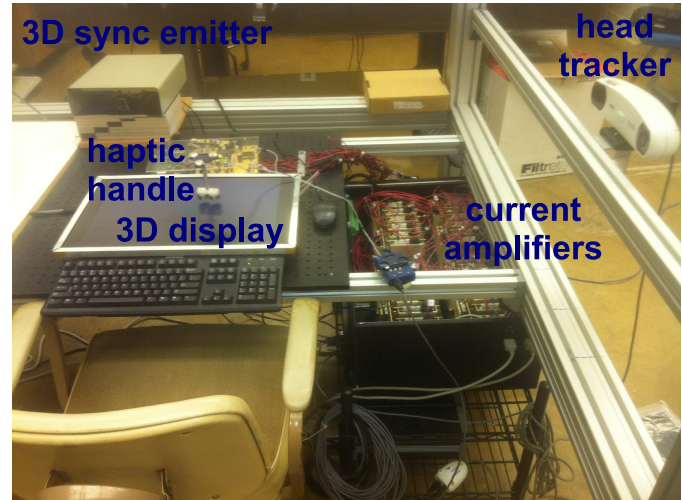


Figure 2: Implemented system

physical simulation software packages include the SOFA framework [24] [25], Bullet Physics, and the PhysX library from NVIDIA. Higher resolution and performance can be obtained by using precomputed deformation modes [26] and 6-DOF haptic rendering including torque feedback as well as force on an interactive instrument can be integrated with simulations [27].

Realistic haptic interaction with dynamic simulated environments typically requires realtime computation at update rates in the range of 1000 Hz. Collision detection, calculation of rigid body contacts [28], and simulation of physical dynamics, must be performed concurrently with the control of the haptic interaction device. Virtual coupling, using a virtual spring and damper to connect an interaction object in the simulated environment with the physical object grasped by the user [29], is a straightforward method to integrate a simulated environment with a haptic interaction device.

Several software packages are freely available for haptic rendering and realtime physical simulation. H3D [30] and Chai3D [31] include driver interfaces for common commercial haptic interface devices such as the Sensable Technologies Phantom [15]. A programming interface is also available with the magnetic levitation haptic interface from Butterfly Haptics LLC [32].

III. IMPLEMENTED MAGNETIC LEVITATION SYSTEM

The motion tracking, magnetic levitation control, haptic rendering, physical simulation, and graphical display in our current system are all executed in real time in separate threads on a single quad-core PC in Linux 2.6. GNU C/C++ was used for all programming. An initial demonstration concept of the system with a simulation of a single paddle instrument and a ball rolling on a plane, an earlier magnet and coil configuration, and a conventional 2D display was demonstrated previously [33]. The current system is shown

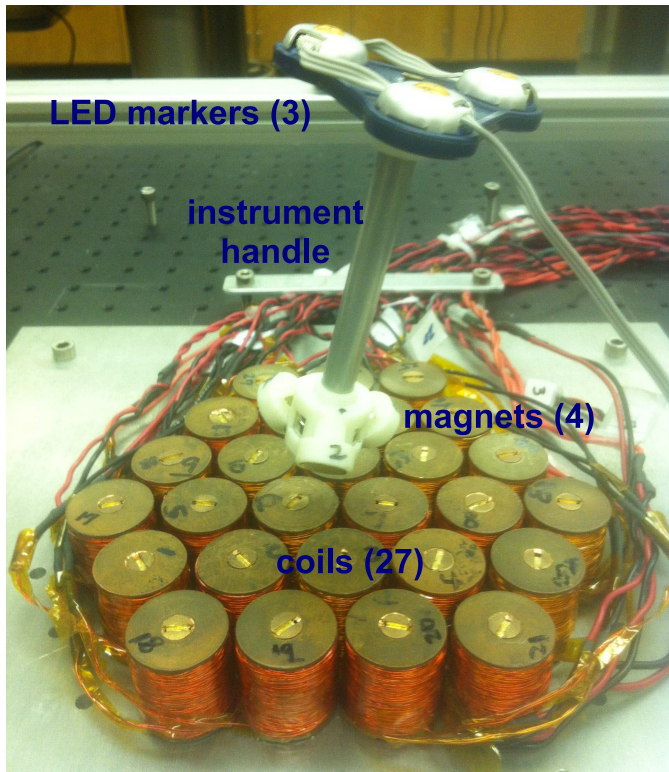


Figure 3: Levitated 4-magnet instrument

in Fig. 2, including the planar 3D display, haptic instrument handle, current amplifiers, and head tracker.

A. Magnetic Levitation Hardware Setup

The motion tracking of the handheld instrument in our system is done using a Northern Digital Optotrak Certus position sensor and three infrared Smart Markers. Motion tracking updates are provided at 860 Hz with a position resolution of approximately 0.01 mm for each marker. Actuation forces and torques are generated by a closely packed array of 27 cylindrical coils, each with 1000 windings, 25 mm diameter, and 30 mm height. Either a two-magnet or four-magnet instrument handle can be used with the system; the two-magnet 125 g instrument can provide greater haptic forces and torques but is more massive and bulky, and the smaller 75 g four-magnet instrument occludes the user's view of the display less due to its compact size. Forces are limited to approximately 4 N due to heating of the actuation coils, although higher momentary peak forces are possible.

The four-magnet instrument is shown in Fig. 3, levitated above the 27-coil array at a height of 30 mm and a tilt angle of 20 degrees. This coil array is underneath the 3D display monitor shown in Fig. 2. The motion tracker for the haptic instrument is mounted on a rigid frame at ceiling level, looking downwards.

B. Design and Control Software

The general design and evaluation methods used in the development of the magnetic levitation system are described in detail in [34]. Electromagnetic modeling of the forces and torques between each magnet and coil was performed using Mathematica [35] from Wolfram Research and Radia [36], a freely available software package developed by the European Synchrotron Radiation Facility.

At each sensor update of the levitation control system, the coil current to levitation force and torque transformation matrix is calculated according to the levitated body position and orientation and the precomputed electromagnetic models, control forces and torques are generated according to proportional and derivative (PD) error gain control laws for each of the total 6 degrees of freedom in translation and rotation, and updated coil currents are calculated using the pseudoinverse of the coil current to force and torque transformation matrix.

IV. CO-LOCATED 3D GRAPHICAL DISPLAY

The NVIDIA 3D Vision package was used with Linux drivers to provide 3D display of the simulated environment. This package uses shutter glasses which are synchronized with the graphics card by an infrared emitter box. A Quadro 4000 graphics card was used with a ViewSonic vm2268 monitor with a 120 Hz update rate. OpenGL and GLUT graphics libraries are used for the 3D graphics rendering.

The case of the monitor was removed and backplane circuit boards and wiring were moved so that the monitor backlight and display could be placed directly on the coil array. The combined thickness is under 10 mm, so that haptic forces and torques can be applied to the handheld instrument up to a vertical height of at least 60 mm. Magnetic fields from the instrument magnets and coil array were not found to interfere with the display, and there are no ferromagnetic components in the display to interfere with the magnetic levitation system. A thin sheet of polycarbonate plastic was fixed on top of the monitor screen for protection from impacts from the magnets and instrument, and an aluminum frame was built to protect the edges of the display.

Head tracking was implemented using a Northern Digital Polaris Vicra and passive reflective markers to produce correct 3D display according to the position of each eye. The spatial position and orientation of the shutter glasses from the positions of four reflective markers fixed to the glasses. Position and orientation data were updated at a 10 Hz rate with a resolution of approximately 0.1 mm for each marker. It would be possible to track both the magnet instrument and the user's head using a single motion tracking system, but this would require using wired infrared markers on the 3D shutter glasses, slowing down the update rate of the magnetic levitation localization due to the additional LED markers on the glasses, and mounting the localizer at least



Figure 4: Shutter glasses with synchronization signal transmitter, reflective markers, and localizer for 3D graphic display with head tracking

3.5 m high so that its sensing volume includes the location of the glasses.

As both the Optotrak and Polaris motion trackers use infrared position sensing, and 3D Vision systems uses infrared communication to synchronize display frames with the shutter glasses, it is necessary to ensure that each infrared system does not interfere with the others. In our system, each set of emitters and receivers are oriented in orthogonal directions and positioned so that each emitter is only visible to its corresponding receiver. The Optotrak sensor is mounted above the table looking down at the LEDs on the instrument, the Polaris is mounted on the side of the table to track the reflective markers on the side of the glasses, and the synchronization emitter is mounted at the front of the tabletop. The synchronization emitter, shutter glasses with reflective position markers, and head tracking localizer are shown in Fig. 4.

The 3D vision system as described produced reasonably convincing 3D graphics but had a number of minor shortcomings. The horizontal position of the monitor resulted in a reduction in brightness observed by the user due to the change in viewing angle. Light reflections from the glossy screen could be distracting, but the room can be darkened to eliminate this problem. The 10-15 Hz update rate of the head tracking system and its communication latency produce a noticeable lag if the user's head moves quickly. The motion tracking reflectors on the side of the shutter glasses are also somewhat cumbersome. Many of the shortcomings of the present head tracking system could be overcome by using a radio-frequency emitter, currently available from NVIDIA, rather than an infrared signal for the shutter glasses synchronization, and using an optical tracker with a quicker update rate.

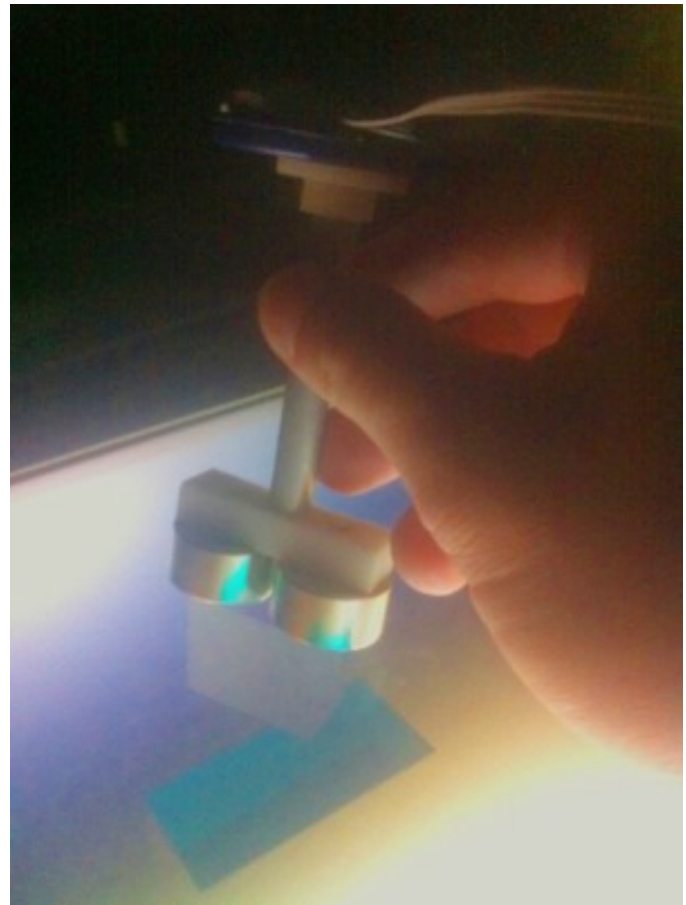


Figure 5: Peg in hole simulation with grasped tool aligned with graphical peg

V. HAPTIC SIMULATIONS

Basic interactive simulations which have been implemented on our system at present include point, edge, and face contacts between simple solid shapes such as square peg-in-hole insertion as shown in Fig. 5, simple dynamic environments including suspended masses and springs, and rolling objects. These simple initial simulations allow the dynamics and contact models of the environments to be modified and adjusted to provide the most realistic haptic interaction while preventing unstable dynamics.

A more sophisticated simulation which involves an instrument contacting a deformable surface is shown in Fig. 6. In this simulation, a virtual extension is added to the actual haptic instrument handle, and the deformation of the surface and reaction forces and torques on the instrument are calculated at the haptic update rate. Damping is added to the internal dynamics of the deformable body and the surface dynamics during contact with the haptic instrument.

The MLHI library and programming interface, originally from Butterfly Haptics LLC, has been adapted for use with our system and can be used for haptic rendering and

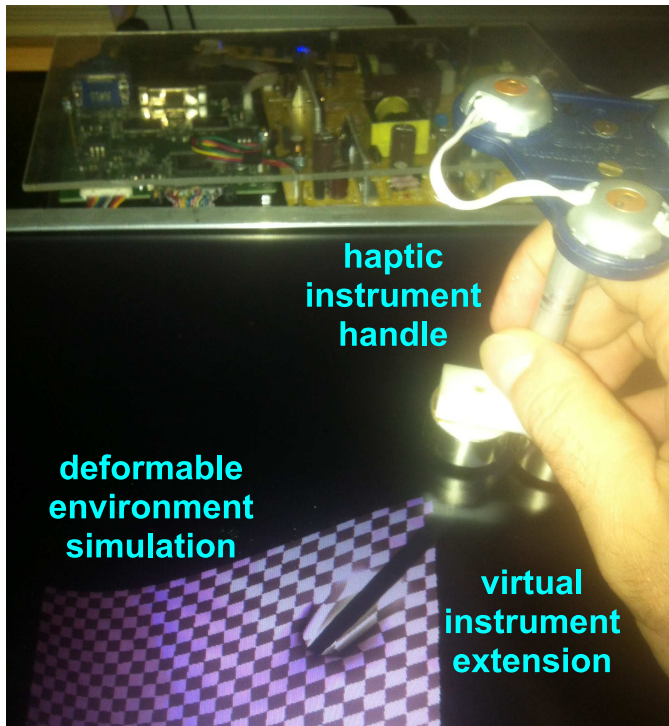


Figure 6: Deformable tissue simulation with grasped tool aligned with graphical scalp

communication between simulation and magnetic levitation threads with a haptic update rate of 1000 Hz. Alternatively, haptic rendering and dynamic simulation calculations can be performed synchronously with the motion tracking at 860 Hz.

A. Software Implementation

Four basic simulations were written to demonstrate the capabilities of the colocated haptic system. All simulations made use of the 3D-parallax, head tracking, and haptic capabilities of the system. All but one made use of the system's 6-DOF position sensing and haptic output capabilities. In all cases, head tracking, 3D display, and haptics were run at full speed, and physics calculations were performed at the haptic update rate.

The simplest simulation consists of a ball hanging on a virtual spring whose other end is moved freely by the user. By moving the maglev handle, the user may swing the ball, experiencing inertial forces as well as seeing 3D parallax effects. No torques are experienced by the user.

A second simulation consists of a paddle manipulated by the user in both translation and orientation, and a ball that the user may bounce on the paddle. This demonstration presents the user with stronger and more variable haptic feedback, this time including gentle torques based upon the location of the paddle-ball contact.

The third simulation consists of a rectangular user-controlled peg and a rectangular hole into which the user may insert the peg. Edge-edge, edge-face, and vertex-face contacts are all possible and result in both forces and torques applied to the haptic instrument handle. While inserted into the hole, haptic feedback is sufficiently stable that the user may safely let go of the instrument handle, leaving the virtual hole walls to support the handle. This demonstration involves stiff haptic feedback in both forces and torques.

The final and most sophisticated simulation consists of a virtual tool controlled by the user and a deformable surface with which the user may interact. The deformable surface is modeled by a hexahedral mass-spring-damper lattice. Deformation and jello-like vibration may be observed by the user on contact with the surface.

While the code for display, physics, and haptic feedback differs from program to program, each demo utilizes a common core of simulation software, which provides for head tracking, 3D-rendering, and timing of physics, graphics, and haptics calculations.

The code for the magnetic levitation system controller was collected into a separate software library, modelled on the MLHI library provided by Butterfly Haptics LLC for their haptic device. This library provides for initiating and shutting down the haptic device, conveying feedback forces to it, and for performing PD control of the haptic device position.

B. Multithreading and Timing

Each of the simulations makes use of multithreading to manage the different timing requirements of graphics, head tracking, and physics/haptic feedback, while still providing low-latency feedback. Each of these tasks is allocated a thread. Communication between threads is performed through data structures stored in global memory.

In order to more easily ensure low-latency feedback, this communication is not synchronized. As no more than one thread writes to a given set of data, the data sets are small, and changes to data between frames tend to be small, error due to threading conflicts is imperceptible to the user.

The program begins by initializing data structures, then launching threads for head tracking and physics. The main thread then assumes the role of running graphics. Haptics is initiated later by the user.

The graphics thread is managed by the freeGLUT library, an open source alternative to the OpenGL utility toolkit. It is responsible for displaying the current state of the simulation data structures, which it does at the display rate of the 3D monitor (120 frames per second). The freeGLUT library uses the same thread to manage keyboard input to the simulation.

Meanwhile, the head tracking thread initializes and repeatedly queries the NDI Polaris for the location of the user's head. If this position can be determined, the translation and orientation of the user's head is calculated and stored in

global memory. This happens at the update rate of the Polaris sensor, which is about 10 Hz.

Finally, a third thread runs the simulation. At the beginning of the program, haptics is not initiated, and so this thread just performs physics calculations, running in a loop at approximately the rate of the maglev controller (860 Hz). When haptics is initiated, a transition is made from running in a loop to running in a callback from the device controller code. Once this has happened, the code runs at the rate of the device controller and haptic rendering is performed in addition to the simulation. When haptics is deactivated a transition is made back to running in a loop.

Upon receiving a command to exit the program, each of the threads shuts down in turn. Next, logging data, if any, is stored, then the program exits.

C. Coordinate System Correspondance

In the colocated haptic system, the size of the display and its location relative to the user's eyes is known. As a result, the apparent locations of virtual objects correspond in a one-to-one fashion to real locations: virtual objects may be considered as embedded in real space. For example, a virtual ball may be thought of as being 2 cm in diameter and located 4 cm below the display. Using information about the location of the user and size of the display, that virtual ball may be rendered on the display so that to the user's eye it appears to be 2 cm in diameter, 4 cm below the center of the display, no matter where the user moves.

Accordingly, virtual units and coordinate systems take on more meaning when used with the colocated system. For simplicity, it was chosen to locate the origin of the virtual coordinates at the center of the 3D display, with coordinate axes aligned to the display's edges, with units of mm.

It is interesting to note that due to the correspondence of virtual and real locations, the simulations' virtual coordinate system is also a coordinate system for the real space surrounding the display. Calculating the location of the user's head in real space also calculates the location of the user's head in virtual space.

D. Deformable Surface Modeling and Simulation

The deformable surface simulation was designed to demonstrate the possibility of sophisticated haptic environments involving non-rigid contacts and complicated geometry. The deformable "landscape" consists of an approximately regular mass-spring-damper lattice whose top side varies in height according to a heightmap. The construction of the landscape proceeds in 3 phases: calculation of the height map, distribution of point masses, and finally, linking neighboring point masses by springs.

The height map consists of a rectangular array of heights, in millimeters, indexed by x and y positions. The heights are either calculated in a pseudo-random fashion or according

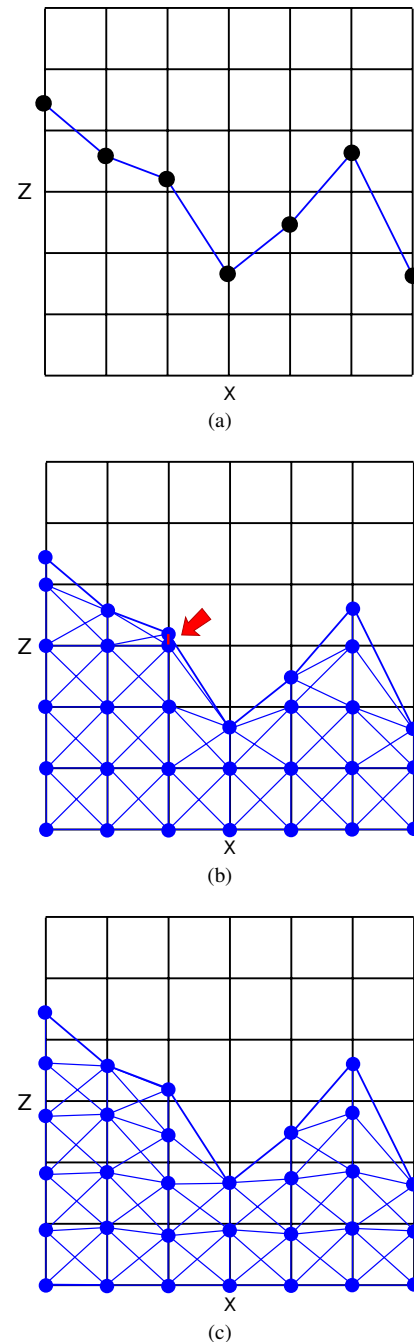


Figure 7: Building the lattice from a heightmap

to a simple function, such as a sine wave. This is enough to determine the surface that the user sees.

The next step is to generate the lattice points beneath the heightmap. To do this, a three dimensional grid is placed over the heightmap. Each point of the grid below the corresponding surface point is made into a point mass of the lattice. A slice of the grid and an example heightmap is pictured in Figs. 7a –7c. In order to avoid unusually short

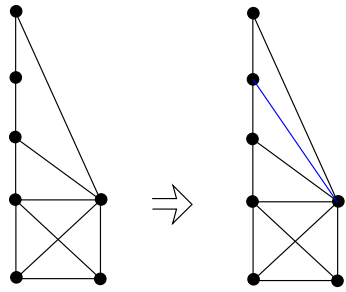


Figure 8: Additional links added when surface vertices are far apart

links (as in Fig. 7b), the masses in each column are respaced uniformly rather than on the points of the grid (Fig. 7c). This means horizontally adjacent vertices in the grid will not be at exactly the same heights in space. All vertices are assigned the same mass, except for the vertices on the bottom and edges, which are assigned infinite mass, to keep them stationary.

The final step is to link the lattice points together. In order to prevent undesirable inversions of the grid, for each grid point, each of the up to 26 horizontal, vertical, and diagonal neighbors of the grid point are linked to it. Adjacent surface vertices that are not already linked together are then linked together. Finally, each of the vertices below a surface vertex but above an adjacent surface vertex are linked to the lower surface vertex. This can be seen in Fig. 8. All of the links are assigned the same stiffness and damping constants.

During physics calculations, a collision detection routine determines the force the user is applying to the point masses. The opposite forces and corresponding torques are added and sent to the haptics device for force feedback. Next, the forces on the point masses due to spring compression and damping are calculated and added to the user's applied force. An Eulerian integration scheme is then used to update the velocity and position of each point mass.

VI. RESULTS

Force and position experimental data in x , y , and z directions obtained during interactive simulations are presented in Figs. 9 and 10. The position data was measured by the position tracking system, and the force data are calculated by the simulations and generated by the coil array of the magnetic levitation system in real time. The commanded forces were shown to be within 0.1 Newtons of force sensor measurements throughout the range of the magnetic levitation system in [19].

The Fig. 9 plots are from a haptic peg-in-hole simulation in which a 25 mm square peg is controlled by the haptic instrument handle and inserted into a 27x54 mm, 10 mm deep square hole. The Fig. 10 plots are from a deformable simulation in which a pointed virtual instrument contacts a deformable object, as shown in Fig. 4. For both cases, haptic

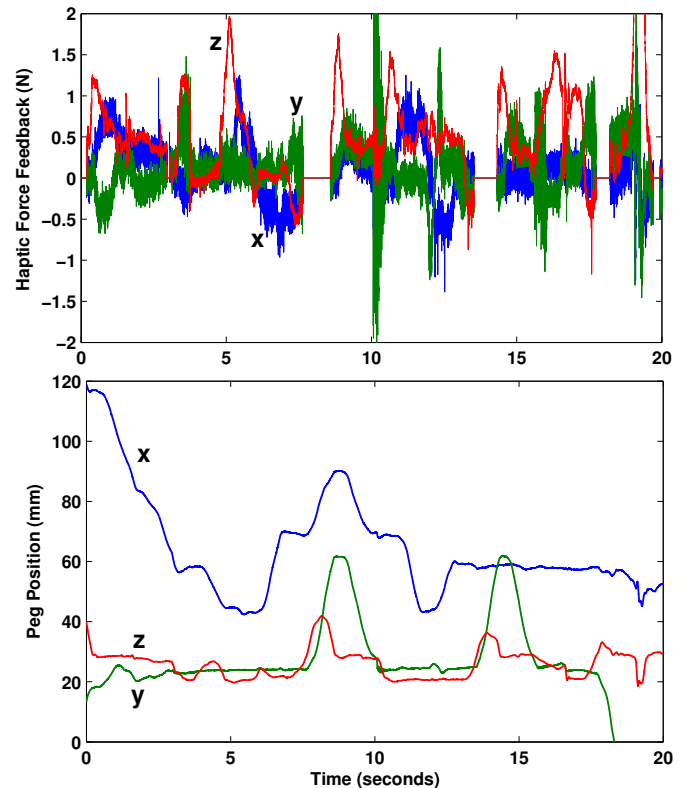


Figure 9: Interactive peg-in-hole simulation data

forces and torques are zero while the instrument is moving freely, contact forces are approximately proportional to the depth of contact, and haptic torques depend on each contact force and the displacement between the contact point and the center of the haptic instrument and simulated tool.

In the peg-in-hole simulation of Fig. 9, the peg is not in contact with the hole or top surfaces at the 8-9 and 14-15 second intervals, the z coordinate is greater than 30, and there is no haptic force feedback. As the peg is moved in and out of the hole, the z position moves between 20 and 30 mm. The x position can vary between approximately 40 and 70 mm while the peg is in the hole, as the hole is more than twice as wide as the peg in the x direction. Non-zero x and y forces are present when the virtual peg is pushed against any of the four sides of the virtual hole. Contact stiffnesses are approximately 0.4 N/mm and the kinetic and static friction coefficients are 0.15 in the simulations.

For the deformable surface of Fig. 10, the probe is moved across the surface during the 12-20 second interval, and the surface is struck with the probe several times in the interval from 8 to 12 seconds. The object was modeled with millimeter-scale surface variations rather than a smooth flat surface. Therefore, this surface texture produces variable vertical (z) forces in response to horizontal (x and y) motions of the instrument tip. Oscillations in both the position

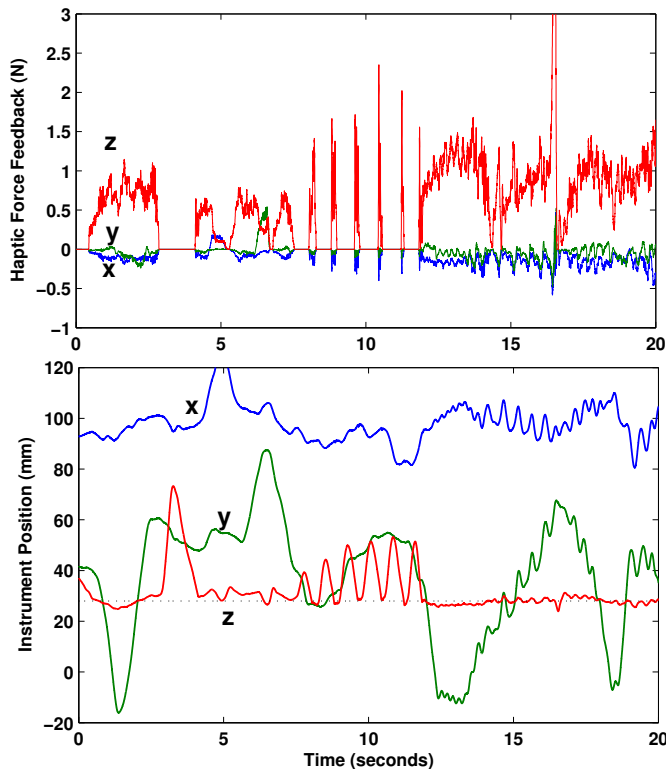


Figure 10: Interactive palpation of deformable surface data

and force data can be seen in the 12-20 second period due to sticking and slipping of the sliding surface contact. Overall the force plots are smoother in the deformable surface simulation than the peg-in-hole simulation due to the compliance and lower friction of the deformable surface.

VII. FUTURE WORK

At present, the magnetic levitation and motion tracking aspects of our system are fully developed, but the interactive environments are at a preliminary stage. We plan to refine the detail and physical realism of the simulated environments to a degree where they are useable and can provide measureable benefits in medical training tasks such as surgery, intubation, and needle driving. User studies will be conducted to evaluate the benefits of colocated haptic and graphical training of simulated medical procedures.

The complexity of the modelled environment and the sophistication of the simulated dynamics can be improved by using the graphics processor for additional numerical computations, as a general purpose graphics processing unit or GPGPU. NVIDIA provides the CUDA [37] programming interface to utilize the parallel processing capabilities of the GPGPU on the graphics cards used, however, the physical simulation programming must be completely reformulated to realize these benefits.

One more planned improvement to be made on the system

is to reconfigure the system to be simpler and more compact. The optical localizer presently in use is over 1.1 m in length and 18 kg and must be fixed at least 1.5 m from the sensed position markers, which necessitates the use of a large rigid frame assembled from aluminum extrusions. Compact localizer systems such as the AccuTrack from Atracsys have specifications with comparable accuracy, update rates, and latency as needed for stable levitation and haptic feedback, yet are much smaller and can be mounted as close as 0.15 m to the position markers on the handheld instrument. It may also be possible to use electromagnetic sensing systems to track the magnet locations [38], however, interference from the actuator coil currents may need to be overcome to realize sufficient positioning accuracy.

VIII. CONCLUSION

Our system is the first to combine high-fidelity haptic interaction through a magnetic levitation coil array with interactive virtual environments displayed in 3D in a co-located manner, where the handheld tool grasped by the user is manipulated in the same tabletop space as the perceived 3D graphics display of the simulated environment. The motion range of the magnetic levitation haptic interface device in both translation and rotation is well suited to human hand motions and tabletop displays.

The operation of the system was demonstrated with 6-DOF haptic interactive simulations with solid objects incorporating rigid-body dynamics and deformable surfaces. Continual increases in the computational speed and capacities available from standard PC hardware, combined with the increasing availability of sophisticated graphics, modeling, and physical simulation programming interfaces, lead to the feasibility of sophisticated interactive medical simulations which could be used with their co-located haptic and graphic interface system.

Our co-located haptic and graphic interface system is novel in that there is no hardware between the user and the display other than the handheld interaction instrument. The 3D environment is displayed close to the surface of the monitor, so there is no conflict between visual convergence and focal ranges. Electromagnetic force and torque actuation is used for haptic interaction rather than a motorized linkage, providing advantages in backdriveability, precision, and response frequency bandwidths.

We have demonstrated the feasibility and function of our system with the basic simulation environments described.

ACKNOWLEDGMENT

This work was supported in part by National Science Foundation grants IIS-0846172 and CNS-0551515, and by the University of Hawaii College of Engineering.

REFERENCES

- [1] P. Berkelman, S. Bozlee, and M. Miyasaka, "Interactive dynamic simulations with co-located maglev haptic and 3D graphic display," in *International Conference on Advances in Computer-Human Interactions*, February 2013, pp. 324–329.
- [2] C. Luciano, P. Bannerjee, L. Florea, and G. Dawe, "Design of the ImmersiveTouch: a high-performance haptic augmented virtual reality system," in *11th International Conference on Human-Computer Interaction*, Las Vegas, August 2005.
- [3] Y. Yokokoji, R. Hollis, and T. Kanade, "WYSIWYF display: A visual/haptic interface to virtual environment," *Presence*, vol. 8, no. 4, pp. 412–434, August 1999.
- [4] P. Olsson, F. Nysjo, S. Siepel, and I. Carlbom, "Physically co-located haptic interaction with 3d displays," in *IEEE Haptics Symposium*, Vancouver, March 2012, pp. 267–272.
- [5] C. Swindells, M. Enriquez, K. MacLeand, and K. Booth, "Co-locating haptic and graphic feedback in manual controls," Computer Science, University of British Columbia, Tech. Rep. TR-2005-28, 2005.
- [6] L. Lin, Y. Wang, Y. Liu, and M. Sato, "Application of pen-based planar haptic interface in physics education," in *International Conference on Computer-Aided Design and Computer Graphics*, September 2011, pp. 375–378.
- [7] H. Noma, S. Yoshida, Y. Yanagida, and N. Tetsutani, "The proactive desk: A new haptic display system for a digital desk using a 2-DOF linear induction motor," *Presence: Teleoperators and Virtual Environments*, vol. 13, no. 2, pp. 146–163, April 2004.
- [8] D. Swapp, V. Pawar, and C. Loscos, "Interaction with haptic feedback and co-location in virtual reality," *Presence*, vol. 10, no. 1, pp. 24–30, April 2006.
- [9] G. Jansson and M. Ostrom, "The effects of co-location of visual and haptic space on judgements of form," in *Euro-Haptics*, Munich, June 2004, pp. 516–519.
- [10] R. L. Hollis and S. E. Salcudean, "Lorentz levitation technology: a new approach to fine motion robotics, teleoperation, haptic interfaces, and vibration isolation," in *Proc. 6th Int'l Symposium on Robotics Research*, Hidden Valley, PA, October 1993.
- [11] S. Salcudean and T. Vlaar, "On the emulation of stiff walls and static friction with a magnetically levitated input-output device," in *ASME IMECE*, Chicago, November 1994, pp. 303–309.
- [12] P. J. Berkelman, R. L. Hollis, and S. E. Salcudean, "Interacting with virtual environments using a magnetic levitation haptic interface," in *Int'l Conf. on Intelligent Robots and Systems*, Pittsburgh, August 1995.
- [13] P. J. Berkelman and R. L. Hollis, "Lorentz magnetic levitation for haptic interaction: Device design, function, and integration with simulated environments," *International Journal of Robotics Research*, vol. 9, no. 7, pp. 644–667, 2000.
- [14] P. Berkelman and M. Dzadovsky, "Extending the motion ranges of magnetic levitation for haptic interaction," in *Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, Salt Lake City, March 2009, pp. 517–522.
- [15] T. Massie and K. Salisbury, "The PHANTOM haptic interface: A device for probing virtual objects," in *Proceedings of the ASME Winter Annual Meeting, Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, Chicago, Illinois, November 1994.
- [16] *Novint Falcon User Manual*, Novint Technologies Inc., 2007.
- [17] S. Grange, F. Conti, P. Rouiller, P. Helmer, and C. Baur, "Overview of the delta haptic device," in *Eurohaptics*, Birmingham, July 2001, pp. 164–166.
- [18] P. Berkelman, "A novel coil configuration to extend the motion range of lorentz force magnetic levitation devices for haptic interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Diego, October 2007.
- [19] P. Berkelman and M. Dzadovsky, "Magnetic levitation over large translation and rotation ranges in all directions," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 1, pp. 44–52, 2013.
- [20] N. J. Groom and C. P. Britcher, "A description of a laboratory model magnetic suspension testfixture with large angular capability," in *IEEE Conference on Control Applications*, Dayton, September 1992, pp. 454–459.
- [21] M. B. Khamesee and E. Shameli, "Regulation technique for a large gap magnetic field for 3d non-contact manipulation," *Mechatronics*, vol. 15, pp. 1073–1087, 2005.
- [22] D. Baraff, "Interactive simulation of solid rigid bodies," *IEEE Computer Graphics and Applications*, vol. 15, pp. 63–75, 1995.
- [23] D. Baraff and A. Witkin, "Dynamic simulation of non-penetrating flexible bodies," in *Computer Graphics (Proc. SIGGRAPH)*, vol. 26. ACM, July 1992, pp. 303–308.
- [24] J. Allard, S. Cotin, F. Faure, P.-J. Bensoussan, F. Poyer, C. Duriez, H. Delingette, and L. Grisoni, "Sofa: an open source framework for medical simulation," in *Medicine Meets Virtual Reality (MMVR'15)*, Long Beach, USA, February 2007.
- [25] M. Marchal, J. Allard, C. Duriez, and S. Cotin, "Towards a framework for assessing deformable models in medical simulation," in *International Symposium on Computational Models for Biomedical Simulation*, London, July 2008.
- [26] J. Barbic and D. James, "Six-dof haptic rendering of contact between geometrically complex reduced deformable models," *IEEE Transactions on Haptics*, preprint published online, June 2008.
- [27] D. James and D. Pai, "A unified treatment of elastostatic and rigid contact simulation for real time haptics," *Haptics-e*, vol. 2, no. 1, 2001.

- [28] C. Zilles and J. Salisbury, "A constraint-based god-object method for haptic display," in *Int'l Conf. on Intelligent Robots and Systems*, Pittsburgh, August 1995.
- [29] J. Colgate, M. Stanley, and J. Brown, "Issues in the haptic display of tool use," in *Int'l Conf. on Intelligent Robots and Systems*, Pittsburgh, August 1995.
- [30] *HAPI Manual*, 1st ed., SenseGraphics Inc., May 2009.
- [31] F. Conti, D. Morris, F. Barbagli, and C. Sewell. *CHAI 3D*. Online: <http://www.chai3d.org>, 2006.
- [32] *Magnetic Levitation Haptic Interface API Reference Manual*, 1st ed., Microdynamic Systems Lab, Carnegie Mellon University, September 2008.
- [33] P. Berkelman, M. Miyasaka, and J. Anderson, "Co-located 3d graphic and haptic display using electromagnetic levitation," in *IEEE Haptics Symposium*, Vancouver, March 2012, pp. 77–81.
- [34] P. Berkelman and M. Dzadovsky, "Novel design, characterization, and control method for large motion range magnetic levitation," *IEEE Magnetics Letters*, vol. 1, January 2010.
- [35] S. Wolfram, *The Mathematica Book*, 5th ed. Wolfram Media, 2003.
- [36] O. Chubar, P. Elleaume, and J. Chavanne, "A three-dimensional magnetostatics computer code for insertion devices," *Journal of Synchrotron Radiation*, vol. 5, pp. 481–484, 1998.
- [37] *Nvidia CUDA Compute Unified Device Architecture Programming Guide 1.0*, Nvidia, 2007.
- [38] C. Hu, M. Li, W. Yang, R. Zhang, and M.-H. Meng, "A cubic 3-axis magnetic sensor array for wirelessly tracking magnet position and orientation," *IEEE Sensors Journal*, vol. 10, no. 5, pp. 903–913, 2010.

Autonomous Load Balancing of Data Stream Processing and Mobile Communications in Scalable Data Distribution Systems

Rafael Oliveira Vasconcelos and Markus Endler
Department of Informatics
Pontifical Catholic University of Rio de Janeiro (PUC-Rio)
Rio de Janeiro, Brazil
rvasconcelos@inf.puc-rio.br, endler@inf.puc-rio.br

Berto de T. P. Gomes and Francisco J. da Silva e Silva
Graduate Program Electric Engineering (PPGEE)
Federal University of Maranhão (UFMA)
São Luís, Brazil
bertodetacio@ifma.edu.br, fssilva@deinf.ufma.br

Abstract—A huge number of applications such as network monitoring, traffic engineering systems, intelligent routing of cars, sensor networks, mobile telecommunications, logistics applications and air traffic control require continuous and timely processing of high volume of data originated from many distributed sources as well as mobile communication and monitoring. The deployment and operation of infrastructures enabling such mobile communication and data stream processing have two key requirements: they must be capable of handling large and variable numbers of wireless connections to the monitored mobile nodes regardless of their current use or locations, and must automatically adapt to variations in the volume of the mobile data streams. This article describes the design, implementation, and evaluation of an autonomic mechanism for load balancing data streams and mobile connections. The autonomic capability has been incorporated into a scalable middleware system based on a Data Centric Publish Subscribe approach - using the OMG Data Distribution Service (DDS) standard - and aimed at real-time and adaptive handling of mobile connectivity and data stream processing for large sets of mobile nodes. Several performance evaluation experiments of the proposed infrastructure are presented, demonstrating its viability and the advantages arising from the use of an autonomic approach to handle the requirements of high variability and scalability.

Keywords-Load balancing, Data Stream Processing, Autonomic computing, DDS, Mobile Communication Middleware.

I. INTRODUCTION

A large number of applications require continuous and timely processing of high-volume of data originated from many distributed sources to obtain real-time notifications from complex queries over the steady flow of data items [1] [2] [3] [4]. This has led to a new computing model called Data Stream Processing [3], focused on sustained and timely filtering, aggregation, transformation and analysis such data streams.

The need to process data streams comes from several application areas, such as network monitoring, traffic engineering systems, intelligent routing of cars in metropolitan areas, sensor networks, telecommunication systems, financial applications and meteorology. Crowd-sourcing applications such as Waze [5], collect data from many distributed mobile devices to infer the actual condition of the routes

(e.g., streets and roads) and guide their users - the drivers - towards the best route. This kind of application requires not only the data fusion from a huge set of mobile devices, but also the processing of this data to infer more complex situations (e.g., the local traffic condition and alternative routes to the driver's destination).

Such applications share the requirement of real-time processing of large flows of sensor data (i.e., context data) produced by hundreds of thousands of client nodes, which may be vehicles, aircrafts, mobile devices, computing devices or smart objects, with embedded sensors. Although some kind of data processing can be performed locally at the client nodes, e.g., simple transformations or classification of sensor data, most other context information about the monitored system as a whole requires parallel data processing by sets of dedicated machines (e.g., clusters of processing nodes). This, in turn calls for load balancing solutions [6] [7] [8] for these clusters.

In order to handle large volumes of data streams in a scalable and self-manageable manner, such systems have to be distributed and autonomous. However, using software and hardware that manages itself requires self-management autonomic capabilities to detect and independently react to run-time problems [9]. Thus, there are several challenges in the field of self-managed and adaptive Data Stream Processing for large-scale mobile systems, since it involves timely communication, scalable processing and context-awareness.

A pillar to build self-manageable systems is Autonomic Computing (AC). The main goal is to build computing systems and applications able to manage themselves, thus minimizing human intervention [10] [11] [12] [13]. In order to accomplish the AC challenges, scientific and technological advances in a wide range of fields and system architectures are required, as well as new programming paradigms [14]. On the other hand, today's mobile communication and data stream processing systems lack autonomic features that are necessary to support the large and variable amounts of data flows envisioned by the massive and ubiquitous dissemination of sensors and mobile devices in our modern society. In particular, the deployment and 24x7 operation

of such mobile data stream processing and communication infrastructures pose two intrinsic technical challenges: they must be capable of (i) handling huge and variable numbers of mobile data connections, and (ii) of automatically adapting to variations in the volume of the mobile data streams.

In order to address these challenges, we have developed a scalable middleware system that supports efficient and adaptive handling of mobile connectivity and data stream processing for thousands of mobile nodes. In this paper, we specifically explain the autonomic load distribution mechanisms implemented in the middleware, and discuss their potential benefits. Experiments with large data stream have demonstrated the low overhead, good performance and the reliability of the proposed solution.

The remainder of this paper is organized as follows: Section II presents an overview of the key concepts and technologies which are used throughout this work: the Data Distribution Service (DDS) for Real-Time Systems standard, the MAPE-K reference model for autonomic systems, and load balancing approaches in middleware. Section III presents the Scalable Data Distribution Layer (SDDL), used as the middleware for mobile communications and the MAPE-SDDL extension, which adds autonomic capabilities to the SDDL middleware. Section IV delves into the proposed *Data Processing Slice Load Balancing* approach for mobile data streams and explains how it was implemented, while Section V describes the detailed evaluation of the implemented system using a prototype application. Section VI reviews related work on load balancing for Publish/Subscribe systems, including DDS, while Section VII discusses the advantages of using an Autonomic Computing for load balancing and the benefits of the load balancing solution proposed by this work. Finally, Section VIII contains some concluding remarks about the central ideas presented in this work and mentions some possible lines of future work on the subject.

II. BACKGROUND

This section presents the main concepts and technologies that are used in our work.

A. Data Distribution Service (DDS)

One of the most promising communication infrastructures for the aforementioned data stream processing applications is the Data Distribution Service (DDS) for Real-Time Systems. DDS is an OMG [15] (Object Management Group) standard for Publish-Subscribe communication that aims to provide an efficient and low-latency data distribution middleware for distributed applications [16] [17]. DDS promotes a fully decentralized P2P (Peer-to-Peer) and scalable middleware architecture based on the Data-Centric Publish-Subscribe (DCPS) model. It also supports a large array of Quality of Service (QoS) policies for communication (e.g., best effort, reliable, ownership, several levels of data persistency,

data flow prioritization and several other message delivery options) [18] [19]. Unlike traditional Publish-Subscribe middleware, DDS can explicitly manage network resources through fine-tuning of its Network Services and use of QoS policies such as Deadline, Latency Budget, Transport Priority, etc, that are critical for implementing real-time and soft real-time systems.

Publishers and Subscribers of a DDS Domain (the collection of nodes pertaining to a single application), which are named Participants, are containers for Data Writers and Data Readers, respectively, which exchange typed data through a common Topic [17]. Pardo-Castellote, Farabaugh and Warren [20] explain that Data Writers and Data Readers are the primary point for a Participant to publish data into a DDS Domain or to access data that has been received by a Subscriber. The DCPS makes it possible to organize its Topics in a relational model, providing support for identity and relations, i.e., for each Topic it is possible to define one or more primary keys, and any number of foreign keys representing, respectively, relationships with other Topics.

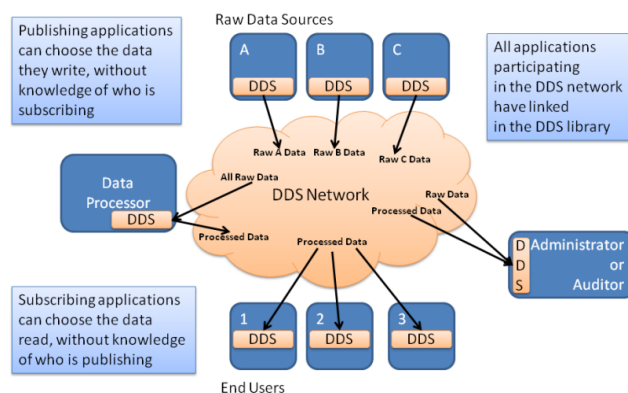


Figure 1. DDS System Architecture [17]

Figure 1 illustrates a hypothetical system that uses DDS as its data distribution middleware. This hypothetical application has some sources of “Raw Data”, a Data Processor that performs some processing on the “Raw Data” to produce “Processed Data”, some End Users that consume the processed data, and an Administrative User performing auditing functions, for instance. DDS supports not only Topic subscriptions, but also content-based subscriptions. The latter are enabled by DDS *Content Filtered Topics*, which holds a *Filter Expression* formed through SQL92 (Structured Query Language). This Filter Expression defines a selective information subscription, i.e., only the topic data that match the Filter Expression are delivered to the Data Reader. An use example of Content Filtered Topic is shown in Figure 2, where a Filter Expression ($\text{Value} > 260$) is applied upon the “Value” field.

The DDS enables applications to filter data based on the content of the data either at the Publisher side (Data

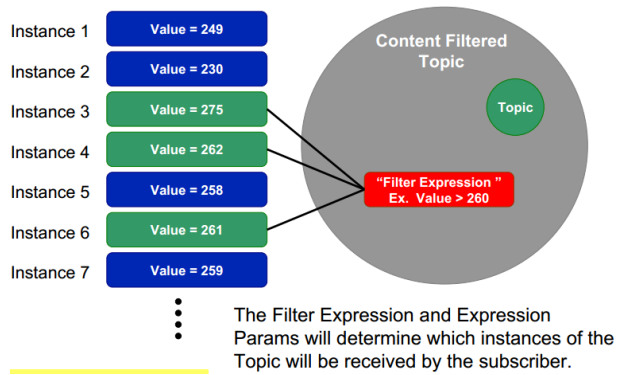


Figure 2. Content Filtered Topic example [21]

Writer) or Subscriber side (Data Reader). By applying filters at the Publisher, some applications can conserve significant network bandwidth by avoiding the network transmission of irrelevant data [22]. Although this capability, for some kinds of application – such as those that have a dynamic and unpredictable number of Publishers and Subscribers – the filtering at the Subscribers is the best choice. In DDS, topics are defined using DDL (Data Definition Language). This language is very similar to the OMG IDL for describing data types. A compiler is then used to translate the DDL type definitions into specific programming language code that is included into an application.

B. MAPE-K Autonomic Architecture

The MAPE-K [23] (*Monitoring, Analysis, Planning, Executing and Knowledge*) model, illustrated in Figure 3, is a general architecture for the development of autonomic software components, which was originally proposed by IBM [9]. This model defines the tasks and the interactions among the main architectural components of autonomic systems. According to MAPE-K, each element in such a system is divided into an autonomic manager and a managed resource.

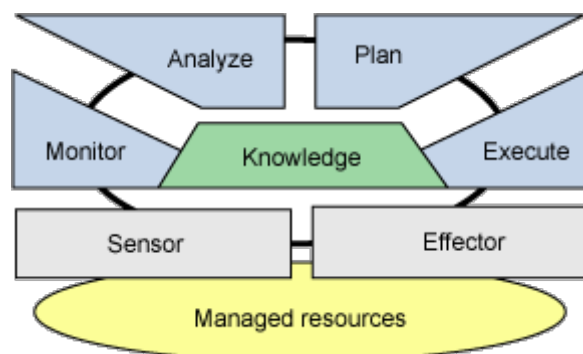


Figure 3. MAPE-K Autonomic Architecture [24]

The managed resource corresponds to the system or some system component providing the business logic that

is to be dynamically adapted as the computing environment changes. The managed resource can be, for instance, a Web server, a database, a software component in a given application (e.g., the query optimizer in a database), an operating system, etc. The autonomic manager performs all the functions comprising the adaptation logic on the managed resource: monitoring, analysis, planning, and adaptation execution. MAPE-K defines two types of access points with the managed resource: sensors and effectors, which are the only means of direct interaction with the managed resource. Sensors are responsible for collecting information from the managed resource. For example, if the managed resource is an application server, this could be for instance, the response times of remote requests from a client. The information collected by the sensors reaches the monitors, where they are interpreted, classified and transformed into higher level information, such as the mean response time distribution for different sorts of requests. This information is then sent to the next stage of the cycle, the analysis and planning phases. This stage produces an action plan, which consists of a set of adaptation actions to be performed by the executor. The effectors are the access points that allow the autonomic managers to perform adjustments or adaptations at the managed resources, such as allocating more/less buffer space or setting some flow control parameter. The decision of which adaptation actions must be applied in a given situation requires a knowledge representation of the computing system, its states and its environment. This knowledge can be represented and processed in different ways (e.g., Ontologies, basic ECA-Rules, machine learning, etc.) and must be shared among the monitoring, analysis, planning and executing components of the autonomic manager.

C. Load Balancing in Middleware

With the widespread and rapid development of Cloud Computing and mobile devices, computational resources have become ubiquitous. Despite the recent technology developments, mobile devices still have more restrictive processing and memory capacity and stringent energy limitations than stationary machines. On the other hand, for several applications one has the option to move some processing tasks from the mobile client side to the server/cluster/cloud side. This shift has several advantages for the application, but also increases the demand for load balancing mechanisms [6] [7] [8] [25], especially at the middleware layer used for communication among the servers and/or cluster nodes.

Load balancing strategies have been classified under a loosely unified set of terms and according to [25], the first classifications came from [26] [27]. Figure 4 depicts the classification proposed by [27]. Following the proposed taxonomy, a load balancing algorithms can be either local or global. Local solutions deal with a single processing node, while global algorithms deal with more than one processing node [25]. A global solution may be divided into static,

when the load balancing algorithm is executed only when there is a new task, and dynamic, which runs the algorithm continuously or periodically. At an operational level, an algorithm may be classified as physically distributed (distributed) or physically non-distributed (centralized). Unlike the centralized approach, in physically distributed load balancing algorithms, the decisions are taken by several nodes. And this decision can be made cooperatively, or non-cooperatively. In the former, the algorithm requires a common agreement among the nodes, while in the latter each node makes a selfish decision. Moreover, according to [28] in the global solution, decisions are made by a single node such as the physically non-distributed defined by [25].

Delving into more details of the load balancing process, a dynamic load balancing algorithm have four main elements that are: (i) Initiation, (ii) Load Balancer Location, (iii) Information Exchange and (iv) Load Selection, shown in Figure 5 [29].

The Initiation policy defines how the current load information is exchanged among the nodes. While in a periodic strategy, information is exchanged at predefined time intervals, an event-driven initiation strategy is based on the local load observation. According to [29], the later strategy better handles load imbalance and has a lower overhead than the periodic strategy when the system load is already balanced.

A designer of load balancing algorithm should choose one of two strategies for the location of the load balancer, i.e., the node in charge of analyzing the system load and deciding whether a load redistribution among the nodes is required. Load Balancer location strategies can be centralized or distributed. Unlike the centralized strategy where a single node evaluates the load of the entire system, a distributed approach has some, or possibility all, nodes responsible for made load balancing decisions.

Because the remaining sub-strategies of the taxonomy shown in Figure 4 and Figure 5 are not of much relevance for this work, we refer to [29] for an in-depth discussion about the characteristics of the other policies. Moreover, the objective of this work is not to propose specific load distribution algorithms, but rather provide general mechanisms that support the implementation of several distributed load balancing algorithms.

III. THE SDDL MIDDLEWARE

A. Overview of the SDDL

The Scalable Data Distribution Layer (SDDL) [19] [30] [31] [32] [33] is a communication middleware that connects stationary nodes running in a DDS Domain and deployed in a cloud to mobile nodes that have an IP-based wireless data connection, as illustrated in Figure 6. Some of the stationary nodes are data stream processing nodes, while others are gateways for the communication with the mobile nodes (MNs). Gateways use the Mobile Reliable UDP (MR-UDP) [19] [32] protocol to maintain a virtual connection

with each MN. The MR-UDP protocol was developed to be robust to short-lived wireless disconnections, IP address changes of the MNs and capable of Firewall/NAT traversal. One of the nodes in the DDS Domain, the Controller, is also a Web Server that can be accessed by a Web browser, for displaying all the MN's current position (or any other node specific information) and for send unicast, broadcast, or groupcast message to the mobile nodes. Figure 6 shows other nodes in SDDL that are Load Balancer, PoA-Manager and Processing Nodes. All nodes showed in Figure 6 will be explained throughout this work.

Taking advantage of DDS' distributed P2P architecture and its highly optimized Real-Time Publish Subscribe wired protocol, SDDL is naturally scalable, i.e., new processing nodes or Gateways can be dynamically added to SDDL's core whenever more MNs have to be served, or new data flow processing is required. In regard to the connections with the MNs, whenever some Gateway is overloaded the data flow to and from a large set of MNs, SDDL is capable of seamlessly migrating a fraction of this set of MNs to a underloaded Gateway. This is possible through a SDDL-internal management node, called the PoA-Manager, which continuously monitors the load of each Gateway - in terms of the number of served MNs - and a Client communication library (CNCLib) at the MNs, which accepts both updates of alternative Gateway addresses and/or commands to reconnect to a new Gateway address, from the PoA-Manager. In spite of the unavoidable mobile disconnection, these handovers between Gateways are very fast and completely transparent to the client applications running on the mobile nodes. On the one side, the messages from the MN are buffered in the CNCLib until the new connection is established, and on the other side, messages addressed to the MN are also temporarily intercepted by a SDDL node and then re-routed to the new Gateway, as soon as it signals the connection establishment.

B. MAPE-SDDL

In order to address general dynamic adaptivity requirements for the SDDL middleware, we decided to extend it with autonomic capabilities. This extension, inspired by the MAPE-K loop, is called MAPE-SDDL. The goal is to support resource monitoring, as well as analysis, planning and execution of dynamic reconfigurations on components of the SDDL middleware. The MAPE-SDDL architecture (at a high level of abstraction) is illustrated in Figure 7. It comprises four services: Monitoring Service (MS), Local Event Service (LES), Analysis and Planning Service (APS), and Control and Executing Service (CES).

1) *Monitoring Service (MS)*: The MS collects data from any SDDL managed resources, such as Gateways and Processing Nodes. The monitoring is applied to properties from these resources, such as: CPU load, amount of memory available, network bandwidth and latency, number of served

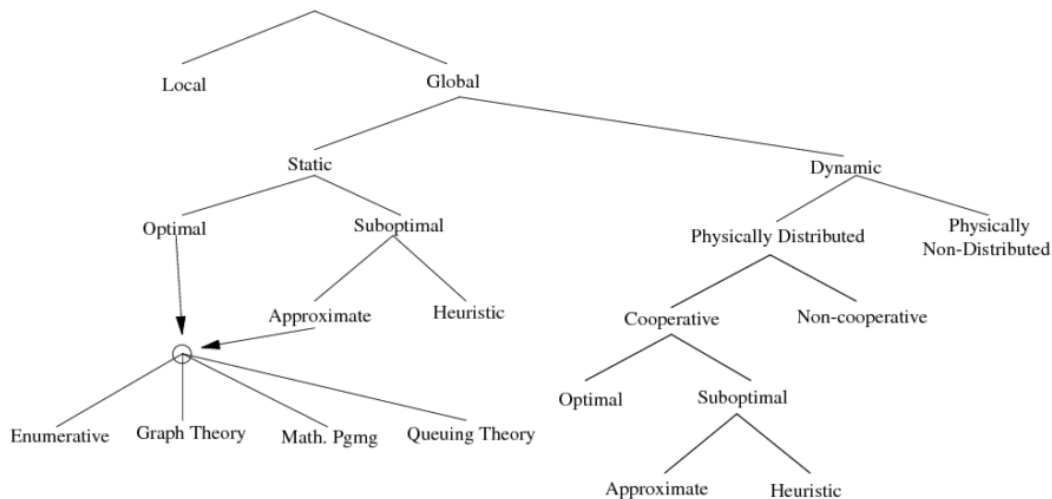


Figure 4. Load balancing hierarchy [27]

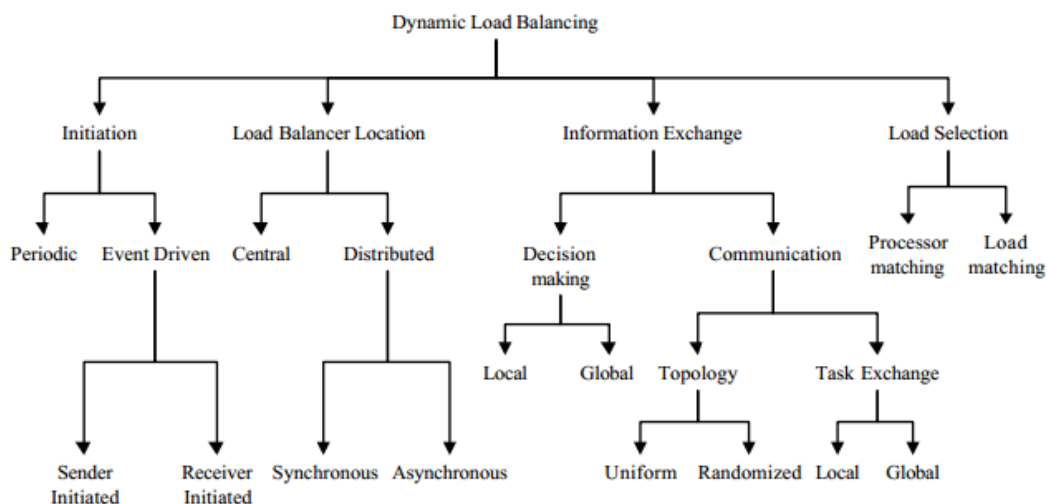


Figure 5. Taxonomy of dynamic load balancing algorithms [29]

MNs (by each Gateway) or number of DPSs assigned to each Processing Node (see DPS concept in Section IV-A). Each Monitor is responsible for a single property. Each property is associated with a set of operation ranges, which are defined by the framework user. For example, one could use the following operation ranges for monitoring the CPU load usage: [0%,30%], [30%,70%] and [70%,100%]. The MS then notifies the LES (Local Event Service) whenever the monitored property switches its operation range, which might indicate a significant change on resource usage. MS and LES are components that run in separate processes, but they are located on the same node. Therefore, unlike other cases in which communication occurs through DDS topics, communication between MS and LES occurs through standart JavaRMI [34].

2) *Local Event Service (LES)*: The LES receives these range change events from the MS and publishes event notifications to subscribed components. Events are occurrences that indicate that a resource availability condition extended itself throughout a specified amount of time, i.e., its duration time. Event evaluation is based on regular expressions written by application developers or operators, as part of each event definition. For an event notification to be triggered, the corresponding expression must remain valid during the specified duration time. This avoids the generation of events when short-lived situations occur (e.g., a CPU load peak on a Processing Node during a few seconds). The publication topic of event notifications is the `Event Notification Topic`, illustrated in Listing 1. This topic is defined by two data structures: the data structure `Property` (line 1) contains the monitored property name (line 3), the

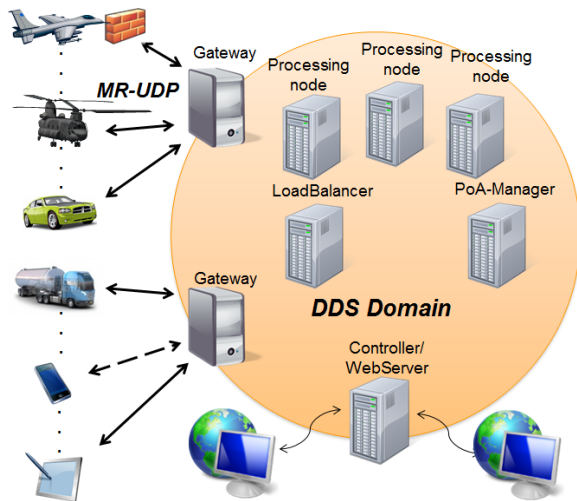


Figure 6. SDDL Architecture

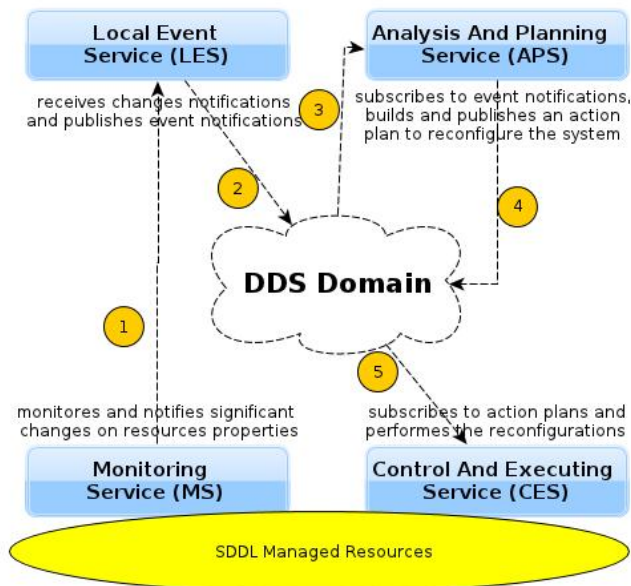


Figure 7. MAPE-SDDL Architecture

property value (line 4), and the timestamp at which the measurement occurred (line 5); the data structure Event Notification (line 8) corresponds to the event itself. The Event Notification contains the event identifier (line 10), the node identifier that triggered the event (line 11), and the set of monitored properties (with their values) that composes the notified event (line 12).

Listing 1. Event Notification Topic (DDL syntax)

```

1 struct Property
2 {
3     string name;
4     double value;
5     string timeStamps;
6 };
7
8 struct EventNotification
9 {
10    string eventID;
11    string sourceID;
12    sequence<Property> propertiesSeq;
13 };

```

3) *Analysis and Planning Service (APS)*: The APS subscribes to the Event Notification topic and analyzes the received notifications identifying eventual problems that requires reconfiguration actions. Mobile connection overload on the Gateways, and unbalanced load between Processing Nodes are examples of problems that are already being diagnosed by the MAPE-SDDL APS. After diagnosis, the APS will compose the dynamic reconfiguration actions to resolve the identified problem, and then build an appropriate action plan. The decision-making algorithm for building the plan is based on user defined rules and uses a rule processing engine. The action plan is a sequence of reconfiguration actions to be executed on SDDL components. The action plan for mobile connectivity management, for instance, takes the form of a mandatory handover request to several mobile nodes (with a new Gateway address list) that is generated by the PoA-Manager, an instance of the APS. Action plans are sent to the CES component through the Action Plan Topic. This topic is defined by two data structures. The data structure Action (line 1) contains the action identifier (line 3), the node identifier that will perform the action (used by CoreDX to filter the DDS message delivery designed for each Processing Node - line 4), and a set of arguments required to perform the action (line 5). In Java, the arguments corresponds to a byte array in order to allow the sending of any serializable object. The data structure ActionPlan (line 8) corresponds to the plan itself, containing the plan identifier (line 10), and the set of actions that comprises this plan (line 11).

Listing 2. Action Plan Topic (DDL syntax)

```

1 struct Action
2 {
3     string actionID;
4     string executorID;
5     sequence<octet> args;
6 };
7
8 struct ActionPlan
9 {
10    string planID;
11    sequence<Action> actionsSeq;
12 };

```

4) *Control and Executing Service (CES)*: Finally, the CES implements the adaptation engine that applies the corresponding reconfiguration actions to SDDL components in response to their availability/load changes. CES is divided into two components: Controller and Executor. The Controller is responsible for managing the execution of the action plan, including the execution order of the reconfiguration actions that must be applied as defined by the APS component. The Executor is responsible for actually executing a given reconfiguration action. Among the dynamic reconfiguration actions currently supported, is the ability of moving DPSs from a Processing Node to another (cf. Section IV-A). The ability of migrating sets of MNs from one Gateway to another (cf. Section III-A) is also implemented by CES, which in this case resides in the mobile Client Lib, that performs the disconnection from one Gateway and the reconnection to the new Gateway.

IV. LOAD BALANCING OF MOBILE DATA STREAMS

A. Proposed Autonomous Approach

This work proposes a load balancing solution for DDS-based systems named Data Processing Slice Load Balancing (DPSLB), illustrated in Figure 8. The key concept of the proposed solution is the Data Processing Slice (DPS), which is the basic unit of load for balancing among server nodes. These nodes will be called Processing Nodes (PNs) throughout the text. The general idea is that each PN has some DPS assigned to it, and that load balancing is equivalent to a redistribution of the total number of DPS among the PNs according to their current load (which is indicated by several metrics, such as CPU and memory utilization).

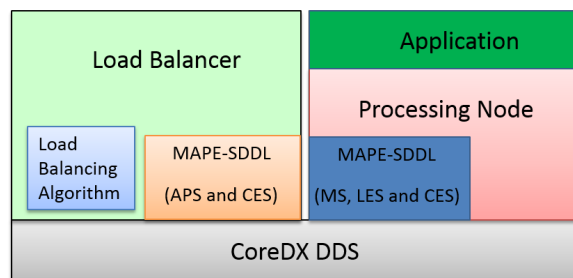


Figure 8. Implementation architecture

The types of DDS nodes that compose the DPSLB approach, showed in Figure 8, are: PNs, which execute the MS, LES and the CES Executor; and the Load Balancer, which executes the APS and the CES Controller of the MAPE-SDDL architecture. The Load Balancer is responsible for monitoring the load of PNs, generating the actions to redistribute the system's workload when an unbalance is detected and controlling the actions executed by PNs to move DPSs between them. The Load Balancer has a module, Load Balancing Algorithm, which may execute any global

algorithm that analyzes the load of the PNs, decides if the system is unbalanced and performs the corrective actions.

The DPSLB solution was designed to Pub/Sub systems that supports content-based subscriptions and are brokerless, i.e., Pub/Sub systems that do not have brokers and employ a fully decentralized P2P architecture such as the DDS standard. The data items produced by the Publisher Client Nodes are delivered directly to the Processing Nodes without the need of centralized elements such as brokers. Thus, the Processing Nodes (PNs) are the Subscribers in charge of processing the data items instead of brokers that route the data items to other elements. It is expected that the proposed solution will be deployed in systems with thousands of Processing Nodes and hundreds of thousands of Client Nodes and a data production rate estimated of dozens of gigabits per second.

In its current conception, the DPS Load Balancing supports applications where each data item is processed independently of any other item. This limitation comes from the way that processing load is distributed among PNs: through the application of disjoint subscription filters. Since a Processing Node does not receive all data items published on the DDS Domain, PN may be unable to process a data item "A" that depends on data item "B" delivered to and processed by another Processing Node. Hence, the proposed load balancing solution is tailored for data-parallel applications, i.e., where each data item is processed independently of other items, and data items can be processed out of order by any Processing Node.

As mentioned, the proposed solution relies on the concept of DPS, or simply, Slice, which represents a percentage of the total system workload being processed by the PNs. Every data item of the data stream (e.g., produced by a mobile node) must be assigned to a single DPS, in order to be processed by some PN. If a data item has no associated Slice, it will not be delivered to a PN for processing. Each Slice is logical identified by a unique numeric ID (identifier), commonly in a range between zero and the total number of defined Slices, minus one. Thus, the DDS Topic carrying application data produced by the publishing nodes has a specific numeric field holding the Slice-ID assigned to each data item.

Unlike Virtual Servers [35] [36] [37], Slices do not behave as new PNs – as this would increase the system overhead since each Virtual Server is a node monitored and managed by the Load Balancer, which increases the CPU, memory and network overheads because more software components are instantiated – but only as a logical partition of the global data volume. The arbitrary assignment of data items to slice IDs enables the choice of load distribution with different granularities (i.e., coarse-grained or fine-grained load distribution). Because the global data item space is partitioned into the set of Slices, a higher amount of Slices allows to split the workload in smaller portions (fine-

grained), and a small amount of Slices means coarse-grained workload distribution. This problem, “balls into bins”, is better explained by Martin Raab and Angelika Steger [38]. The number of DPS is also a upper bound for the maximum quantity of PN, since each PN needs at least one Slice to get involved into the DPSLB.

The Attribution Function is responsible for choosing a valid DPS for each produced data item. The DPSLB solution requires the Attribution Function to be a very low cost function, since it has to compute/choose a DPS for each produced data item, and this data will probably be produced at a very high rate. This function may be a hash function applied to a field of the data item, to the data producer's ID, or a random value. A good candidate function for this is modulo operator (remainder of division). Attribution Function does not have to ensure that the data items are uniformly distributed over the total set of Slices since the workload can be balanced by re-arranging the number of Slices assigned to each PN.

Figure 9 illustrates an Attribution Function that consists of a hash function that applies the modulo operator to the Sensor ID field, in a configuration with ten Slices. Hence, each data item computed by the Attribution Function is assigned to one Slice. As shown, Sensor IDs 21, 1, 52 and 19 are mapped to Slice IDs 1, 1, 2 and 9, respectively. The Attribution Function must be called before the data item is published in the DDS Domain.

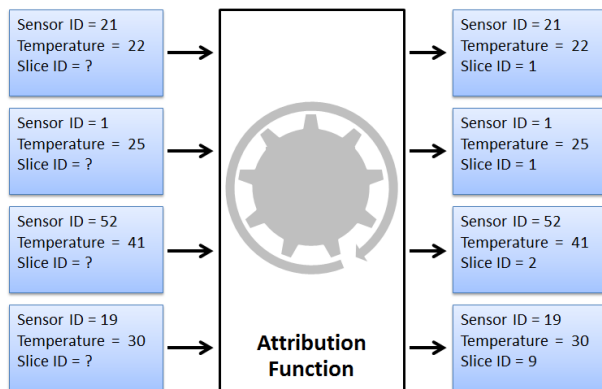


Figure 9. An example of Assignment Function applied upon data item

In our context, Load Balancing is the process of moving Slices from a PN to another. The process is started when the Load Balancer detects a load unbalance of the system and decides that some DPS should be moved to a different PN to reach a better performance. During this process both PNs involved, i.e., the Slice-giving and the Slice-taking PN, must work in a coordinated manner so to guarantee that all data items are processed, and only by one of the PNs.

The Load Balancer plays the role of coordinator of the reconfiguration actions to be executed in the Load Balancing Process, which are effectively executed by the CES

Executor component running in the overloaded and the underloaded PNs. The algorithm within the Load Balancer has to inform which are the Slice-giving and the Slice-taking PNs and how many Slices should be moved among these PNs, thus starting the Load Balancing Process. The Load Balancing Algorithm is a generic module that can be implemented using many algorithms. This module is notified about new PNs that are able to join the DPSLB solution and called when the Load Balancer needs to analyze the system workload. After being called, the algorithm has to classify the PNs and inform how many Slices should be moved from overloaded nodes (Slice-giving) to underloaded nodes (Slice-taking). With this information, the Load Balancer is able to generate and send the corresponding commands to PNs.

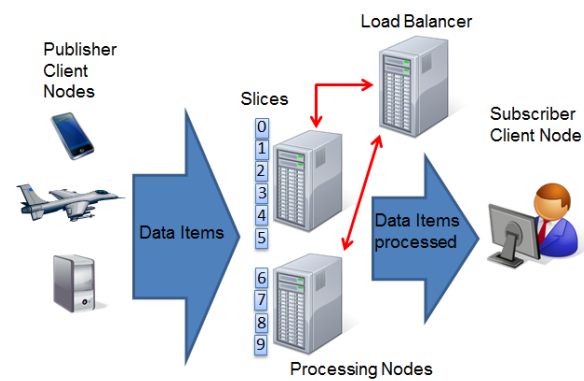


Figure 10. Interactions between clients, PNs and Load Balancer

Figure 10 illustrates the interactions between the nodes that compose the DPSLB Solution. Data items produced by Publisher Client Nodes are processed by PNs and Subscriber Client Nodes receive the processed data from PNs. The Load Balancer interacts only with PNs: both to gather their current workload and to send the load distribution actions to the corresponding PNs (depicted as red arrows in Figure 10). Figure 11 shows the redirection of the data stream when DPS-5 is moved from PN A to PN B. During the Load Balancing Process both PNs receive the data items of DPS-5, but initially none of them will process the data from this DPS. Instead, they store these received data in their local caches. Then, PN A sends its cached items to B. After receiving A's cached items, PN B has to identify the data items that appear in both caches and then generate a Merged Cache, which contains all data items of DPS-5 without duplicates. Finally, DPSLB layer on B is able to notify application about the data items in the Merged Cache. The actions executed during the Load Balancing Process ensure that there is neither data item loss nor data item processed more than once.

If there are more than two PNs involved in the Load Balancing Process, the Load Balancer starts one Load Balancing Session for each pair of Slice-giving and Slice-taking PNs.

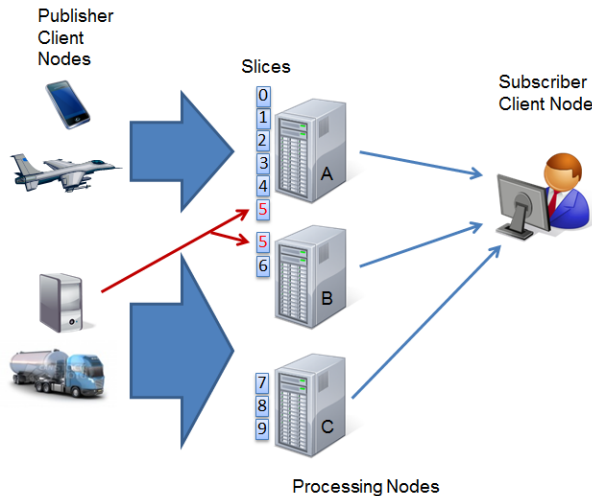


Figure 11. Data flow during Load Balancing Process

For instance, if there are one Slice-giving PN (PN A) and two Slice-taking PNs (PNs B and C), the Load Balancer starts one Load Balancing Session for PNs A and B and after this Load Balancing Session finishes, it starts the second one with PNs A and C. Each Load Balancing Session involves only two PNs. The Load Balancing processes permits not only use new PNs to increase the system's resources but also to reduce it when some PNs are idle, which enables the system have an elasticity of resources. To do so, all Slices assigned to an idle PN should be moved to another PN before the idle PN can leave the system.

B. Implementation

1) *Load Balancer*: Roughly speaking, the Load Balancer receives event notifications through MAPE-SDDL, analyzes them, possibly generates an action plan and sends the corresponding commands through MAPE-SDDL to the Processing Nodes involved in the defined reconfiguration actions. An action plan is generated and sent in response of a detection of unbalance load.

The Load Balancing Algorithm executes the logic for analyzing the system load, deciding which will be the Slice-giving and Slice-taking PNs and how many Slice should be moved from the first to the latter. The Load Balancing Algorithm must implement the `LoadBalancingAlgorithm` interface that consists of two methods: `onNewProcessingNode()` and `analyzeLoad()`. The method `onNewProcessingNode()` is called when the Load Balancer detects that a new PN arrived in the system and `analyzeLoad()` is called every time that a new event notification is received from a PN. This last method returns a collection of Slice-Movement objects, which contains the Slice-giving and Slice-taking PN and how many Slices should be moved to the Slice-taker.

2) *Processing Node*: The PN is the managed resource from the perspective of the MAPE-K model used in the developed DPSLB prototype. In addition to the data processing, which is intrinsically determined by the application build upon DPSLB, each PN periodically verifies its monitored properties and, depending of their operation ranges, notifies the LES that evaluates these values against the specified expression. Hence, LES eventually sends a event notification, which holds all monitored data, to the Load Balancer through APS. The current version of this prototype periodically checks the PN's monitored properties every two seconds and then sends a event notification to the Load Balancer.

When a PN receives a data item (or a sample in DDS jargon), it checks whether the data item is assigned to a Slice that it is responsible for. If this is the case, the PN notifies the application through the `onNewData()` method.

3) *CES and Load Balancing Process*: CES is the adaptation engine that enables PNs to receive actions for moving Slices as a consequence of a load redistribution action plan. The actions supported by the PN are: `addSlice`, `removeSlice`, `updateSliceState` and `sendCacheToNode`. `RemoveSlice` is used to set a Slice to the Not In Use state, which means that data items assigned to it can be discarded by the PN because another PN is processing these data items. On the other hand, `addSlice` changes a Slice to the Available state, meaning that the PN is responsible for processing the data assigned to the Slice. Therefore, the `addSlice` and `removeSlice` actions do not actually add or remove a Slice, but only change the Slice state. The action `updateSliceState` changes the Slice state to In Load Balancing Session, which will be hereafter explained.

During a Load Balancing Process, both Slice-giving and Slice-taking PNs should update the state of the involved Slices to In Load Balancing Session. After the Slice-giving PN updates and removes the Slices, the Slice-taking PN can proceed with the update action. The specific sequence of actions sent by the Load Balancer to move a DPS between two PNs are: (i) Update the DPS's state at A to In Load Balancing Session in order to cache the new data items received, (ii) Add the DPS at B with In Load Balancing Session state in order to start caching the data items, (iii) Remove DPS at A to inform that A can discard the new data items received, (iv) Update DPS's state at B to Available to inform that B can process the new data items and (v) Send cache from A to B. After this, B will generate and process the Merged Cache, and A will continue to process the data of its other Slices. The add and remove actions determine if the corresponding data items are delivered or not, respectively, to a node in a DDS Domain. This is possible by a dynamic adjustment of the subscriber filters.

The data items of a Slice cache are sent through a DDS

Topic named *CacheTopic*. The *CacheTopic* carries fields to inform the Slice ID, Slice-giving PN ID, Slice-taking PN ID and the data items, which are serialized into a byte array. The Slice-taking PN, after receiving a *CacheTopic* sample, deserializes the data items and gets its Slice local cache. With both local and remote caches, the Slice-taking PN uses a Java Set in order to generate the Merged Cache, which is a set that has no duplicated data items. After generating the Merged Cache, the Slice-taking PN removes it from the local cache and delivers the data items to the application through the data items' application data reader listeners.

V. EVALUATION

Initial dynamic adaptation tests performed with the MAPE-SDDL middleware have already shown encouraging performance results. Regarding MAPE-SDDL's connectivity load balancing, we did the following test: We initially connected 600 simulated mobile nodes (MNs) to one Gateway, and then activated a new 'empty' Gateway. After a while, the PoA-Manager identified a load unbalance, and requested half of the MNs to migrate simultaneously to the new Gateway. At this bulk handover, all 300 MNs were able to reconnect at the new Gateway in less than 750 ms and none of the data items produced regularly (every 10 seconds) by each of the MNs was lost.

In order to evaluate the DPSLB solution and its implementation, we also developed a prototype application that utilizes the DPSLB prototype for balancing of its data processing load. This prototype application consists of clients that publish color images into the DDS domain, and PNs that receive the images, convert them to grayscale and, thereafter inform the corresponding client about completion of the image processing. Both communication paths happen through two DDS Topics.

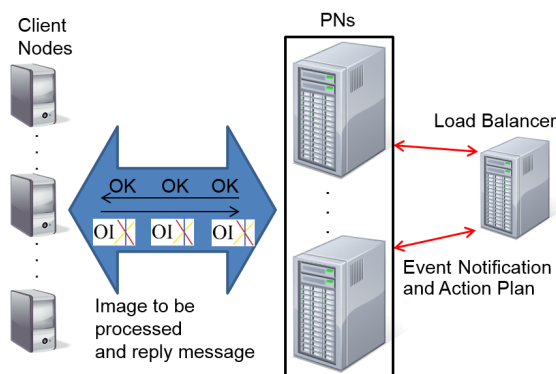


Figure 12. Deployment of the evaluation application

Figure 12 illustrates the deployment of the prototype application used for evaluation. Clients publish images through the *ClientTopic* and PN servers reply with completion notifications published into the *ServerTopic*,

which are shown in Figure 13. The *ClientTopic* has the fields: *sliceId* (required by DPSLB to produce the merged cache); *id* of the data item; *senderId* to identify the client; *timestamp* to inform the data item creation time and message, that carries the serialized image. The *ServerTopic* holds fields: the data item *id*, *timestamp*, *senderId* and message, which carries the reply message, a serialized Java String, such as "Processed". Although this message could as well carry the result image (grayscale), this application prototype sends only a "OK" message, since the content and size of the reply message is irrelevant for evaluating the DPSLB solution. The Load Balancer analyzes the load of PNs and, transparently to the application, balances their image processing workload. It is important to stress that there is no communication, neither directly nor indirectly, between clients and the Load Balancer. Hence, the load generated by clients does not affect the Load Balancer, only the PNs.

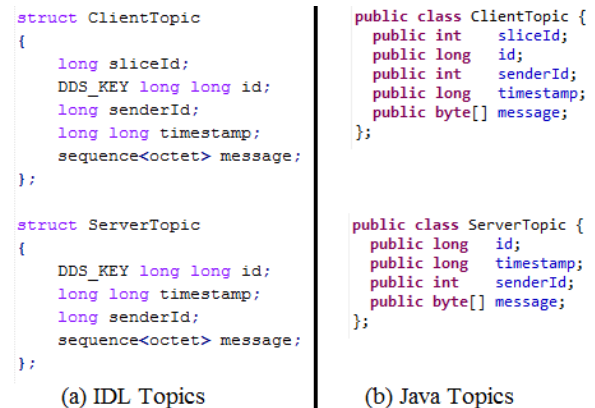


Figure 13. Evaluation application topics

Since the image processing done by this evaluation has no restrictions with the delivered order and dependency between each image that is processed (i.e., there is no relationship between the images published by the same client), this application may not be classified as data stream processing. However, the processing done by the application layer is totally independent of the Processing Node layer since it just delivers the data items to the application layer. Moreover, this processing task demands high CPU utilization and serves to validate that the DPSLB solution is able to effectively distribute the load among the PNs without result in data item loss/duplication.

DPSLB prototype was tested with data/image publication rates starting from 160 (1.4 MB/s) up to 1,365 (10 MB/s) data items per second. The Attribution Function of choice was the modulo operator applied on the *id* field, and the number of available slices was chosen to be 10. The setup used for the experiment, as illustrated in Figure 14, was the following: 5 PNs, one Load Balancer and a Client simulator deployed on three physical machines (PMs) executing in a

LAN with bandwidth of 100 Mbps. Each PN executed on a dedicated virtual machine (VM) running the Ubuntu 12.04 32-bit Operating System, configured to use one CPU core and 512 MB. The three physical machines had following configurations: Intel i5 4 x 2.66 GHz, 8 GB DDR3 1333 MHz running Windows 7 64-bit; Intel i5 4 x 3.1 GHz, 8 GB DDR3 1333 MHz running Fedora 15 64-bit; and Intel Dual-Core 4 x 2.66, 8 GB DDR2 667 MHz running Mac OS X 10.7.5. The chosen virtualization product was Oracle VM VirtualBox since it is free and has cross-platform support.

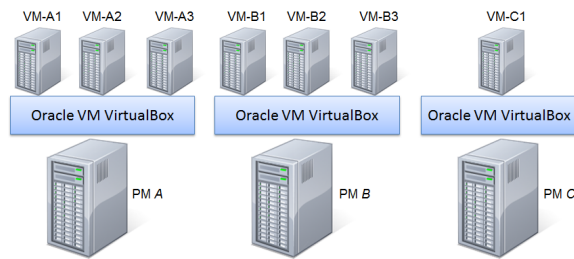


Figure 14. Deployment of the Virtual Machines

The Load Balancer and PN on VM-A1 were initiated before the evaluation starts. At initial time 0s (second zero) the Client Node was started with a data production rate of 1.4 MB/s. After 18 seconds the second PN on VM-A2 was added to the system; at 25s the third PN on VM-B1 was detected by the Load Balancer; at 35s the fourth PN on VM-B2 arrived and at 45s a fifth PN on VM-B3 joined the system. Finally, at 59s the data produced by the Client Node was increased from 1.4 MB/s to 10 MB/s. The evaluation was finalized at time 85s.

A. Throughput

The throughput metric – expressed in data items per second (DI/s) – was used to demonstrate that an increase of the set of PNs leads to an increase of the system's processing capacity, as expected. This metric was collected at the client side and the throughput in an instant of time represents the amount of reply messages received at the specified instant of time from all PNs.

Figure 15 shows that the system throughput increases by a nearly equal amount whenever a new PNs arrives at the system. The vertical red lines indicate the point of time when a new PN joined the system, and the green arrow indicates the time when the data production rate was increased to 10 MB/s. The throughput started from 40 DI/s and reached up to 323 DI/s at 72s. With five PNs and a data production rate of 1.4 MB/s, the system was able to process up to 245 DI/s and the mean was 240 DI/s. Finally, when the client node augmented its production rate to 10 MB/s, at 59s, the throughput experienced a fall to 163 DI/s and after 6s reached 283 DI/s. From 66s until the end of the evaluation, the throughput had an average of 317 DI/s. Immediately after

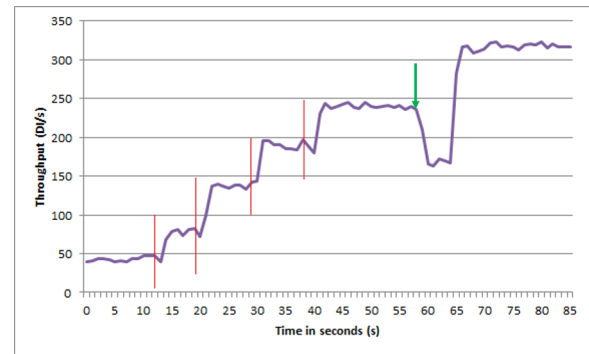


Figure 15. Throughput over the time of the experiment (DI/s X seconds)

a new PN joins the system, the throughput suffers a small retraction and after 1 second the system achieves a higher throughput level.

The decrease of the throughput at 59s may be explained by the fact that the network and the DDS middleware had to deal with a sudden burst of the data production rate. In order to achieve better throughput and reduce the CPU and network overheads, DDS can aggregate many samples (a.k.a. data items) into a single packet and send this single packet, instead of sending many small data samples, which helps to increase the latency to send the data items and consequently decrease the throughput. Another noteworthy issue is that in this test the data production rate almost reached the theoretical network bandwidth of 100 MB/s.

It is important notice that the throughput grows almost proportionally to added processing capabilities of the PNs. Specifically in this evaluation, the major capability is CPU speed, as image processing requires most resources in CPU throughput.

B. CPU Usage

The CPU usage shows that, using the modulo operator as Attribution Function, the DPSLB solution effectively achieves an even distribution of the data items over the PNs and that this data flow drives to an equal increase of the CPU usage (expressed in percentage (%)). The CPU usage was collected at each PN.

Analyzing Figure 16, it is possible to notice that as soon as the PN becomes active, its CPU usage goes up to a value higher than 90% and all PNs have a small CPU usage difference from each other. As expected, the system's load (mean load) is increased whenever a new PN arrives in the system and starts processing data items. When the data production rate was increased, at instant 59s, the system load fell from 92% to 85%, together with the throughput, but after 6s went again up to 97%.

This momentary decrease on the CPU usage probably shares the same explanation given for the throughput dip: a sudden burst of data traffic. The fall on the CPU usage

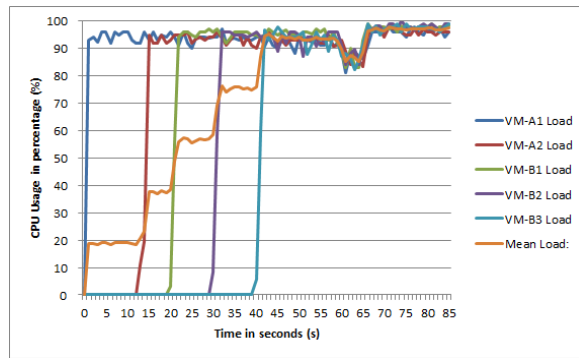


Figure 16. CPU usage over the time of the experiment (% X seconds))

suggests that it was caused by a bottleneck at the network and DDS communication layer.

C. Round-trip Delay

The round-trip delay (RTD), or Round-trip Time, is measured in seconds (s), and encompasses the time interval from the instant a client sends a data item until it receives an acknowledgment informing that the data item was successfully processed by a PN. The RTD was collected at the client side and the RTD in an instant of time that represents the mean RTD of all data items processed by all PNs at the specified instant of time. An increase of the RTD may indicate that the system is receiving more data items than it is able to process.

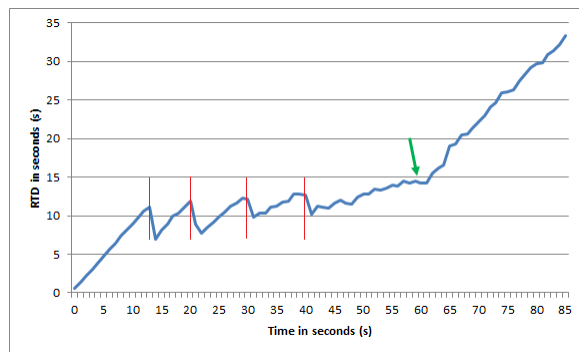


Figure 17. Round-trip Delay over the time of the experiment (RTD in seconds X time in seconds)

The RTD during the experiment is shown in Figure 17, where the vertical red lines indicate when a new PN joined the system, as those in Figure 15, and the green arrow indicates the time when the data production rate was increased to 10 MB/s. This chart reveals that the data production rate is higher than the processing capacity of the system since the RTD increases. It also shows sudden drops of the RTD whenever new PNs arrived on the system. This phenomenon can be explained by the fact that a new PN has no data items on its queue, so that the first data items

it processes have a low RTD, which in turn helps to reduce the mean RTD. But after a while the data items are queued also at the new PN because it is not able to process them at the rate that they are delivered, and thus, the RTD keeps increasing.

In spite of the steady increase of the mean RTD, it is possible to observe from Figure 17 that, after instant 40s, the RTD begins to have a smoother increase: i.e., from 0s to 5s, where there was one PN, the RTD increased by approximately five seconds, while between 40s and 60s, when all 5 PNs had joined the system, the RTD increased by less than five seconds. But starting at 59s, as a result of the increase of the rate of published data items, the RTD started again rising faster than in the interval between 40s and 60s, which again is due to the insufficient processing capacity of the system against the high rate of data item production.

D. Overhead

To assess the Load Balancing overhead, we compared the throughput and the mean round-trip delay of the same image processing application in two configurations: using the DPSLB solution and without Load Balancing support. The overhead of the DPSLB solution was expressed by percentages (%) of the throughput loss, and the mean RTD increase, respectively. To evaluate the DPSLB overhead, 10,000 data items were produced with a data production rate of 1,150 data items per second (DI/s).

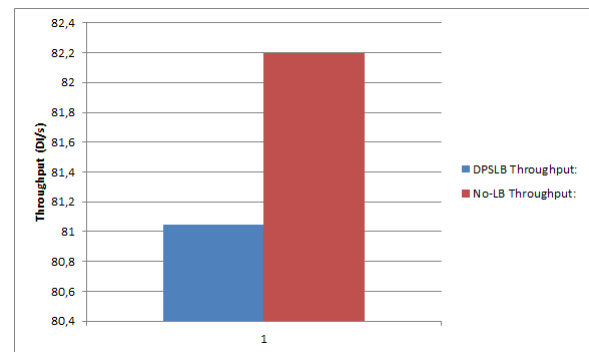


Figure 18. Mean throughput (DI/s) comparison among DPSLB solution and another without Load Balancing)

The application using DPSLB was able to process 81.044 DI/s and the application without any load balancing support, was able to process 82.194 DI/s, as shown in Figure 18. These numbers show an overhead of 1.4% introduced by the DPSLB implementation. Regarding to the RTD, shown in Figure 19, the application using DPSLB had a mean RTD of 60.45 seconds, while the application without DPSLB delivered a mean RTD of 59.51 s. This difference represents an increase of 1.58% on the RTD.

The mean time required to complete a Load Balancing Process with a data production rate of 10 MB/s and ten slices was 454 ms. While in Load Balancing Process, the DPSLB

Table I
IMPACT OF THE LOAD BALANCING PROCESS ON RTD AND THROUGHPUT

| | 2 PNs | 3 PNs | 4 PNs | 5 PNs |
|--------------------------|----------|----------|----------|----------|
| RTD before | 10.610 s | 11.159 s | 12.335 s | 12.856 s |
| RTD during | 11.115 s | 11.944 s | 12.145 s | 12.652 s |
| RTD after | 6.940 s | 8.856 s | 9.838 s | 10.227 s |
| Throughput before | 48 DI/s | 83 DI/s | 142 DI/s | 189 DI/s |
| Throughput during | 40 DI/s | 72 DI/s | 143 DI/s | 180 DI/s |
| Throughput after | 68 DI/s | 100 DI/s | 195 DI/s | 231 DI/s |

prototype resulted in a mean CPU overhead of 1.4% when analyzing the CPU usage on the PN that gives some slices to another. We believe that the overhead introduced by the DPSLB solution has a low cost when compared with the benefits that it introduces. Both throughput loss and RTD increase are lower than 1.6%, which seems a reasonable overhead in return of Load Balancing support.

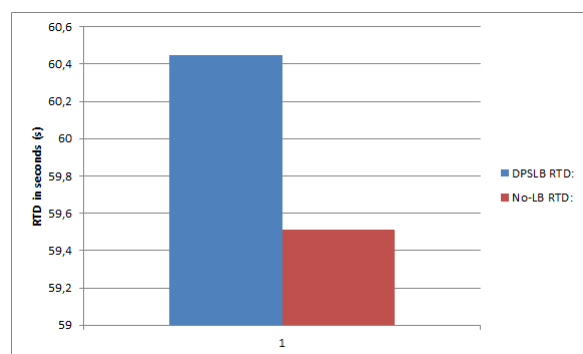


Figure 19. Mean Round-trip Delay comparison among DPSLB solution and another without Load Balancing)

The Load Balancing Process Overhead tries to capture the impact of the Load Balancing Process on the system's throughput and mean RTD. The RTD and throughput before, during and after the Load Balancing Process are shown in Table I. The columns 2, 3, 4 and 5 PNs show the number of PNs participating in the Load Balancing Process.

When the second, third, fourth and fifth PNs arrived, the RTD was increased by 4.76% and 7.035% and decreased by 1.54% and 1.58%, respectively. The mean of the RTD overhead for all these four load balancing situations was therefore 2.167%. When a Slice-Taking PN receives Slices from two Slice-Giving PNs, the Slice-Taking PN has a Load Balancing Session for each Slice-Giving PN, which are sequentially executed. Thus, as soon as a Load Balancing Session is over, the Slice-Giving PN keeps running normally and the Slice-Taking PN is able to process data items that are assigned to the Slices received from the Slice-Giving PN. This behavior allows the PNs to start processing data items as soon as Load Balancing Session is completed and, hence, do not contribute to an increase the RTD.

Table II
LOAD BALANCING PROCESS OVERHEAD FOR DIFFERENT THE NUMBERS OF SLICES AND DATA ITEM PRODUCTION RATES

| | 10 Slices | 100 Slices | 1,000 Slices |
|-----------------|-----------|------------|--------------|
| 1.4 MB/s | 401 ms | 422 ms | 432 ms |
| 4 MB/s | 406 ms | 433 ms | 450 ms |
| 10 MB/s | 454 ms | 479 ms | 491 ms |

Analyzing the throughput versus the arrival of new PNs, the throughput was decreased by 16.667%, 13.253% and 4.762% when the second, third and fifth PNs joined the system, and increased by 0.704% when the fourth PN joined, which represents a mean overhead of 8.494%. However, there is a trend towards lower overheads as more PNs join the system since the overhead starts by 16.667% till 4.762% when the second and fifth PNs joined the system, respectively. The higher throughput when the fourth PN arrived may have occurred because the Load Balancing Process involved only a single PN that was already active, which could help to maintain the throughput almost stabilized.

In order to measure the influence of the number of Slice and the data production rate on the Load Balancing Process performance, the number of Slices available was increased from 10 to 100 and 1,000 and the data production rate from 1.4 MB/s to 4 MB/s and 10 MB/s. From Table II it is possible notice that the data production rate has a higher impact on the overhead than the number of Slices. When the data production rate was increased by a factor of 10, the time required to complete the Load Balancing Process increased by in 13.217%. On the other hand, by increasing 10 and 100 times the number of Slices, this only augmented the Load Balancing Process time by 5.237% and 7.73%, respectively. This behavior suggests that the network saturation has a greater impact on the Load Balancing Process overhead than the increase of the number of Slices.

VI. RELATED WORK

There is much research and development of autonomic load balancing in middleware for distributed systems, but to the best of our knowledge, there is no other work that leverages the benefits of the MAPE-K model for dynamic adaptiveness in DDS-based systems, and more specifically, proposes a load balancing approach for mobile data stream processing that is reliable, efficient and flexible.

A common load balancing solution applied on Web Servers, cloud computing and clusters is based on centralized dispatcher [39] [40] [41] [42] [43] [44] [45] where all data stream or requests go through the dispatcher, which chooses one server node to process a set of the data stream or to accept the client request. It is important stress that this approach has a centralized load balancer that is a bottleneck and is not reasonable on Pub/Sub systems.

While in Pub/Sub systems nodes with the same subscription receive the same data, in the proposed load balancing solution PNs – which are homogeneous and have the same subscription – do not receive the same data, instead each data is delivered to a single PN in order to produce a data stream flow. To do so, the proposed solution manages how data are routed by DDS to PNs, which is simpler than routing problem in Pub/Sub systems. However, the novel proposal, [46] [47] [6] propose load balancing mechanisms for distributed systems, either for the routing layer or the data processing layer.

The work by Cheung et al. [46] has developed a load balancing mechanism to balance the subscription load among brokers on the Padres Pub/Sub system [48], where publishers or subscribers may freely migrate among brokers. While [46] focuses on the routing layer for a broker-centered Pub/Sub system and clients (publishers or subscribers) are impelled to change their *brokers* for data flow load balancing, DPSLB solution is based on DDS' P2P architecture for balancing the load among PNs.

REVENGE [47] is a DDS-compliant infrastructure for news dispatching among mobile nodes and that is capable of transparently balancing the data distribution load within the DDS network. In the same way, [47] only load balances the routing substrate, while MAPE-SDDL is able to load balance the mobile connections via PoA-Manager and PNs via Load Balancer in the DPSLB.

In [6], a non-coordinated load balancing approach that relies on *magnetic fields* is proposed: the idea is that underloaded nodes attract data from overloaded nodes. In an completely opposite way, the Load Balancer in DPSLB performs the MAPE-K tasks of Analysis, Planning and Execution, and carefully synchronizes the re-allocation of Data Processing Slices from one PN to another. This has the advantage of a more efficient and reliable load balancing, but the drawback of the dependability of the Load Balancer.

One of the most remarkable differences between [46] and this work is that Cheung and Jacobsen work with heterogeneous client nodes that have to receive all data that match their subscriptions. In a apposite way, the DPSLB works with homogeneous server nodes that have the same subscription but should not receive the same data. While [46] focuses on balancing the Brokers's load by migrating clients (publishers or subscribers) to other Brokers, the proposed solution by this work relies on DDS' P2P architecture for balancing the data flow processing load among PNs rather than balancing the subscription and dissemination loads, which is transparently done by DDS and SDDL.

Similarly to Cheung and Jacobsen's work, in REVENGE [47] load balancing is focused only in the routing of subscriptions and notifications, rather than load balancing the data processing load. Unlike REVENGE, this work proposes a solution to balance the load on PNs so as to enable the deployment of new services that require a great amount of

computational resources. To achieve fault tolerance and a better load balancing on the routing layer, REVENGE works with the concept of multi-domain communication and "hot copies" of the routing substrate. Our work neither works with multi-domains nor have capabilities to support fault tolerance on its load balancing.

Magnetic Field [6] approach is a decentralized and not coordinated load balancing where there is not a Load Balancer. Thus, the nodes communicate with each other to build the magnetization network. Our work, on the other hand, is a coordinated and global load balancing where Load Balancer is in charge of gathering load information about nodes and making and managing the load redistribution actions among nodes. Differently from the message attractions in magnetic fields, our approach selects a single PN that must process each data (message). To do so, the proposed solution manages the data routing done by DDS.

While Calsavara and Lima Jr's approach [6] relies on the attraction of messages through the magnetization relationship, this works proposes an approach that relies on slices, which is expected to provide an efficient load balancing system that is able to directly delivery messages (data) to the appropriate nodes on DDS-based systems.

The main advantage of employ coordinated and global load balancing is that better analyzes and decisions can be done since the Load Balancer is able to gather information about all PNs after take any decision. Since the Load Balancer act also as a coordinator for the PNs during a Load Balancing Process, there is no need for complex autonomic algorithms and leader election at the PNs to decide which actions should be executed by each PN to realize the load balancing without conflicts. The clear disadvantage of this approach is that the Load Balancer may be a point of failure and bottleneck for the system's scalability when there are hundreds of thousands of PNs since the Load Balancer has to analyze the load of all PNs.

Although there are many other load balancing approaches that are found both in academia and industry, none of them explores the capability of the node to receive all data published in a DDS Domain without the need of Brokers or a central dispatcher. In order to effectively realize a processing load balancing in a DDS Domain, a possible approach is simply managing the subscription filters so to control the data stream processing. Therefore, all data routing and its optimization is responsibility of DDS.

VII. DISCUSSION

This paper proposed a novel approach to load balancing mobile connections and data streams based on the MAPE-K model, which entails several advantages that go far beyond a simple boost of performance. Most current load balancing methods are quite inflexible, since they always make same sorts of decision, without considering that the system may require different load distribution approaches depending

on the current state of data stream processing (high/low load) or the state of the infra-structure (e.g., node failures, communication failures, re-organization, etc.). Moreover, by being disassociated from any Autonomic architecture, traditional load balancing mechanisms fail to incorporate self-monitoring, self-analysis and self-adaptation behavior as response to changes in the execution environment.

Since our load balancing mechanism is structured according to the MAPE-K model it is able to deliver sustained load balancing performance under various conditions. For example, based on the information collected by MS, CES is capable of adjusting parameters of the load balancing algorithm directly (i.e., parametric adaptation). For other changes of conditions, CES may even substitute the load balancing algorithm by a more effective one. Adaptation can also be used to circumvent failures of PNs and Gateways. For these cases, the APS can choose the best parameters, algorithms or techniques for handling outage of failed elements, and recovery actions. Finally, it is also possible to implement a machine learning technique in the APS, which would allow the load balancing mechanism to anticipate future change demands, and thus react in a more effective and efficient way. This knowledge about past behavior and adaptation performance of the system would have to be represented and analyzed by the adaptation logic, which is a feature made possible by the MAPE-K model.

Furthermore, our load balancing approach for Data Stream Processing is targeted at DDS-based systems, which support fully decentralized system architectures. It is a generic solution, transparent to the SDDL applications, able to route data streams to PNs with low overhead and is inherently scalable, i.e., it supports large numbers of nodes and large-volume data streams. Also, since PNs can dynamically join or leave the system during operation, DPSLB supports seamless variations of computational resources.

The DPSLB Solution works with any type of application object/message and is totally transparent to the application developers, who must only inform which DDS Topics are subject to load balancing by DPSLB. They can still customize their applications with the DDS QoS policies of their choice. For example, fault tolerance can be achieved by deploying replicas of a PN responsible for some data slices, and using the *Ownership* DDS QoS policy to dynamically switch between the redundant output flows produced by the PN replicas.

Finally, the DPSLB also supports customization since several load balancing algorithms (i.e., implemented in APS of MAPE-SDDL) can be applied in the Load Balancer. For example, one could deploy an algorithm which seeks the uniform load distribution, or otherwise, use another algorithm which tries to minimize the system's global energy consumption.

VIII. CONCLUSION AND FUTURE WORK

The need for remote monitoring and high performance processing of large mobile data streams in a timely manner is becoming common to many systems such as Intelligent Transportation Systems, Fleet Management and Logistics, and integrated Industrial Process Automation.

The main contribution of this work is the development of a novel approach to load balancing that has two main novelties: its autonomic behavior based on MAPE-K model and the use of DDS as its communication infra-structure. The underlying middleware, MAPE-SDDL, supports not only load balancing of mobile connections among different Gateways but also node balancing of data stream processing across multiple PNs. To the best of our knowledge MAPE-SDDL is the first middleware that has developed an autonomous load balancing approach tailored to DDS-based systems. Preliminary performance evaluations have shown encouraging results what motivate us to continue the development of SDDL and its autonomic extensions.

By being disassociated from any autonomic reference model, traditional load balancing mechanisms fail to incorporate self-* properties, which are the pillars for the development of more adaptive and scalable systems. Moreover, most of the traditional load balancing approaches are not well suited for high-throughput mobile communication and data stream processing systems, as they are not based on a communication layer with real-time communication capabilities. On the other hand, our load balancing approach was specially designed for decentralized systems based on the DDS standard, and hence is capable of fulfilling application requirements such as real-time and high throughput data communication and processing, scalability and fault tolerance.

The evaluation has yielded encouraging performance result, which motivate us to proceed with the development of SDDL's adaptivity and load balancing capabilities. In particular, we could check that during the Load Balancing Process there was neither any data item loss nor duplication. It could also be noticed that the addition of new PNs effectively enhances the system's processing capacity and does not drive to an rise of the overhead. For example, the overall throughput could be augmented from 40 DI/s for one PN, to 323 DI/s, when five PNs are used. While new PNs help to increase the throughput, more PNs reduce the RTD – or at least to reduce the growth – because the load is divided across the available PNs since the system is able to promptly process more data items in the same unit of time. Most importantly, the proposed DPSLB solution allows a stream processing system to scale in the number of PNs with acceptable overhead. The overhead introduced by the DPSLB prototype represents only 1,4% and 1,58% of the throughput and RTD, respectively, which is a low cost compared to the benefit of having a load balancing mechanism.

With a data stream rate of 10 MB/s and 1.000 Slices, the DPSLB prototype was able to move 500 Slices from a single PN to a second PN in less than 500 ms (milliseconds), which is fast enough for many stream processing applications. And regarding load balancing of mobile node (MN) connections, MAPE-SDDL is able to migrate 300/600 MNs from one Gateway to another in less than 750 ms.

As future work, we are planning the design, implementation and evaluation of other Attribution Functions and load balancing algorithms, both for data stream processing and connectivity load distribution, as well as developing support for general state transfers among the managed resources (PNs or Gateways) during Load Balancing Process. New Attribution Functions would be valuable to study how the DPSLB solution behaves in face of uneven Attribution Function and heterogeneous PN resource capacities. In particular, we believe that collecting statistical data about the data items received in each Slice, and the corresponding workload associated to each Slice will certainly enable much better Load Balancing Algorithms. Thereby, they would be able to take into account the different data stream rates assigned to each Slice, and to decide which specific Slices, not only how many, should be moved among the PNs.

We also plan to design and implement a new component in MAPE-SDDL, called Distributed Event Service (DES), which will allow the detection of composite events made of basic events from different event sources (e.g., distributed PNs). DES is required for cases where the decision to reconfigure the system must consider the combination of events detected by several LES at distributed resources. For example, it could be used to detect the overload in a group of PNs, instead of the overload detection at individual PNs. The DES will thus enable the load balancing mechanism to be driven by a global perspective on a distributed group of Nodes.

ACKNOWLEDGMENT

“This work is partly supported by project Mobile InfoPAE, CNPq scholarships n°. 310253/2011-0 and 140966/2013-7, and FAPEMA.”

REFERENCES

- [1] R. O. Vasconcelos, M. Endler, B. d. T. P. Gomes, and F. J. d. S. e. Silva, “Towards Autonomous Load Balancing for Mobile Data Stream Processing and Communication Middleware Based on Data Distribution Service,” in *ICAS 2013: The Ninth International Conference on Autonomic and Autonomous Systems*, Lisbon, 2013, pp. 7–13.
- [2] R. O. Vasconcelos and M. Endler, “A Dynamic Load Balancing Mechanism for Data Stream Processing on DDS Systems,” M.Sc Thesis, Departamento de Informatica, PUC-Rio - Pontificia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2013.
- [3] M. Stonebraker, U. Çetintemel, and S. Zdonik, “The 8 requirements of real-time stream processing,” *ACM SIGMOD Record*, vol. 34, no. 4, pp. 42–47, Dec. 2005.
- [4] A. Margara and G. Cugola, “Processing flows of information,” in *Proceedings of the 5th ACM international conference on Distributed event-based system - DEBS '11*. New York, New York, USA: ACM Press, 2011, p. 359.
- [5] Waze, “Free GPS Navigation with Turn by Turn - Waze,” 2013. [Online]. Available: <http://www.waze.com/>. [Accessed Dec. 18, 2013].
- [6] A. Calsavara and L. A. P. Lima Jr., “Scalability of Distributed Dynamic Load Balancing Mechanisms,” in *ICN 2011 The Tenth International Conference on Networks*, no. C, 2011, pp. 347–352.
- [7] M. Randles, D. Lamb, and A. Taleb-Bendiab, “A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing,” in *2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops*. IEEE, 2010, pp. 551–556.
- [8] Q. Zhang, L. Cheng, and R. Boutaba, “Cloud computing: state-of-the-art and research challenges,” *Journal of Internet Services and Applications*, vol. 1, no. 1, pp. 7–18, 2010.
- [9] IBM, “An architectural blueprint for autonomic computing,” *IBM White Paper*, 2006.
- [10] M. C. Huebscher and J. a. McCann, “A survey of autonomic computing—degrees, models, and applications,” *ACM Computing Surveys*, vol. 40, no. 3, pp. 1–28, Aug. 2008.
- [11] S. Hariri, B. Khargharia, H. Chen, J. Yang, Y. Zhang, M. Parashar, and H. Liu, “The Autonomic Computing Paradigm,” *Cluster Computing*, vol. 9, no. 1, pp. 5–17, Jan. 2006.
- [12] R. Sterritt, “Autonomic computing,” *Innovations in Systems and Software Engineering*, vol. 1, no. 1, pp. 79–88, Mar. 2005.
- [13] J. Kephart and D. Chess, “The vision of autonomic computing,” *Computer*, vol. 36, no. 1, pp. 41–50, Jan. 2003.
- [14] M. Parashar and S. Hariri, “Autonomic computing: An overview,” in *In Proceedings of the 2004 international conference on Unconventional Programming Paradigms (UPP'04)*, J.-P. Banâtre, P. Fradet, J.-L. Giavitto, and O. Michel, Eds. Le Mont Saint Michel: Springer-Verlag, Berlin, Heidelberg, 2005, pp. 247–259.
- [15] OMG, “Object Management Group,” 2013. [Online]. Available: <http://www.omg.org/>. [Accessed Dec. 18, 2013].
- [16] G. Pardo-Castellote, “OMG data-distribution service: Architectural overview,” *ICDCSW '03 Proceedings of the 23rd International Conference on Distributed Computing Systems*, 2003.
- [17] C. Tucker, “What can DDS do for You? Learn how dynamic publish-subscribe messaging can improve the flexibility and scalability of your applications.” *OMG Whitepapers: Data Distribution Service Portal*, 2013.

- [18] M. Xiong, J. Parsons, J. Edmondson, H. Nguyen, and D. C. Schmidt, "Evaluating the Performance of Publish/Subscribe Platforms for Information Management in Distributed Real-time and Embedded Systems," *OMG Whitepapers: Data Distribution Service Portal*, 2010.
- [19] L. David, R. Vasconcelos, L. Alves, R. André, and M. Endler, "A DDS-based middleware for scalable tracking, communication and collaboration of mobile nodes," *Journal of Internet Services and Applications (JISA)*, vol. 4, no. 1, p. 16, 2013.
- [20] G. Pardo-Castellote, B. Farabaugh, and R. Warren, "An introduction to DDS and data-centric communications," 2005. [Online]. Available: http://www.omg.org/news/whitepapers/Intro_To_DDS.pdf. [Accessed Dec. 18, 2013].
- [21] G. Pardo-Castellote, "DDS Tutorial – Part II - Hands On," 2009. [Online]. Available: http://www.omg.org/news/meetings/GOV-WS/pr/rte-pres/DDS_Tutorial_RTEW09.pdf. [Accessed Dec. 18, 2013].
- [22] T. O. Computing, "Learn About How it Works: Take the CoreDX DDS Tour Twin Oaks Computing, Inc," 2012. [Online]. Available: http://www.twinoaksc computing.com/coredx/dds_tour. [Accessed Dec. 08, 2013].
- [23] J. O. Kephart and D. M. Chess, "The vision of autonomic computing," *Computer*, vol. 36, pp. 41–50, January 2003.
- [24] I. Naick, "Make autonomic computing a reality with IBM Tivoli. Using IBM Tivoli Provisioning Manager and IBM Tivoli Intelligent Orchestrator to create an on demand environment," *IBM White Paper*, 2004.
- [25] A. K. Y. Cheung, "Dynamic Load Balancing in Distributed Content-based Publish/Subscribe," Ph.D. dissertation, M.Sc Thesis, Graduate Department of Electrical and Computer Engineering, University of Toronto, 2006.
- [26] N. Shivaratri, P. Krueger, and M. Singhal, "Load distributing for locally distributed systems," *Computer*, vol. 25, no. 12, pp. 33–44, Dec. 1992.
- [27] T. Casavant and J. Kuhl, "A taxonomy of scheduling in general-purpose distributed computing systems," *IEEE Transactions on Software Engineering*, vol. 14, no. 2, pp. 141–154, 1988.
- [28] D. Grosu and A. Chronopoulos, "Algorithmic Mechanism Design for Load Balancing in Distributed Systems," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 77–84, Feb. 2004.
- [29] A. Osman and H. Ammar, "Dynamic load balancing strategies for parallel computers," in *International Symposium on Parallel and Distributed Computing (ISPDC)*, 2002.
- [30] R. O. Vasconcelos, L. Silva, L. Alves, and M. Endler, "Scalable Data Distribution Layer - Overview, Use Instructions and Download," 2012. [Online]. Available: <http://www.lacrio.com/sddl/>. [Accessed Dec. 18, 2013].
- [31] R. O. Vasconcelos, L. David, L. Alves, R. André, and M. Endler, "Real-time Group Management and Communication for Large-scale Pervasive Applications," Rio de Janeiro, 2012, Monografias em Ciência da Computação - MCC 05/2012, Dep. de Informática, PUC-Rio, ISSN 0103-9741.
- [32] L. David, R. Vasconcelos, L. Alves, R. André, G. Baptista, and M. Endler, "A Large-scale Communication Middleware for Fleet Tracking and Management," in *Salão de Ferramentas, Brazilian Symposium on Computer Networks and Distributed Systems (SBRC 2012)*, Ouro Preto, 2012.
- [33] M. Endler, R. O. Vasconcelos, L. David, R. André, and L. Alves, "A DDS-based middleware for scalable tracking and communication of wireless-connected mobile nodes in a WAN," Rio de Janeiro, 2012, Monografias em Ciência da Computação - MCC 06/2012, Dep. de Informática, PUC-Rio, ISSN 0103-9741.
- [34] Oracle, "Getting Started Using Java RMI," 2013. [Online]. Available: <http://docs.oracle.com/javase/6/docs/technotes/guides/rmi/hello/hello-world.html>. [Accessed Dec. 18, 2013].
- [35] F. Dabek, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica, "Wide-area cooperative storage with CFS," in *Proceedings of the eighteenth ACM symposium on Operating systems principles - SOSP '01*. New York, New York, USA: ACM Press, 2001, p. 202.
- [36] A. Rao, K. Lakshminarayanan, and S. Surana, "Load balancing in structured P2P systems," in *Proceedings of IPTPS*, 2003, pp. 68–79.
- [37] L. Xia, H. Duan, X. Zhou, Z. Zhao, and X.-W. Nie, "Heterogeneity and load balance in structured P2P system," in *2010 International Conference on Communications, Circuits and Systems (ICCCAS)*. IEEE, Jul. 2010, pp. 245–248.
- [38] M. Raab and A. Steger, "“Balls into Bins”—A Simple and Tight Analysis," in *Proceedings of the Second International Workshop on Randomization and Approximation Techniques in Computer Science (RANDOM '98)*, pp. 159–170, 1998.
- [39] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg, "Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services," *Performance Evaluation*, vol. 68, no. 11, pp. 1056–1071, Nov. 2011.
- [40] C.-C. Yang, C. Chen, and J.-Y. Chen, "Random Early Detection Web Servers for Dynamic Load Balancing," in *2009 10th International Symposium on Pervasive Systems, Algorithms, and Networks*. IEEE, 2009, pp. 364–368.
- [41] V. Suresh, D. Karthikeswaran, V. Sudha, and D. Chandraseker, "Web server load balancing using SSL back-end forwarding method," *Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on*, pp. 822–827, 2012.
- [42] Z. Zhang and W. Fan, "Web server load balancing: A queueing analysis," *European Journal of Operational Research*, vol. 186, no. 2, pp. 681–693, Apr. 2008.

- [43] D. C. Shadrach, K. S. Balagani, and V. V. Phoha, "A Weighted Metric Based Adaptive Algorithm for Web Server Load Balancing," in *2009 Third International Symposium on Intelligent Information Technology Application*. IEEE, 2009, pp. 449–452.
- [44] T. C. Chieu, A. Mohindra, A. a. Karve, and A. Segal, "Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment," in *2009 IEEE International Conference on e-Business Engineering*. IEEE, 2009, pp. 281–286.
- [45] A. Corsaro, "DDS in SCADA, Utilities, Smart Grid and Smart Cities," 2012. [Online]. Available: <http://www.slideshare.net/Angelo.Corsaro/dds-in-scada-utilities-smart-grid-and-smart-cities>
- [46] A. K. Y. Cheung and H.-A. Jacobsen, "Load Balancing Content-Based Publish/Subscribe Systems," *ACM Transactions on Computer Systems*, vol. 28, no. 4, pp. 1–55, December 2010.
- [47] A. Corradi, L. Foschini, and L. Nardelli, "A DDS-compliant infrastructure for fault-tolerant and scalable data dissemination," in *The IEEE symposium on Computers and Communications*. IEEE, June 2010, pp. 489–495.
- [48] G. Li and H.-A. Jacobsen, "Composite subscriptions in content-based publish/subscribe systems," in *Proceedings of the ACM/IFIP/USENIX 2005 International Conference on Middleware (Middleware '05)*, G. Alonso, Ed. Grenoble, France: Springer-Verlag New York, Inc., 2005, pp. 249–269.

Comparison of Simultaneous Measurement of Lens Accommodation and Convergence in Stereoscopic Target with Sine Curve Movement

Takehito Kojima, Kazuki Yoshikawa, Masaru Miyao
Graduate School of Information Science
Nagoya University, Nagoya, Japan
{tkojima45, kazu.kc.yl}@gmail.com, miyao@nagoya-u.jp

Tomoki Shiomi
Tokushima Labour Standards Inspection Office
Tokushima, Japan
srw_107130_nano@yahoo.co.jp

Abstract—Recently, many advances have been made in 3D technology. However, the influence of stereoscopic vision on human sight remains insufficiently understood. "Accommodation convergence discrepancy theory" states that when a person views stereoscopic images, a visual discrepancy occurs because convergence focuses at the position of the virtual object, while lens accommodation is fixed on the screen. It is widely accepted in the field that this is the main reason for visual fatigue caused while viewing 3D images. However, we have not found such a mismatch in experiments with young subjects. The aim in this study was to compare the fixation distance of accommodation and convergence in viewing real objects and 3D video clips. We measured accommodation and convergence in subjects who watched both real objects and 3D video clips with similar movements. From the result of this experiment, we found that no discrepancy exists in viewing either 3D video clips or real objects. We argue that the symptoms that occur when viewing stereoscopic vision may not be due to a discrepancy between lens accommodation and convergence. To compare the accommodative response and amplitude in different age groups, we fit the experimental results to the operation of a sine curve.

Keywords- accommodation; convergence; simultaneous measurement; stereoscopic vision; depth of field; sine curve fitting

I. INTRODUCTION

Investigations of the influences of stereoscopic vision on the human body are essential in order to ensure safe and comfortable viewing of virtual 3D objects. In a previous study with associates, we verified that both convergence and lens accommodation are connected with the motion of the virtual object when viewing 3D images [1][2][3][4].

On the other hand, when viewing stereoscopic images, people sometimes feel visual fatigue, 3D sickness, or other discomfort [5]. And one of the main theories for the cause of this visual fatigue is still the "accommodation convergence discrepancy theory". According to this theory, when viewing 3D images lens accommodation remains fixed on the screen while convergence moves to the position of the virtual object [6][7].

The relationship between accommodation and convergence is one factor that enables humans to see one object with both eyes. Toates [8][9] said that the proximity of the target appears to cause vergence, and that

accommodation, to be a specific accommodative effort, is associated with innervation to vergence. Accommodation and vergence are mutually interacting control systems. It is possible under normal conditions for accommodation to depend on convergence to a certain extent.

Convergence occurs when an image is captured differently with both eyes (parallax). The recent methods of 3-dimensional images, for example, liquid crystal shutter systems, lenticular systems, and polarized filter systems have improved to make it easier for human convergence. The latest technology in this area in visual 3D production has shown many improvements focusing on convergence [10][11].

We conducted these experiments and discussion shown below based on our previous work.

1. Lens accommodation was measured and compared with both a real object and a 3-D image.

2. More than 100 subjects were divided into four age groups and were tested. We applied the lens accommodative response data from these subjects to fit the operation of a sine curve. Then we summarized the fitting of the sine curve by age group. After this, we evaluated the accommodative ability or the delay in accommodative response by age.

3. Rejection of the "accommodation convergence discrepancy theory" leads to the question of whether the subjects saw blurred images when they focused on the virtual object instead of display. One reason for not seeing any blurring would be the existence of depth of field. When subjects watch the target, the pupil is contracted by the luminance. It is advantageous in order to obtain a deep depth of field [12][13].

II. MATERIALS AND METHODS

Explanation of the instrument used for the experiment and the experiment method was shown below.

A. Instruments

1) WAM-5500

We used the WAM-5500 by Shigiya Machinery Works, Ltd. in this experiment (Figure 1a).

This instrument can measure the refractive value (accommodation value) of a single eye when the subject gazes at a target of a given distance. It can measure pupil diameter continuously and monocular accommodation and pupillary diameter at a sampling interval of 0.2 seconds. The WAM-5500 has also been used in investigations of eyestrain

and transient myopia [14][15] based on accommodative values. Moreover, the WAM-5500 has been used in investigations of lens accommodation response under near work conditions and visual discomfort over a year [16][17], and its reliability was found to be sufficient.

2) EMR-9

The EMR-9 Eye Mark Recorder (NAC Image Technology Inc.) can measure the binocular scan paths (eye movements) using the pupillary/corneal reflex method. The resolution for eye movement is 0.1 degrees, with a measurement range of 40 degrees and sampling rate of 60 Hz. The convergent focus distance can be easily calculated from the obtained binocular eye movement data based on the calibration for 9 points (3×3), as shown in Figure 1b.

3) WMT-1

The WMT-1 by Shigiya Machinery Works, Ltd. is a target movement viewing system. It consists of a movable plate (about 1 m in length), a control device, and software that can be connected to a PC, and is the same as that of a numerical control (NC) robot.

By combining the WAM-5500 and WMT-1, we built a system to measure the accommodation value for a moving visual target. By connecting with the PC and having controls from exclusive communication software (WCS-1) (Figure 1c), it was possible to ascertain the position information for visual targets at 0.01-second intervals. Accommodation was measured continuously and pupil diameter was measured at 0.2 seconds intervals.

B. Methods

1) Experiment I. Simultaneous Measurement of Accommodation and Convergence for a 3D Video Clip in Diopter Sine Drive and Step Drive

For the simultaneous measurement of accommodation and convergence, we combined the WAM-5500 and EMR-9 and connected them with a link cable. We set the start times of the two data collecting devices.

The images used were from OLYMPUS Advanced POWER 3D™, which is a CG 3-dimensional video. The images were created using the stereo image fabrication technique from OLYMPUS Memory Works, Ltd. This

technique involves the use of two cameras showing a background image, and two cameras showing an object in motion so that the views are superimposed (Figure 2).

In a previous study, we found that the reaction of the subject's lens accommodation with OLYMPUS Advanced POWER 3D shows a nearer value to natural vision than conventional 3D images [2].

Fujine et al. [18][19] suggested that the viewing distance should be a minimum of three times the absolute display height. We decided to follow this recommendation as part of our procedure. The specification of the display and the 3D image are shown in tables I and II).

For the first 10 seconds, subjects viewed a white circle in the center of a black screen. Then a moving sphere appeared, going back and forth between 1.0D (1 m) and 2.0D (50 cm). The subjects used binocular vision, and simultaneous measurements of accommodation and convergence were made with the WAM-5500 and the EMR-9, respectively.

There were three patterns of movement. The first pattern was a sine curve drive in a 10-second period for 30 seconds. The second pattern was a sine curve drive in a 2.5-second period for 10 seconds. The third pattern stopped for 10 seconds at distances of 1.0 D (1 m), 1.5 D (67 cm), and 2.0 D (50 cm) from the front of the eye of the subject (step drive). The order of precise stoppage was 1.0 D, 1.5 D, and 2.0 D.

The low-screen brightness was 12.7 cd/m² and the high-screen brightness was 70.4 cd/m². We measured the luminance in a white part of the sphere through a circular polarized filter and a dichroic mirror on the WAM-5500.

2) Experiment II. Simultaneous Measurement of Accommodation and Convergence for a Real Object in Sine Curve Drive and Step Drive

A Rubik's Cube was used as the real visual target because of its ease of recognition as a geometric form. This visual target was fixed to the movable plate of the WMT-1, and the PC controlled the movement of the target forward and backward. During measurement, the subjects were instructed to gaze at the visual target.

The PC recorded various data, including the time code from the measurement start time, the position information on

(a) WAM-5500



(b) EMR-9



(c) WMT-1 and WAM-5500



Figure 1. Instruments used in the experiments: (a) WAM-5500, (b) EMR-9, and (c) WMT-1 and WAM-5500.



Figure 2. Movement of 3D images

TABLE I. SPECIFICATIONS FOR DISPLAY

| | |
|--------------|------------------------------------|
| System | Circularly Polarizing Filter (CPF) |
| Manufacturer | Mitsubishi |
| Model | RDT233WX-3D |
| Screen Size | 23-Inch |
| Resolution | 1920x1080 |
| Refresh Rate | 60Hz |

TABLE II. SPECIFICATIONS FOR 3D IMAGES

| | |
|--|--|
| Viewing Distance | 1.0 m (1.0 D) |
| Interpupillary Distance Setting | 60 mm |
| 3D Data Format | Side by Side |
| Stereoscopic Effect (Popping) | In Front of Eye 50 cm (2.0 D) Parallax Angle 3.5° |
| Stereoscopic Effect (Retraction) | In Front of Eye 1.0 m (1.0 D) Parallax Angle 0° |
| Background | 1.21 m (0.83 D) |
| Video Brightness | High Luminance |
| Screen Luminance (cd/m ²) (Over the grass) | 70.4 |

the visual target from the WMT-1, the accommodation value and the pupil diameter from the WAM-5500. The WMT-1 was used in two movement patterns (the diopter-sine drive and the step-drive). The diopter-sine drive includes two different periods (10 and 2.5 seconds) of movement, the same as in Experiment I. The step drive suspended a real object for 10 seconds at three points: 1.0 D, 1.5 D, and 2.0 D, the same as in Experiment I.

3) Experiment III. Measurement of Accommodation for Diopter Sine Drive of Real Object and Fitting to a Sine Curve

We measured accommodative change while the subjects gazed at a moving object (the Rubik's Cube). At this time, the subjects were asked to gaze at the center of the Rubik's Cube. The moving object oscillated between 1.0D and 2.0D from the front of the eye of the subject.

There were two patterns of movement, as in Experiment I. The first pattern was a sine curve drive with a 10-seconds period for 30 seconds. The second pattern was a sine curve drive with a 2.5-seconds period for 10 seconds.

4) Subjects

For Experiments I and II, the subjects were seven individuals from the age of 21 to 47 years old who participated in the simultaneous measurement of accommodation and convergence when viewing the 3D video clip and real object.

For Experiment III, the subjects were 135 individuals from the age of 17 years old to 85 years old who participated in the accommodation measurement with the real object.

The subjects were divided into the following four groups: young (n=40, 17-29 years old), young-middle age (n=23, 30-44 years old), middle-aged (n=37, 45-64 years old), and the elderly (n=34, aged 65 and over).

The crystalline lens loses elasticity with age and its refractive power also decreases. The clinically measured amplitude of accommodation, which includes both the true dioptric change in the power of the eye and ocular depth-of-focus, decreases fairly steadily from about 13 D at the age of 16 to 2 D at the age of 50 and thereafter [20][21]. Therefore, these groups were divided according to visual function characteristics, especially accommodative ability. For example, the young group had sufficient accommodative power. The young-middle age group had somewhat weak accommodation power and does not suffer from presbyopia. They can clearly see close objects 20 to 30 cm from their eyes without much effort. The middle-aged group had mild difficulty in seeing near objects because of presbyopia.

In this group, some individuals wore glasses for near-sighted issues and others did not. The elderly group had severe presbyopia, so they generally wore convex glasses.

5) Method of Analysis for Experiment III

We selected samples that successfully measured two periods or more. Each datum was fitted to a sine curve. The processed data were averaged for each age group.

In the following equations, α represents the average, κ_0 represents half the amplitude, and κ_1 represents delay.

$$y = \alpha + \kappa_0 \cdot \sin(36 \cdot x + \kappa_1) \quad (1)$$

$$y = \alpha + \kappa_0 \cdot \sin(144 \cdot x + \kappa_1) \quad (2)$$

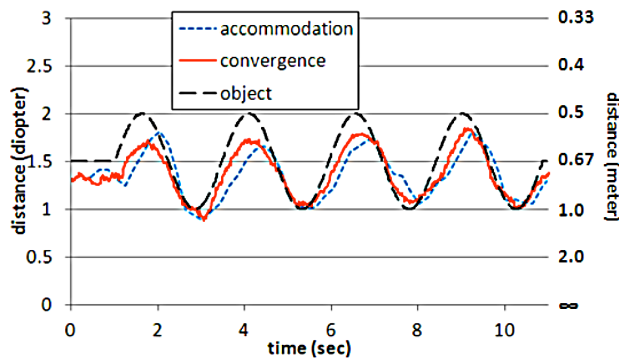


Figure 3. Simultaneous measurement values for accommodation and convergence with a real object (24-year-old male)

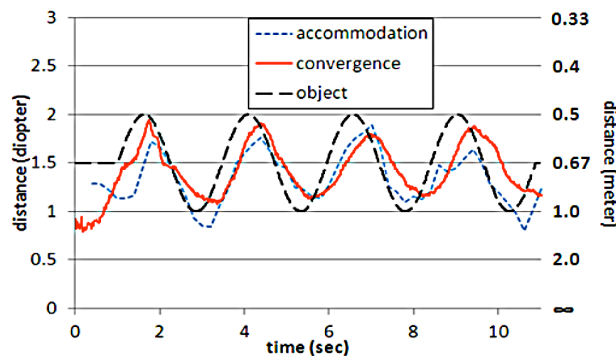


Figure 4. Simultaneous measurement values for accommodation and convergence with 3D images (24-year-old male)

Since the movement of the visual target was a sine curve, each subject's measurement data was fitted to the sine curve.

The Rsquare.exe of the software used a curved reliance panel with a least-squares method for fitting to a sine curve.

The amplitude and delay of the measurement data were totaled and averaged for every group, and modeling based on equations (1) and (2) which were performed for the 10-seconds period (1) and 2.5-seconds period (2).

6) Technical Limit of the Measuring Instrument

The sampling frequency of the WAM-5500 during operation is 5 Hz. This value is not sufficient to measure the frequency of accommodative reaction, especially in 2.5-second period movement. In the lower sampling rate, when measuring rapid reciprocating motion as a 2.5-seconds period, careful operation is required to obtain accurate measurements. (According to H. Anderson et al., the delay of an accommodative reaction is about 0.3 seconds [22]).

However, the time resolution for the accommodation value of WAM-5500 (0.2 seconds) is considered relatively low. Furthermore, since the pupil diameter becomes smaller with age, the measurement success rate with elderly subjects is reduced. Therefore, the number of samples decreases.

The data in this study were restricted to subjects in whom measurements were possible under the following conditions: visual performance is high and pupil diameter is sufficiently large.

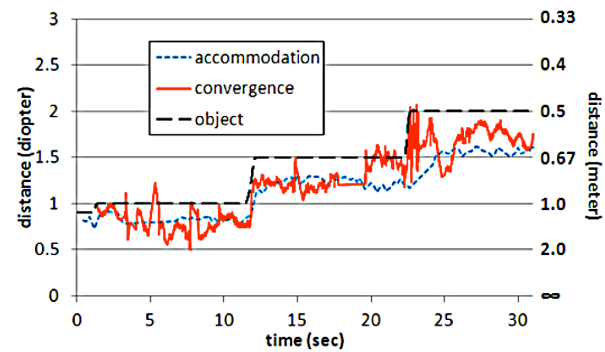


Figure 5. Simultaneous measurement values for accommodation and convergence with a real object in step movement (23-year-old male)

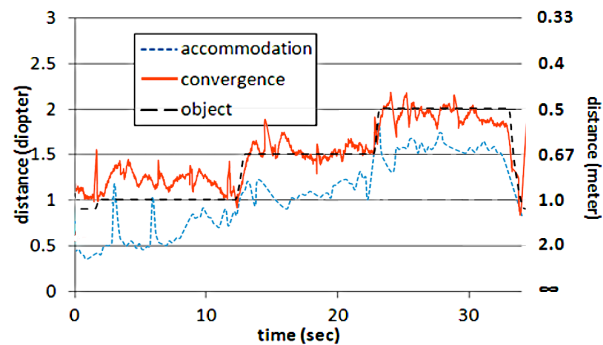


Figure 6. Simultaneous measurement values for accommodation and convergence with 3D images in step movement (24-year-old male)

III. RESULTS

The experiments I and II compared the real object with the 3-D image, and experiment III was based upon the age group.

A. Experiments I and II: Comparison of Accommodation and Convergence about Real and 3D Objects

1) Sine Curve Drive

In the case of the young subjects, convergence and accommodation were similar and synchronized for the movements of both real and 3D image objects. The convergence values were in agreement with the position of the visual target in a bright environment. The accommodative values were in a similar position to convergence or slightly distant from the visual target (Figures 3 and 4).

2) Step Drive (1.0D, 1.5D, and 2.0D)

Figure 5 shows the values of the simultaneous measurements of accommodation and convergence with the real object during the step drive. Figure 6 shows the simultaneous measurement values for accommodation and convergence with 3D images during the step drive.

In the case of the younger subjects, accommodation and convergence were similar for the step movements of both the real and 3D image objects. The convergence values were in agreement with the position of visual targets (real and virtual targets). Accommodative values were slightly further away from the visual target.

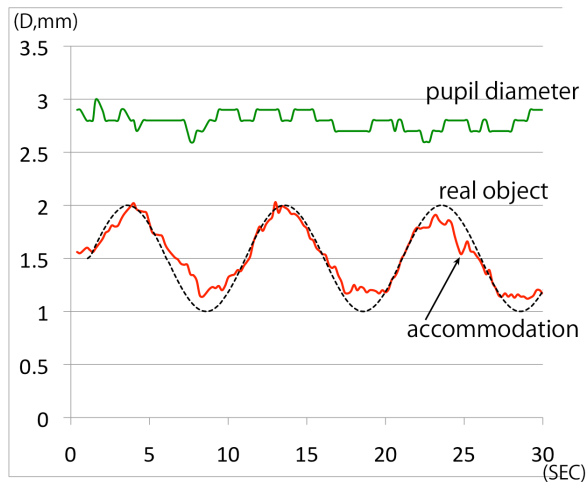


Figure 7. Typical example of younger subject, 10 seconds period (23-year-old male)

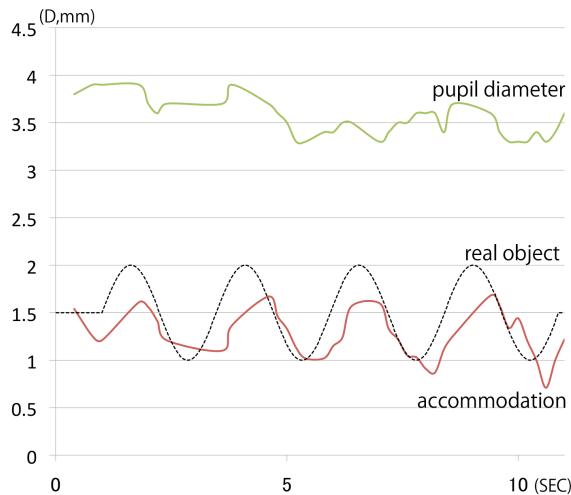


Figure 8. Typical example of younger subject, 2.5 seconds period (24-year-old male)

With both the real object and 3D images, lens accommodation had focused at a place 0.3-0.4D distant from the position of the visual target [4][21]. In step movement with 3D images, the value for accommodation moved clearly away from the visual target.

B. Experiment III: Comparison between Age Groups of Lens Accommodation while Gazing at Sine Curve Movement of a Real Object

1) Younger Subjects (17-29 years of age)

Figures 7 (10 seconds period) and 8 (2.5 seconds period) show the results for accommodation and pupil diameter in the younger subjects for sine curve real object movement.

The accommodation and pupil diameter values in 15 of 40 subjects are superimposed and averaged in Figure 9 (10 seconds period). Figure 10 shows an analysis of 20 of the 40 subjects (2.5 seconds period).

Figure 7 shows the results for a subject (23 years old, male) who viewed a visual target during a period of 10 seconds.

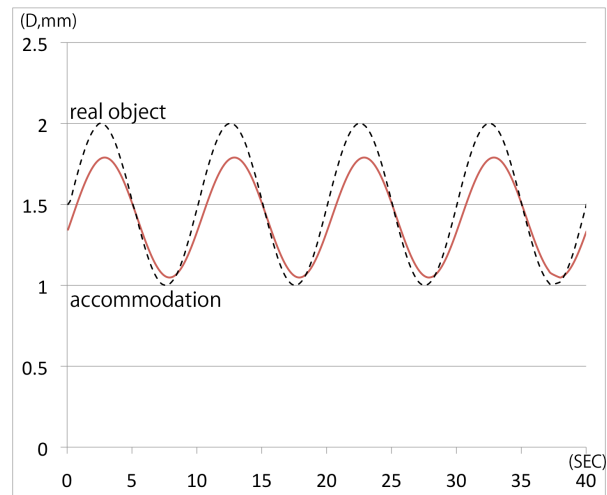


Figure 9. Younger subject fitting results, 10 seconds period

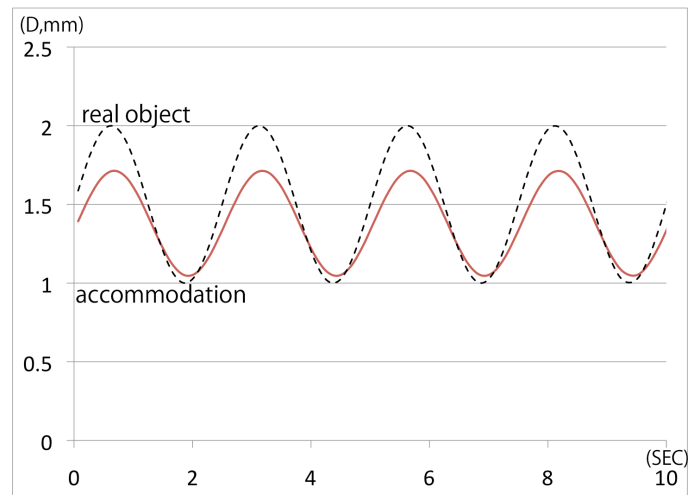


Figure 10. Younger subject fitting results, 2.5 seconds period

The values for accommodation matched those with the real object movement. On the other hand, the pupil diameter showed little variation, with a mean value of 2.8 mm.

Figure 8 shows of the results for a different subject (24 years old, male) who viewed visual target with a period of 2.5 seconds.

The values for lens accommodation were synchronized with the movement of the visual target. The visual target of the real object moved back and forth from 2.0 D (50 cm) to 1.0 D (1 m). The mean for the lens focus was recorded from 1.80 D (56 cm) to 0.87 D (1.15 m). The pupil diameter showed a slight variation with a similar phase to the sine curve movement of the visual target.

As explained in the Methods section, we used Rsquare.exe software, which performs a curved reliance panel using a least-squares method. It was fit to a sine curve.

Figure 9 shows the fitting results for the young subjects with a period of 10 seconds.

We superimposed the data of the 15 cases that were successfully measured out of 40 people.

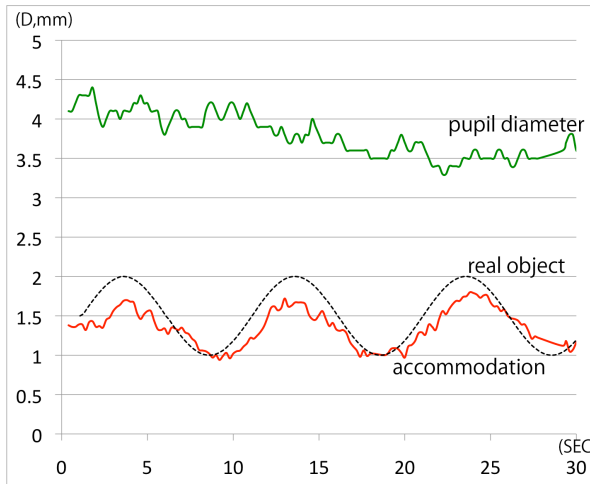


Figure 11. Typical example of young middle-aged subject, 10 seconds period (41-year-old male)

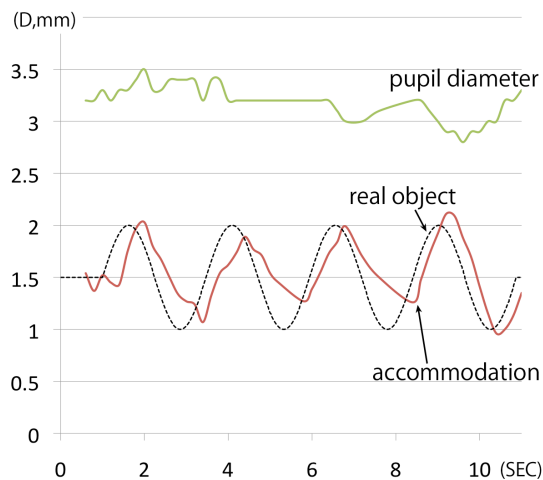


Figure 12. Typical example of young middle-aged subject, 2.5 seconds period (36-year-old female)

The calculated values of the sine curve fitting showed a movement of between 1.05 D (95 cm) and 1.80 D (56 cm). The value of the amplitude is reduced to 75% of the visual target, and the average value of the sine curve was reduced approximately 0.075 D. The delay was about 0.4 seconds for the visual target.

Figure 10 shows the fitting results for the young subjects with the period of 2.5 seconds. We superimposed data of the 20 cases that were successfully measured out of 40 people. The calculated values of the sine curve fitting were between 1.05 D (95 cm) and 1.70 D (59 cm). The value of the amplitude was reduced to 65% of the amplitude of the visual target (1.0-2.0 D), and the average value of the sine curve was reduced approximately 0.13 D. The delay was about 0.2 second against visual target.

2) Young Middle-aged Subjects (30–44 years old)

Figures 11 (10 seconds period) and 12 (2.5 seconds period) show the results of the accommodation and the pupil diameter for sine curve real object movement of the young middle-age subjects.

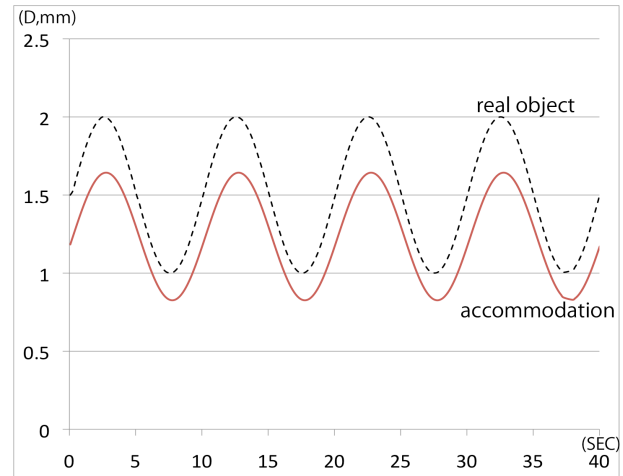


Figure 13. Young-middle subject fitting results, 10 seconds period

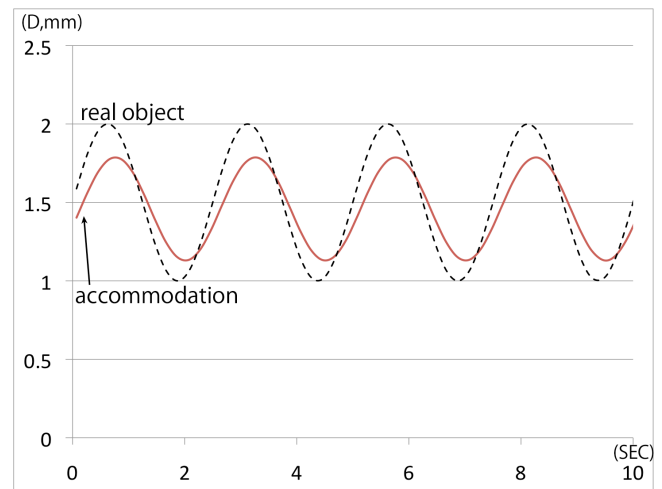


Figure 14. Young-middle subject fitting results, 2.5 seconds period

The accommodation and pupil diameter values in 10 subjects out of 23 were superimposed and averaged.

Figure 11 shows the example of one subject (41 years old, male) who viewed visual target with a 10-second period for 30 seconds. The values for accommodation were partially matched with the real object movement.

The values for accommodation were between 0.94 D (1.06 m) and 1.80 D (56 cm). The average value of the sine curve was reduced approximately 0.29 D. On the other hand, the pupil diameter showed little variation and had a mean value of 3.8 mm.

Figure 12 shows the results from a subject (36 years old, female) who viewed the visual target with a 2.5-second period for 10 seconds. The lens accommodation values were synchronized with the movement of the visual target.

The visual target of the real object moved back and forth from 1.0 D (1 m) to 2.0 D (50 cm). The mean lens focus (accommodation) was recorded from 0.96 D (1.04 m) to 2.11 D (47 cm). The pupil diameter showed no relation to the visual target. A characteristic reaction was seen during the 4th period.

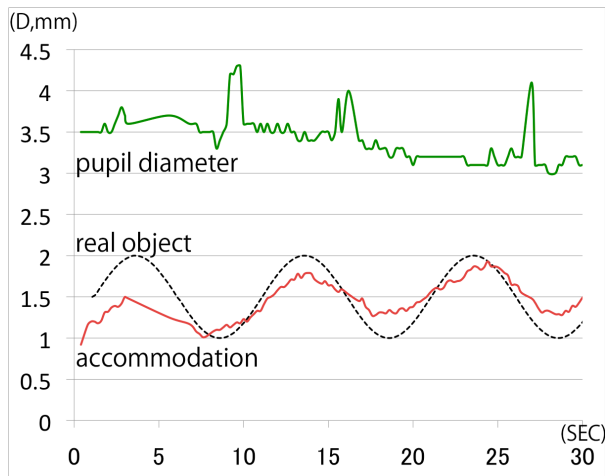


Figure 15. Typical example of middle-aged subject, 10 seconds period (46-year-old female)

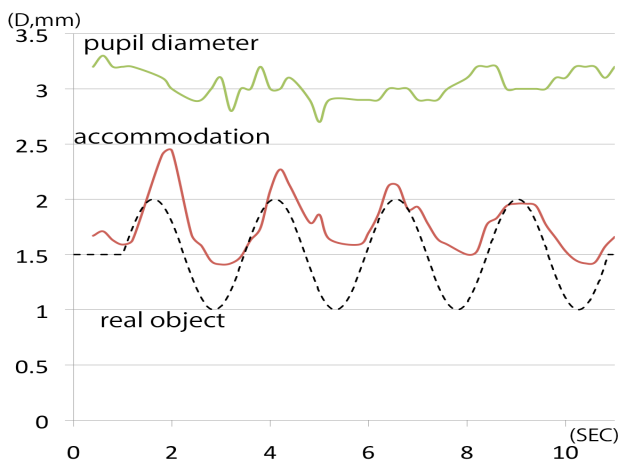


Figure 16. Typical example of middle-aged subject, 2.5 seconds period (46-year-old female)

The pupil seemed to be constricted pupil like in a near reaction.

Figure 13 shows the fitting results for the young middle-age subjects with the period of 10 seconds. We averaged the data of 10 cases in which measurements were successful.

The calculated values of the sine curve fitting moved back and forth between 0.83 D (1.20 m) and 1.64 D (61 cm). These values were reduced to 75 % of the amplitude of the visual target, and the average value of the sine curve was reduced approximately 0.27 D. The delay was about 0.3 seconds against the movement of the visual target.

Figure 14 shows the fitting results for the young middle-aged subjects with the period of 2.5 seconds. We averaged the data for 10 of 23 cases in which the measurements were successful. The calculated values of the sine curve fitting moved back and forth between 1.13 D (88 cm) and 1.79 D (56 cm). The value of the amplitude is reduced to 75 % of the visual target. The delay was about 0.3 seconds against the movement of the visual target.

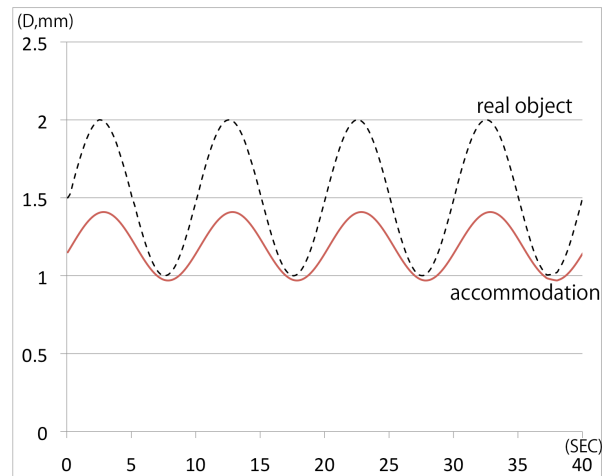


Figure 17. Middle-aged subject fitting results, 10 seconds period

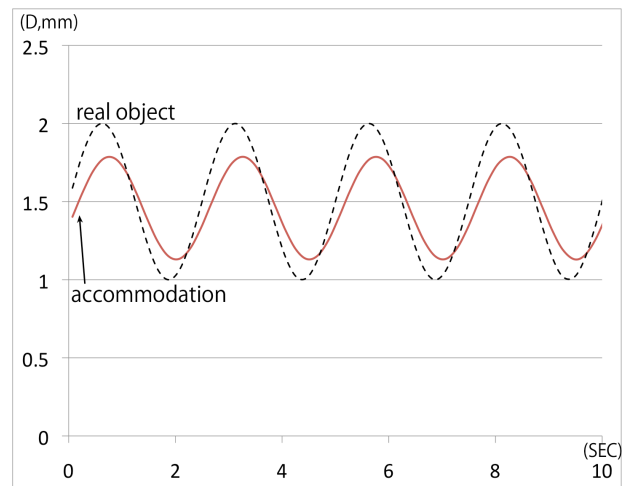


Figure 18. Middle-aged subject fitting results, 2.5 seconds period

3) Middle-aged Subjects (45–64 years old)

Figures 15 (10 seconds period) and 16 (2.5 seconds period) show the accommodation and pupil diameter results for sine curve real object movement for the middle-aged subjects.

The accommodation and pupil diameter values for 9 of 37 subjects were superimposed and averaged, as shown in Figures 15 (10 seconds period) and 16 (2.5 seconds period). Figure 15 shows an example of one subject (46 years old, female) who viewed the visual target with the period of 10 seconds. The values of accommodation were partially matched with the movement of the real object. The values of accommodation of the back and forth movement were between 1.20 D (83 m) and 1.80 D (56 cm). The pupil diameter showed little variation, with a mean value of 3.4 mm. Figure 16 shows an example of another subject (46 years old, female) who viewed the visual target for the period of 2.5 seconds. The values of lens accommodation were synchronized with the movement of the visual target.

The values of lens accommodation were synchronized with the movement of the visual target.

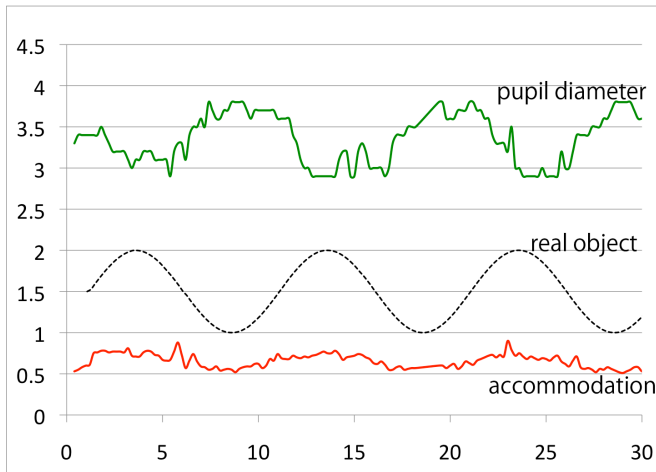


Figure 19. Typical example of elderly subject, 10 seconds period (72-year-old female)

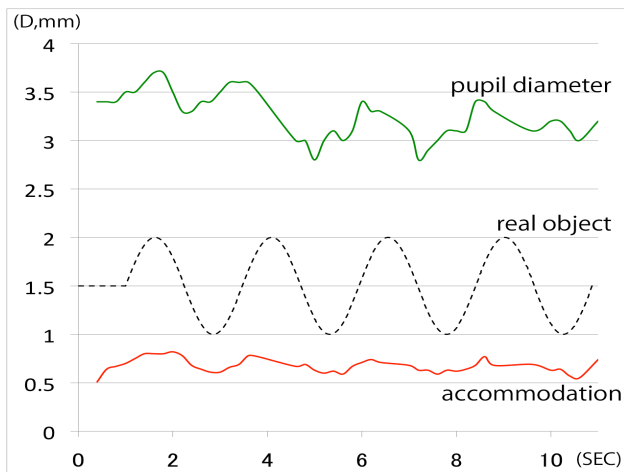


Figure 20. Typical example of elderly subject, 2.5 seconds period (72-year-old female)

The real object visual target moved back and forth from 1.0 D (1 m) to 2.0 D (50 cm). The mean lens focus (accommodation) for this movement was recorded from 1.5 D (67 cm) to 2.44 D (40 cm). The delay was about 0.2 seconds. The pupil diameter was nearly unrelated to the visual target. There seemed to be some pupil constriction that was a slight reflective reaction. This subject was near-sighted at about -1.25 D. The values of lens accommodation were synchronized with the movement of the visual target. The real object visual target moved back and forth from 1.0 D (1 m) to 2.0 D (50 cm). The mean lens focus (accommodation) for this movement was recorded from 1.5 D (67 cm) to 2.44 D (40 cm). The delay was about 0.2 seconds. The pupil diameter was nearly unrelated to the visual target. There seemed to be some pupil constriction that was a slight reflective reaction. This subject was near-sighted at approximately -1.25 D.

Figure 17 shows the fitting results for the middle-aged subjects for the period of 10 seconds. We superimposed the data for the 9 cases that were successfully measured among the 37 subjects.

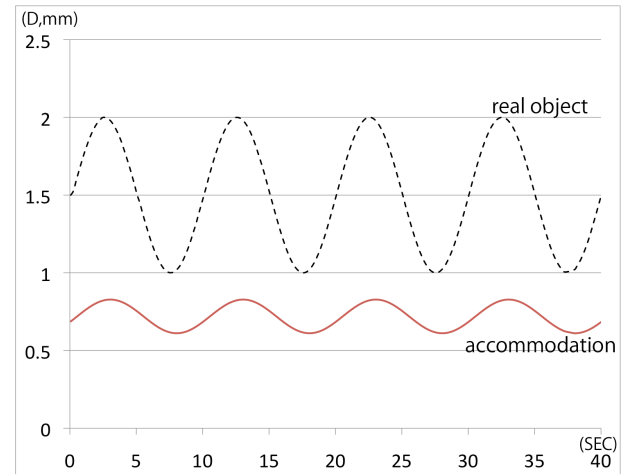


Figure 21. Elderly subject fitting results, 10 seconds period

| TABLE III. FITTING RESULTS | | |
|----------------------------|--------|---|
| Age | Period | Formula |
| 17 - 29 | 10 | $y=1.42+0.37 \times \sin(36 \times t-14.18)$ |
| | 2.5 | $y=1.38+0.33 \times \sin(144 \times t-7.84)$ |
| 30 - 44 | 10 | $y=1.23+0.41 \times \sin(36 \times t-9.72)$ |
| | 2.5 | $y=1.46+0.32 \times \sin(144 \times t-19.80)$ |
| 45 - 64 | 10 | $y=1.19+0.22 \times \sin(36 \times t-12.74)$ |
| | 2.5 | $y=1.22+0.27 \times \sin(144 \times t-10.64)$ |
| 65 - | 10 | $y=0.72+0.11 \times \sin(36 \times t-19.68)$ |
| | 2.5 | $y=0.80+0.14 \times \sin(144 \times t+15.04)$ |

The calculated values of the sine curve fitting showed that the back and forth movement was between 0.97 D (1.03 m) and 1.41 D (71 cm). The value of the amplitude was reduced to 44% of the visual target. The value of accommodation was reduced approximately 0.6 D on the near-point side. The delay was about 0.4 seconds against the movement of the visual target.

Figure 18 shows the fitting result for the middle-aged subjects with the period of 2.5 seconds. We superimposed data of the 9 cases that were successfully measured among the 37 subjects. The calculated values of the sine curve fitting showed back and forth movement between 0.95 D (1.05 m) and 1.50 D (67 cm). The value of the amplitude was reduced to 75 % of the visual target. The delay was about 0.2 seconds against the movement of the visual target.

Generally, people in their 40's suffer from presbyopia and need to use glasses. Therefore, almost all the subjects of this group were considered to be affected by presbyopia.

4) Elderly Subjects (Age 65 or More)

Figures 19 (10 seconds period) and 20 (2.5 seconds period) show the typical accommodation and pupil diameter results for sine curve real object movement in the elderly

subjects. The accommodation and pupil diameter values for 4 out of 34 subjects were superimposed and averaged, as shown in Figure 21 (10 seconds period).

Figure 19 shows a typical subject (72 years old, female) who viewed the visual target for a period of 10 seconds. The values for accommodation were almost unchanged for the real object movement. The values for accommodation showed a very weak back and forth movement between 0.47 D (2.13 m) and 0.90 D (1.11 m). The pupil diameter showed typical synchronization with the distance of the visual target. When the target was close to the subject, the pupil diameter was about 2.8 mm. When the target moved away, the pupil diameter was about 3.8 mm. The elderly subjects seemed to compensate for their poor accommodative power by using extreme pupil constriction to capture close targets clearly.

Figure 20 shows a typical subject (72 years old, female) who viewed the visual target for the period 2.5 seconds. The values of lens accommodation were almost unchanged with the movement of the visual target. The mean of the lens focus (accommodation) showed a back and forth movement from 0.59 D (1.69 m) to 0.82 D (1.22 m). The delay was about 0.9 seconds. The pupil diameter showed typical synchronization with the distance of the visual target. When the target was close to the subject, the pupil diameter was about 2.8 mm. When the target was far away from the subject, the pupil diameter was about 3.6 mm. These subjects also used extreme pupil contraction to compensate for the reduction in their lens accommodation ability.

Figure 21 shows the fitting results of the elderly subjects for the period of 10 seconds. We superimposed data of the 4 cases that were successfully measured among the 34 subjects. The calculated values of the sine curve fitting showed a back and forth movement between 0.61 D (1.64 m) and 0.83 D (1.20 cm).

The value of the amplitude was reduced to 22 % of the visual target. The value of accommodation was reduced approximately 1.17 D on the near-point side. The delay was about 0.9 seconds against the movement of the visual target.

IV. DISCUSSION

This section presents a discussion of experiments I and II (simultaneous measurement of accommodation and convergence) followed by a summation of experiment III (fitting to a sine curve).

A. Experiments I and II: Comparison of Simultaneous Measurement Results of Lens Accommodation and Convergence

Hoffman et al. stated that there is an inconsistency between accommodation and convergence, and they said that lens accommodation in viewing 3D images should be fixed at the position of the display [6]. However, they used a very short viewing distance (30 cm) that produced a small depth of field. Shibata et al. also reported an inconsistency between accommodation and convergence [23]. Their experimental stimuli were random dot stereograms depicting sinusoidal depth corrugations. They used a unique test with a spatial-frequency modulated depth stimulus of small amplitude.

The amplitude was small (peak–trough disparity = 4 arcmin), and spatial frequency was high (1, 1.4, and 2 cpd). Their stimuli were displayed on two static image planes, spaced 1.2 D apart. However, these two studies did not actually measure accommodation and convergence in their subjects simultaneously. In contrast, we used the Power 3D™ (Olympus Memory Works, Corp.) for the stimulus in this experiment. This technique involves the use of two cameras showing a background image, and two cameras showing an object in motion, so that the views are superimposed. It is able to show multiple focal planes corresponding to different focal lengths and convergence angles. It presents a very natural dynamic in the movement of the image in consideration of the natural human eye. Therefore, in our experiment, accommodation for the artificial 3D image closely followed the virtual position of the moving target, as if the image were a real moving object.

Other researchers have reported that an accommodation-convergence discrepancy can create problems such as eyestrain and visual discomfort [8][9][24][25].

However, in this experiment, we found no mismatch in accommodation and convergence, at least in the younger subjects participating in the study.

According to our previous studies, accommodation does not agree strictly with a real object (or with a virtual image) but does agree with a position slightly behind the object [4][26]. Our past studies have shown that the accommodation gap behind the object in younger subjects was within 0.4 D. The gap in the present experiment was also in this range. When subjects viewed 3D video clips in this study, both accommodation and convergence nearly agreed with the virtual position of the 3D video clips.

Experiment III: Amplitude of Accommodation According to Age with Real Object, and Average Delay

The data from the subjects were classified into four groups. For the 10 seconds period, the groups were young: 15/40, young middle-aged: 10/23, middle-aged: 9/37, and elderly: 4/34. For the 2.5 seconds period, they were young: 20/40, young middle-aged: 10/23, middle-aged: 9/37, and elderly: 3/34. Figures 9,10,13,14 and 21 show the fitting of the data to the sine curve for each subject, and the average of each amplitude and delay. In general, the amplitude for accommodation becomes smaller with age and the delay of accommodative response becomes longer [27][28]. The amplitude changes and becomes significantly smaller beginning in middle age. However, the delay for accommodative response was nearly the same at 0.3 seconds for all groups except the elderly group. The elderly group showed a notable delay of 0.9 seconds in accommodative response.

B. Relation between Depth of Field and Blurring

Patterson [29] reported that the accommodation convergence conflict should be a problem only in near-eye displays, and that it likely would not occur under most stereo display viewing conditions because of the depth of field [30].

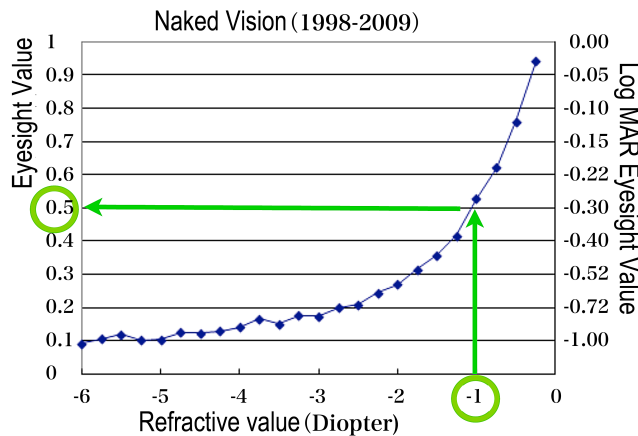


Figure 22. Acuity of fifth grade elementary school student without astigmatism

Two factors that affect a person's perception of depth of field are pupil size and resolution. A person's depth of field changes as the pupil diameter decreases linearly with an increase in luminance [31][32]. The pupil diameter will be slightly over 6 mm for a luminance level of 0.03 cd/m^2 and near 2 mm for a luminance level of 300 cd/m^2 . For each millimeter of decrease in pupil diameter, the depth of field increases by about 0.12 D [33][29]. The depth of field is also affected by the spatial resolution. Ogle and Schwartz [34] found that the total depth of focus increased by approximately 0.35 D per 0.25 as the arcmin increased in the angular target size. They showed that the total depth of focus was an average of 0.66 D for a 1.0-arcmin target and 2.0 D for a 2-arcmin target. In our experiment, the screen was set at 1.0 D (1 m) from the subject and the object emerged to the point of 2.0 D. Most of our subjects accommodated at 0.4 D behind the object at 1.6 D (93 cm). Typically, a perfect match for accommodation and convergence in such a case would be at 2.0 D (50 cm); however, most individuals would show lens accommodation at 0.4 D, which is the boundary point of the depth of field. The usual TV screen has a brightness of 300 cd/m^2 . If the illumination occurs on an indoor screen, the diameter of a pupil will be about 2.0 mm, and the depth of field will be about $\pm 0.5 \text{ D}$. None of the subjects in this study commented on any blurring.

C. Relation between Refractive Power and Blurring

When lens accommodation moves forward to the position of a virtual object when viewing a 3D image, the object displayed on the screen appears the same as the image, as if the person had myopia. Therefore, if the character is not too small it can be viewed satisfactorily [34]. For example, it is considered that if you focus the lens accommodation to a virtual object "popping out" 50 cm from the display at a viewing distance of 1 m, the image shown on the screen can be seen the same as an image viewed with a decimal visual acuity of 0.5 (refer to Figures 22 and 23).

Patterson [29] stated that the interval of the depth of field was on the order of 1.0 D on average.

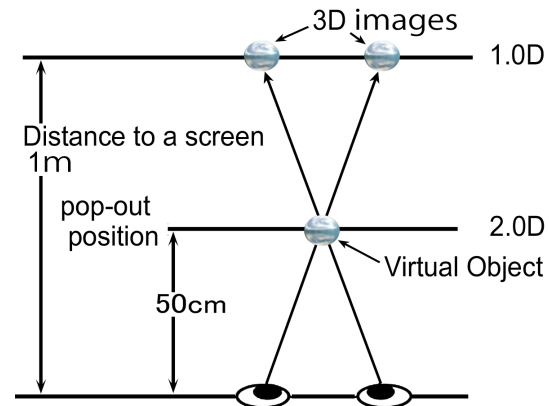


Figure 23. Lens accommodation focused at the pop-out position (50 cm apart from the display)

Therefore, when a subject's gazing point is at 0.5 m, the range of total depth of field would be from a distance of about 0.1 m in front of a fixed point to about 0.17 m behind the fixed point. For a fixed distance of 1 m, the interval of the depth of field would be from a distance of about 0.33 m in front of the point to about 1.0 m behind the visual point. For a fixed distance of 2.0 m, the interval range of the depth of field would be from about 1 m in front of the point to an infinite distance behind the fixed point. Wang et al. [35] also showed that the depth of field increased with age because of the constriction in pupil diameter. According to these authors, the typical depth of field values for young observers was approximately 0.8 D to 1.2 D. In our present study, none of the subjects reported blurred images. This might be because the target was set in the depth of field range when subjects were viewing 3D images.

V. CONCLUSION

In younger subjects, the real object and 3D image were interlocked with the movement of the object, without large deviation between accommodation and convergence, and the focus position changed.

Since the elasticity of the crystalline lens is lost with age and accommodative power decreases, there is a discrepancy between accommodation and convergence in the middle-aged and the elderly, even if it is a natural vision state. 3D images on a screen can be seen without much blurring even if the accommodation focus moves to the position of the virtual object when the 3D image visual target is popping out. Subjects can recognize an object with little or no blurring despite the separation of accommodation from the screen because the position of their focus is within the depth of field. Patterson [29] and Patterson and Silzars [36] proposed that the eyestrain and viewing discomfort that accompany the viewing of stereo displays comes from a high level of conflict between the presence of binocular parallax in the display and the absence of motion parallax.

In the future, we would like to study in more detail how this high level of conflict may contribute to visual fatigue, 3D sickness, or other discomfort in people who view 3D images.

ACKNOWLEDGMENT

This research was partially supported by a grant from JSPS Kakenhi (B) Numbers 24300046 and 23300032.

REFERENCES

- [1] T. Shiomi, T. Kojima, K. Uemoto, and M. Miyao, "Comparison of Simultaneous Measurement of Lens Accommodation and Convergence in Viewing Natural and Stereoscopic Visual Target," ACHI 2013, pp. 147-150.
- [2] S. Hasegawa, A. Hasegawa, M. Omori, H. Ishio, H. Takada, and M. Miyao, "Stereoscopic Vision Induced by Parallax Images on HMD and Its Influence on Visual Functions," Virtual and Mixed Reality - New Trends, Lecture Notes in Computer Science, vol. 6773, 2011, pp. 297-305.
- [3] H. Hori, T. Shiomi, T. Kanda, A. Hasegawa, H. Ishio, Y. Matsuura et al., "Comparison of accommodation and convergence by simultaneous measurements during 2D and 3D vision gaze," FORMA Special Issue, 2011.
- [4] M. Miyao, S. Ishihara, S. Saito, T. Kondo, H. Sakakibara, and H. Toyoshima, "Visual accommodation and subject performance during a stereographic object task using liquid crystal shutters," Ergonomics, vol. 39, no. 11, 1996, pp. 1294-1309.
- [5] M. Lambooi, W. IJsselstein, M. Fortuin, and I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: a review," J. Imaging Sci. Technol. 30201-1-30201-14, 53, 03, 2009.
- [6] D. M. Hoffman, A. R. Girshick, K. Akeley, and M. S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," J. Vis., vol. 8, no. 3, 2008, pp. 1-30.
- [7] K. Ukai and P. A. Howarth, "Visual fatigue caused by viewing stereoscopic motion images: background, theories, and observation," Displays, vol. 29, no. 2, 2008, pp. 106-116.
- [8] F. M. Toates, "Accommodation function of the human eye," Physiol. Rev., vol. 52, 1972, pp. 828-863.
- [9] F. M. Toates, "Vergence eye movements," Doc. Ophthalmol., vol. 37, 1974, pp. 153-214.
- [10] A. Cho, T. Iwasaki, and K. Noro, "A study on visual characteristics binocular 3-D images," Ergonomics, vol. 39, no. 11, 1996, pp. 1285-1293.
- [11] R. Sierra, F. Uchio, N. Iguchi, and H. Taki, "Improving 3D imagery with variable convergence and focus accommodation for the remote assessment of fruit quality," SICE-ICASE Int. Joint Conf., 2006, pp. 3554-3558.
- [12] W. N. Charman and H. Whitefoot, "Pupil diameter and the depth-of-field of the human eye as measured by laser speckle," Int. J. Optics, vol. 24, no. 12, 1977, pp. 1211-1216.
- [13] S. Marcos, E. Moreno, and R. Navarro, "The depth-of-field of the human eye from objective and subjective measurements," Vision Res., vol. 39, no. 12, 1999, pp. 2039-2049.
- [14] C. Tosha, E. Borsting, W. H. Ridder 3rd, and C. Chase, "Accommodation response and visual discomfort," Ophthalmol. Opt., vol. 29, 2009, pp. 625-633.
- [15] E. Borsting, C. Tosha, C. Chase, and W. H. Ridder 3rd, "Measuring near-induced transient myopia in college students with visual discomfort," Amer. Acad. Opt., vol. 87, no. 10, 2010, pp. 760-766.
- [16] C. Chase, Tosha C, Borsting, E, Ridder, and W.H. 3rd., "Visual discomfort and objective measures of static accommodation," Optom. Vis. Sci., vol. 86, no. 7, 2009, pp. 883-889.
- [17] E. Borsting, E. Chase, C. Tosha, C. and W. H. Ridder 3rd, "Longitudinal study of visual discomfort symptoms in college students," Optom. Vis. Sci., vol. 85, no. 10, 2008, pp. 992-998.
- [18] T. Fujine, Y. Kikuchi, M. Sugino, and Y. Yoshida, "Real-life in-home viewing conditions for flat panel displays and statistical characteristics of broadcast video signal," Jpn. J. Appl. Phys., vol. 46, no. 3B, 2007, pp. 1358-1362.
- [19] T. Fujine, Y. Yoshida, and M. Sugino, "The relationship between preferred luminance and TV screen size," Proc. SPIE 6808, 68080Z-1-12, 2008.
- [20] M. Dubbelman, G. L. Van der Heijde, H. A. Weeber, and G. F. J. M. Vrensen, "Changes in the internal structure of the human crystalline lens with age and accommodation," Vision Res., vol. 43, no. 22, 2003, pp. 2363-2375.
- [21] C. Ramsdale and W. N. Charman, "A longitudinal study of the changes in the static accommodation response," Ophthalmic Physiol. Opt., vol. 9, no. 3, 1989, pp. 255-263.
- [22] H. Anderson, A. Glasser, R. Manny, and K. Stuebing, "Age-Related Changes in Accommodative Dynamics from Preschool to Adulthood," Investigative Ophthalmology & Visual Science, 51, 2010, pp. 614-620.
- [23] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks, "The zone of comfort: predicting visual discomfort with stereo displays," J. Vis., vol. 11, no. 8, 2011, pp. 1-29.
- [24] R. Patterson, M. D. Winterbottom, and B. J. Pierce, "Perceptual issues in the use of head-mounted visual displays," Hum. Factors, 2006, vol. 48, no. 3, pp. 555-573.
- [25] M. Velger, "Helmet-Mounted Displays and Sights," Boston: Artech House, 1998.
- [26] Y. Otake, M. Miyao, S. Ishihara, M. Kashiwamata, T. Kondo, H. Sakakibara, and S. Yamada, "An experimental study on the objective measurement of accommodative amplitude under binocular and natural viewing conditions," Tohoku J. Exp. Med., vol. 170, 1993, pp. 93-102.
- [27] G. Heron, W. Charman, and C. Schor, "Dynamics of the accommodation response to abrupt changes in target vergence as a function of age," Vision Res., vol. 41, 2001, pp. 507-519.
- [28] G. Heron, W. N. Charman, and C. M. Schor, "Age changes in the interactions between the accommodation and vergence systems," Optometry Vision Science, vol. 78, no. 10, 2001, pp. 754-762.
- [29] R. Patterson, "Human factors of stereo displays: An Update," Journal of SID, vol. 17, 12, 2009, pp. 987-996.
- [30] F. W. Campbell, "The depth of field of the human eye," Int. J. Optics, vol. 4, no. 4, 1957, pp. 157-164.
- [31] I. E. Loewenfeld, "The Pupil: Anatomy, Physiology and Clinical Applications," Ames: Iowa State University Press, 1993.
- [32] P. Reeves, "The response of the average pupil to various intensities of light," J. Opt. Soc., vol. 42, 1920, pp. 35-43.
- [33] K. N. Ogle and J. T. Schwartz, "Depth of focus of the human eye," J. Opt. Soc. Am., vol. 49, 1959, pp. 273-280.
- [34] G. Smith, "Relation between Spherical Refractive Error and Visual Acuity," Optometry Vision Science, vol. 68, no. 8, 1991, pp. 591-598.
- [35] B. Wang and K. J. Ciuffreda, "Depth of focus of the human eye: Theory and clinical applications," Surv. Ophthalmol., vol. 51, no. 75, 2006, pp. 75-85.
- [36] R. Patterson and A. Silzars, "Immersive stereo displays, intuitive reasoning, and cognitive engineering," J. SID, vol. 17, no. 5, 2009, pp. 443-448.

Automobile Driving Interface Using Gesture Operations for Disabled People

Yoshitoshi Murata and Kazuhiro Yoshida
Faculty of Software and Information Science
Iwate Prefectural University
Takizawa, Japan

y-murata@iwate-pu.ac.jp, kyoshida@ipu-office.iwate-pu.ac.jp

Abstract— A steering operation interface has been designed for disabled people that uses right and left gesture operations. A questionnaire survey on gestures made with appendages had shown that gestures other than right and left ones were not suitable for driving a car. The interface incorporates both non-linear and semi-automatic steering control. Experiments using gyro sensors and a driving simulator demonstrated that driving operation using the foot, forefinger, wrist, or lower arm after training was close to conventional steering wheel operation. Sufficient practice in using the proposed interface should therefore enable users to achieve steering control close to that achieved with a steering wheel.

Keywords—automobile driving interface; disabled people; gyro sensor; gesture operation; appendage operation; driving simulator

I. INTRODUCTION

Disabled people generally want to stand on their own two feet, and achieving mobility is an important step in doing this. One way for them to enhance mobility is by driving automobiles to which driving-assistance devices have been attached. However, there has been a lack of development of new automobile driving interfaces that would enable disabled people, especially people with arm and wrist disabilities, to drive cars. Hence, we are designing a new steering operation interface for disabled people that is operated by gestures. We developed a prototype control device that used a gyro sensor, evaluated it by using a driving simulator and a skillful participant, and presented it at the Association for Community Health Improvement (ACHI 2013) [1].

The first auxiliary device for people with arm and wrist disabilities, the original of Honda's Franz system [2], was developed in the 1960s. A car is operated with only the feet in this system. Since the steering wheel is turned by pumping the pedals, its operation is not intuitive.

The autonomous car and the brain controlled car are ideal solutions for disabled people. Autonomous cars have been developed by many automobile manufacturers in addition to those by Google [3][4]. They need a very detailed 3D-map and many sensors to detect pedestrians, other cars, and obstacles around them. Therefore, their manufacturing costs must be expensive. Brain controlled cars have also been developed by researchers including those by automobile manufacturers [4][5]. A skillful driver for the brain control interface can indicate several kinds of commands. The

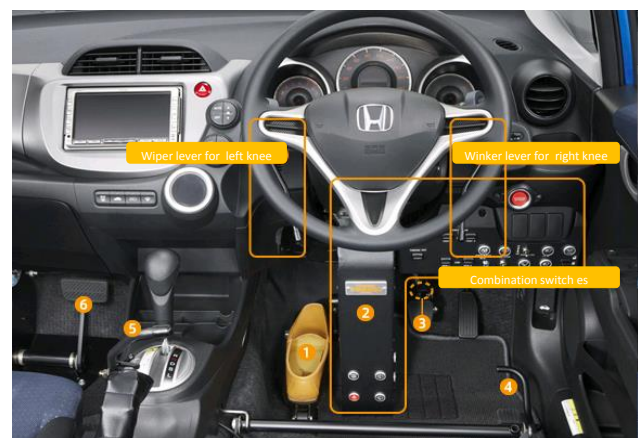
control is not accurate for letting the brain controlled cars operate within open traffic. Autonomous car technologies are needed for brain controlled cars to drive on usual roads. That is to say, brain controlled cars also need detailed 3D-maps and many sensors, and they are expensive. A current realistic solution for disabled people is to control steering with bodily appendages they can move.

The steering wheel in the system developed by Wada and Kameda was controlled with a joystick, and the brake and accelerator were controlled with another joystick [6][7]. This system has aided many disabled people, but strength is needed to operate the joysticks. Moreover, the levers onto which the joysticks were fixed had to be customized for the hand positions of individual users.

In any case, mechanical devices such as these lack flexibility and have to be customized for users. Hence, they are inherently expensive.

The on-going shift from hydraulic to electronic driving interface systems (e.g., steering and braking) means that systems combining computer chips with sensors can now be used to easily control these driving interfaces. Candidate sensors include Kinect sensors and gyro sensors.

In this paper, we verified the results we presented in ACHI 2013 by checking them in experiments, and investigated what movements by appendages drivers found to be natural by administering questionnaires.



- | | |
|-----------------------|----------------------------|
| (1) Steering pedal | (4) Selection bar for feet |
| (2) Steering box | (5) Side brake for knee |
| (3) Brake lock button | (6) Sub-brake for exercise |

Figure 1. Honda's Franz system

After related work is discussed in Section II, we will describe the driving simulator we developed to evaluate our proposed driving interface in Section III. Gestures, i.e., movements by appendages assigned to various functions are explained in Section IV and the driving interface equipment we developed is presented in Section V. The experimental evaluation we conducted is described in Section VI. The key points are summarized and future work is mentioned in Section VII.

II. RELATED WORK

Since the purpose of this study is to design a steering operation interface for disabled people that is operated by gestures, we introduce an advanced driving interface for people who have difficulty moving their arms and/or hands. We also introduce sensors that support driving a car by gesturing.

A. Driving Interface for Disabled people

The Franz system used by Honda is aimed at people who have difficulty moving their arms and hands. The user operates a car with only his or her feet [2]. It was originally implemented in a Honda Civic in 1982, which was the first vehicle to introduce the Franz system in Japan. It has now been implemented in a Honda Fit.

The steering wheel is turned right or left by pumping a steering pedal (see Fig. 1). The transmission is shifted into drive by lifting the selection bar, into reverse by pushing it down, and into park by pushing it further down. The turn signals and windshield wipers are operated by turning levers with the right and left knees. Power windows and lights are controlled by flipping switches up or down with the right foot or knee.

Wada and Kameda developed a car driving interface for people who do not have enough strength to control a steering wheel, accelerator pedal, or brake pedal. They used joysticks instead of a steering wheel and pedals. Steering, braking, and acceleration in the initial version [5] were controlled with one joystick. Two joysticks are used in the latest version shown in Fig. 2 [6]. The joystick on the right controls the steering and that on the left controls acceleration and braking. The relationship between the angle of the steering wheel and the angle of the joystick is a polyline, as seen in Fig. 3. This means that a driver can sensitively control the steering wheel around a neutral position and can turn the wheel quickly when making a wide turn. People who can freely move their hands can drive automobiles with this device.

However, such mechanical devices must be customized to fit individual users' disabilities and physical form.

B. Sensors for gesturing

Several driving interfaces using Kinect sensors have been developed. A user can drive a virtual car in a simulated world with the "Air Driving" interface developed by Forum8 by moving his or her hands and feet in front of a sensor [8]. Since there must be at least 50 cm between the sensor and the appendage that is gesturing, it cannot be used in actual cars. Rahman et al. developed an interface for car audio operation that used a Kinect sensor [9]. Although this

interface has been demonstrated in an actual car, its use as a driving interface (e.g., steering and braking) has not been investigated.

Döring et al. developed a multi-touch steering wheel that could not only control steering but also the car audio [10]. However, users with arm disabilities had trouble operating it.

Other examples of using acceleration sensors and/or gyro sensors as gesture operation interfaces include those in video games and home appliance remote controls [11].

Unfortunately, there were not existing sensors for gesturing to control a steering wheel, accelerator pedal, and a brake pedal in an actual car.



Figure 2. Wada and Kameda's joystick driving interface

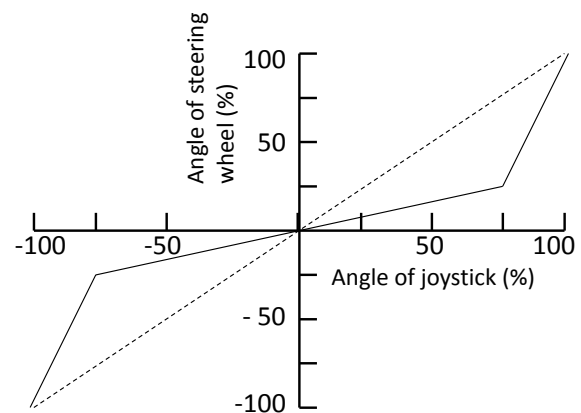


Figure 3. Relationship between angle of joystick and angle of steering wheel

III. DRIVING SIMULATOR

Before evaluating the proposed driving interface in an actual car, we evaluated it in a driving simulator to avoid traffic accidents. We introduce our developed driving simulator in this section.

A. Driving course

As one of our ultimate aims is to help disabled people obtaining a driver's license, we design a driving simulator to not only measure driving operability, but also exercise when driving. The three main issues with the driving simulator are:

- (1) The feel of driving has to be similar to that of a real car.
- (2) It has to be possible to measure the position of the car within driving lanes.
- (3) It has to be easy to choose various driving courses from real roads.

We extract road data from maps such as Google Maps. The creation tools prepare 3D roads from extracted data, as shown in Fig. 4 [12]. We use OpenGL [13] as the 3D program interface and develop a program using the “glut”, “sdl” [14], “glew” [15], and “OpenAL” tools [16].

First, the simulation program has to find a direction perpendicular to the parametric curve that expresses the center line of the road to create the width of an approximated road, and it then calculates the coordinates of a point shifted to the right or left of the center line, as shown in Fig. 5.

A tangential angle of an arbitrary point on the curve can be calculated as

$$\theta = \tan^{-1} \frac{dy}{dx}. \quad (1)$$

The point of the road edge is a position that shifts to the road width from an arbitrary point on the curve. The point of the road edge can be calculated as

$$\begin{aligned} x_r(t) &= W_r \cdot \cos(\theta - \frac{\pi}{2}) + x(t) \\ y_r(t) &= W_r \cdot \sin(\theta - \frac{\pi}{2}) + y(t) \\ x_l(t) &= W_l \cdot \cos(\theta + \frac{\pi}{2}) + x(t) \\ y_l(t) &= W_l \cdot \sin(\theta + \frac{\pi}{2}) + y(t). \end{aligned} \quad (2)$$

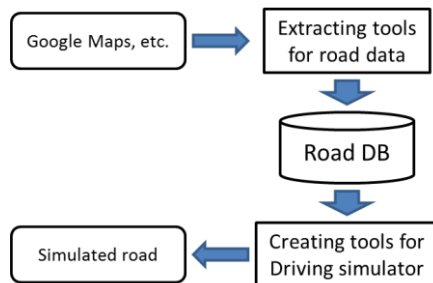


Figure 4. Outline for creating driving simulator road

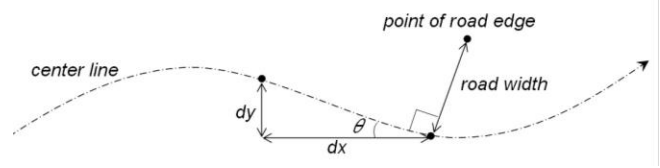


Figure 5. Method of creating 3D roads

3D road polygons are created by changing three-dimensional spline curve parameter “ t ” from zero to one, calculating many points on the road edge, and storing these points in a vertex array. Road center lines and lane lines are created by changing value W in Eq. (2).

When the width or the number of lanes at a curve's parameter, $t = 0$, differs from the width or number of lanes at $t = 1$, the simulation program finds that the road has a right- or left-turn-only lane. When a road has a right- or left-turn-only lane, the road width has to be gradually increased. The simulation program in our system calculates a smooth curve that expresses increments in the width of the lane. We used the sigmoid function to increase the width. The sigmoid function is a monotonic increase function and has one inflection point. Therefore, it is suitable for expressing a right- or left-turn-only lane.

The polygon for a crossing consists of all curve functions that connect the crossing. The calculated curve function's parameter “ t ” changes from zero to one in the same way as for a road, and a crossing polygon is created.

Fig. 6 has examples of a 3D road environment created by the simulation program according to this method.



(a) Example of straight road



(b) Example of crossing

Figure 6. Examples of created 3D roads

B. Motions of car

Two motions are simulated: gyration and acceleration [17][18].

1) Gyration

Steady gyrating motion is applied to the car under three main assumptions.

- The movement of the car is broadside motion of a rigid body. That is, the car is rigid and free of distortion.
- The speed is constant throughout each curve.
- The characteristics of the tires on the right are the same as those on the left.

The radius, R , of gyrating movement is given by the following equation, in which V is the running speed and δ is the steering angle.

$$R = (1 + CV^2) \frac{1}{\delta}. \quad (3)$$

The C is given by the following equation, in which the mass of the car is m , the cornering force on the front tires is K_f , that on the rear tires is K_r , the wheel base is l , and the distances between the car's center of gravity and the front and rear axles are l_f and l_r .

$$C = -\frac{m}{2l^2} \frac{l_f K_f - l_r K_r}{K_f K_r}. \quad (4)$$

Each parameter is set to produce driving characteristics similar to those of an actual car. The cornering force is controlled by adjusting the radius of the gyrating movement, i.e., the larger the radius, the stronger the cornering force.

2) Acceleration

The acceleration, A_c , of an actual car depends on the engine torque, the transmission gear ratio, the tire radius, the vehicle weight, and the engine speed. The engine speed depends on the degree to which the accelerator pedal is pressed.

Air resistance R_a and rolling resistance R_r are considered to be the total running resistance.

$$R_a = \frac{1}{2} C_d \rho S V^2, \quad (5)$$

where C_d is the aerodynamic coefficient, ρ is the fluid density of air, and S is the total surface area of the car.

$$R_r = C_{rr} mg, \quad (6)$$

where m is the mass of the car, C_{rr} is the rolling coefficient, and g is the gravitational acceleration. The resulting acceleration, A , is given by

$$A = A_c - (R_a + R_r). \quad (7)$$

C. Simulation display

There is an example view seen through the windshield in Fig. 7. The upper right shows the position of the car on the course. The operation monitoring tool we developed to facilitate operation is shown in Fig. 8. It helps the driver recognize the angle of the sensor from the angle of the steering wheel and the angle of the toes in case of rolling the ankle. It also displays the degree to which the accelerator or brake pedal has been pushed.



Figure 7. Example view through front window

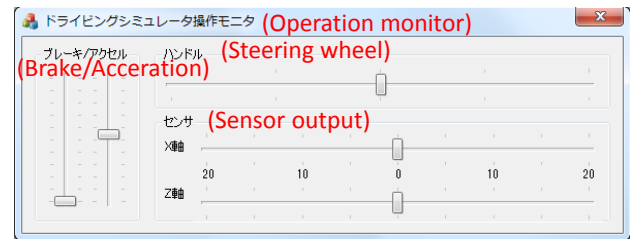


Figure 8. Operation monitoring tool

D. Measured data

Nine data items are measured.

- 1) Steering angle
- 2) Running speed
- 3) Distance driven and driving time
- 4) Position of car on course
- 5) Distance between left of car and left lane marker line
- 6) Distance between right of car and right lane marker line
- 7) Degree to which accelerator pedal was pushed
- 8) Depth to which brake pedal was pushed
- 9) Angle of car relative to driving direction

IV. GESTURES (MOVEMENTS BY APPENDAGES) FOR OPERATION

Here we describe the requirements for steering operation and control schemes that satisfy them. We then describe gestures for each body part on which a sensor is attached.

A. Operating functions

It is necessary to have door open/close, window open/close, wiper on/off, and turn signal on/off functions to drive an automobile in addition to the basic operations of steering, braking, and accelerating. Moreover, since automobiles typically have an audio system, a navigation system, and a climate control system, a driver should be able to operate these systems as well. Other than for the basic operating functions, a fine degree of control is not needed for the operating functions—they can generally be controlled by flipping a switch, as in Honda's Franz system. Moreover, voice-command control systems like that used by Samsung's Smart TV [19] could also be used. Of the basic operations requiring a fine degree of real-time control (steering, braking, and accelerating), we focus on steering, which requires the finest degree of control. The results from our research should easily be able to be transferred to braking and accelerating.

B. Steering operation requirements

Steering an automobile by moving bodily appendages should produce the same results as manually turning the wheel. Given this basic requirement, we derived four specific requirements.

- 1) *The automobile should be able to be steered within \pm about 500 degrees from the neutral position.*
 - *There should be a fine degree of steering control around the neutral position.*
 - *Steering should be quick when making a wide turn.*
- 2) *The driver should be able to keep the vehicle within the lane on both straightaways and curves of various radii at a normal driving speed.*
- 3) *The driver should be able to drive stably, and not zigzag, on straightaways.*
- 4) *The driver should be able to traverse a curve while keeping the steering wheel at a position fixed immediately before entering the curve and then exit the curve into a straightaway by gradually returning the steering wheel to the neutral position.*

C. Steering control

The steering wheel in an actual automobile can be turned about three complete revolutions from wheel lock to wheel lock ($\sim 1080^\circ$). In contrast, the movable angle of a joint angle is about $20\text{--}90^\circ$, which is much less than that of a steering wheel. Hence, it is impossible to control steering with a joint angle because it is not the same as that of a steering wheel.

We thus introduce **the non-linear steering control** and **the semi-automatic steering control**. The direct operation angle and automatic steering angle are determined, as outlined in Fig. 9, which illustrates steering control with a foot and an ankle. The driver operates using the non-linear steering control within the direct operation angle. Although Wada and Kameda used a polyline function for their steering control with a joystick, we used a non-linear function ($y = x^n$). We set $n = 3$ on the basis of our experimental results, which are described in Section VI. The steering angle increases automatically when it is beyond the direct operation angle. The rate of increase depends on the speed of

the car; the faster the car moves, the lower the rate. The driver can stop further increases in the steering angle by lifting his or her toes (about 20° for the case in Fig. 9). The driver can return the steering angle to the neutral position by lowering his or her toes. Drift error is canceled by carrying out this operation while the car is running straight.

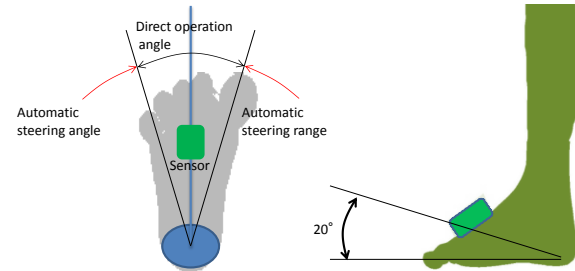


Figure 9. Example of controlling steering by foot

D. Sensor attachment and actions

We measure the car control characteristics for several actions: rolling the ankle, moving the forefinger, moving the wrist, rolling the lower arm, moving the lower arm backward and forward, and moving the upper arm backward and forward. The motions of appendages to turn a car to the right or left may differ for individuals. For example, someone may move his or her finger down to turn a car to the right, while another may move his or her finger up to turn to the right. Hence, we obtained information about different motions by individuals by administering questionnaires before measuring the car control characteristics. There were 29 participants. The results we obtained from the questionnaires are summarized in Table I. Most people chose the same action for motions that led to the right or left, such as rolling the lower arm. However, the number of people who chose alternative motions was roughly the same for motions that did not lead to the right or left such as moving his or her fingers up or down. For example, 86% of people chose rolling their right lower arm to the right to turn a car to the right. However, 52% of participants chose "up" and 48% of them chose "down" for moving their left finger up or down.

We predicted that there would be opposite relations before the questionnaires were administered between moving the right lower arm forward or backward and moving the left lower arm backward or forward, and moving the right upper arm forward or backward and moving the left upper arm backward or forward. Nevertheless, there were not extensive opposite relations, but slightly opposite relation in the results obtained from the questionnaires.

The positions of the sensors and the motions of appendages are as follows.

[Rolling ankle]

We considered using knee turning and knee movements to move gyro sensors. However, as these movements produce a narrow angle of movement, we roll the ankles. The sensor is placed on top of the foot, as shown in Fig. 9. The sensor moves when a foot are pivoted right or left on the heel.

TABLE I. RESULTS FROM QUESTIONNAIRES FOR MOTION OF TURNING CAR TO RIGHT

| Left hand | | Turn to right | Right hand | | Turn to right |
|-----------|----------------|---------------|------------|-----------------|---------------|
| Finger | Up | 15 | Finger | Up | 18 |
| | Down | 14 | | Down | 11 |
| Finger | Right | 26 | Finger | Forward (Right) | 27 |
| | Forward (Left) | 3 | | Left | 2 |
| Wrist | Up | 16 | Wrist | Up | 16 |
| | Down | 13 | | Down | 13 |
| Wrist | Right | 24 | Wrist | Right | 27 |
| | Left | 5 | | Left | 2 |
| Lower arm | Forward | 11 | Lower arm | Forward | 21 |
| | Backward | 18 | | Backward | 8 |
| Lower arm | Right | 25 | Lower arm | Right | 28 |
| | Left | 4 | | Left | 1 |
| Upper arm | Forward | 17 | Upper arm | Forward | 14 |
| | Backward | 12 | | Backward | 15 |

[Moving forefinger]

There are two movements for a forefinger. The first is when the back of the hand is raised upward, the forefinger can be moved up and down, the second is when the back of the right hand is toward the right; the right forefinger can move right or left. About half the participants for the former motion chose the up direction to turn a car to the right, and the rest chose the down direction, as summarized in Table I. This means that about half of all people may make operational mistakes in the first. However, about 90% of people chose the same operation to turn a car to the right in the second. This means that most people will not make operational mistakes. Therefore, we choose the latter motion to drive a car with the forefinger. A sensor is placed on the second joint of the forefinger and is moved as shown in Fig. 10.

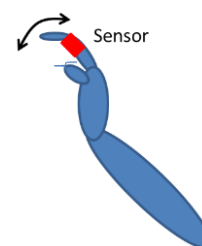


Figure 10. Moving forefinger (top view)

[Moving wrist]

There are two motions for the wrist, which are the same as those for the forefinger. We choose a motion when the back of the hand moves forward to the right and the hand moves right or left for the same reason as that for the forefinger.

The sensor is placed on the back of the hand and is moved as shown in Fig. 11.

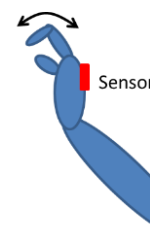


Figure 11. Moving wrist (top view)

[Moving lower arm]

We considered two motions for the lower arm. The first was moving the lower arm forward or backward, and the second was rolling the lower arm right or left. Most people will not make operational mistakes in rolling the lower arm right or left (the latter case) from the results in the questionnaires. In contrast, ~30% of people may make operational errors in the former case. However, since the ratio is less than that for the forefinger and wrist cases, we measure both their control characteristics.

The sensor is placed on the lower arm and is rolled as shown in Fig. 12.

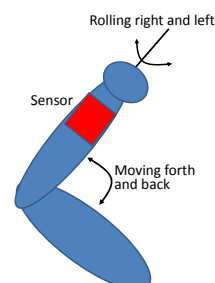


Figure 12. Rolling and moving lower arm (top view)

[Moving upper arm backward and forward]

There is not big difference between the number of participants who chose to move their right lower arm forward to turn a car to the right and the number who chose to move it backward. Therefore, this gesture is not basically suitable for the driving interface. Nevertheless, we measure it this time.

The sensor is placed above the elbow and is swung as shown in Fig. 13.

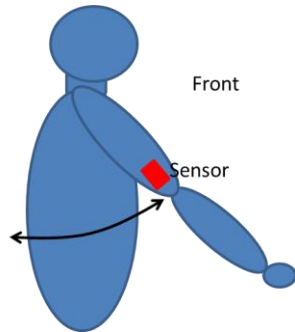


Figure 13. Moving upper arm (side view)

V. DRIVING INTERFACE EQUIPMENT

In this section, we describe various types of sensors, the type we used and its preprocessing, and the system configuration for the driving interface and its connection to a driving simulator.

A. Candidate sensors or devices

We considered five types of sensors or devices to control steering.

- Kinect sensors
- Video cameras
- Rotary encoders
- Gyro sensors
- Strain gauges

It is problematic to use Kinect sensors or video cameras because the units have to be attached to cars, and the locations for possible attachment are limited. Moreover, there must be at least 50 cm between a Kinect sensor and the gesturing appendage, which greatly limits the possible locations for attachment, as was previously mentioned.

Rotary encoders require the use of a mechanical adapter to measure the joint angle of fingers, elbows, or ankles.

Since gyro sensors are not only affected by the joint angle but also vehicle motion, they must be attached to vehicles to eliminate this effect. Moreover, gyro sensors have drift error that increases cumulatively and it is very difficult to completely remove this cumulative error. However, gyro sensors can very flexibly be attached to bodily appendages.

Strain gauges do not have drift error and are not affected by vehicle motion. Therefore, they are better suited to measuring joint angles than gyro sensors, when it is possible to attach them to joints. We plan to investigate their usefulness in future work.

B. Used sensor and its preprocessing

We used gyro sensors to evaluate gesture operations as the initial stage of our research regardless of various problems such as the influence of movements of a car to apply them to a real car. The main reasons for this were that gyro sensors are very flexible and can be attached to bodily appendages and moving joints, and the simulated car did not physically move. We used a practical 3-axis accelerometer system (ATR Promotions, WAA-010 [20]) as the sensor terminal. Not only a gyro sensor but also an accelerometer sensor, a terrestrial magnetic sensor, and a Bluetooth unit are mounted on it.

The drift phenomenon in gyro sensors is a problem for driving simulators. Angular velocity under static conditions is not zero but some other value. Example data of angular velocity and its integrated data (angle) on the WAA-006 are presented in Fig. 14. The value for angular velocity is small and varies around zero. However, since its characteristics of distribution are not normal, the value for the angle remains plus or minus for a long period. A moving average filter and 1st function adaptation are applied to reduce the drift phenomenon. The slope and the intercept of the 1st adapting function are derived from various moving averaged angular velocity data, e.g., 200 samples, under static conditions. They are automatically renewed to the latest data. Decisions under static conditions are determined from moving averaged angular velocity data that are within some threshold level. We could choose the average of the 10 largest angular velocity data under static conditions in this paper. The final output data of angular velocity are offset by the adapted 1st function. There is an example of compensated data under static conditions in Fig. 15. They remained zero under static conditions. When a gyro sensor begins to move, angular velocity data are offset by the latest adapted 1st function. Angular velocity data and their integrated data that are angle data are provided in Fig. 16 when a gyro sensor is being moved by foot.

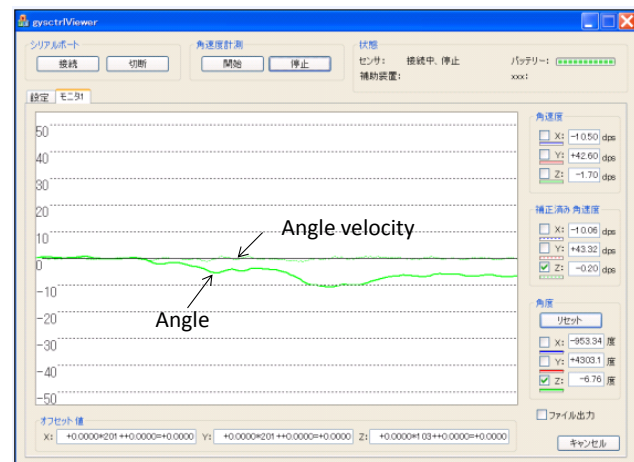


Figure 14. Uncompensated data for WAA-010 under static conditions

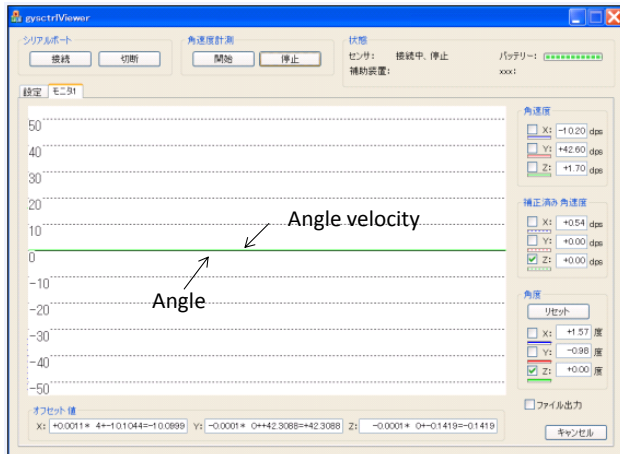


Figure 15. Compensated data of WAA-010 in static condition

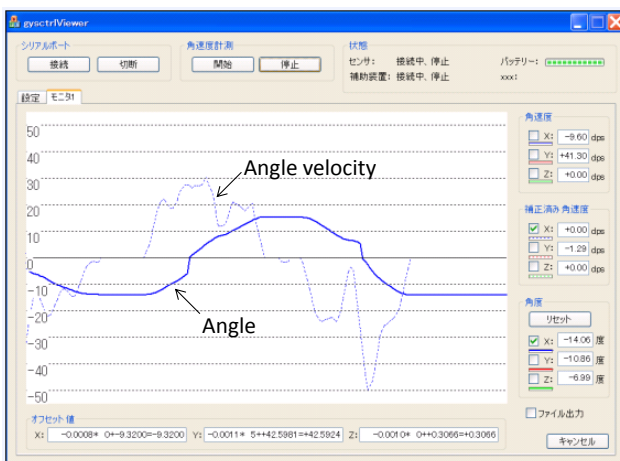


Figure 16. Angle velocity and angle data while moving WAA-010

C. Sytem configuration

We designed the driving interface we propose that will not only be used for operating a simulated car but also a real car. Therefore, the driving interface unit must be connected to a driving simulator and a real car. We designed our driving interface unit to comprise a wireless gyro sensor terminal and a PC. We developed an angle data convertor and a driving simulator and mounted them on a PC. Since the driving interface unit should be tested with an immersive driving simulator before it is applied to a real car, we also designed it to be connected to an immersive driving simulator. Therefore, the data convertor was comprised of a serial communication library, a sensor control library, an angle conversion library, and an immersive driving simulator communication library. The relationship between libraries had the layered structure shown in Fig. 17. The serial communication library provided communication functions for the Bluetooth unit. The sensor control library sent commands to control the sensor terminal WAA-010

such as the sampling rate. The angle conversion library transformed data received from the sensor terminal to a data format to enable steering control. This library contained integration that converted the angular velocity to an angle and the drift compensation function described in the previous paragraph. Since the immersive driving simulator at our university did not have a movable pedestal, this unit did not have a gyro sensor that cancelled the movements of the car. There is an example dialog box for setting the parameters in Fig. 18. It is possible to monitor output data and to set up connection parameters and compensation parameters for drift errors. The practical parameters are as follows;

Connection:

- Serial port number (COM10)
- Sampling period (5 ms)
- Average number of sampled data (5)

Drift error compensation:

- Number of data obtained for initial data (200).
- Windows size for moving average (5).
- Number of array lists for storing larger values (10).

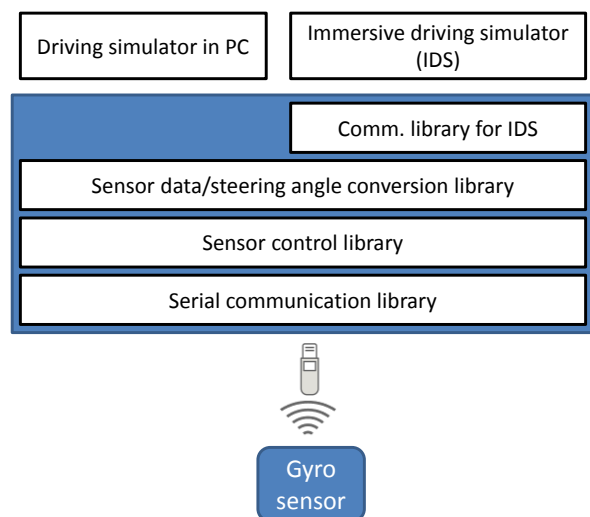


Figure 17. Configuration of the driving interface equipment

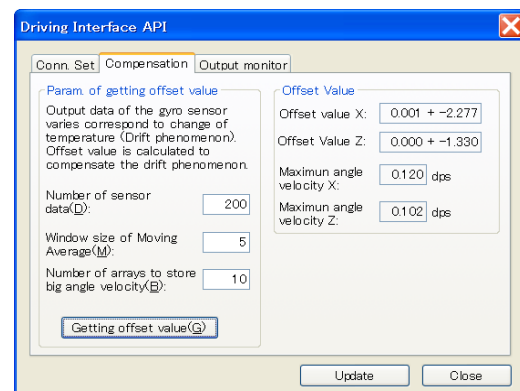


Figure 18. Dialog box for setting parameters

The values within parentheses are examples of each parameter. The offset values X and Z and the maximum angular velocities X and Z under static conditions were calculated from the above data. These data are presented on the right of the dialog box.

VI. EVALUATION

We evaluated the ability of a driver to keep within the lane and to drive stably without zigzagging on straightaways while using the proposed driving interface. Since potential users likely have difficulty moving their arms and hands, we first measured fundamental data for the foot. We then measured data for other parts of the body.

A. Evaluation issues

We calculated the ratio of lane departure (RLD) and the standard deviation of the driving gap (SDDG) to analyze performance against the 2nd and 3rd steering operation requirements described in Section IV-B;

- As can be seen from Fig. 19(a), lane departure means that one or more of the tires run on or across a lane marker line. RLD is the ratio between the distance driven and the distance during, which the car left the lane.
- As we can see from Fig. 19(b), the driving gap is the distance between the lane center and the car's center line. A value of zero means that the car is centered in the lane. SDDG is calculated using the values obtained for the car running on a straight portion of the course.

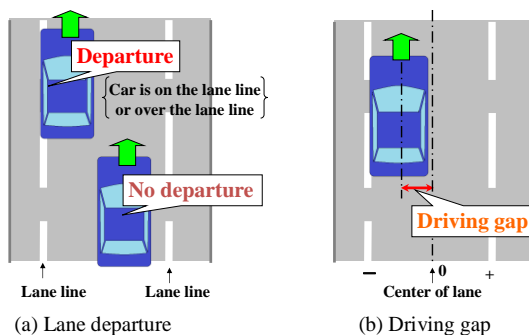


Figure 19. Lane departure and driving gap

B. Test course

As one of our ultimate aims is to help disabled people obtaining a driver's license, we use a driving route based on a typical course at a driving school (Fig. 20) to measure RLD and SDDG. It is comprised of a rectangular outer course, two crank-shaped courses, two S-shaped courses, and two parallel parking courses. The outer course is 300×120 m and had a corner radius of 20 m. A 3.3-m-wide driving lane runs in each of the courses.

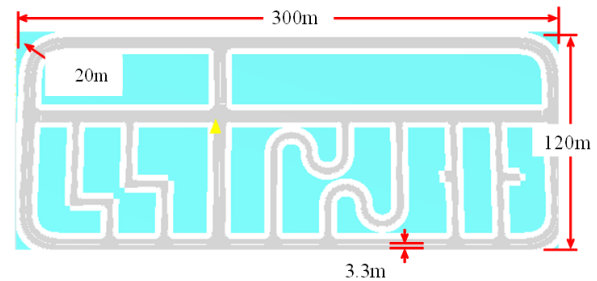


Figure 20. Driving course for evaluations

C. Results for operation by foot

Since the first objective of this research is to develop a driving interface for people with disabled arms and fingers, we focus on foot-controlled steering, as illustrated in Fig. 9.

A non-linear function ($y = x^n$) is used to control steering within the direct operation angle. We measured the position of a car on the course and calculated the RLD and SDDG for $n=1-4$ in $y = x^n$. We also measured and calculated the same data using a steering-wheel-type game controller for comparison. The measured and calculated data used for the non-linear function are listed in Table II for one of the four participants, who was person Y who had a great deal of experience driving a car using his foot with the proposed driving interface. These data were measured and calculated in September 2012. Since the details of the experiment by Wada and Kameda have not been published, we could not compare the precision of our control function with theirs.

TABLE II. MEASURED AND CALCULATED DATA FOR NON-LINEAR CONTROL FUNCTION FOR SKILLFUL PARTICIPANT

| | $y=x$ | $y=x^2$ | $y=x^3$ | $y=x^4$ | $y=x$ | Game str. wheel |
|----------------------|-----------------|---------|---------|---------|----------------|-----------------|
| Operating body part | Left feet | | | | | |
| DOA* | $\pm 20^\circ$ | | | | $\pm 15^\circ$ | |
| SWA** | $\pm 180^\circ$ | | | | $\pm 30^\circ$ | |
| Distance driven (km) | 7.91 | 7.91 | 7.91 | 7.91 | 2.39 | 2.37 |
| Ave. speed (km/h) | 26.6 | 26.3 | 25.5 | 27.3 | 14.3 | 30.4 |
| RLD (%) | 0.15 | 0.38 | 0.24 | 0.91 | 9.8 | 0 |
| SDDG (m) | 0.21 | 0.24 | 0.27 | 0.29 | 0.15 | 0.11 |

*DOA: Direct operation angle

**SWA: Corresponding steering wheel angle
(Measured and calculated in Sep. 2012)

We initially thought that a driver could easily operate the car by using the semi-automatic steering control. However, the RLD and SDDG were much worse than those with the game controller when the direct operation angle (DOA) was $\pm 15^\circ$ and the corresponding steering wheel angle (SWA) was $\pm 30^\circ$. We observed that it was very difficult to drive the car using the semi-automatic steering control during typical driving maneuvers. Hence, we changed DOA to $\pm 20^\circ$ and the

corresponding SWA was $\pm 180^\circ$. It is possible to drive a car through most of the corners in normal driving within these parameters. Although neither of the two participants in this experiment negotiated the corners of the rectangular outer course smoothly at $\text{DOA} = \pm 15^\circ$ and $\text{SWA} = \pm 30^\circ$, as seen in Fig. 21, both of them could negotiate the same corners at $\text{DOA} = \pm 20^\circ$ and $\text{SWA} = \pm 180^\circ$. We concluded that controlling the car with the semi-automatic steering control was not suitable for normal driving except for parking and traversing the crank- and S-shaped courses. Detailed data traversing the crank- and S-shaped course will be given later.

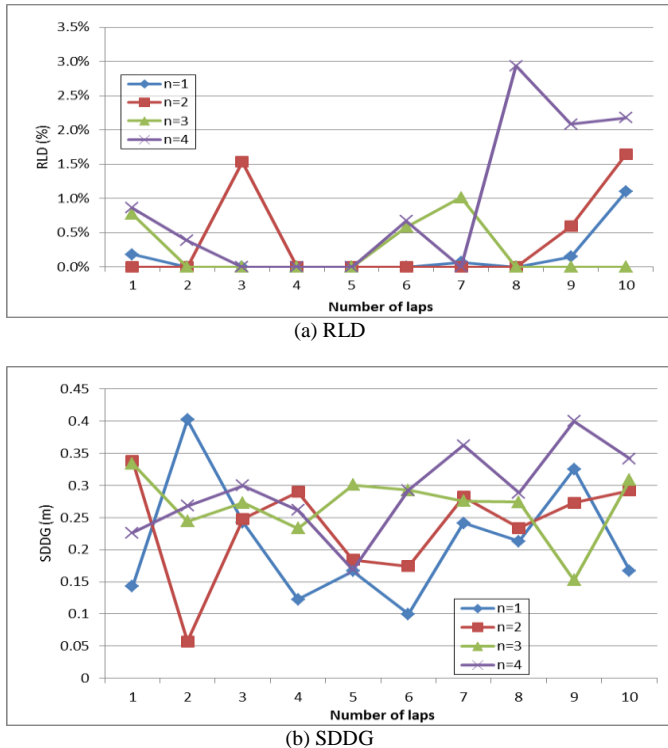
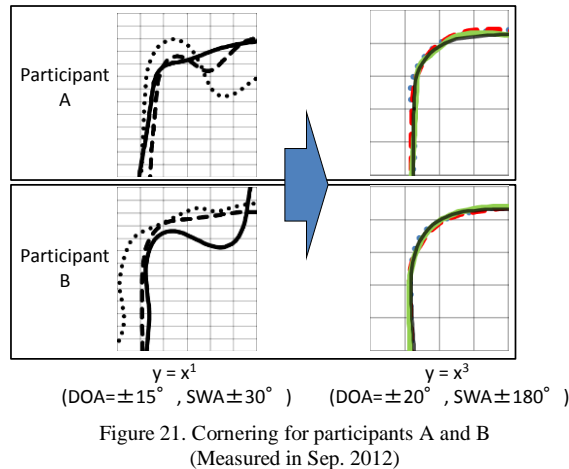


Figure 22. Car control characteristics against the number of laps
(Measured in Oct. 2012), * Average speed: 26.4 km

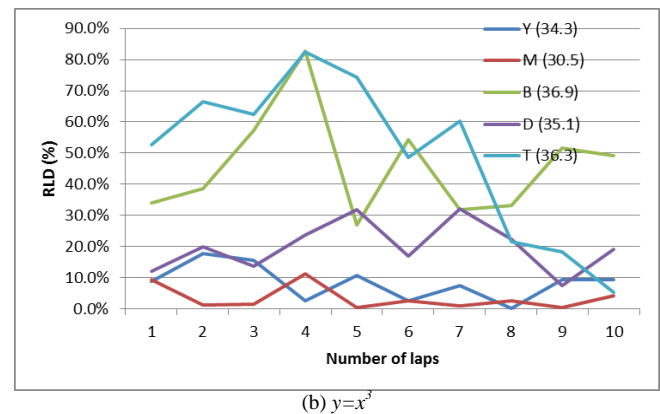
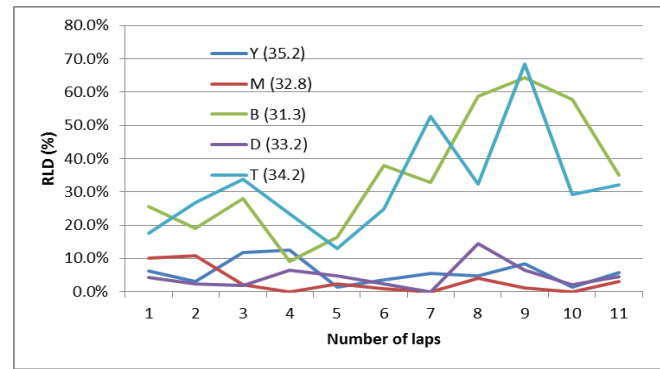


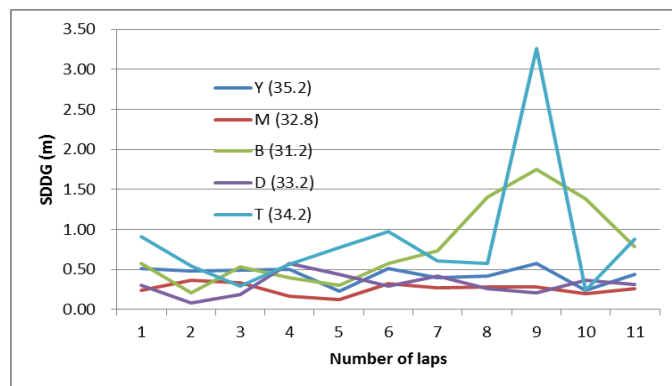
Figure 23. RLD against the number of laps (Measured in Jul. 2013)
* Average speed for each participant: written within parentheses

We considered which non-linear function was best to control the car. RLD and SDDG are the smallest for $n = 1$, as listed in Table II. However, the difference in values between $n = 1$ and $n = 3$ is negligible. Since the driving interface needs to enable a car to be driven for a long time, we measured 10 laps of the outer course for each function. The data are plotted in Fig. 22. There is not a big difference for the number of laps and the non-linear function in SDDG, but RLD is the smallest for $n=3$ during the last three laps. Since these data were measured and calculated for participant Y in Oct. 2012, we measured and calculated RLD and SDDG for four other participants in addition to participant Y at $y = x$ ($\text{DOA} = \pm 20^\circ$, $\text{SWA} = \pm 180^\circ$) and $y = x^3$ ($\text{DOA} = \pm 20^\circ$, $\text{SWA} = \pm 180^\circ$) in July 2013. The measured and calculated data for RLD and SDDG in each lap are plotted in Figs. 23 and 24. The data including those on participant Y measured and calculated in July 2013 are rather worse than those measured in Oct. 2012. Participants B and T, especially, could not drive the car well. RLD and SDDG become better in later laps at $y = x^3$, and become worse in later laps at $y = x$. B and T's data clearly reveal this tendency, and their values come close to the values of the other three participants. Although data for the other three participants have the same tendency, the changes in them are small. This tendency is the same as the data for participant Y measured and calculated in Oct. 2012. Participant Y's data in Oct. 2012 are outstanding. RLD during the last three laps is zero percent. Yet his data in Jul.

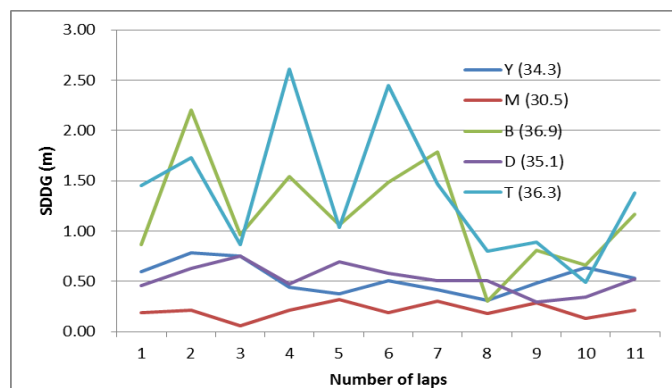
2013 are worse than his data measured and calculated 10 months earlier. Their values are roughly the same as those for participants M and D.

Although the quantity of data is insufficient, these issues suggest that most drivers can drive cars well with their feet, but they have to continue to train (i.e., drive) to retain high levels of skill. The non-linear function of $y=x^3$ may be better than $y=x$.

We calculated the RLD for the crank- and S-shaped course for four participants as well in Oct. 2012. The measured and calculated data are listed in Table III. The values are their averages. Traversing the crank- and S-shaped course is more difficult than traversing the rectangular course. The driver had to use both the non-linear and the semi-automatic steering control. Driving precision for these two courses differed greatly. As shown in Figs. 25 and 26, the precision of participant S is very close to that with the game steering wheel while that of participant C substantially diverged from it. This indicates that performance with the steering operation interface we propose should approach that with a steering wheel as the amount of practice and experience increases.



(a) $y=x$



(b) $y=x^3$

Figure 24. SDDG against the number of laps (Measured in Jul. 2013)

* Average speed for each participant: written within parentheses

TABLE III. MEASURED AND CALCULATED DATA FOR CRANK- AND S-SHAPED COURSES

| Operation device | Crank-shaped course | | S-shaped course | |
|---------------------|---|--------|-----------------|--------|
| | Game wheel | Sensor | Game wheel | Sensor |
| DOA | $\pm 20^\circ$ | | | |
| SWA | $\pm 180^\circ$ | | | |
| Cont. function | $y = x^2$ and semi-automatic steering control | | | |
| Distance driven (m) | 239.7 | 234.6 | 370.8 | 377.4 |
| Ave. speed (km/h) | 9.7 | 8.3 | 15.0 | 12.9 |
| RLD-Ave. (%) | 17.2 | 22.9 | 8.9 | 16.8 |
| RLD-Max. (%) | 26.6 | 42.8 | 21.2 | 40.8 |
| RLD-Min. (%) | 6.6 | 8.5 | 0 | 1.2 |

(Measured in Oct. 2012)

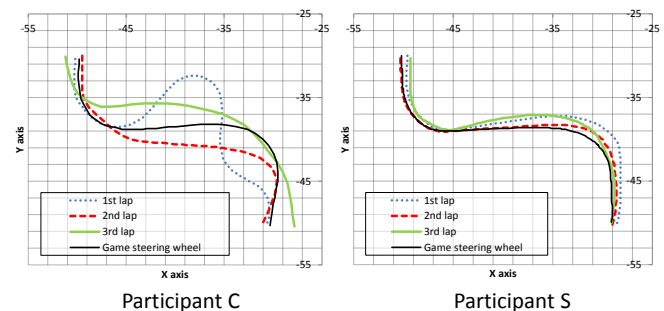


Figure 25. Crank-shaped course performance for participants C and D (Measured in Oct. 2012)

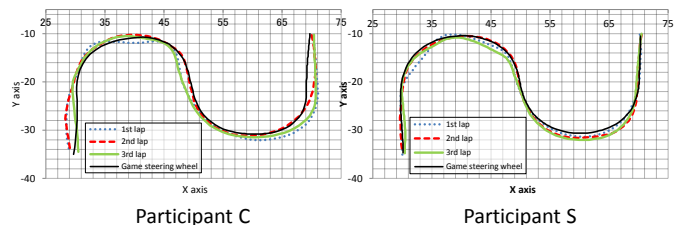


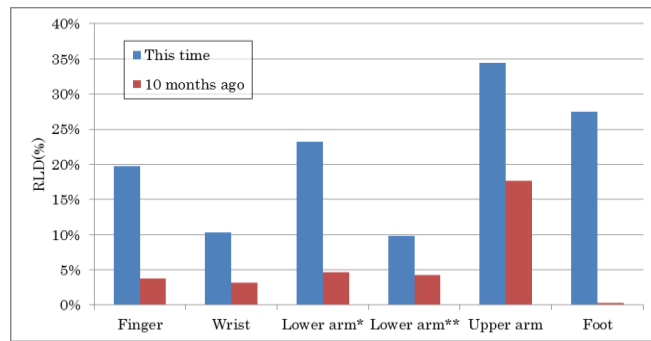
Figure 26. Precision on S-shaped course for participants C and D (Measured in Oct. 2012)

D. Results from other appendages

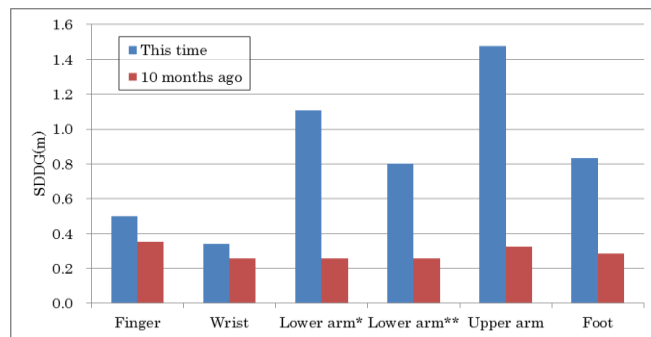
We measured and calculated the RLD and SDDG for steering control using the forefinger, wrist, lower arm, and upper arm in addition to the foot to examine to what extent the proposed driving interface could be applied to various types of disabilities. The data are presented in Fig. 27. There are two kinds of bars in each graph. One was measured for participant Y in Oct. 2012, and the other was measured for four participants that did not include Y in Aug. 2013. Each data was an average of driving three laps of the outer course. The movements are illustrated in Figs. 15–18. They drove three laps of the outer course.

Y's data for each appendage in Oct. 2012 are better than the data in Aug. 2013, especially Y's RLD for the foot in Oct. 2012 is very small. The reason why Y's feet data are very good is that Y trained for long periods and had a great deal of skill in controlling the car with his feet. His lengthy training and exceptional skills must have affected driving with his other appendages because their data were rather good.

There are large differences between appendages in Aug. 2013. Since this was the first time for participants who took part in the experiment to drive a simulated car with their appendages, ease of driving with each appendage for each participant clearly became apparent.



(a) RLD for several body parts



(b) SDDG for several appendages

Figure 27. Car control characteristics for several appendages

* Moving lower arm backward and forward

** Rolling lower arm right or left

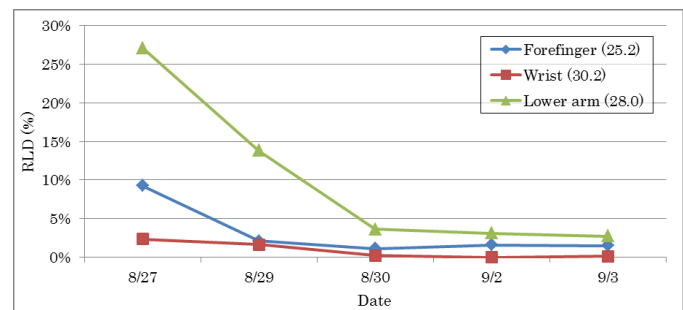
*** Average speed: 24.8 – 33km

The data for the upper arm is the worst because in both experiences it must have been difficult to precisely move the upper arm (and shoulder).

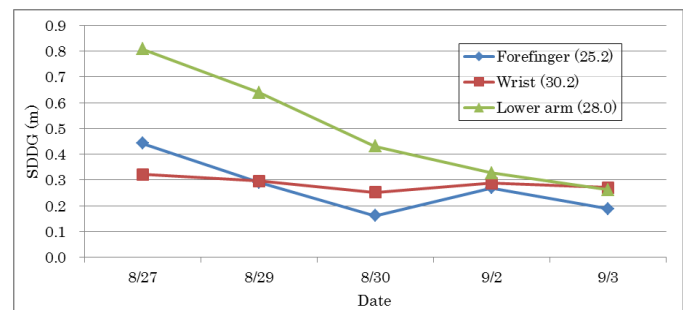
In any case, measured and calculated data except for the data for Y's feet in Oct. 2012 given in Fig. 27 are not sufficient to drive a car. At least, driving characteristics RLD and SDDG with each appendage must be closer to that with the game wheel. The data for participant Y measured in Oct. 2012 shows that more practice should enable users to achieve steering control closer to that with a steering wheel. Therefore, we measured driving characteristics RLD and SDDG against the training period. Moving forward or backward with the upper arm and the lower arm are not

suitable for driving a car from Table I. RLD and SDDG for both of them are not good enough in Fig. 27. We chose turning right or left with the forefinger and wrist, and rolling right or left with the lower arm. There was one participant for each gesture. The data are provided in Fig. 28. We measured the data after 30 minutes practice each day. Each data was an average of driving three laps of the outer course. The car control characteristics data improved in initial three day (90 minutes). Especially, the participant for rolling the lower arm improved in driving a car dramatically. RLD and SDDG became less than 3% and 0.3m after 5 days practice. But, SDDGs of the participant for the forefinger and the wrist on 2 September became worse than those on 31 August. The reason is that 31 August and 1 September were weekends, and the patients could not practice for two days.

Most people should be able to drive a car well by a few hours of practice regardless of irrespective of their body characteristics. But, they have to continue to train to retain high level of skills.



(a) RLD



(b) SDDG

Figure 28. Car control characteristics against the number of days for practice (Measured in Aug. and Sep. 2013)

* Average speed for each participant: written within parentheses

VII. CONCLUSION

The steering operation interface we proposed for disabled people uses gesture operation. Questionnaires on gestures made with appendages indicated that gestures not leading to the right or left were not suitable for driving a car. The interface incorporated both non-linear and semi-automatic steering control. Simulated experiments using foot control and gyro sensors indicated that semi-automatic steering control was only suitable for parking and traversing crank- or s-shaped courses, and that non-linear steering control ($y=x^3$)

was better than linear steering control ($y=x$) for typical driving maneuvers. Data on worse drivers (participants in the experiment) revealed that they could not remain stable in later laps at $y=x$, but they improved in later laps at $y=x^3$. Driving operations by training the forefinger, wrist, and lower arm (rolling) were close to that achieved with a steering wheel. More practice in using the new interface should enable user to achieve steering control that is closer to that with a steering wheel.

We plan to develop a prototype of a control device using strain gauges instead of gyro sensors to avoid influencing car movements and to evaluate driving operations with it. We also plan to evaluate our proposed interface in an actual car.

ACKNOWLEDGMENT

This research has been conducted as part of the Iwate Strategic Research Foundation. We would like to express our appreciation to students in Murata-Lab, Iwate Prefectural University for taking data with our experiment.

REFERENCES

- [1] Yoshitoshi Murata, Kazuhiro Yoshida, Kazuhiro Suzuki, and Daisuke Takahashi, "Proposal of an Automobile Driving Interface Using Gesture Operation for Disabled People," IARIA ACHI 2013, March 2013.
- [2] Development of Honda's Franz System Car;
<http://world.honda.com/history/challenge/1982franzsystemcar/index.html>, December 2013.
- [3] Erico Guizzo, "How Google's Self-Driving Car Works," IEEE Spectrum, February 26, 2013.
- [4] Autonomos Labs, <http://www.autonomos.inf.fu-berlin.de/>, December 2013.
- [5] Daniel Göhring, David Latotzky, Miao Wang, and Raul Rojas, "Semi-Autonomous Car Control," Intelligent Autonomous System 12, Springer, pp. 393-408, 2013.
- [6] Joystick Driving System: allows wheelchair users to drive a car;
http://www.youtube.com/watch?v=EvMii_a7qi4, December 2013.
- [7] Masayoshi Wada and Fujio Kameda, "A joystick car drive system with seating in a wheelchair," IEEE IECON '09, pp. 2163-2168, November 2009.
- [8] FORUM8 Air Driving and RoboCar
<http://www.youtube.com/watch?v=LMr2dyfAzt0>, December 2013.
- [9] A.S.M. Mahfujur Rahman, Jamal Saboune, and Abdulmoteleb El Saddik, "Motion-path based in car gesture control of the multimedia devices," ACM DIVANet '11, Proceedings of the first ACM international symposium on Design and analysis of intelligent vehicular networks and applications, pp. 69-75, November 2011.
- [10] Tanja Döring, Dagmar Kern, Paul Marshall, Max Pfeiffer, Johannes Schöning, Volker Gruhn, and Albrecht Schmidt, "Gestural interaction on the steering wheel: reducing the visual demand," ACM CHI '11, Proceedings of the 2011 annual conference on Human factors in computing systems, pp. 483-492, May 2011.
- [11] Yoshitoshi Murata, Nobuyoshi Sato, Tsuyoshi Takayama, and Shinetsu Onodera, "A Gesture-based Remote Control for Finger Disabled People," IEEE, GCCE 2012, pp. 411-415, October 2012.
- [12] Shinya Saito, Yoshitoshi Murata, Tsuyoshi Takayama, and Nobuyoshi Sato, "An International Driving Simulator: Recognizing the Sense of a Car Body by the Simulator," Workshops in AINA 2012, W-FINA-S12.1, pp. 254-260, March 2012.
- [13] OpenGL – The Industry Standard for High Performance Graphics, <http://www.opengl.org/>, December 2013.
- [14] Simple DirectMedia Layer, <http://www.libsdl.org/>, December 2013.
- [15] The OpenGL Extension Wrangler Library, <http://glew.sourceforge.net/>, December 2013.
- [16] OpenAL Soft, <http://kcat.strangesoft.net/openal.html>, December 2013.
- [17] Masato Abe, "Automotive Vehicle Dynamics - Theory and Applications," Tokyo Denki University Press, 2008.
- [18] Giancarlo Genta, "We apply the steady gyrating movement to a car under following assumptions," World Scientific Publishing, 1997.
- [19] Samsung SMART-TV,
<http://www.samsung.com/us/2012-smart-tv/>, December 2013.
- [20] ATR Promotion WAA-010, <http://www.ATR-p.com/sensor10.html>, December 2013.

Business Process Modelling for Measuring Quality

Farideh Heidari

System Engineering Section
TPM Faculty
Delft University of Technology
Delft, The Netherlands
f.heidari@tudelft.nl

Pericles Loucopoulos

Manchester Business School
University of Manchester
Manchester, the United Kingdom
pericles.loucopoulos@mbs.ac.uk

Frances Brazier

System Engineering Section
TPM Faculty
Delft University of Technology
Delft, The Netherlands
f.m.brazier@tudelft.nl

Abstract—Business process modelling languages facilitate presentation, communication and analysis of business processes with different stakeholders. This paper proposes an approach that drives specification and measurement of quality requirements and in doing so relies on business process models as representations of business processes. The approach is presented in the form of a conceptual model and its application is demonstrated for a simplified version of a business process. However, communication becomes a challenge in cross-organizational business processes where multiple business process modelling languages are being practiced which calls for an abstraction as an integration of concepts of these business process modelling languages. In this paper, a business process integrating meta-model is presented as an abstraction of concepts of seven mainstream business process modelling languages. Attaining such level of understanding and specifying business processes fosters specification and measurement of quality requirements.

Keywords- *Quality requirements, Quality specification, Quality measurement, Business process, Business process modeling, Business process integrating meta-model.*

I. INTRODUCTION

Today, new businesses are demanding enterprises to understand the behaviour of the business systems and its influence on the development of information systems that supports their operation. Rapid business and organizational changes, knowledge intensity of goods and services, the growth in organizational scope, and information technology have intensified organizational needs for such understanding. The approach of describing organizations in terms of business processes not only helps organizations to be more responsive to the business and organizational changes but also helps them in the development of information systems. Moreover, the collective ability of business processes to achieve its requirements is central to achieving high-performance organizations.

Attaining a level of understanding and specifying business processes is a challenge, which calls for business process modelling [1]. Business Process Modelling is currently not only of core importance to the development of software systems [2, 3] but also in presenting, analysing

and improving business processes [4] within enterprises. Business process models are domain specific conceptual models that support presentation and integration of business processes' requirements within an enterprise. Linking business process quality requirements to business process concepts enables IT and business experts to define their requirements collaboratively at a common abstract level during the earliest stage of design and development of information systems. In addition to quality requirements, annotation of business process models with related information artefacts using domains' vocabulary leverages different concepts (e.g., goals, rules, patterns, motivation, etc.) into the scope of business process [5, 6].

Business process modelling provides the realization and the presentation of business processes in different levels of abstraction from individual concepts (e.g., activity), to composition of concepts (i.e., sub-processes), and to the business process as a whole. The motivation of this paper is to show how business process modelling can be deployed for quality specification and measurement of business processes in different levels of abstraction. The focus is not on the evaluation of the models themselves as it is assumed that the models are well-formed and syntactically and semantically correct.

Quality evaluation of business processes in the context of their models is not a straightforward task. Many different business process-modelling approaches have been developed, each with their own specific business process modelling languages designed to meet a specific business requirement. The proliferation of business process modelling techniques is realized as a notorious problem for business process management [7]. Standardization has been discussed for more than ten years, none of the proposals is commonly accepted as de facto standard in the industry [7]. In practice, multiple business process languages are often being used within one and the same enterprise. A systematic realization and representation of concepts and relationships between the concepts of different business process modelling languages in a business process meta-model is essential [8]. This meta-model is universal and language independent abstraction of

the concepts of today's mainstream business process modelling languages. This paper provides a brief introduction on an integrating business process meta-model and its application. Besides, evaluation of business process in through its model is addressed in this paper. The levels of business process concepts are considered for specification and measurement purpose. The approach is exemplified using a real-life business process in an industrial case.

The paper is organized as follows: Section II presents a brief summary of related works. Section III elaborates on business process integrating meta-model, and its application and it introduced an enriched version of it with quality factors. Section IV introduces the approach to specification and evaluation of quality of business processes' concepts. Section V illustrates the proposed approach for a business process and instantiate the framework for it. The paper concludes in Section VI with a number of observations, reflections and suggestions for future work.

II. RELATED WORK

Quality has been the topic of research in several closely related disciplines such as requirement engineering, software engineering, workflow analysis, industrial engineering, system dynamics and discrete event simulation [1].

Different levels of granularity can be considered for realizing and measuring quality in an enterprise involving many organizational layers from the very general, i.e., organization-wide quality to concepts of business processes. The analysis of the current state-of-the-art reveals variations in specification and measurement of requirements. The plethora of approaches has led to compare the existing approaches based on a set of criteria.

Synoptically, investigation of the most relevant approaches in following aspects will be considered in this section: (A) the way they are being practiced (i.e., methodology e.g., systematic or ad hoc), (B) representation of business process and quality requirement (modelling and language dependency), (C) generalizability of the approach (i.e., application scope e.g., generic vs. specific), (D) measurement method (e.g., quantitative vs. qualitative) is conducted in this section.

A. Methodology

While "focus of work", "required inputs", "expected outputs" and a "set of phases", "technique used" and possibly "support tool" are prescribed with details in an approach, the approach is considered to be systematic in terms of methodology (e.g., [14], [9], [10], [11], [12], [1], [13]); otherwise the approach is considered to be ad-hoc in terms of methodology (e.g., [14], [10],[15]).

Wolter et al. [16] deploys a method to assign elements of their security model to a process model. Capturing quality dimensions of a business process in the form of a framework are considered by Heravizadeh et al. [17]. A framework for

evaluation of business process quality is introduced by Kedad et al. [18]. A requirements engineering framework with the aim of allowing active stakeholder participation is introduced by Donzelli et al. [11]. Pourshahid et al. [19] introduces a framework to measure and align processes and goals subjectively. In their work, key performance indicators (KPI) are added to user requirement notation (URN) together with explicit goals for each business process. A scenario-based methodology and a toolset for business process modelling and analysis is introduced by Glykas [12]. The approach defines and measures KPIs in qualitative as well as quantitative manner.

Heidari et al. [13] proposes a systematic approach in the form of meta-models and method steps to capture and evaluate quality of individual concepts of business processes, considering non-functional requirements defined by stakeholders. The evaluation results are compared with the quality objectives derived from non-functional requirements. They identify quality dimensions of performance, efficiency, reliability, recoverability, permissibility and availability for corresponding business process concepts and introduce objective and quantitative formulae for evaluating them.

The approach by Firesmith [10] proposes a checklist of questions over which defects in software-intensive system architectures would be realized. Measurement is included in the structure although the process toward the measurement is not discussed. In a theoretical attempt, Lohrmann et al. [15] provides a definition for business process quality and introduce business process quality model. There are no details provided on how the measurement should be conducted.

B. Modelling and language dependency

Modelling is concerned with the way an approach represents a business process. The consideration here is the use of formal or semi formal languages in the representation (e.g., [8], [17], [14], [9], [20]) Language dependency examines this fact if an approach's focus is on a specific language. Language independent approaches are not tied to any specific modelling languages (e.g., [18], [21], [13], [1], [8],[22]).

Heidari et al. [13] identifies quality metrics and factors for business process having its model as a given. Their approach is language independent and considers different concepts of business process (i.e., event, input, output and activity).

Heinrich et al. [23] uses the quality characteristics and attributes of processes. They distinguish on the basis of the ISO/IEC standard for software quality [14] to enhance BPMN. Saeedi et al. [14] proposes a set of quality requirement factors for BPMN concepts. Role Activity Diagram notation is considered for representation of business processes by Aburub et al. [20]. The strategic rationale for the choice of business processes to be specified in BPMN

models and described in terminology familiar to business people are considered by Decreus et al. [9].

Heidari et al. [8] introduces a business process meta-model as an integration of concept of seven business process modelling techniques. The meta-model is enriched with quality related information (i.e., quality factors). The result presented as a quality-oriented meta-model encompassing quality factors of throughput, cycle time, timeliness, cost, resource efficiency, cost efficiency, maturity, recoverability, security and availability.

C. Application scope

The application aspect is concerned with the target of the approach. Generic approaches can be applied to all or most situations (e.g., [22], [18]). Specific approaches (e.g., [16], [20]) are dedicated to a particular class or application or business sector.

From a managerial point of view, TQM and ISO quality standards act as general guidelines that are applicable to all types of organization regardless of the types of product and service they offer, organization size, turnover, location, type of industry, etc. The literature on business process performance measurements systems such as [24] and [25] try to provide generic guidelines for developing business process performance measurement systems. Quality tools and techniques [26] and tools for process improvements such as Kaizen Blitz, Poka-Yoke and process simulation [27] are generic and try to assist stakeholders in improving their processes and presenting output of quality measurements without realizing quality dimensions, factors and metrics.

Wolter et al. [16] focuses on security requirements. Aburub et al. [20] introduces an approach in remodelling business processes for identification and inclusion of Non-Functional Requirements (NFRs) for a specific case. With the focus on quality of business process model, the approach by Said-Cherfi et al. [21] considers ontologies in a number of specific domains.

Powel et al. [28] and An et al. [29] focus on the area of supply chain management and production and define quality dimensions, factors and metrics for a specific situation. Similar to the efforts in system dynamics, discrete-event simulation is used for simulation and analysis purposes and introducing the quality dimensions and factors for a specific situation [30-32].

D. Measurement method

This aspect is concerned with the degree to which an approach can support quantitative measures, that is, the results of an evaluation based on objective and quantitative metrics (e.g., [33] in the area of workflow and web-services, [13] measuring individual concepts of business process), or qualitative measures, that is, the results based on analysis requiring individual judgement and interpretation (e.g., [34]).

Cardoso et al. [33] proposes an approach for estimation of workflow properties (e.g., execution cost, execution time, and reliability), using the properties of activities constituting block-structured process models containing sequences, XOR blocks, AND blocks, and structured loops. Reduction of patterns is the common technique of analytical models in quantitative approaches.

Heidari et al. [13] identifies quality dimensions of performance, efficiency, reliability, recoverability, permissibility and availability for corresponding business process concepts and introduce objective and quantitative formulae for evaluating them in business process concept level (i.e., activity, input, output, event).

Most of managerial approaches (e.g., TQM [26], ISO standards [35], business performance measurement [24], quality tools and techniques [27]) provide generic guidelines and assistance for organizations, while the realization of the quality factors and metrics are not in their scope.

To conclude, there are variations in methodology, in the specification approaches used for presentation, the target application and measuring method of these quality approaches. The majority of approaches discussed above are based on the assumption that a formal language (e.g., BPMN) is used to describe business processes, the majority of which use one representation scheme. A few are language independent. Some provide a systematic way of working and some are generic enough to be applied in generic situations. Some are quantitative and some focus qualitative method of measurement. The approach introduced in this paper in some ways complements and in others extends existing approaches by emphasizing a well-structured way for specification and objectively measurement of business processes, which is generic in application and language independent. The desire of this paper is to provide an approach, which is systematic and well-structured, generic enough and not tied to a specific domain or situation, quantitative and objective, and while considering the formal expression of business processes, it is not tied to any particular business process modelling language.

III. BUSINESS PROCESS INTEGRATING META-MODEL

During requirement analysis, an important consideration is to understand current business processes. In this effort, business process modelling plays a key role. There are currently many business process modelling languages being advocated or practiced [36]. A closer inspection however, shows that there are many similarities and a great deal of convergence. From a theoretical perspective, it is vital to have a clear understanding of the semantics of these approaches, their overlaps, differences and similarities. Only then does it become possible to systematically and objectively understand the similarities and limitations of different approaches.

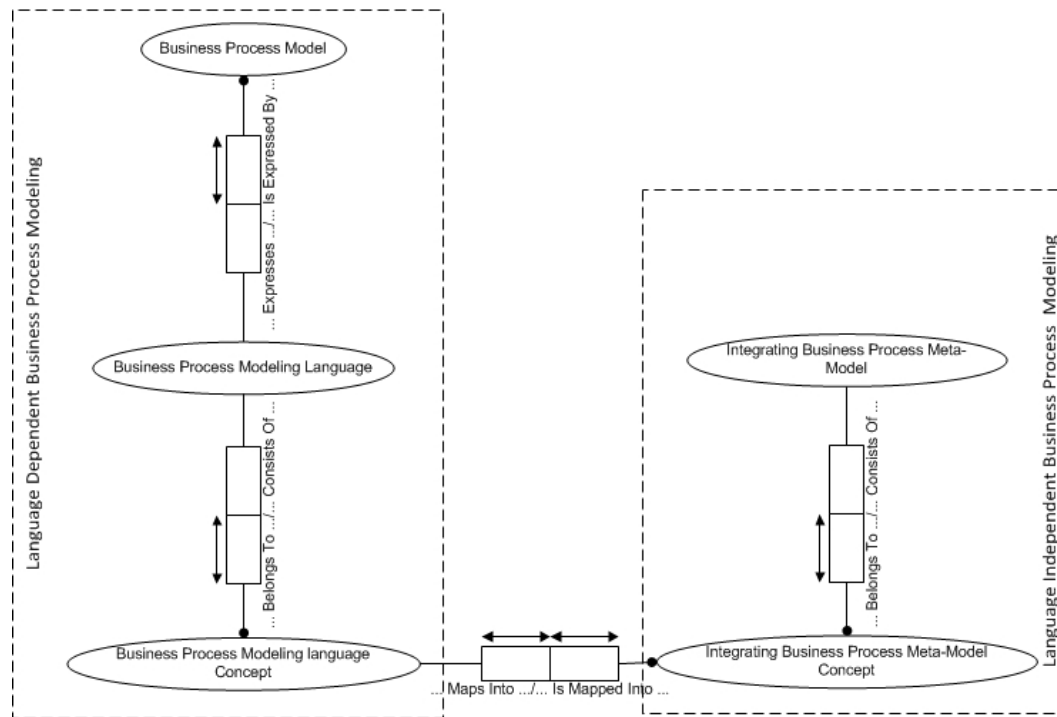


Fig. 1. Principles of the integrating meta-model development

Synchronization between requirements and business process models requires a common basis. This common basis can be presented in the form of a meta-model confined to requirement aspects. Such abstraction is to provide an explicit representation of knowledge, that can be understood by both computers and people [37]. Therefore, having an abstraction as the basis for quality specification and measurement of any business processes, using a corresponding model, would go a long way to integrating the field (i.e., business process modelling approaches) in the form of a meta-model and to facilitating a more systematic way of treating quality.

This paper argues that there is a need to view modelling structures through a lens that focuses on the semantics of concepts and relations and their ability to express different aspects of a business process rather than on syntax of the language used. This meta-model provides a language-independent business process ontology and is open to further extensions.

A meta-model is an explicit model of the constructs and rules needed to build specific models within a domain of interest. A valid meta-model is an ontology, as its constructs and rules represent entities in a domain. The formalism of a generic purpose modelling language (GPML) (e.g., UML class diagram) provides the ontology description. An ontology makes knowledge explicit expressing the concepts and relationships between them in a language close to the natural language; fostering an “understanding bridge”

between business and IT experts [2]. In Siau et al. [38] meta-modelling is classified as positivism in epistemology and realism in ontology. In essence, a meta-modelling approach aims to be independent of an observer’s appreciation of the modelling methods. In comparison to other approaches for describing modelling languages such as graph grammars, meta-models offer an intuitive way to specify modelling languages [39].

A business process integrating meta-model represents an abstraction of business process concepts, which is universal and not dedicated to one single business process modelling language. The approach in formation of the meta-model is presented as a conceptual model (Fig. 1). In this paper, a business process is analysed for its quality through its model, which obviously would be expressed in some business process modelling languages. Building a domain ontology includes the task of defining basic concepts and structures that are applicable in the target domain [40]. In this approach the concepts of different business process modelling techniques are going to be integrated to form the meta-model. In an example, Heidari et al. [41] proposes a meta-model resulted from integration of following languages: Business Process Modelling Notation (BPMN), Integrated Definition for Function Modelling (IDEF0 and IDEF3), Role Activity Diagram (RAD), Unified Modelling Language Activity Diagram (UML-AD), (Structured Analysis and Design Technique (SADT), and Event-driven Process Chain (EPC). The meta-model of each language is created and the

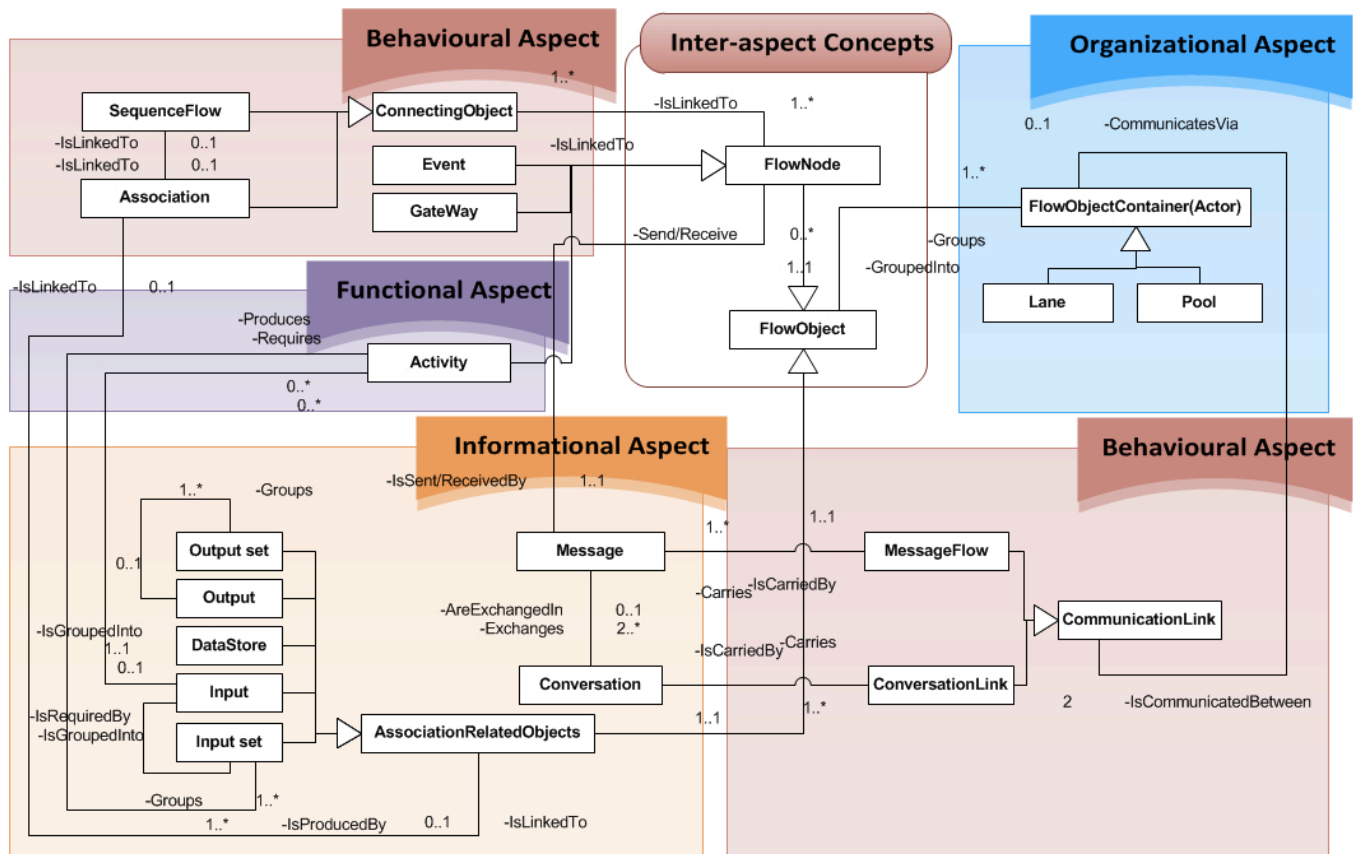


Fig. 2. The overview of the integrating business process meta-model in relation to different aspect

semantics of each individual concept are identified in this approach.

Each business process modelling language has its own and rich underlying semantics. As concepts from different languages do not perfectly match semantically, the unified framework first categorizes concepts into different aspects of a business process namely: *functional*, *behavioural*, *organizational* and *informational* in the form of a taxonomy. In this taxonomy for example role (RAD), organizational unit (EPC), swimlane (BPMN), partition (UML AD) are grouped together in the organizational aspect as they can serve the purpose of representing executors of an activity.

Fig. 2 demonstrates a general view of the business process meta-model in terms of the main concepts and relationships between different aspects. The proposed business process integrating meta-model represents an abstraction of business process concepts, is universal and not dedicated to one single business process modelling language. The business process integrating meta-model clarifies the exact relationships between the concepts. Moreover, it provides an adequate semantics specification prohibiting invalid interpretations by experts in different domains. Transforming these explicit syntactic relationships into a machine-readable language like Web Ontology Language

(OWL) provides the option of direct implementation. The ontology also provides an abstraction upon which elicitation, definition and documentation of requirements can happen.

One of the applications of business process ontology as a repository is as a reference to support explication of requirements. An ontology can describe both functional and non-functional requirements [2]. One of the applications of a business process ontology confined to quality aspects is that stakeholders can define their desired requirements in a higher level (meta-model) rather than in specific business process model, that covers just one situation. A business process ontology, enriched with the desired requirements, can act as a reference model for future enriched business processes generations (Fig. 3).

Fig. 3 provides a partial view of the instantiation of the meta-model and incorporation of requirements to the business process concepts (meta-level) and instances (Model). Fig. 3 shows that the desired requirements not only can be incorporated into the business process concepts in the meta-level but in instantiations using Protégé, Protégé provides direct, objective and straightforward incorporations of requirements. This facilitates not only communication between different stakeholders but also provides a guideline independence of the developer appreciations. This can

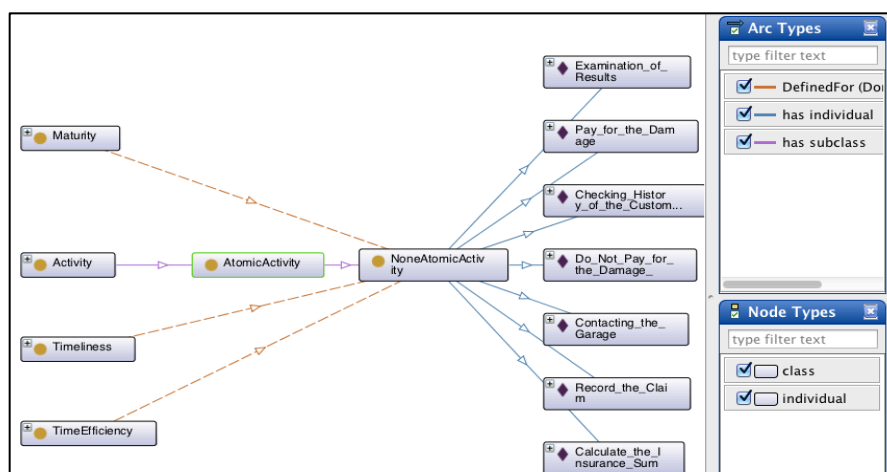


Fig. 3. An example of the instantiation and incorporation of quality requirements

fosters efficiency, integrity, consistency, and reusability and reduces human mistakes, etc. The business process integrating meta-model can also act as a repository. This repository can have several applications: (a) to represent models created via deploying any of the constructing modelling languages as its instantiations, (b) to be a reference between multiple business process modelling approaches of the same project, (c) to provide the basis for a repository of emerging business process models irrespective of the language used, (d) to be extended to a knowledge

base, (e) to facilitate direct implementation, and (f) to be a reference model fostering incorporation of stakeholders' requirements.

The quality factors can be realized for specific concepts of the integrating business process meta-model. Heidari et al. [13] introduces quality factors of throughput, cycle time, timeliness, cost, resource efficiency, cost efficiency, time efficiency, reliableness, failure frequency, time to failure, time to recover, maturity, authority, time to shortage, time to access and availableness for their related business process

Table I. Quality Dimensions and Factors

| Dimension | Factor | Business Process Concept | | | |
|----------------|---------------------|--------------------------|--------|-------|----------|
| | | Event | Output | Input | Activity |
| Performance | Throughput | X | X | X | |
| | Cycle Time | | | | X |
| | Timeliness | | | X | X |
| | Cost | | | X | X |
| Efficiency | Resource Efficiency | | | | X |
| | Time Efficiency | | | | X |
| | Cost Efficiency | | | X | X |
| Reliability | Reliableness | | | | X |
| | Failure Frequency | | | | X |
| Recoverability | Time to Failure | | | | X |
| | Time to Recover | | | | X |
| | Maturity | | | | X |
| Permissability | Authority | | | X | X |
| Availability | Time to Shortage | | | X | |
| | Time to Access | | | X | |
| | Availableness | | | X | |

concepts namely, event, output, input and activity (Table I).

The integrating business process meta-model can be enriched formally with these quality factors. The resulting meta-model is shown in Fig. 4. Note that this is a subset of the entire integrating meta-model, focusing on those concepts that are related to the quality aspects. Business process concepts are shown in white classes and quality factors are shown in grey classes. The meta-model enriched with quality factors can be used for quality modelling and business process redesign and can help practitioners to consider quality requirements of a business process at the earliest stage of system development.

IV. THE MODEL-BASED SPECIFICATION AND MEASUREMENT APPROACH

This paper proposes an approach to the specification and the measurement of quality requirements for business process concepts. This approach considers quantitative metrics for business processes in its specifications and measurement. The conceptual framework of the approach is

The “conceptual framework” encompasses a set of concepts that link requirements to specific business process concepts, their factors and corresponding metrics. Requirements can be classified into functional and non-functional requirements of a business process [42].

Functional requirements refer to the ability of the business process to deliver qualified products and services as well as the ability of the outcome to fulfil its functional expectations [43]. Glinz [44] offers a set of classification rules for distinguishing between functional requirements (FRs) and non-functional requirements (NFRs) in system engineering. In this classification, the concept of non-functional requirements can be defined as: requirements about timing, processing or reaction speed, input volume or throughput as well as specific quality of business process concepts as a whole reflected in those ended in “-ilities” namely: reliability, security and availability etc. This paper considers the notion of non-functional requirements by [44] as quality.

Fig. 5 depicts that “Quality Requirements” (e.g., capturing customer data must be most of the time without failure), are associated with a “Business Process” (e.g., accepting client). A “Business Process” is responsible for fulfilling a set of “Quality Requirements”. A “Business Process” consists of numerous “Business Process Concepts”. A “Business Process Concept” (e.g., capturing customer data) belongs to a “Business Process” (e.g., accepting client).

A “Quality Requirement” is expressed by a “Stakeholder” (e.g., company manager) and is operationally queried by a set of “Quality Questions” (e.g., what is the percentage of the time that the execution is without failure out of the whole time of execution?). Operationally querying

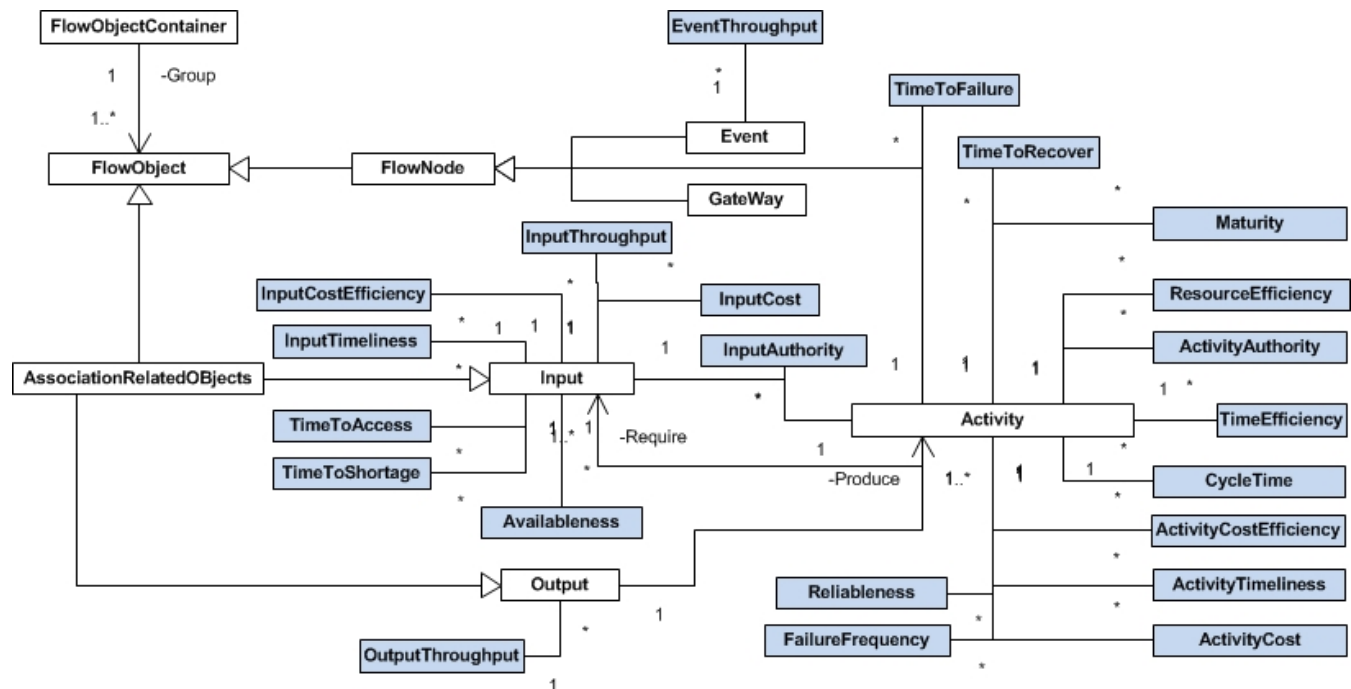
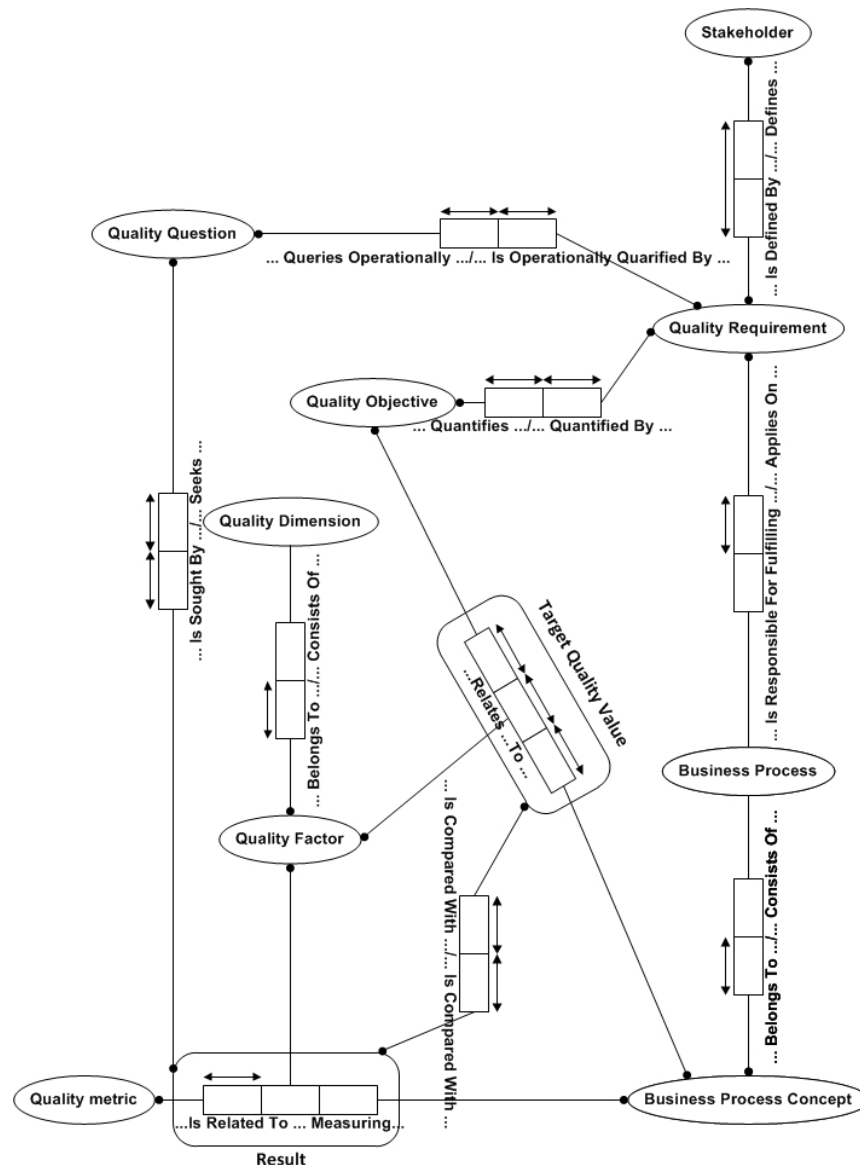


Fig. 4. The business process integrating meta-model enriched with quality factors

Dimension” representing an aspect of business process concept quality e.g., performance, efficiency and those colloquially referred to “-ilities”.

The “Quality Objective” (e.g., more than 95% of the time, execution should be without failure) is a way of quantitatively analysing the “Quality Requirement” and is defined as some “Target Quality value” (e.g., $\geq 95\%$), which is shown as objectified relationship of three concepts namely those of “Quality Factor” (e.g., maturity) for particular “Business Process Concept” (e.g., capturing customer data) and a corresponding “Quality Metric” (e.g., $M(a)=[TF(a)/TF(a)+TR(a)]*100$).



The gap between “Quality Objective” and the observed current performance through “Quality Question” is shown in the relationship of “Target Quality Value” and “Result”. Several “Quality Metrics” can be associated to a single “Quality Factor” as there might be several ways for evaluating it. Different stakeholders can indicate different quality metrics based on their needs [18].

The conceptual model guides the systematic application within the approach. Specifically, the process is described as an algorithm in Pseudocode (Fig. 6).

The contribution of this framework is in the establishment of a set of conceptual structures that are independent of the descriptive languages, or applications. Applicability of the framework is illustrated for an example of business process in the next section.

V. DEMONSTRATION OF APPLICABILITY IN AN EXAMPLE

The applicability of the approach this paper proposes is demonstrated for a simplified version of business process, namely “Accepting clients” from an anonymous enterprise. The business process is known to this enterprise.

First, in a more visual way, the instantiation of the

conceptual framework is provided (Fig. 7). Fig. 7 illustrates not only the business process in terms of a model but also provides examples of the related elements for quality specification and measurement considering the business process concepts. Later, the conceptual framework is instantiated formally (Fig. 8) to demonstrate its application relates to the example in the form of an ORM model.

As can be observed from Fig. 7, there are different departments/roles involved in the process. The process trigger is the arrival of a request to accept a client. To accept the client, a set of activities is performed in a predefined order. Some related quality factors are shown in Fig. 7 namely, time to recover, time to failure, maturity, authority, timeliness, cycle time, and throughput. Quality factors are assigned to the business process concepts via dashed lines as shown. For the matter of distinction, quality factors are shown in a separate box below the example. The “business process” is presented via applying BPMN as a “Business process modelling language” supported by a business process meta-model e.g., [8].

Fig. 7 shows that the quality requirement of “Capturing client data should be executed more than 95% of the time without failure”, is associated with the business processes concept of “capturing client data”; this concept belongs to

ALGORITHM.

BEGIN

A stakeholder defines Quality requirements in natural language

FOR each quality requirement

Define the business process referenced in the quality requirement;

Define a quantified expression as quality objective;

FOR each quality objective

Determine the business process concept to which quality is referred;

Define quality factor for this concept;

Define the metric to be applied to this quality factor;

ENDFOR

Query the quality of the business process as a question;

FOR each question

Identify the business process concept being queried;

Identify quality factor for this concept;

Apply the metric to be applied to this quality factor;

Obtain result of measurement;

ENDFOR

Compare the result of measurement with quality objective;

Define degree of satisfying quality objective;

ENDFOR

Return the results to stakeholder;

END

Fig. 6. Algorithm for quality specification and measurement

business process of “accepting client”. The Requirement is expressed by a the “company manager” as the stakeholder and is operationally queried by questions of “What is the percentage of the time that the execution is without failure out of the whole time of execution?” The quality factor “maturity” can be measured by a quality metric expressed as follows:

$$M(a)=[TF(a)/TF(a)+TR(a)]*100 \quad (1)$$

Where “a” denotes the “Activity”, TF(a) is the “Time to Failure” and TR(a) is the “Time to Recover”.

Applicability of the proposed approach is also demonstrated via instantiation of the conceptual framework (Fig. 8) with regards to the example. The instantiation is focused on the quality requirement of “capturing client data should be executed more than 95% of the time without failure”. Instances are introduced as “roles” in “fact tables”. Information in the fact tables is in line with the example described earlier and provided in Fig. 7.

VI. CONCLUSION AND FUTURE WORK

This business process integrating meta-model as an

abstraction provides an explicit specification of the shared conceptualization and understanding of enterprises between IT and non-IT experts. This paper focuses on a specific modelling approach, that of Business Process Modelling, and the use of a meta-model for modelling and evaluating quality aspects of business processes. Specification and measurement of requirements based on concepts in business process meta-model fosters communication between experts.

Many different approaches have been developed, each with their own specific business process modelling languages (BPML) designed to meet specific business requirements. In cross-organizational business processes and heterogeneous organizations where multiple business process modelling languages are deployed, there is a need for a unified and integrating view to ease communication and foster understandability.

This paper proposes an approach that drives specification and measurement of quality requirements. This paper assumes that the quality of a business process can be defined by the degree to which pre-defined properties of pre-defined concepts identified within a business process are linked to stakeholder requirements. The methodological stance of the

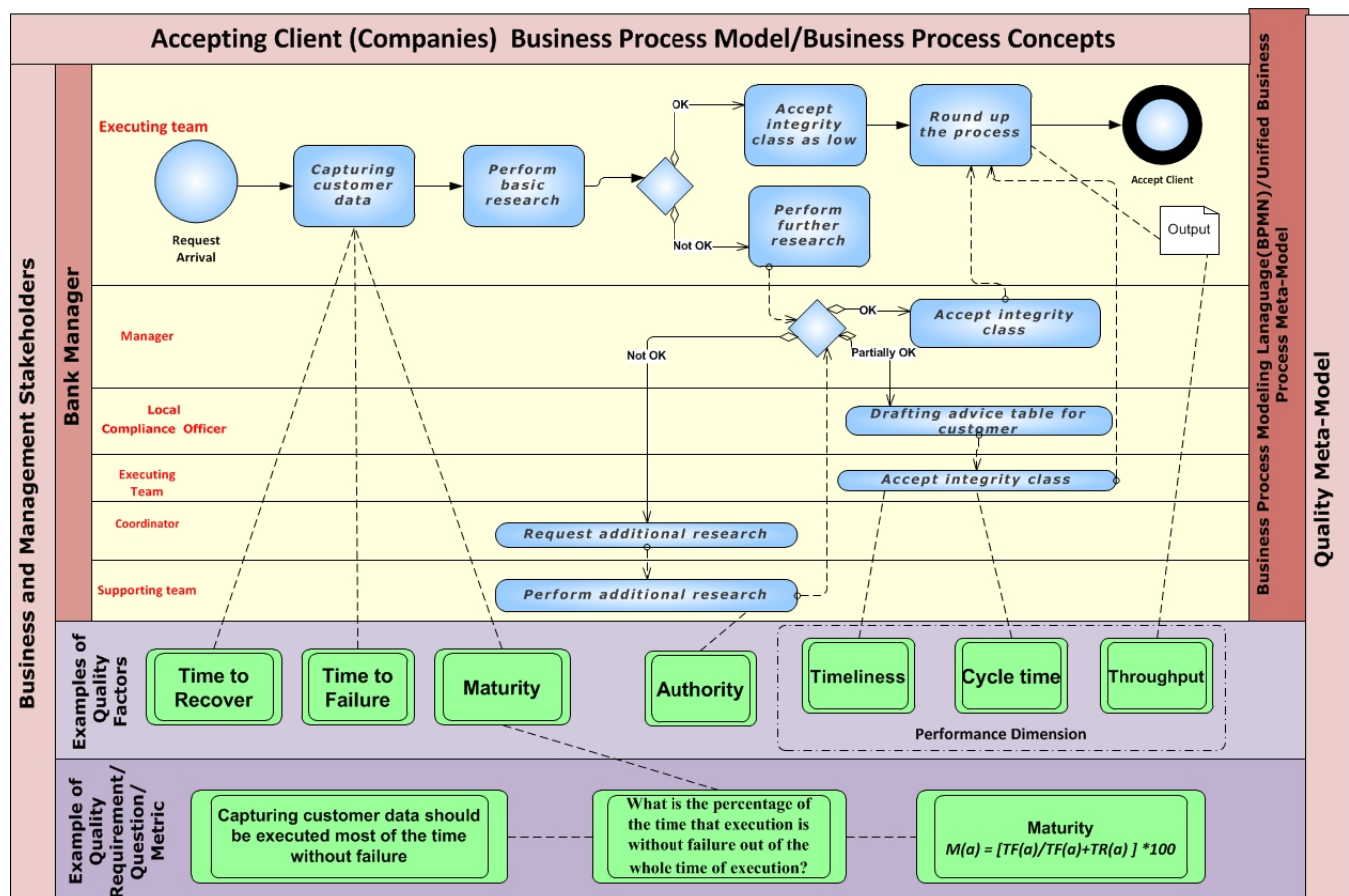


Fig. 7. Business process “Accepting Client” and examples of quality factor, requirement, question and metric

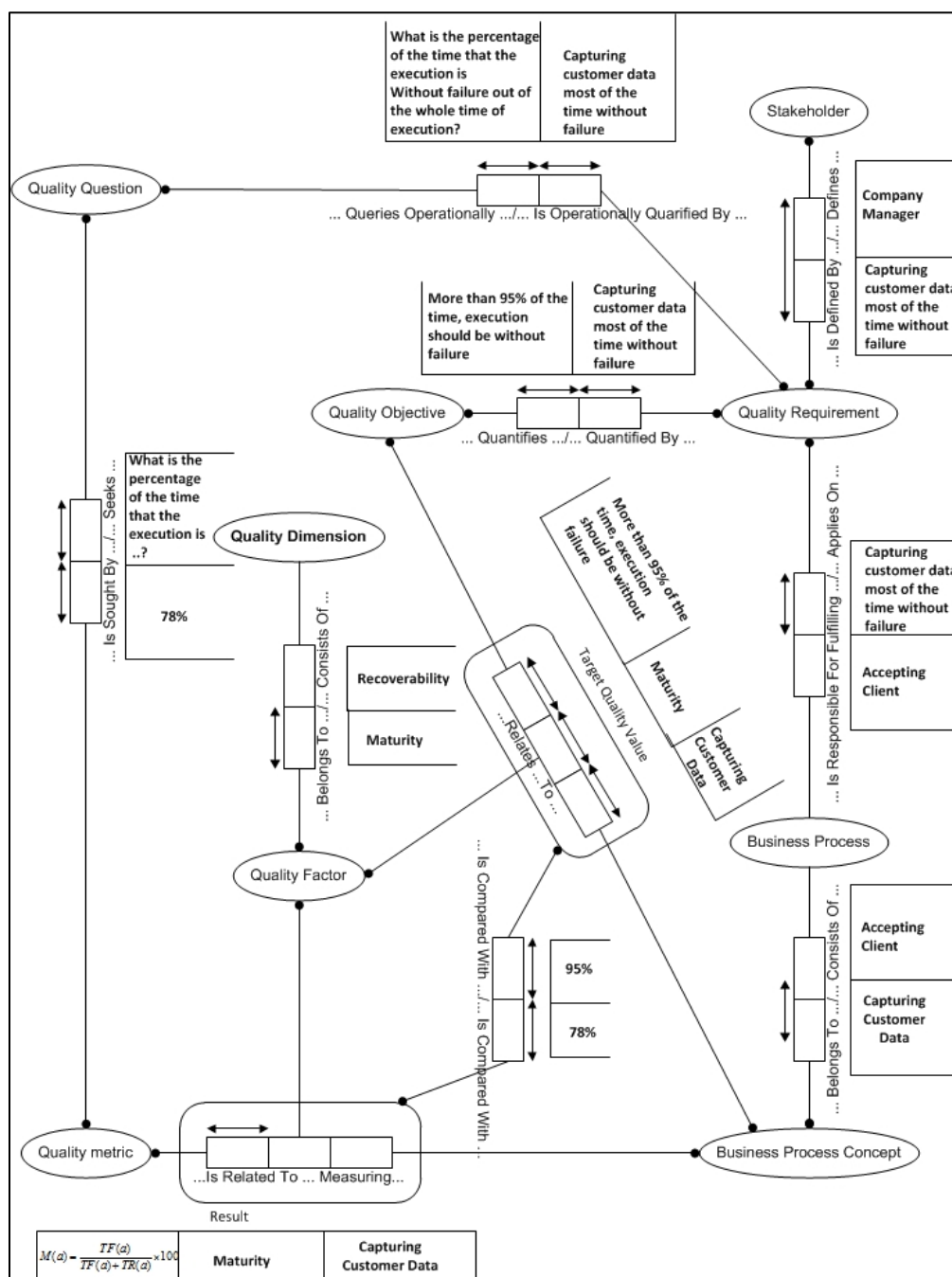


Fig. 8. The conceptual framework instantiation for “accepting client” business process (quality requirement: capturing client data should be most of the time without failure)

approach proposed in this paper is systematic. Stakeholder goals are transformed to objectified components, a quality objective and a quality question, that are directly linked to quality factors for a pre-defined business process concept. The approach relies on *formal* expressions of business processes in business process models. At the same time, it is *independent* of any language. The utility of the approach is *generic*, i.e., applicable to any application and within any domain.

The approach is systematic and provides methodological means to specify and measure requirements for business processes. In line with the items introduced in Section IV, a set of phase is prescribed with details on the “way of working” within each phase, i.e., (a) focus of work, (b) required inputs, (c) expected outputs, (d) techniques used and (e) support tools. On the basis of these four criteria, the

approach is considered to be systematic in that there is an identifiable and generic approach to business process quality computation (the focus), analysing business processes given the stakeholders' requirements (the input), deriving a quality evaluation (the output), using objective formulae for such a measure (the technique) and finally having a formal representation scheme for automated support (the tools).

The outcomes of this research are beneficial in the areas of business and management, requirement engineering, software engineering, business process modelling and service-oriented architectures. In the areas of requirement engineering and software engineering, these results make it possible for practitioners to consider quality requirements at the earliest stage. In the area of process modelling, the outcomes lead to a quality-driven modelling and redesigning. Moreover, qualified business processes have direct impacts on the quality of web-services.

This paper establishes a strong framework upon which different methodological and technological developments may emerge such as an enhancement of existing business process modelling tools with a simulation component, the development of a workbench for analysing measured qualities and the development of further cases on an industrial basis [46]. Future research will focus on extensions and developments both in theoretical and practical perspectives. Exploring possibilities to enhance existing industrial business process-modelling tools with quality evaluation extensions is currently subject of research. Also, strategic modelling approaches such as system dynamics are to be coupled to business process modelling using parametric definitions according to quality criteria and experimenting with 'what-if scenarios' thus giving stakeholders an early view of the impact of their choices, on the behaviour of a business process [47]. In addition to the specification and measurement of quality requirements for individual business process concepts, there is a need for measuring requirement fulfilment by a business process as a whole or a part of a business process. There is a need for an approach that can foster objective evaluation of the degree to which a quality requirement for a business process is achieved based on the achievements of its individual concepts.

REFERENCES

- [1] F. Heidari, P. Loucopoulos, and F. Brazier, "Ontology for Quality Specification in Requirements Engineering," in *Fourth International Conference on Models and Ontology-based Design of Protocols, Architectures and Services (MOPAS 2013)*, Venice, Italy, 2013, pp. 7-12.
- [2] D. E. Jenz, "Business processes ontologies: speeding up business process implementation," Jenz & Partner GmbH 2003.
- [3] A. W. Scheer, *Business Process Engineering*, Second ed. Berlin: Springer-Verlag, 1994.
- [4] M. Havey, *Essential Business Process Modelling*. O'Reilly, USA, 2005.
- [5] P. Loucopoulos and E. Katsouli, "Modelling Business Rules in an Office Environment," *SIGOIS Bulletin*, vol. 13, pp. 28-37, 1992.
- [6] E. Kavakli and P. Loucopoulos, "Experiences with Goal-Oriented Modelling of Organisational Change," *IEEE Transactions on Systems, Man and Cybernetics - Part C*, vol. 36, pp. 221-235, 2006.
- [7] J. Mendling, G. Neumann, and M. Nuttgens, "A comparison of XML interchange formats for business process modelling," in *Workflow handbook*, L. Fischer, Ed., ed: Future strategies Inc., 2005, pp. 185-198.
- [8] F. Heidari, P. Loucopoulos, and Z. Kedad, "A quality-oriented business process meta-model," in *International Workshop on Enterprise & Organizational Modelling and Simulation, CAiSE/EOMAS 2011, Lecture Notes in Business Information Processing (LNBIP)*, 88, 2011, pp. 85-99.
- [9] K. Decreus and G. Poels, "A Goal-Oriented Requirements Engineering Method for Business Processes," in *CAiSE Forum*, Berlin Heidelberg, 2010, pp. 29-43.
- [10] D. Firesmith, "Quality Requirements Checklist," *Journal of Object Technology* vol. 4(9), pp. 1-8, 2005.
- [11] P. Donzelli and P. Bresciani, "Improving Requirements Engineering by Quality Modelling – A Quality-Based Requirements Engineering Framework," *Journal of Research and Practice in Information Technology*, vol. 36(4), pp. 277-294, 2004.
- [12] M. M. Glykas, "Effort Based Performance Measurement in Business Process Management," *Knowledge and Process Management*, vol. 18(1), pp. 10-33, 2011.
- [13] F. Heidari and P. Loucopoulos, "Quality Evaluation Framework (QEF): Modeling and Evaluating Quality of Business Processes," *International Journal of Accounting Information Systems*, vol. in press, p. in press, 2013.
- [14] ISO/IEC, "Software engineering - Product quality - Part 4: Quality in use metrics," vol. ISO/IEC TR 9126-4:2004(E), ed. Switzerland: ISO, 2004.
- [15] M. Lohrmann and M. Reichert, "Understanding Business Process Quality," in *Business Process Management, Theory and Applications. Studies in Computational Intelligence* ed Berlin Heidelberg: Springer 2013, pp. 41-73.
- [16] C. Wolter, M. Menzel, A. Schaad, P. Miseldine, and C. Meinel, "Model-driven business process

- security requirement specification," *Journal of Systems Architecture*, vol. 55, pp. 211-223, 2009.
- [17] M. Heravizadeh, J. Mendling, and M. Rosemann, "Dimensions of Business Processes Quality (QoBP)," in *Proceedings of the 6th International Conference on Business Process Management Workshops (BPM Workshops)*, Milan, 2008, pp. 80-91.
- [18] Z. Kedad and P. Loucopoulos, "Considering quality factors for business processes during requirement engineering," in *Fifth International Conference on Research Challenges in Information Science (RCIS)*, 2011, pp. 1-9.
- [19] A. Pourshahid, D. Amyot, L. Peyton, S. Ghanavati, P. Chen, M. Weiss, *et al.*, "Toward an integrated user requirement notation framework and tool for business process management," in *3rd Int. MCE Tech Conf. on eTechnologies*, Montréal, Canada, 2008, pp. 3-15.
- [20] F. Aburub, M. Odeh, and I. Beeson, "Modelling non-functional requirements of business processes," *Information and Software Technology*, vol. 49, pp. 1162-1171, 2007.
- [21] S. Si-Said-Cherfi, S. Ayad, and I. Comyn-Wattiau, "Aligning Business Process Models and Domain Knowledge: A Meta-modeling Approach " in *Advances in Databases and Information Systems, AISC 186*, Berlin, Heidelberg, 2013, pp. 45-56.
- [22] P. Loucopoulos and F. Heidari, "Evaluating Quality of Business Processes," in *Modelling and Quality in Requirements Engineering-Modelling and Quality in Requirements Engineering, Essays Dedicated to Martin Glinz*, N. Seyff and A. Koziol, Eds., ed MV-Wissenschaft, Munster: 61-73, 2012, pp. 61-73.
- [23] R. Heinrich, A. Kappe, and B. Paech, "Modeling Quality Information within Business Process Models," *Proceedings of the SQMB'11 Workshop, TUM-II104, Karlsruhe (Germany), February 21th*, pp. 4-13, 2011.
- [24] P. Kueng, "Process Performance Measurement System: a tool to support process-based organizations," *Total Quality Management*, vol. 11, pp. 67-85, 2000.
- [25] D. Heckl and J. Moormann, "Process performance management," in *Handbook on Business Process Management I*, J. vom Brocke and M. Rosemann, Eds., ed Berlin: Springer, 2010, pp. 115-135.
- [26] B. G. Dale, D. T. van der Wiele, and J. van Iwaarden, *Managing Quality* vol. Fifth. Singapore: Blackwell Publishing Ltd, 2007.
- [27] H. Eriksson and M. Penker, *Business Modeling with UML : Business Patterns at Work*. New York, USA: Wiley Publishing, 2000.
- [28] S. G. Powell, M. Schwaninger, and C. Trimble, "Measurement and control of business processes," *System Dynamics*, vol. 17, pp. 63-91, 2001.
- [29] L. An and J. J. Jeng, "On developing system dynamics model for business process simulation," *Proceedings of the 2005 Winter Simulation Conference ,Orlando, FL, USA*, pp. 2068-2077, 2005.
- [30] E. W. East, J. C. Martinez, and J. G. Kirbya, "Discrete-event simulation based performance quantification of web-based and traditional bidder inquiry processes," *Automation in Construction*, vol. 18, pp. 109-117, 2009.
- [31] S. R. Nidumolua, N. M. Menonb, and B. P. Zeigler, "Object-oriented business process modeling and simulation: A discrete event system specification framework," *Simulation Practice and Theory*, vol. 6, pp. 533-571, 1998.
- [32] V. Hlupic and S. Robinson, "Business process modeling and analysis using discrete-event simulation," in *Proceedings of the 1998 winter simulation conference*, 1998, pp. 1363-1370.
- [33] J. Cardoso, A. Sheth, J. Miller, J. Arnold, and K. Kochut, "Quality of service for workflows and web service processes," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 1(3), pp. 281-308, 2004.
- [34] S. Adam, N. Reigel, and J. Doerr, "The role of quality aspect for the adequacy of business process and business information systems," *International Journal of Business process Integration and Management*, vol. 4, pp. 124-133, 2009.
- [35] ISO, "BS EN ISO 9000," ed. Switzerland: ISO, 2000.
- [36] W. J. Kettinger, J. T. C. Teng, and S. Guha, "Business process change: A study of methodologies, techniques, and tools," *MIS Quarterly*, vol. 21, pp. 55-80, 1997.
- [37] S. Fan, L. Zhang, and Z. Sun, "An ontology based method for business process integration " in *International Conference on Interoperability for Enterprise Software and Applications*, China, 2009, pp. 135-139.
- [38] K. Siau and M. Rossi, "Evaluation techniques for systems analysis and design modelling methods – a review and comparative analysis," *Information system journal*, pp. 1-20, 2007.
- [39] H. Fill and P. Burzynski, "Integrating ontology models and conceptual models using a meta modeling approach," in *11th International Protege Conference*, Amsterdam, Netherlands, 2009.
- [40] H. S. Na, O. H. Choi, and J. E. Lim, "A method of building domain ontologies based on transformation of UML models " *Proceedings of*

- 4th Int'l Conf. on Software Engineering Research, Management and Applications, pp. 332–338, 2006.
- [41] F. Heidari, P. Loucopoulos, F. Brazier, and J. Barjis, "A Meta-Meta-Model for Seven Business Process Modelling Languages," in *15th IEEE Conference on Business Informatics*, Vienna, Austria 2013, pp. 216-221.
- [42] P. Loucopoulos, J. Sun, L. Zhao, and F. Heidari, "A Systematic Classification and Analysis of NFRs " in *19th Americas Conference on Information Systems (AMCIS 2013)* Chicago, USA 2013.
- [43] P. Loucopoulos and R. E. M. Champion, "Knowledge-Based Support for Requirements Engineering," *Information and Software Technology*, vol. 31, pp. 123-135, 1989.
- [44] M. Glinz, "On Non-Functional Requirements," in *15th IEEE International Requirement Engineering Conference*, 2007, pp. 21-26.
- [45] V. Peralta, V. Goasdoue-Thion, Z. Kedad, L. Berti-Equille, I. Comyn-Wattiau, S. Nugier, *et al.*, "Multidimensional Management and Analysis of Quality Measures for CRM Applications in an Electricity Company," in *Proceedings of the 14th International Conference for Information Quality (ICIQ)*, Potsdam, Germany, 2009.
- [46] D. Karagiannis, "A Business Process Based Modelling Extension for Regulatory Compliance," in *Multikonferenz Wirtschaftsinformatik , MKWI 2008*, München, 2008, pp. 1159-1173.
- [47] P. Loucopoulos, K. Zografos , and N. Prekas, "Requirements Elicitation for the Design of Venue Operations for the Athens 2004 Olympic Games," in *Proceedings of 11th IEEE International Requirements Engineering Conference*, Monterey Bay, California, U.S.A, 2003, pp. 223-232.

Machine Learning Methods Applied on Long Term Data Analysis for Rain Detection in a Partial Discharge Sensor Network

Leandro H. S. Silva, Sergio C. Oliveira

Polytechnic School of Pernambuco
University of Pernambuco
Recife - PE, Brazil
{lhss, scampello }@ecompp.poli.br

Eduardo Fontana

Department of Electronics and Systems
Federal University of Pernambuco
Recife-PE, Brazil
fontana@ufpe.br

Abstract — Partial discharges (PD) on the surface of high voltage insulators are directly related with the accumulation of pollution. A complete partial discharge sensor network was previously developed and has been in operation for approximately three years. This system records the PD activity, classifying it into four levels. As the PD activity is influenced by the weather conditions, each sensor system in the network also measures the one-hour average temperature and relative humidity. Also a fuzzy inference system was developed to extract the flashover occurrence risk level based on the partial discharge activity recorded. However, a strong rain event can wash insulators almost instantaneously, in turn decreasing the risk level. For a correct interpretation of the results it is important to properly analyze the weather data to detect the rain occurrence. This paper presents a comparison among three machine learning techniques for rain detection from humidity and temperature data, namely, Naïve Bayes Classifier, Support Vector Machines and Multilayer Perceptron Neural Network. These are trained on data gathered by meteorological stations located nearby the PD sensors and used in conjunction with the data obtained by the Sensor Network. Studies on the generalization training power and long term data analysis on sensor data are performed and presented.

Keywords- *Partial discharges; rain detection; pattern recognition; leakage current; insulators; sensor network.*

I. INTRODUCTION

High-voltage transmission lines are affected by many problems. One of them is the pollution accumulated on the surface of the insulators that support the conducting cables in the towers. When combined with high relative humidity the pollution layer becomes a conductive layer. A leakage current flows through this conductive layer causing irregular heating and then humidity evaporation, creating thin dry bands. The increase of electric charges in the boundaries of dry bands causes high electric fields that produce partial discharges (PD) across dry bands [1], [2]. The PD phenomenon can increase in rate and intensity until a complete discharge, known as flashover, bypasses all insulators, in turn causing an outage in power transmission [3].

One way to avoid flashover events is to remove the pollution layer deposited over the insulator string by periodic

washing. However, this is a high-cost operation and failures may occur during the procedure.

Aiming to assist the decision regarding the need for maintenance, a sensor network was previously developed to detect and classify partial discharges according to their frequency of occurrence and intensity [4]. This system comprises an optical sensor coupled to an optical fiber, which transmits the leakage current signal [5] to an electronic processing module, which has also a temperature and a humidity sensor [6]. The collected data are transmitted via satellite and stored in a database.

A fuzzy inference system has been developed in order to extract the flashover risk occurrence. The risk level is incremented and decremented according to the level of partial discharge activity considering its intrinsic relation with relative humidity [7]. The use of a fuzzy system has the advantage of being able to represent uncertainties of natural language.

However, on strong rain events the insulators are washed ceasing the risk of flashover. This almost instantaneous risk variation is not reflected on the fuzzy risk level. This work aims to develop a system capable of detecting the instantaneous cleaning of the insulator by strong rains, based on the available humidity and temperature data. Rain detection would make the fuzzy risk classification system more precise and turn the maintenance schedule more robust, reducing costs due to unnecessary washes.

Common electronic rain sensors are only capable of detecting rain in a small surface and are not capable of quantifying the event [8]. Electromechanical rain sensors are capable of easily detecting and quantifying rain. Nevertheless, when installed in outdoor environments this kind of sensor accumulates water, in turn attracting infestation by wasps or bees. The presence of these insects increases the risk for operators of the power transmission company and increases the failure rate of the rain sensor itself once the hives might block the mechanical parts of the sensor.

Temperature and humidity data gathered by the sensor network exhibits a daily regular pattern. This pattern is changed by rain events and a new rain pattern starts to occur. So, a pattern recognition system can be applied to detect the insulator washing by rain. A pattern recognition prototype system was developed based on the reliable data obtained from the Brazilian Institute of Meteorology (INMET)

database. This database has humidity and temperature information as well as the amount of rain precipitation per hour.

This paper compares three well know machine learning algorithms applied for the task of rain detection: Naïve Bayes Classifier, Support Vector Machine (SVM) and Artificial Neural Network Multilayer Perceptron (ANN MLP) [1]. Based on previous results [1], the MLP was applied in a data set gathered by the partial discharge sensor network and visual inspections were carried out to ensure empirically the success of the proposed rain detection strategy.

II. SATELLITE SENSOR SYSTEM NETWORK

The sensor network is composed by six monitoring nodes and it has been in operation for three years in the Northeast region of Brazil. Each node is composed by an optical sensor, an electronic processing module and a satellite transmission modem [4], as illustrated in Fig. 1.

Each hour the sensor node transmits the partial discharges activities, average temperature and average humidity. The partial discharge activity is classified into four current ranges named N1 to N4, which are related to current pulses larger than 5, 10, 20 and 40 mA, respectively [4].

The information gathered by each sensor is organized into two 64-bit packets and transmitted via satellite each half hour. After reception the data are stored in a database. The access to this database is provided by a system called ADECI (a portuguese acronym for Electric Performance Evaluation on Insulator Strings). Only identified employees of CHESF (the generation and distribution company in the Northeast region of Brazil) can access the information.

III. DATA SETS AND RAIN PATTERN

The temperature and humidity have an almost regular daily behavior. During the day, the temperature is high and the humidity is low; at night the temperature falls down and the humidity goes up. During rain events this behavior is modified because the rain causes an immediate increase in humidity and decrease in temperature. This behavior can be seen in Fig. 2 – during rain events, that start to occur beyond the dashed line time point, the temperature falls down and the humidity goes up. This behavior is better observed in heavy rain events.

The INMET meteorological stations data is organized as daily 24 string data containing average temperature and humidity as well as the amount of rain precipitation in millimeters per hour. Linear interpolations were used to complete the series on every data missing less than 5 consecutive hours. When the time period of the missing data was larger than 5 hours, data for the full day were excluded from the database.

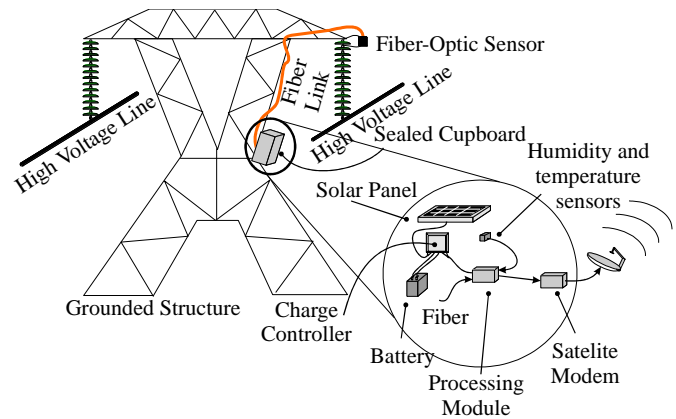


Figure 1. Sensor node for partial discharge monitoring.

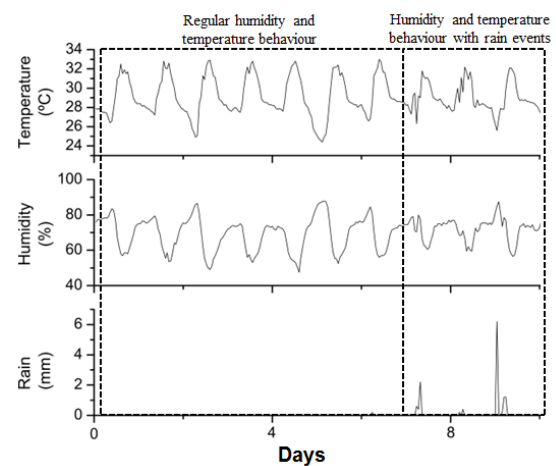


Figure 2. Plots of temperature and humidity patterns and the respective amount of rain.

The INMET database was used to train each detection rain model for further use on the ADECI database. Fig. 3 shows the sensor network topology and each node of the nearest INMET meteorological station. Although each sensor node has a near INMET station, the distance between them is about tens of kilometers and a rain event in the INMET station does not imply a similar occurrence in the nearest sensor location.

The data set was organized on day-long vectors as shown in Table I. Parameters T0 to T23 represent the temperatures uniformly distributed in 24 hours, as well as U0 to U23 represent the corresponding humidity values. If the day has a total rain precipitation larger than 1 mm, the day is classified as rainy. Otherwise it is classified as *no-rain*.

TABLE I. DATA SET ATTRIBUTES AND CLASS.

| Attributes | | | | | | Class |
|------------|-----|-----|----|-----|-----|------------------|
| T0 | ... | T23 | U0 | ... | U23 | [rain / no-rain] |

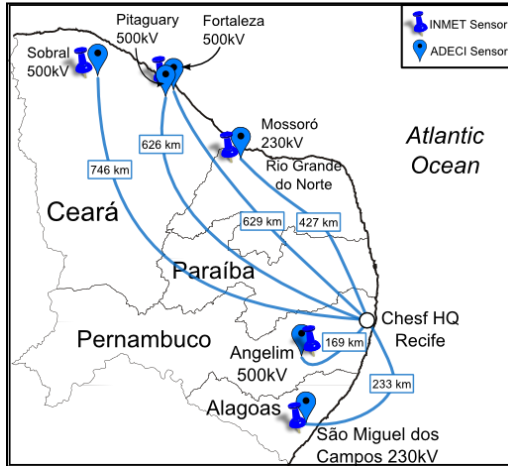


Figure 3. Sensor node and INMET station location.

IV. APPLIED TECHNIQUES

A. Naïve Bayes Classifier

A Naïve Bayes Classifier [9] is a supervised-learning statistical technique. A vector x represents m features (x_1, x_2, \dots, x_m) . In this work, each dimension of vector x comprehends an attribute of the database. The a posteriori probability of having rained in a specified day can be calculated using Bayes theorem as

$$P(\text{rain}|x) = \frac{P(\text{rain})P(x|\text{rain})}{P(x)}. \quad (1)$$

In (1), $P(x)$ is the probability of x occurring in the data set and $P(x|\text{rain})$ is the likelihood probability of x occurring in the *rain* class.

By using the naïve assumption, i.e., the attributes are conditionally independent, the likelihood probability of $P(x|\text{rain})$ is

$$P(x|\text{rain}) = \prod_{i=1}^m P(x_i|\text{rain}). \quad (2)$$

It means that under the naïve assumption, the conditional distribution over the *rain* class can be expressed as

$$P(\text{rain}|x) = \frac{1}{Z} P(\text{rain}) \prod_{i=1}^m P(x_i|\text{rain}), \quad (3)$$

where Z , the evidence, is a scaling factor dependent only on the features of the x vector.

All the Naïve Bayes Classifier parameters (the class prior and feature probability distributions) can be approximated with relative frequencies from the training set. In this work the continuous values associated with each class were considered to have a Gaussian distribution.

B. Multilayer Perceptron Neural Network

The ANN MLP [10] is an artificial neural network whose architecture is based on multiple layers of neurons: an input layer, one or more hidden layers and an output layer. The number of hidden layers can be changed depending on the application.

Each neuron can be seen as an element with inputs, weights, one activation function and the output signal. The output signal of each neuron is given by

$$y_j = f \left(\sum_{i=1}^n x_{ji} w_{ji} \right), \quad (4)$$

where y_j is the output signal of the j -th neuron, x_{ji} is the i -th entry of the j -th neuron, w_{ji} is the i -th weight of the j -th neuron and f is the activation function. In this work the sigmoid function was used as activation function [10]. The signal is propagated from the input layer to the output layer – where the classifier result is available.

The training of an MLP consists on adjusting the weights. The objective is to train the MLP network to achieve a balance between the ability to respond correctly to the input patterns used for training and the ability to provide good results for other similar inputs, i.e., train the network to be capable of performing generalization. For this task, the classic backpropagation algorithm was used to realize the training of the neural network [10].

C. Support Vector Machine

The SVM [11] is a statistically robust learning method in which the training process consists of finding an optimal hyperplane which maximizes the margin between two classes of data in the kernel induced feature space.

Given an input data of n samples x_i ($i = 1, \dots, n$) classified into two classes, each one of the classes associated with labels are $y_i = +1$ for the positive class (*rain*) and $y_i = -1$ for the negative class (*no-rain*), respectively. For linear data, it is possible to determine the hyperplane

$$f(x) = xw + b = 0, \quad (5)$$

where w is an M -dimensional vector and b is a scalar. This separating hyperplane should satisfy the constraints

$$\begin{aligned} x_i w + b &\geq 1, \text{ if } y_i = +1 \\ x_i w + b &\leq -1, \text{ if } y_i = -1 \end{aligned} \quad (6)$$

Furthermore, as the SVM searches for an optimal hyperplane, the margin width between the support vectors and the optimum hyperplane must be maximized, as shown in Fig. 4. The margin is calculated as

$$2 \cdot d = \frac{2}{\|w\|}, \quad (7)$$

so $\|w\|$ must be minimized.

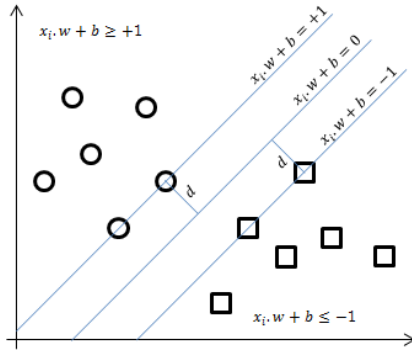


Figure 4. Support Vectors and separating hyperplane.

There is also the introduction of positive slack variables ξ_i , to measure the distance between the margin and the vectors x_i , which means that some mistakes can be tolerated. The optimal hyperplane separating the data can be obtained by solving the optimization problem

$$\min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^M \xi_i, \quad (8)$$

subject to

$$y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0 \quad (9)$$

The constraints aim to put the instances with positive label at one side of the margin of the hyperplane, and the ones with negative labels at the other side. Parameter C represents the cost, which is a positive constant specified by the user.

The optimization problem of the SVM is usually solved by introducing the lagrangian multipliers α_i , transforming the problem on the dual quadratic optimization.

The SVM method can also be used to classify nonlinear problems. By using a nonlinear mapping function, called Kernel function, the original data are mapped into a high-dimensional feature space, where the linear classification is possible. There are different Kernel functions used in SVMs, such as linear, polynomial, sigmoidal and Gaussian RBF. The selection of the better Kernel function is very important, since this function will define the feature space in which the training set examples will be classified [11].

V. METHODOLOGY

A. Experiments to setup parameters

At first, some experimental arrangements were made in order to evaluate the best setup parameter for the ANN MLP and for the SVM approaches.

For the ANN MLP, the number of hidden layers was limited in two. The tested topologies are shown in Table II. There are two MLP output neurons, one indicates the *rain* class and the other indicates the *no-rain* class. The validation set, necessary to avoid overfit was generated by selecting randomly 30% of the normalized complete data set.

TABLE II. EXPERIMENTAL ARRANGEMENT FOR MLP.

| Neuron quantity | |
|--------------------|---------------------|
| First hidden layer | Second hidden layer |
| 10 | 0 |
| 20 | 0 |
| 30 | 0 |
| 40 | 0 |
| 5 | 5 |
| 10 | 10 |
| 20 | 20 |
| 30 | 30 |

For the SVM, four kernel functions were tested: radial basis, linear, sigmoid and polynomial. For each kernel function the C parameter assumed, respectively, the values 1, 5, 10 and 30. The ε parameter was fixed at 0.001. And for the Naïve Bayes Classifier a Gaussian distribution function was assumed.

The test method for all experiments was the stratified cross-validation 5-fold. For the MLP the experiment was repeated twenty times. The Coruripe INMET database (near São Miguel dos Campos in the map of Fig. 3) was used to evaluate the best setup parameter for each technique.

The metrics used to compare the three techniques are the TP (True Positive) rate and the F-Measure. The F-Measure is an accuracy evaluation that considers the precision generating an overall score about the classifier. For this application, the TP of *no-rain* class is a very important measure, and this rate must be maximized. A false positive for the *rain* class will cause a decrease of the risk level of a flashover and the prediction system can miss the flashover event because of this false positive rain detection.

B. Experiments to evaluate the training applied to other data bases

With the best setup parameters, all three techniques were trained with the data from Coruripe INMET station and the trained models were applied in all others INMET stations.

The main objective was to evaluate if a training performed on one station could be applied to another one. The geographic limits of the training and the influence of the climate were also investigated.

C. Results on ADECI data

The trained models were applied on ADECI databases aiming to verify if the rain detection was satisfactorily. The analysis of these experiments could not be measured quantitatively because the ADECI data does not include the rain information. Instead careful visual inspections were made to identify the temperature and humidity behavior changes in order to qualitatively verify the results obtained. Those visual inspections will be better described in section VI-C.

VI. RESULTS

A. Evaluation of setup parameters

Table III presents the results for the Naïve Bayes Classifier. There are no parameters to adjust on this

classifier. The numbers in the table indicate that the Naïve Bayes Classifier achieves TP rates over 0.5 for both classes. However, the FP (false positive) rate of the *no-rain* class is still high for the application (the FP for the *no-rain* class is 0.227). The high result of FP *no-rain* is a bad issue as it can lead to unnecessary maintenance action for insulators wash.

Table IV presents the results for all ANN MLP topologies experimented.

TABLE III. EXPERIMENTAL ARRANGEMENT FOR NAÏVE BAYES CLASSIFIER.

| TP rate <i>rain</i> | TP rate <i>no-rain</i> | F-Measure <i>rain</i> class |
|---------------------|------------------------|-----------------------------|
| 0.807 | 0.798 | 0.746 |

TABLE IV. RESULTS FOR ANN MLP.

| Topology (as in Table II) | TP rate <i>rain</i> class | TP rate <i>no-rain</i> class | F-Measure <i>rain</i> class |
|---------------------------|---------------------------|------------------------------|-----------------------------|
| 10, 0 | 0.802 (0.047) | 0.873 (0.020) | 0.790 (0.016) |
| 20, 0 | 0.793 (0.049) | 0.877 (0.022) | 0.788 (0.016) |
| 30, 0 | 0.784 (0.049) | 0.878 (0.021) | 0.783 (0.016) |
| 40, 0 | 0.784 (0.049) | 0.878 (0.021) | 0.783 (0.016) |
| 5, 5 | 0.810 (0.050) | 0.866 (0.025) | 0.791 (0.016) |
| 10, 10 | 0.810 (0.051) | 0.869 (0.024) | 0.792 (0.017) |
| 20, 20 | 0.814 (0.049) | 0.867 (0.022) | 0.793 (0.016) |
| 30,30 | 0.812 (0.051) | 0.867 (0.022) | 0.793 (0.018) |

In order to choose the best topology for the ANN MLP, statistical tests were made. With the Shapiro Wilk test [12] all samples follow the normal distribution, and with the F test, all samples have the same variance. Complying with these assumptions, the T-Student test was applied to evaluate the best topology with statistical significance. The result of the T-Student test proves that there is no statistical difference between the topologies. So, the topology with fewer neurons in one layer was chosen. As shown in the highlighted cells in Table IV the results of the ANN MLP were better than those of the Naïve Bayes Classifier.

Table V presents the results for the SVM. In this table, only the best results for each kernel function are presented.

As the SVM classifier presents a unique solution, the set of parameters that resulted on the highest F-Measure was chosen (Radial Basis kernel function and C equals 10.0).

The results obtained with training and execution of the classifiers within the same database show that the rain pattern recognition is possible.

B. Training Generalization

A complete investigation about the best data training set is presented in Figs. 5 to 9. In these figures each INMET station was used as the data training set and the trained classifiers were tested on all stations, including that selected for training.

Fig. 5 exhibits the results obtained using Sobral INMET station as the training data set. All three classifiers were able to achieve good results of the True Positive *no-rain* patterns but the True Positive rates for the *rain* pattern were too low and unacceptable. It might be due to the few rain patterns on the INMET Sobral database. Instead, the small generalization power of Sobral station, all three classifiers present acceptable results when evaluated on the Sobral database itself.

TABLE V. RESULTS FOR SVM.

| Kernel Function | C | TP rate <i>rain</i> | TP rate <i>no-rain</i> | F-Measure <i>rain</i> |
|-----------------------|----|---------------------|------------------------|-----------------------|
| Linear | 1 | 0.758 | 0.896 | 0.781 |
| | 5 | 0.754 | 0.880 | 0.767 |
| | 10 | 0.754 | 0.880 | 0.767 |
| Polynomial (3 degree) | 1 | 0.256 | 0.973 | 0.393 |
| | 5 | 0.575 | 0.929 | 0.676 |
| | 10 | 0.643 | 0.916 | 0.717 |
| Radial Basis | 1 | 0.720 | 0.910 | 0.766 |
| | 5 | 0.749 | 0.889 | 0.777 |
| | 10 | 0.758 | 0.902 | 0.785 |
| Sigmoidal | 1 | 0.671 | 0.921 | 0.741 |
| | 5 | 0.744 | 0.905 | 0.778 |
| | 10 | 0.754 | 0.905 | 0.784 |

Fig. 6 exhibits the results obtained using Fortaleza INMET station as the training data set. Likewise as the training with the data of Sobral, this training with Fortaleza data exhibits good results for the *no-rain* class for all three classifiers on all locations, but Sobral. For the *rain* class, acceptable results were achieved on all databases and classifiers, except for Mossoró. The SVM and MLP classifiers produced better results for both *rain* and *no-rain* classes.

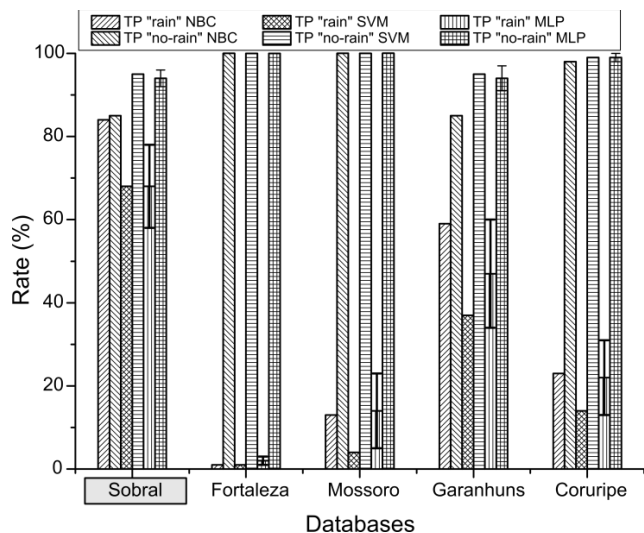


Figure 5. True positive rates for all nodes with Sobral data set training.

Using the Mossoró INMET station as the training database, only acceptable results for the *no-rain* class were obtained. Results for the Mossoró station itself, presented in Fig. 7, are worse than those of Fig. 6, obtained by training with the Fortaleza database.

For the Garanhuns and Coruripe INMET stations the best results obtained are with the respective databases. Figs. 8 and 9 clearly show that all three classifiers yielded excellent results for both the *rain* and *no-rain* classes when the trained classifiers were applied to the corresponding database.

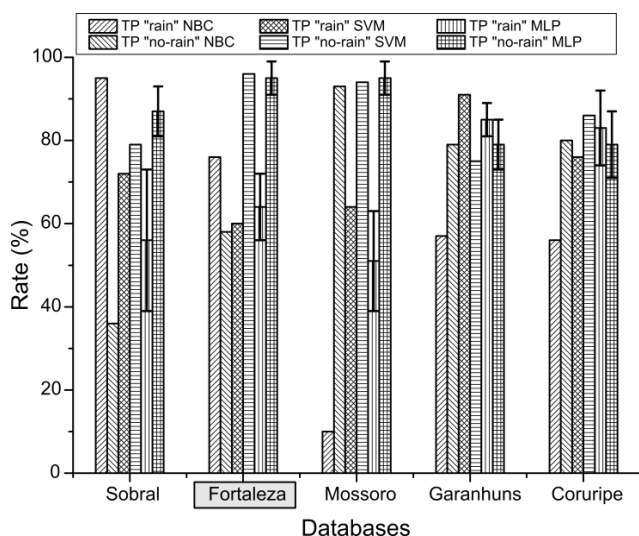


Figure 6. True positive rates for all nodes with Fortaleza data set training.

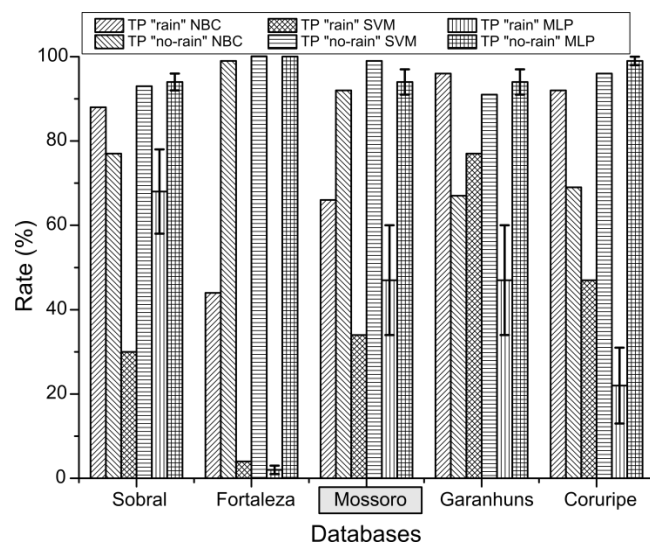


Figure 7. True positive rates for all nodes with Mossoró data set training.

Comparing all databases, the best option to predict the rain event in a given node of the sensor network is to use the database of its nearest INMET station.

Given that only the SVM and MLP classifiers presented good results, both of them could be chosen for the following analysis.

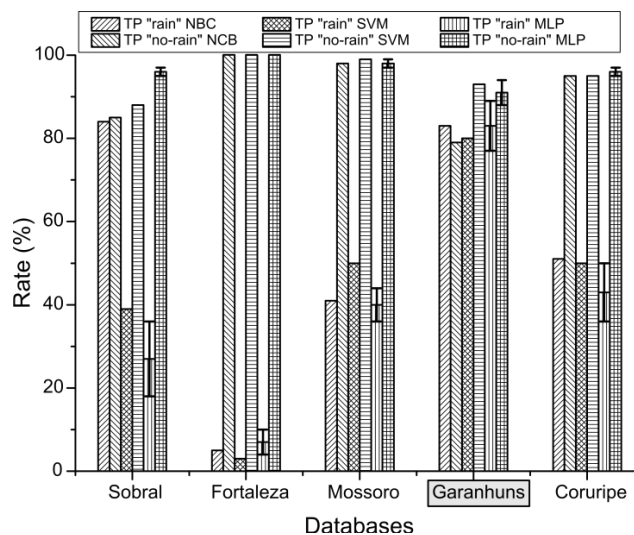


Figure 8. True positive rates for all nodes with Garanhuns data set training.

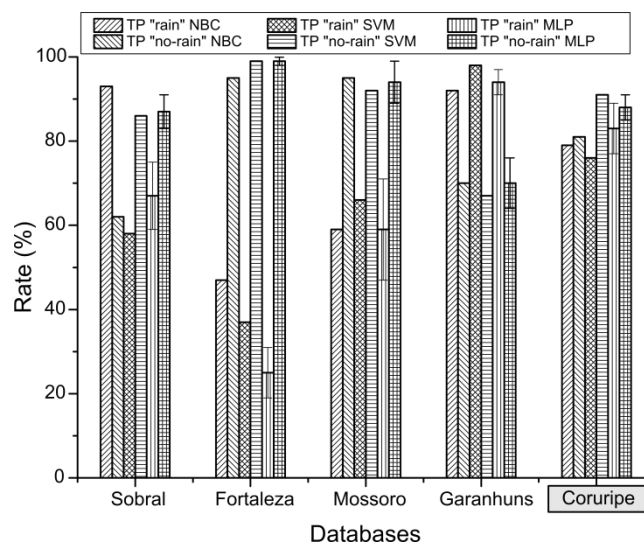


Figure 9. True positive rates for all nodes with Coruripe data set training.

C. Preliminary Evaluation on the ADECI data

The ADECI data does not include the information about the amount of rain, so, a visual analysis was made in order to verify results. In this preliminary investigation, the ANN MLP classifier was tested on the ADECI databases.

Fig. 10 shows the result of the ANN MLP classifier trained with the data from Coruripe INMET station and applied to the São Miguel dos Campos sensor system. The result of the ANN MLP is a binary neuron indicating class *rain* (one) and *no-rain* (zero). As can be seen in Fig. 10, the rain pattern was successfully recognized in some data subsets. The visual analysis of the rain pattern matches with the previous patterns in Fig. 2. Fig. 11 shows the result of the same ANN MLP applied in the Mossoró sensor system.

There are some possible rain events not successfully recognized. These events are marked in Figs. 10 and 11.

However, for every rain detected, the visual analysis of temperature and humidity suggests a rain event.

If a rain is not properly detected, as indicated in Figs. 10 and 11, the risk level will not be reset. If the risk level before the rain event was high enough to require schedule maintenance, this maintenance will happen, even with the insulator rain wash, causing an unnecessary spending by the electric company. But some rain events were detected, and in these cases the maintenance schedule could be reprogrammed with this new information. It is not possible to quantify the rain detection efficiency but visually it is possible to verify that 6 out of the 9 rain events in Fig. 10 were properly detected.

Furthermore, during a rain event, naturally there is an increase in the activity rate, mainly in the N1 range as can be seen in the last rain event, marked in Fig. 10. This activity increase causes an increment in the risk level leading to wrong interpretations. With the proper rain detection strategy, the activity increase can be related to the rain event and the risk level is not increased.

Regarding the rain events that were not recognized in the data of Figs. 10 and 11, the general visual analysis suggests that the false negative rain rate was higher for the Mossoró records relative to those of São Miguel dos Campos. The efficiency decrease observed in the Mossoró station suggests that it decreases with distance, indicating that one single model cannot be used to analyze all network nodes.

D. Long term ADECI data Analysis

After the training generalization described in Section VI-C, the classifiers were applied on long term ADECI data from Mossoró and São Miguel dos Campos. For the Mossoró ADECI data analysis, the ANN MLP classifiers trained on Mossoró and Coruripe INMET stations were used.

Fig. 12 shows only seven rain patterns recognized when the MLP trained on INMET Mossoró was applied on the Mossoró ADECI data. Fig. 13 presents the results obtained by the MLP trained on Coruripe INMET and applied on the ADECI data of Mossoró. The first 20 days are typically not rainy days and the pollution deposition was registered as activities on ranges N1 to N4. A small electric activity variation, which could be explained by a rain event, was observed. Only the MLP trained on Coruripe station was able to detect this possible event. Between days 25 and 275 there is no significant electric activity registered on ranges N1 to N4. This is because many of the rain events, shown in Fig. 13, clean the insulators.

In the last 100 days of the experiment, strong rain events are detected by both MLP results of Figs. 12 and 13. It is easy to reinforce the results shown in Figs. 7 and 9 that the best station (at this moment) for detecting rain events occurring at the Mossoró node is the Coruripe INMET station. The Mossoró INMET station data has 574 *no-rain* examples against only 70 *rain* examples; while the Coruripe INMET station data has 422 *no-rain* examples and 211 *rain* examples. In other words, the Coruripe INMET station is a more balanced database.

The severe weather conditions near Mossoró with high temperatures and low relative humidity damaged the

humidity sensors. Even after its replacement on day 240, approximately six month later the sensor was damaged again. Even with the malfunction of the humidity sensors the ANN MLP classifier was able to detect the rain pattern recorded by the temperature sensor in Mossoró. It means that the daily pattern recorded by the temperature sensor alone contains enough information to allow inferring the occurrence of rain events. For further experiments, a feature selection can be used to reduce the amount of data presented to the classifier.

Loss of data transmission were expected below 2% as specified by the satellite link providers [4]. But sequential losses were observed for periods longer than a few days. The causes of these large losses remain unknown, but strong rain events can jam the satellite transmitting signals producing these outages. In spite of these losses, future sensor system firmware updates will reduce these effects with the incorporation of data delivery checks and requests for package retransmissions.

Both ANN MLP classifiers trained on Mossoró and Coruripe INMET stations were applied on the long term ADECI data of São Miguel dos Campos. Results are presented in Figs. 14 and 15 for a period almost two years long. After the initial humidity sensor calibration on the first 70 days the sensor behaves very well for almost 600 days. After this period of continuous work the humidity sensor appears to exhibit malfunction yielding unreliable records.

Again, periods of no data transmission were observed, some of which occurring simultaneously with the periods of data loss verified in the Mossoró sensor system. Given that the distance between the two stations is approximately 600 km, and the cities have significant differences in climate, these transmission losses probably occurred due to an overall outage of the satellite link.

During the first 50 days of experiment, few rain events were observed, as indicated in Figs. 14 and 15. This period corresponds to the dry season and the absence of rainy days is reinforced by the activities registered by the sensor network on ranges N1 to N4.

The period between days 50 and 350 corresponds to the rainy season and a large number of rain events were observed, as indicated in the plots of Figs. 14 and 15. Again the absence of PD activities during this period reinforces the presence of many rainy events as correctly detected by the ANN MLP classifier.

Approximately one year after the sensor installation, a new dry season initiated near day 350. The electric activity on the polluted insulators increases again (ranges N1 to N4) and few rain events were detected. The classifier trained on Mossoró INMET station detected only two rain events between days 400 and 450, against four rain events detected by the classifier trained on Coruripe INMET station. The two last rain events, observed only on Fig. 15, are the most important difference of Figs. 14 and 15. These two rains match with the activity decrease, meaning that these rains washed the insulators naturally preventing partial discharges to grow to dangerous values.

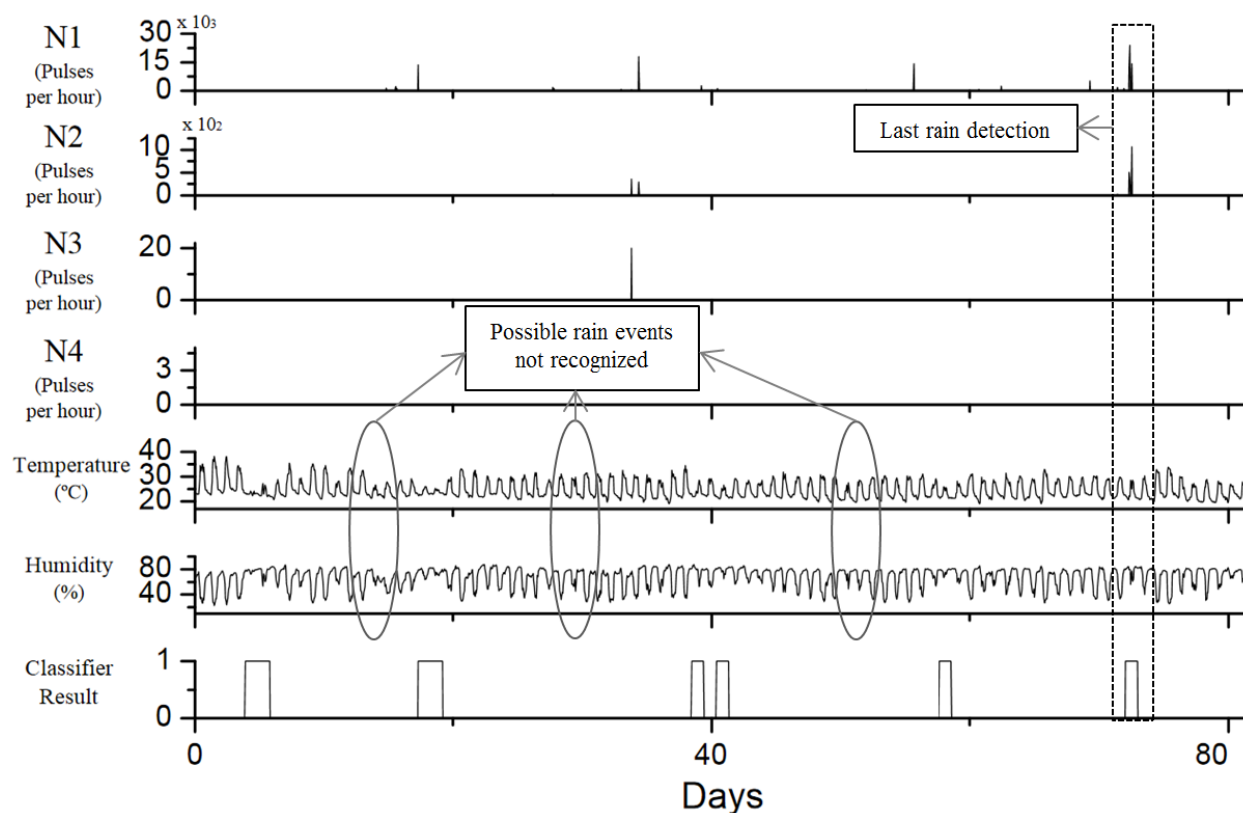


Figure 10. ADECI registers from *São Miguel dos Campos* sensor system and rain detection by ANN MLP classifier for 80 days.

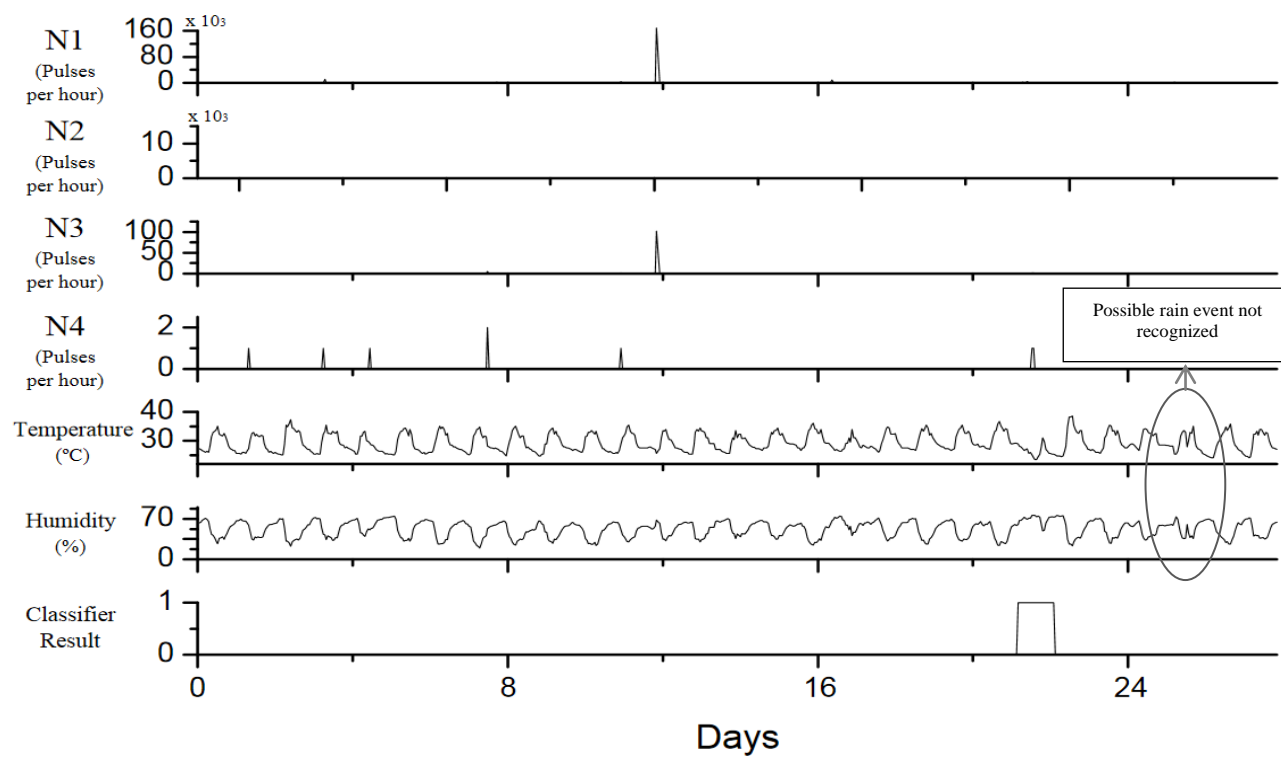


Figure 11. ADECI registers from *Mossoró* sensor system and rain detection by ANN MLP classifier for 28 days.

Comparing Figs. 14 and 15, it is possible to infer that the data of Fig.15 exhibit a better correlation between the detected rain events and the activities recorded on ranges N1 to N4. This visual analysis on the ADECI data of São Miguel dos Campos confirms the more efficient training data set of Coruripe INMET station as expected by results of Figs. 7 and 9.

These initial analyses on the ADECI data from São Miguel dos Campos and Mossoró visually shows that a balanced database for training the MLP classifier is more important than the climate differences. Due to that, some algorithms to balance the input data for the classifiers must be used in future works.

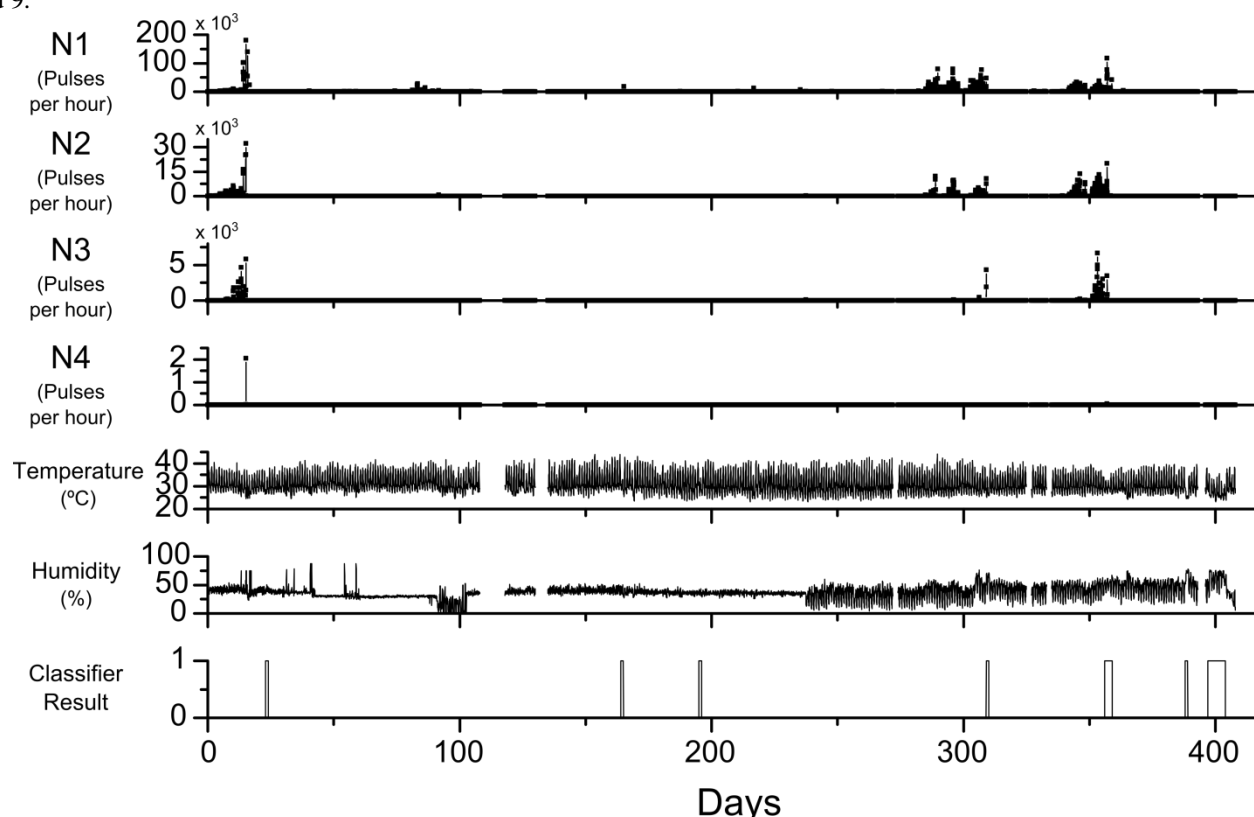


Figure 12. Application of ANN MLP classifier trained and applied on Mossoró INMET and ADECI data, respectively.

VII. CONCLUSION AND FUTURE WORK

This work presented a successful attempt to detect rain from the analysis of relative humidity and temperature data obtained from Chesf's sensor network. Results show that it is possible to detect rain events and use them to improve the flashover risk classification.

The initial tests were performed on the reliable data from INMET meteorological stations in the Northeast Region of Brazil. Three techniques employed, namely, Naïve Bayes Classifier, ANN MLP and SVM, presented acceptable results when tested on data from the same database. However, when the classifiers were trained with data from one station and applied to a distinct station, only the SVM and ANN MLP classifiers presented acceptable results. Given the different climates between the station sites, the generalization ability of the classifier is an important feature. Since SVM and ANN MLP presented similar results, only the ANN MLP was used as the classifier to be used with the ADECI data.

In the initial tests the ANN MLP trained with the São Miguel dos Campos INMET station was applied in data sets from the ADECI database. Data gathered by ADECI from two sensor locations in the network stations were used to evaluate the ANN MLP. The rain pattern was successfully recognized in this database, with some false negatives. Long term studies were performed on the ADECI data recorded by the Mossoró and São Miguel dos Campos sensor systems. The two closest INMET stations chosen for direct and cross tests confirmed that the best option is to perform training on a database from a nearby INMET station.

The results of this work can help improve the maintenance schedule system of the electric utility company. Without a rain detection attribute, when a rain event occurs, the sudden PD activity increase can give a false indication of risk of flashover. With the addition of the rain detection attribute, this effect will not be taken into account and after the rain event, the predicted risk of flashover can be reset because the insulators were washed.

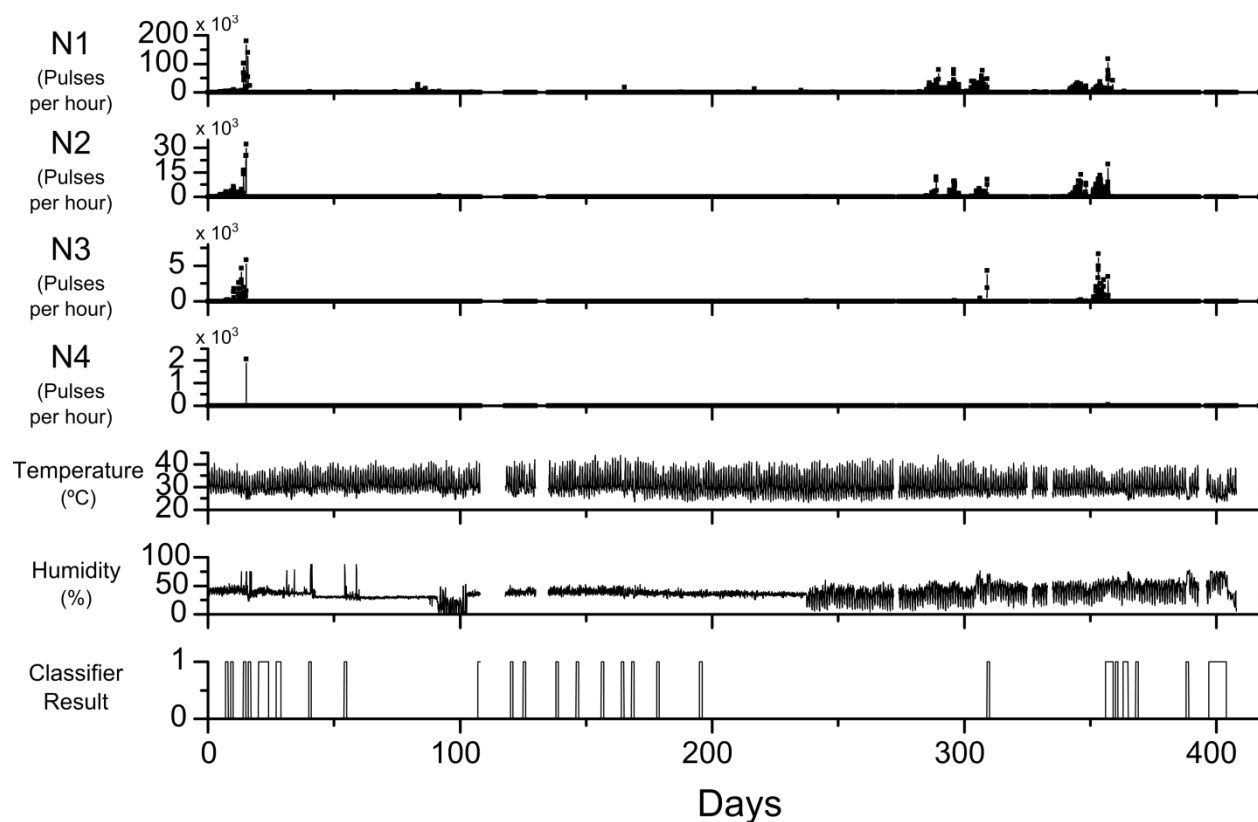


Figure 13. Application of ANN MLP classifier trained on *Coruripe* INMET and applied on *Mossoró* ADECI data.

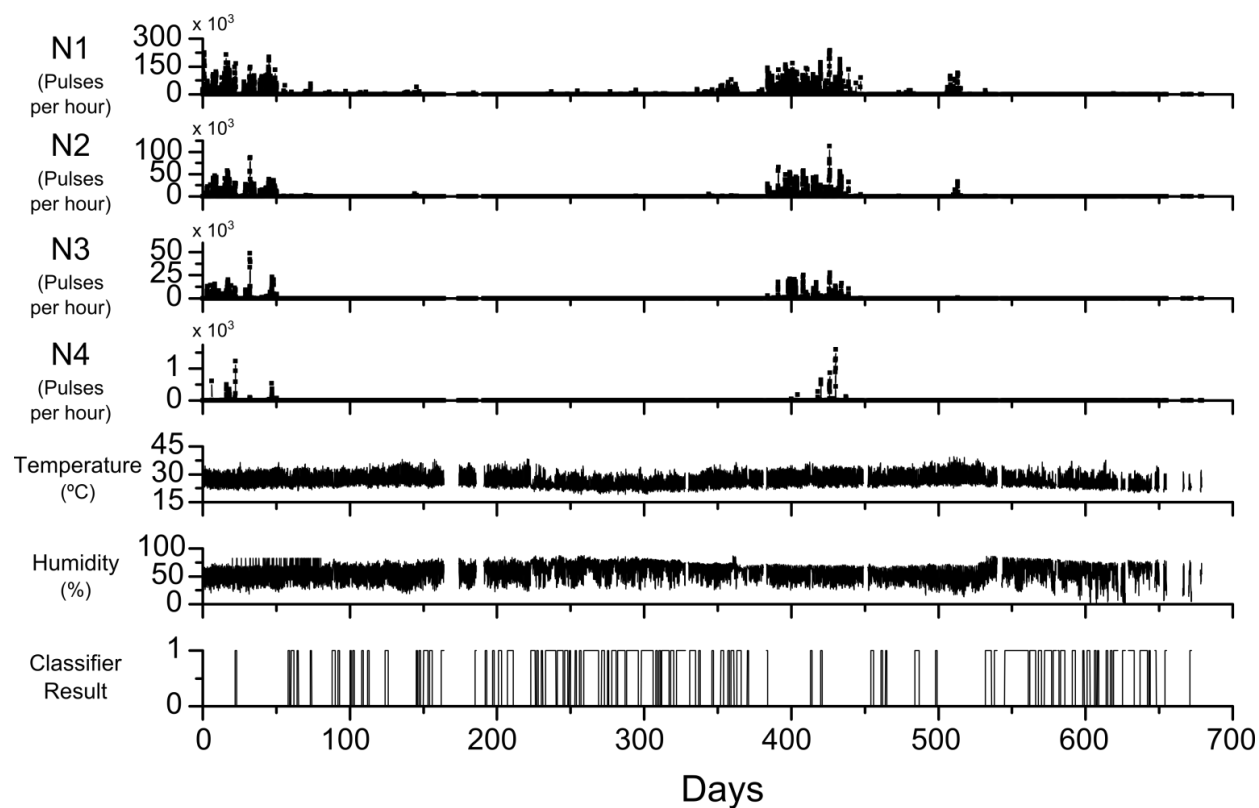


Figure 14. ANN MLP classifier trained on *Mossoró* INMET and applied on *São Miguel dos Campos* ADECI data.

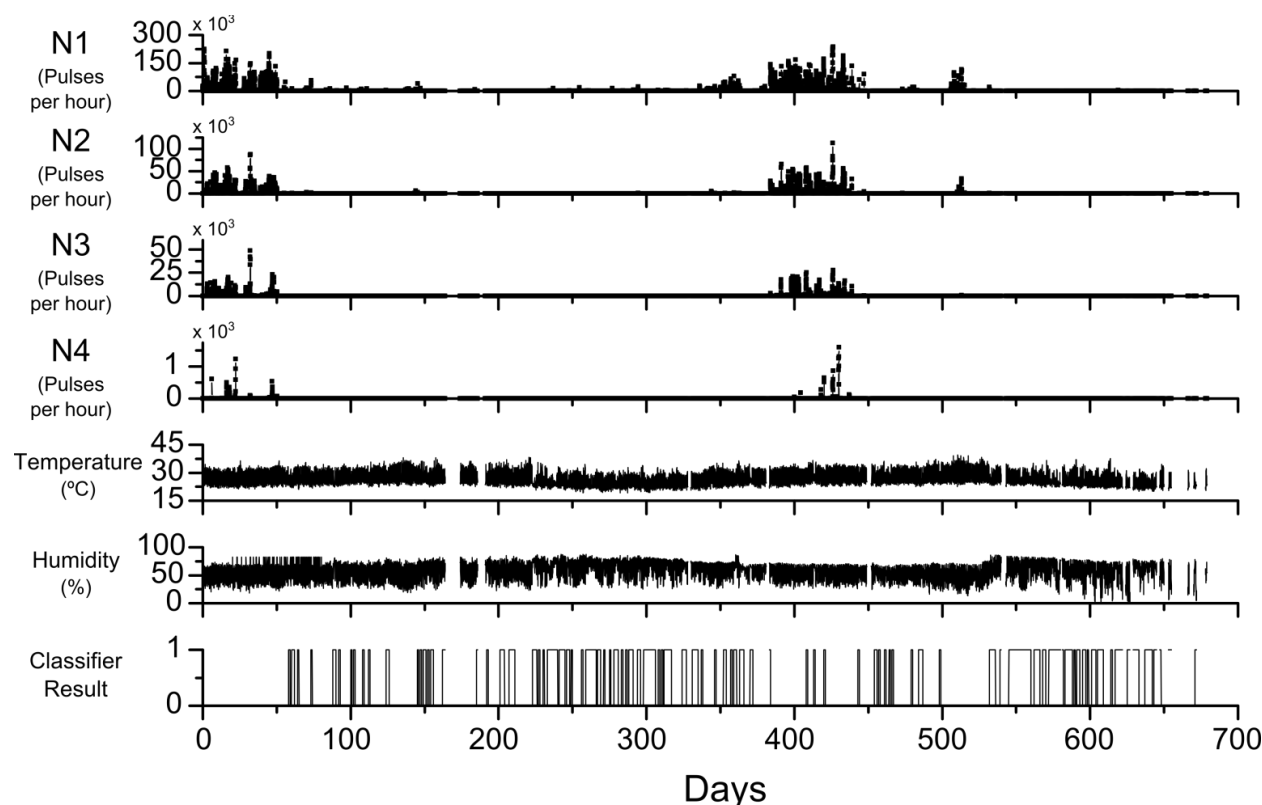


Figure 15. ANN MLP classifier trained on *Coruripe* INMET and applied on *São Miguel dos Campos* ADECI data.

Future work on the use of machine learning methods for rain detection will be directed to the evaluation of the threshold for rain precipitation to allow classifying or not a given day as rainy. Data sets from different locations will also be used in order to test the influence of climate characteristics on the proposed approach for rain detection and define the geographical boundary within which the same model can be applied.

Another improvement on the flashover risk prediction system is to use more than one classifier in order to prevent false positive and negative results.

ACKNOWLEDGMENT

The authors thank CHESF.

REFERENCES

- [1] L. H. S. Silva, S. C. Oliveira, and E. Fontana, "Evaluation of Machine Learning Methods in a Rain Detection System for Partial Discharge Data Analysis," *INTELLI 2013: The Second International Conference on Intelligent Systems and Applications*, pp. 176–183, 2013.
- [2] M. G. Danikas, "The definitions used for partial discharge phenomena," *IEEE Transactions on Electrical Insulation*, vol. 28, no. 6, pp. 1075–1081, 1993.
- [3] E. O. Abdelaziz, M. Javoronkov, C. Abdeliziz, G. Fethi, and B. Zohra, "Prevention of the interruptions due to the phenomena of the electric insulators pollution," *Control, Communiation and Signal Processing, 2004. First International Symposium on*, pp. 493–497, 2004.
- [4] E. Fontana, J. F. Martins-Filho, S. C. Oliveira, F. J. M. M. Cavalcanti, R. A. Lima, G. O. Cavalcanti, T. L. Prata, and R. B. Lima, "Sensor Network for Monitoring the State of Pollution of High-Voltage Insulators Via Satellite," *IEEE Transactions on Power Delivery*, vol. 27, no. 2, pp. 953–962, Apr. 2012.
- [5] E. Thalassinakis and C. G. Karagiannopoulos, "Measurements and interpretations concerning leakage currents on polluted high voltage insulators," vol. 421, pp. 421–426, 2003.
- [6] E. Fontana, S. Oliveira, F. J. M. M. Cavalcanti, R. Lima, J. f. Martins-Filho, and E. Meneses-Pacheco, "Novel Sensor System for Leakage Current Detection on Insulator Strings of Overhead Transmission Lines," in *2006 IEEE PES Power Systems Conference and Exposition*, 2006, vol. 21, no. 4, pp. 2255–2255.
- [7] H. O. de Lima, S. C. Oliveira, and E. Fontana, "Flashover risk prediction on polluted insulators strings of high voltage transmission lines," in *2011 11th International Conference on Intelligent Systems Design and Applications*, 2011, pp. 397–401.
- [8] K. N. Choi, "Omni-directional rain sensor utilizing scattered light reflection by water particle on automotive windshield glass," in *2011 IEEE SENSORS Proceedings*, 2011, pp. 1728–1731.
- [9] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience; 2 edition (October 2000), 2000, p. 654.
- [10] S. Haykin, *Neural Networks: A Comprehensive Foundation*. PTR Upper Saddle River, NJ, USA; Prentice Hall, 1994.
- [11] A. Christmann and I. Steinwart, *Support Vector Machines*, Springer. New York, NY: Springer New York, 2008.
- [12] G. K. Kanji, *100 Statistical Tests*, 3rd editio. London: SAGE Publications Ltd, 2006, p. 256.

Implementation of a Map-Reduce based Context-Aware Recommendation Engine for Social Music Events

Wolfgang Beer

Software Competence Center Hagenberg GmbH
Softwarepark 21
4232 Hagenberg, Austria
Email: wolfgang.beer@scch.at

Christian Derwein, Sandor Herramhof

Evntogram Labs GmbH
Leonfeldner Strasse 328
Linz, Austria
Email: [chris, sandor]@evntogram.com

Abstract—In our modern ubiquitously connected world the amount of ever available product and service information within our daily lives is exploding. Powerful client devices, such as smartphones and tablets allow the users to get access to an unlimited amount of information on every product or service available. As the amount of available information on products by far exceeds the users time to examine and filter detailed pieces of information in every situation, we expect that client-centric and context-aware information filtering is one of the thriving topics within the next years. A popular approach is to combine context-awareness with traditional recommendation engines in order to evaluate the relevance of a large amount of items for a given user situation. The goal is to proactively evaluate the situation of a user in order to automatically propose relevant products. Within this work we describe a general approach and the implementation of a software framework that combines traditional recommendation methods with a variable number of context dimensions, such as location or social context. The main contribution of this work is to show how to use a MapReduce programming model for aggregating the necessary information for calculating fast context-aware recommendations as well as how to overcome a typical cold start problem. The use-case at the end of this work evaluates the practical benefit of our general framework to introduce a client-centric, MapReduce-based recommendation engine for real-time recommending music events and festivals.

Keywords—context awareness, context aware recommendation, decision support, recommendation system.

I. INTRODUCTION

Today, the world is annotated by petabytes of digital product and service information distributed across many different ubiquitously accessible global data repositories. Users of various applications and services are constantly submitting additional information or feeding the data repositories with their preferences and experiences. Smartphones and tablets act as a window for viewing and receiving this annotated information as well as to give the users an input device in order to collect additional information. E-business, marketing and e-commerce is profiting a lot by this pervasive use of additional product and customer information. Global marketplaces, such as eBay, Amazon, Apple iTunes or Google Play, offer millions of different products and services in hundreds of categories. These

categories span a wide spectrum of product families from traditional hardware to software and mobile apps, eBooks, electronics, video and music streaming or even food. The huge amount of permanent available information makes it difficult or even impossible for users to manually select a relevant subset. As a human user is not able to review all available information, the selection of this subset is of crucial importance for both, the human consumer as well as for the information publishers. The most common real world scenario is a human user searching for a product or service and a huge number of companies offering information on their specific offer. Recommendation engines are one available technique to overcome this information overload and to automatically select a subset of relevant information for a human user. According to the huge number of products available in global marketplaces and the consumers limited time and motivation to check all similar products, recommendation engines provide the necessary rating and pre-filtering for human consumers. Recommendation systems, such as the product recommendation at Amazon or eBay, are already present for several years. Without traditional recommendation systems, the consumer soon gets lost within the huge amount of available products. In a previous work we proposed a general method for combining different context-dimensions along with our general context-model to describe people along with their music interests [1]. In order to solve that problem, global marketplaces soon recognized the need for transparent product recommendation within their systems. In 2006, the Netflix Prize competition was initiated with a 1 million dollar prize for achieving a ten percent or more improvement of Netflix's video recommendation algorithm. The training set that Netflix published for the price competition contained around 100 million ratings from about 500.000 anonymous customers on 17.000 videos. The contest attracted 48.000 competing teams from 182 different countries. The winning team (BellKor) from AT&T Research Labs (made up of Bob Bell and Chris Volinsky, from the Statistics Research group in AT&T Labs, and Yehuda Koren) was able to improve the performance of Netflix's recommendation algorithm by 8.43 percent. So it is obvious that traditional

recommendation systems play an important role in modern consumer markets. While recommendation methods for traditional item recommendation, such as Slope One recommendation or Matrix Factorization, have been widely addressed within the last decade, many interesting aspects of client-centric recommendation systems have not been within the focus by the recommendation research community so far. Bell et al. identified several such research aspects during their work on the Netflix prize competition [2]. One of these aspects is to address the client-centric view on recommendation systems, in terms of evaluating and including the consumer's actual context during the recommendation process. Client-centric recommendation system approaches, such as implementations on smartphones and mobile devices need to focus on the user's demands in a tight relation to the users actual situation. For any mobile user the context-dimensions time, location, weather, activity and companions play a major role in any decision. Bell et al. also identified that a combination and blending of several quite simple recommendation approaches often result in excellent recommendations. In this work, we will present the implementation of a software framework that uses a MapReduce programming model approach for distributed data aggregation for blending of multiple context-dimensions. The framework is built on top of a MongoDB noSQL distributed database and uses the Apache Mahout recommendation framework for designing new context-aware and customizable client-centric recommendation models.

The remainder of this work is structured as follows: Section II gives a short overview on state-of-the-art in recommendation systems, music and event recommendation engines, map reduce data aggregation and related work on how to introduce context-awareness in recommendations. Section III focuses on the requirements a general framework for context-aware recommendation systems has to fulfill. Section IV gives an abstract overview on our approach for introducing context-information in traditional recommendation methods and Section V defines a practical software architecture and implementation of our approach. Section VI explains some evaluation results that were collected during the test phase. Section VII concludes with an application case study that introduces context-aware recommendation in the domain of social music and festival events. The last Section VIII discusses general findings, conclusions as well as further research activities.

II. STATE OF THE ART

The importance of context-awareness in human-centered computing systems has been discussed by various different research communities, including ubiquitous and pervasive computing, mobile computing, e-commerce and e-business, information retrieval and filtering, marketing and management as well as within several engineering disciplines. Through the massive increase of hardware capabilities in combination with cheap broadband access in consumer electronics, such as mobile phones and tablet PCs, the need for context-related information filtering is dramatically increasing too. To discover

and evaluate the context and situation of a mobile user is a key challenge within the smart filtering of relevant information out of a huge information space. The term context-aware software was first used in the Xerox PARC research project PARCTAB in 1994 [3]. In this work, the term was specifically dedicated to software that is able to adapt according to its actual location, the collection of nearby people, hosts and accessible devices. Also the possibility to track the changes of context information over time, in other words to store historic situations, was mentioned. Over the years, different research groups enriched this basic definition of context and context-aware software. Brown et al. [4] widened the scope of context information to temperature, time, season and many other factors. Due to the fact that theoretically the number of context information factors is unlimited, the definition of context by Anhind K. Dey is one of the most commonly used:

Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves. [5]

This definition of context specifies that context contains any kind of information about an entity in order to understand its situation. Often the term context is limited to location information and location-awareness, but in recent years context also is enriched with the social network of a user. Collecting and evaluating the social dimension of context related to a specific user goes hand in hand with a detailed discussion about privacy and security. An interesting fact about the above definition of context is that Dey identifies three base classes that classify all objects: person, place and object. This kind of classification has practical reasons but is also fixed to a location-dependent view of context information. Over the last ten years several architectures and implementations of software middleware frameworks were published that emphasized the aggregation and interpretation of context-information. In our basic research work from 2003 we already proposed the possibility to use Event-Condition Action (ECA) rules to model the context of an entity [6]. The basic idea behind most of the research activities within context-acquisition, processing and interpretation is to use a user's context information in order to filter relevant information (e.g., on products, services, locations) from the huge collection of available information. An approach that tries to solve the same challenge is to use recommendation methods and algorithms to select a subset of information that seems to be relevant for a user. These recommendation approaches have long tradition within global marketplaces. Traditional recommendation systems take a set U of users and a set of products (items) P , which should be recommended to a user. A recommendation system then provides a utility function f that measures the relevance of a product out of set P to a given user. This utility function f ($f : U \times P \rightarrow R$, where R is an ordered set of numbers) assigns a rating to each item (or even to a compound set of items) in a way that captures the relevance or preference

for a specific user. The objective of recommendation systems is to find or learn this utility function f . Function f is used to predict the relevance of items out of P and of new appearing items with similar attributes. In the literature different approaches exist for finding a function f by using an available dataset. Traditional recommendation approaches are distinguished into two major strategies: content filtering and collaborative filtering.

A. Content Filtering

The content filtering approach creates profiles for each item and user, in order to characterize and compare its nature [7]. Each profile contains a specific set of attributes, which can be used to compare objects. For example, a restaurant could have a cuisine attribute, describing the type of food it offers, a location attribute, a vegetarian tag, and so on. A recommendation function f chooses items that are similar to items the user has already chosen or rated before. The utility function compares the user's profile and calculates the similarity of a user profile with the available items. Therefore, the user profile allows the recommendation engine to create a list of items that could fit to a given user profile. Many implementations of this approach additionally refer to Linked Data information, such as RDF stores and Semantic Web repositories, to classify and search systematically for related information.

B. Collaborative Filtering

In collaborative filtering approaches, the recommendation function chooses items that were preferred by other users with similar attributes. Collaborative filtering approaches depend on either explicit or implicit user ratings of items. By rating different items a user can feed explicit ratings into the recommendation engine, while implicit feedback is collected by the system through the analysis of the users behavior (previous purchases, navigation path, search terms, etc.). Collaborative filtering is domain-free, which means that it can be applied to any application area and to different data aspects, which could be hard to formulate into an explicit profile. Collaborative filtering is more accurate than content filtering [7], but has the challenge of starting without any initial data sets (cold start problem). It is not directly possible to address new users or objects where the system has no initial data set available. Popular collaborative filtering methods are neighborhood methods and latent factor models. The Pearson's correlation coefficient $sim(u, v)$ is often used to calculate the popular neighborhood method kNearest Neighbor, in order to measure the similarity between the target user u , and a neighbor v . Within the Pearson's correlation the symbol \bar{r}_u corresponds to an average rating of user u and P denotes the set of products or items.

$$sim(u, v) = \frac{\sum_{i \in P} (r_{u,i} - \bar{r}_u)(r_{v,i} - \bar{r}_v)}{\sqrt{\sum_{i \in P} (r_{u,i} - \bar{r}_u)^2} \sqrt{\sum_{i \in P} (r_{v,i} - \bar{r}_v)^2}} \quad (1)$$

Another method uses association rules to explicitly model the dependency and similarity of items. A rule could for example state that if a customer buys item A in combination with item B, then the engine should also recommend to buy item C. One of the most widespread methods for calculating latent factors is matrix factorization, which is described in detail in [7]. Most of the modern recommendation systems use a combination, a so called hybrid approach, of content filtering and collaborative filtering approaches to further improve the accuracy of recommendations. Beside these traditional approaches for implementing recommendation algorithms, several groups are working on the challenge of customizing recommendations and to build flexible recommendation queries. REQUEST: a query language for customizing recommendations was published by Adomavicius et. al. in 2011 [8], which promotes a custom query language to build flexible and customized recommendation queries based on multidimensional OLAP-cubes. Contributions have been made by research groups that built various application scenarios for context-aware recommendation systems, ranging from recommendation of sights within the tourism domain [9], restaurants [10], or even people (e.g., glancee.com).

Calculating recommendations out of a huge amount of distributed data sets also means to handle these distributed calculations within acceptable performance. Typical data sets for traditional recommendation systems consist of millions of ratings and products as well as of hundred thousands of users. This amount of data require strong processing power and a data aggregation method that can cope with distributed and parallel processing of data sets. A popular and stable framework for processing distributed data sets is Apache Hadoop, which builds upon the HBase and implements a software framework that supports data-intensive distributed applications. The underlying HBase is an open source, non-relational, distributed database modeled after Google's BigTable approach and is written in Java.

III. FRAMEWORK REQUIREMENTS

This section discusses general requirements for implementing a framework that supports the design of context-aware recommendation systems. To discuss all requirement in detail would exceed the scope of this work, so we focus on several requirements that had a high priority for our use-case in Section VII.

A. Flexible and dynamic customization

A client-centric view on the recommendation process, demands for a flexible user interface to enable the customization and fine tuning of recommendation impact factors for non-technical users. So the users should be able to control the learning and recommendation process at a most fine grain level, while the configuration and presentation should be on an abstract and understandable level. The user should be able to specify a variable number of impact factor dimensions and even to add custom defined impact factors. The framework

should normalize all the chosen impact factors and automatically provide a list of recommended items that is sorted according to the weighted sum of normalized impact factors.

B. Temporal aspect

Temporal aspects [11] deal with the change of the context and with the change of the content profiles over a timeline. A recommendation framework has to consider the fact that the importance of specific datasets may change over time. It makes a big difference, if a person has bought an item yesterday or 10 years ago. A general framework has to cope with this varying impact.

C. Transparency

To raise the users' confidence in recommendations, it is of crucial importance to give immediate and transparent feedback on recommendations. The recommendation framework has to provide a human understandable explanation for a given recommendation set. Sundaresan, from eBay research, published a great article about the 6 questions you have to address during the design and implementation of recommendation engines [12] (What, Where, When, Why, Who and How). He also points out that recommendation engines that address the transparency aspect (the Why question), offer a better conversion rate in e-commerce applications. There are several user studies that clearly show that addressing the transparency aspect improves the performance of recommendation engines [13].

D. Performance

The performance of the calculation and delivery of recommendations for a user is one of the most critical non-functional requirements. The acceptance of a user much depends on whether the information is shown at the right time. This is even more important for sensing the context and delivering the recommendation results to a mobile user, as especially this environment is changing a lot within a quite short period of time. Recommendations that consider the location and activity of a user have to react in time to provide recommendations in the specific situation, when a user expects them. As actual recommendation approaches harvest and analyze a huge amount of data, the requirement for performance during the distributed data retrieval and processing is critical for every implementation.

E. Quality

As users are implicitly benchmarking recommendation engines according to the quality of recommendations they are able to provide, it is necessary for a general framework to provide a standard approach for evaluating the quality of recommendation engines. A framework has to provide implicit and explicit quality evaluations, which means that the framework constantly evaluates the quality of results by using test data sets, as well as to explicitly ask the users for quality feedback.

IV. APPROACH

Within the scope of this work a general approach for the implementation of context-aware recommendation systems is presented. This approach mainly proposes to introduce a map-reduce programming model for processing large context information data sets with a parallel and distributed cluster of noSQL databases. The combination of highly dynamic context information with traditional recommendation algorithms puts high demands in particular on the performance of calculations as well as on the performance of data aggregation. Especially within the process of combining and aggregating raw sensor information, to gain abstract context information, the efficiency and performance of clustered data aggregation is a critical aspect. A general approach for context-aware recommendation systems has to define the impact of context related, dynamic information on the recommendation process. Compared to the traditional recommendation approaches, which were already discussed in Section II, we combine these traditional collaborative filtering approaches with user related context-information. This also means that for each individual application scenario there exists a quite specific collection of context aspects that offer high relevance for the recommendation of entities in a certain situation. While the location information might not be relevant for recommending books in an online bookshop, it is of crucial importance for the recommendation of nearby restaurants. Within our proposed approach, each dimension of a given context, such as location, weather or companions is represented through an impact function. An impact function defines the influence of one dimension of a given context on the overall relevance within the recommendation process. All impact functions are of the given form $f_i : U \times P \times C \rightarrow R$, where U represents the set of users, P the set of products and C a dimension of a given context (e.g., location, number of nearby friends, etc.). The weighted sum of all normalized impact functions results in an overall relevance for a product p , a given user u and context c , where w_i represents a weight that the end user defined for a specific dimension of the context:

$$f(u, p, c) = \sum_{i < P} f_i(u, p, c) w_i(c) \quad (2)$$

A general framework for context-aware recommendation systems has to offer the basis for customizable recommendation engines that consist of a variable and dynamic set of impact functions that can either be predefined by the framework (e.g., aspects such as distance, user ratings, history, friends, ...), or explicitly defined by users. Furthermore, the users are able to dynamically define and adapt the weight of different impact functions according to their preferences, which is shown in Figure 1. As Figure 1 shows, the customized recommendation system within this example contains six different impact functions. Each of these impact functions calculates the relevance



Fig. 1. General approach for a weighted combination of context with collaborative filter dimensions

of a product for a given user and context dimension. As it was already mentioned, a context dimension could be the distance, friends (companions), product category, ratings, or prices. This multidimensional approach is not limited to a fixed set of impact functions, but can be enhanced by including additional context dimensions. Therefore, a designer of a domain specific recommendation system has to provide a domain-specific set of additional functions, in order to improve the quality of recommendations for the users in different application domains. The radar chart in the lower left corner of Figure 2 visualizes the weight an individual user defined for a given set of impact functions. Each user is able to specify his own, very personal weight of each context or collaborative filter dimension. The example setting shown in Figure 2 defines a typically high weight for the distance between the user and a given event (here given by 30% impact), 5% impact for the price of an event and moderate weight for the artist (20%) performing an event, companions and friends coming along (15%) and the collaborative filtering result (20%). By giving the user the possibility to define his own personal weights it is possible to modify the recommendation result on a quite fine granular level. An important aspect of this approach is also the ability to calculate the impact of each context dimension by using completely different strategies. While the impact of location and distance could be calculated through a simple spatial query, the rating impact function could be implemented as traditional collaborative filtering approach. These individual impact functions are then calculated by using a map-reduce programming model approach. The huge amounts of raw data sets are collected within several parallel map steps and aggregated in subsequent reduce steps until the result is fully available. The following section describes in detail the overall system architecture as well as the implementation details of this approach.

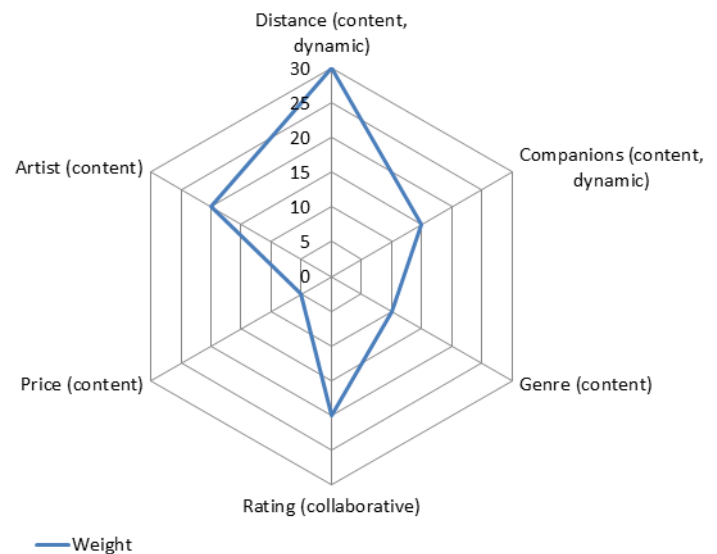


Fig. 2. User defined weight for available context and collaborative filter dimensions

V. IMPLEMENTATION

The general software architecture for building a context-aware recommendation system is derived into a typical client-server architecture model. This client-server architecture uses a MapReduce programming model, as it was already mentioned in the general approach in Section IV along with several critical subsystems. As it is shown in Figure 3, the server defines all necessary subsystems for data access and third-party information retrieval, user interfaces for manual content selection and correction, as well as the context-sensitive

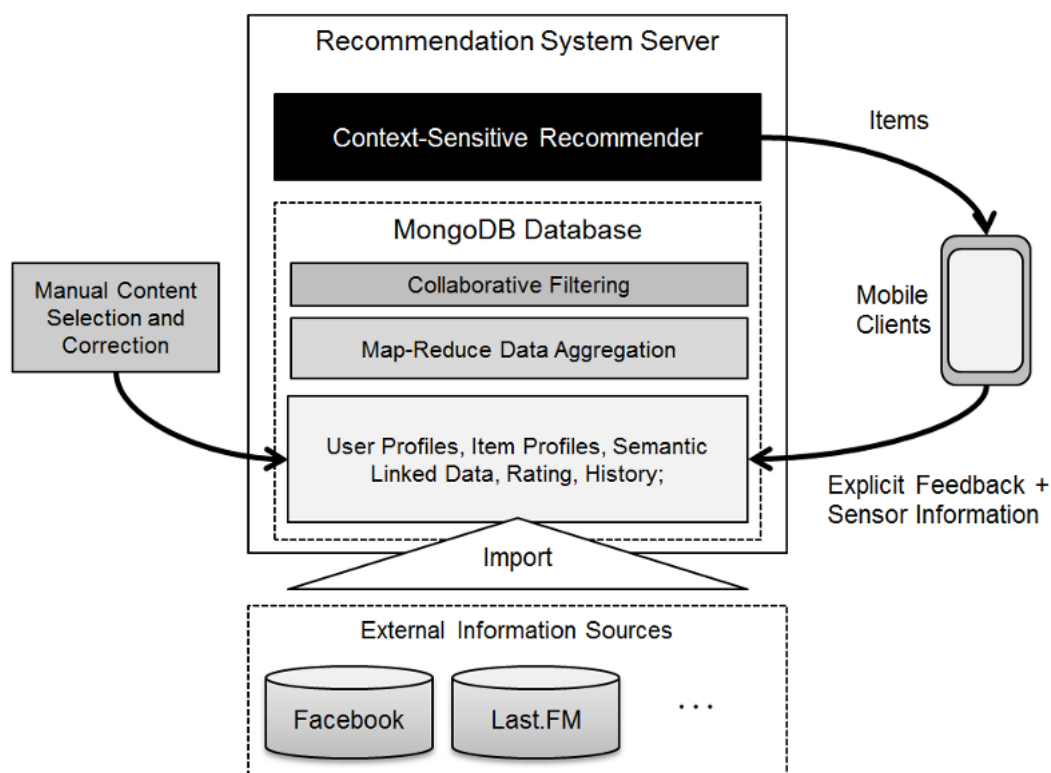


Fig. 3. General software architecture for the implementation of a context-aware recommendation system

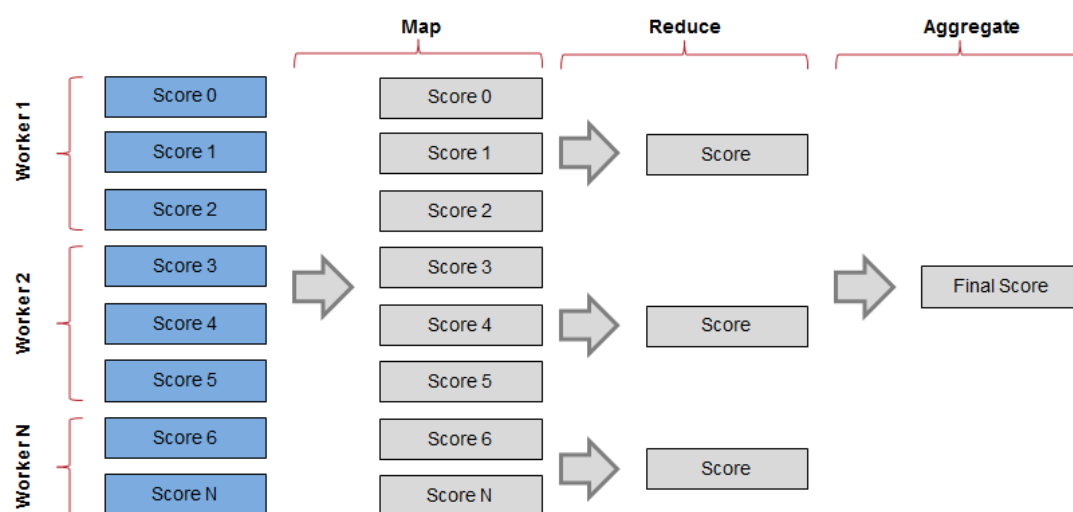


Fig. 4. The MapReduce data aggregation process

recommender. Main part of the server implementation is the management of raw context information along with bound user and item profiles. Semantic data that is used to calculate the similarity and relation between given users or items is stored by using linked data repositories. All additional semantic data can be accessed by using semantic web standards and query languages, such as RDF and SPARQL. The purpose of accessing these sources of semantic information is to receive additional item-based similarity measurements that are used in combination with traditional collaborative filtering result. External sources of semantic information, such as Facebook or Last.FM, are either directly imported and duplicated, or directly accessed through a defined service interface. The decision if an external information source is either imported or directly accessed depends on the third-parties' service level agreements. On top of the management of raw context and profile information the map-reduce data aggregation layer is responsible for collecting and aggregating these raw information into abstract context-information. Within the map-reduce layer several individual map-reduce processes are calculating normalized context-dimensions that are combined to a common rating table between users and items. The MapReduce programming model defines two fundamental steps: Map and Reduce. During a parallel Map step all distributed databases collect the available data sets. Following code shows a typical structure implementing a Map-function within the noSQL database MongoDB:

```
function()
{
    emit(
    {
        user_id : this.user_id,
        item_id : this.ref_id
    },
    {
        score : this.score
    }
    );
}
```

Within this Map function all database sets are collected that contain a score between users and items and emitted as intermediate result. The Reduce step gathers all these intermediate results and calculates an overall score for all user item relations, as it is shown within the following example:

```
function(key, values)
{
    var total = 0;
    for ( var i = 0;
        i < values.length;
        i++ )
    {
        total += values[i].score;
    }
}
```

```
return { score : total };
};
```

After the last Reduce step has been performed, the resulting data set is organized as a sparse score matrix between individual users and available items. The collaborative filtering layer on top uses traditional recommendation engines to fill the missing gaps within this sparse user-item score matrix. Typical algorithms used within the collaborative filtering layer are slope one recommendation or matrix factorization methods. Figure 4 visualizes the stepwise, distributed MapReduce process for parallel data aggregation.

A. Cold Start Problem

Another important implementation detail is how to handle and avoid the initial cold-start problem that is typical for collaborative filtering solutions. Within our approach a combination of content based filtering with aggregated substitute ratings is used to fill the gap of missing explicit user ratings. As the relevance of social events for a person is very much related to the actual distance, the recommendation engine first uses a sorted list of distances in combination with selected indirect factors that are extracted from the users contexts. As the initial system not only lacks of a large number of explicit user ratings but also of a large number of users, additional user information is gathered from connected Facebook profiles. These passive user profiles are not considered as active participants of the system but act as a critical mass for calculating hidden factors within given user/item ratings. According to the given software architecture and sensor availability, we selected the following indirect factors for calculating an aggregated substitute rating:

Library

The smartphone of the user contains a given music artist. Some mobile platforms also provide some simple statistics about the last time the user listened to that artist or how often.

Forecast

A forecast is given if the user expects to visit an event and shares that information with his friends.

Checkin

Similar to other widespread location-aware social platforms, like Foursquare, a Checkin is explicit information that the user arrived at a given event.

View

Several different view statistics record the users history of reading artist, event or venue information.

An aggregated combination of these indirect rating factors is then used to fill the sparse rating matrix. The metamorphosis between indirect ratings and explicit user ratings is a continuous process in which explicit user ratings iteratively replace the substitute values whenever available.

As a result of the collaborative filtering process, the server database contains a matrix of given ratings, user and product profiles as well as additional semantic information. The recommendation module is responsible for combining the different

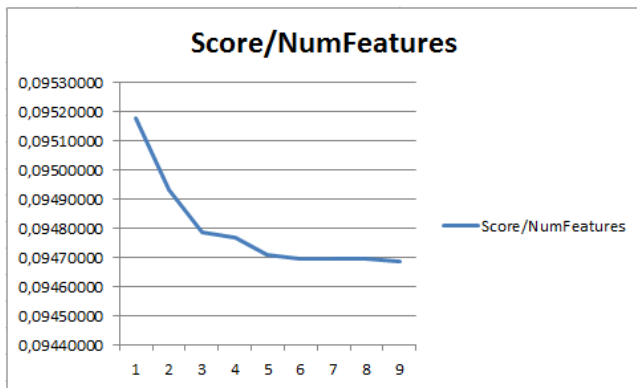


Fig. 5. Evaluation score by a varying number of features

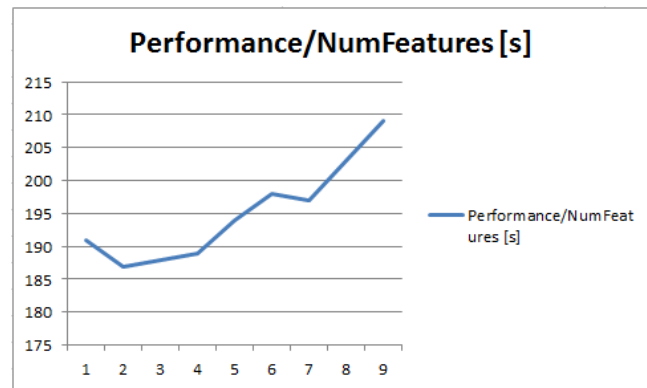


Fig. 6. Evaluation of performance by a varying number of features

dimensions of the recommendation approach in Section IV and to communicate the resulting ratings to the clients. The client-server communication is implemented as a lightweight REST (REpresentational State Transfer) service approach. On the client-side, a local application is visualizing the resulting list of recommendations and is collecting the necessary context information in combination with the user's feedback on the given recommendations.

VI. EVALUATION

Within this evaluation section some detailed results in terms of recommendation quality and performance are visualized. The quality of recommendation results within this work is evaluated by calculating a quality score by using an average absolute difference evaluation method (mean average error score) that divides the available ratings into 70% training data and 30% evaluation data. The collaborative filtering method used within this evaluation is based on a matrix factorization projection of given users and event items onto a feature space. Figure 5 shows the evaluation score with varying number of features (starting with one feature to nine features). Our evaluation showed that increasing the number of features above an amount of five does not significantly improve the overall recommendation score. At this stage of our evaluation data, the evaluation score itself does not provide any useful information as most of the ratings are binary ratings automatically extracted from passive user profiles. So the range of the ratings are between zero and one.

Beside the evaluation of how the number of features improves the recommendation score we also evaluated the performance of the overall calculation. The performance tests were done on a standard Windows 7 Laptop device with 4GB RAM and a Intel Core i5 64bit CPU with 1.70GHz. Figure 6 shows that the overall calculation of the weighted recommendation ratings by using the discussed MapReduce programming model is performed within around 190 seconds. The database contains 408,634 passive and active user profiles and 22,901 different events (items). After the MapReduce programming model aggregated the scores, as it was shown in Section V the resulting number of ratings is 399,905.

The evaluation of this framework shows that it is a valid concept and approach for implementing a context-sensitive recommendation engine that uses a MapReduce programming model in combination with collaborative filtering. The evaluation of the quality of recommendations does not provide any significant results as the data set does not contain enough explicit user ratings so far.

VII. USE-CASE: EVNTOGRAM

The following use-case was selected out of a running project in cooperation with EVNTOGRAM, which is a platform operator for personalized and context-sensitive recommendation of music events. The philosophy of EVNTOGRAM is to analyze the users' music favorites and activities, as well as their social interaction, in order to offer personalized and context-aware recommendations for events, specifically in the domain of music events, such as concerts and music festivals. A recommendation approach, explained in Section IV and Section V, helps to include various context-dimensions into the calculation of the relevance of an event for a given user. EVNTOGRAM records these context-dimensions, such as the users' activities, social interaction, music listening habits and individual ratings, in order to sort a list of music events according to the calculated relevance, as it is shown in Figure 7. In a first prototype EVNTOGRAM is trying to find out, which subset of context-dimensions is providing good recommendations for the users. In that sense 'good' means the feedback the user is providing for a given ordering of items. After the initial release of the EVNTOGRAM client app for the platforms Android and iOS recommendation calculations were performed for around 500 active and more than 400,000 passive user profiles. Passive user profiles represent profiles that come from external sources, such as Facebook and help to overcome the so called cold start problem. The operation of the EVNTOGRAM platform within the last month returned one critical user review concerning the degree of transparency for event recommendations. The criticism is related to the quite unclear process of proposing events by our recommendation engine to the users. A detailed explanation of the recommendation process would help the user to understand why a rating

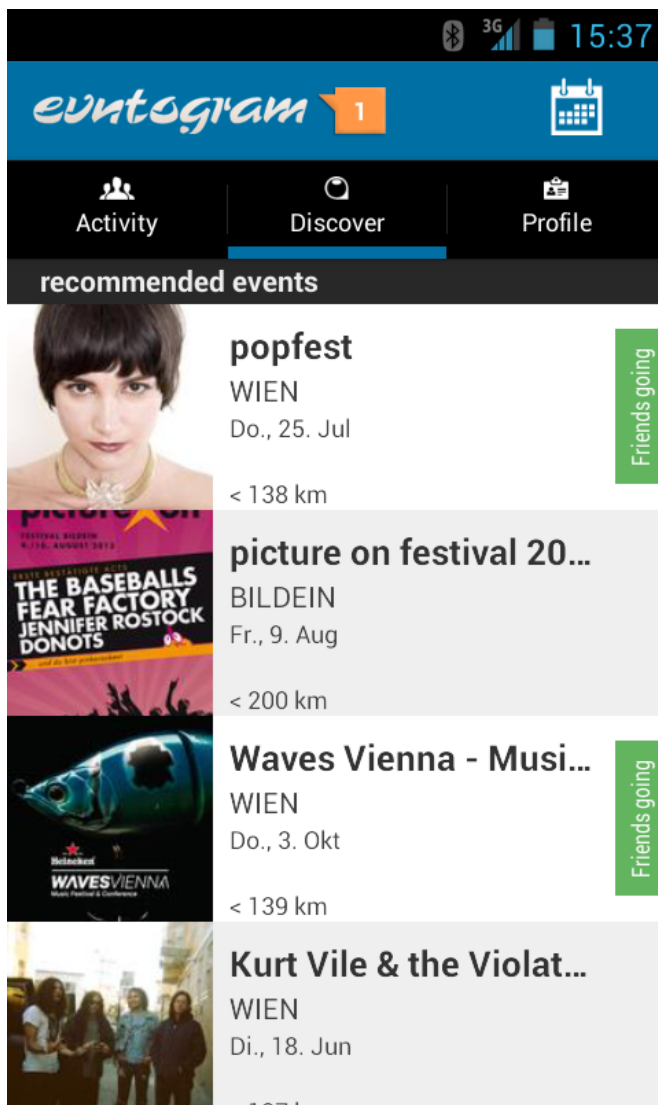


Fig. 7. List of recommended Music Events delivered by the EVNTOGRAM Android Client

for an event was calculated.

VIII. CONCLUSION

In this work, we propose a general MapReduce-based approach, as well as software architecture for the implementation of context-aware recommendation systems. The approach as well as the framework offers high flexibility according to the definition and configuration of new context dimensions in form of impact functions, which influence the recommendation of items for given users. The framework is domain-free, which means that this approach can be implemented and adapted for different application domains. The context-aware recommendation of items of all kind, ranging from products in e-commerce to activities and services in sport and fun will get much attention in future software development. A customization of a domain-specific recommendation engine on top of our proposed approach could be implemented

with reduced development effort, as it is mainly reduced to a simple selection of context dimensions. We think that a general framework for designing and implementing such recommendation systems for different application domains is of great importance. The next steps within our work will be to gather empirical feedback from the community within the given use-case of recommending music related events and to improve the degree of transparency for the recommendation process.

REFERENCES

- [1] W. Beer, W. Hargassner, S. Herramhof, and C. Derwein, "General framework for context-aware recommendation of social events," in *Proceedings of the Second International Conference on Intelligent Systems and Applications (INTELLI)*. IARIA, 2013, pp. 141–146.
- [2] R. M. Bell, Y. Koren, and C. Volinsky, "The bellkor solution to the netflix prize," accessed: 31/01/2013. [Online]. Available: <http://www2.research.att.com/volinsky/netflix/ProgressPrize2007BellKorSolution.pdf>
- [3] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in *Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on Mobile Computing Systems and Applications*. IEEE, 1994, pp. 85–90.
- [4] P. J. Brown, J. D. Bovey, and X. Chen, "Context-aware applications: From the laboratory to the marketplace," *IEEE Personal Communication*, vol. 4, no. 5, pp. 58–64, Oct. 1997.
- [5] A. Dey and G. Abowd, "Towards a better understanding of context and context-awareness," in *CHI 2000 Workshop on The What, Who, Where, When, and How of Context-Awareness*, 2000.
- [6] W. Beer, V. Christian, A. Ferscha, and L. Mehrmann, "Modeling context-aware behavior by interpreted eca rules," *Euro-Par 2003 Parallel Processing*, pp. 1064–1073, 2003.
- [7] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.
- [8] G. Adomavicius, A. Tuzhilin, and R. Zheng, "Request: A query language for customizing recommendations," *Info. Sys. Research*, vol. 22, no. 1, pp. 99–117, Mar. 2011.
- [9] W. Beer and A. Wagner, "Smart books: adding context-awareness and interaction to electronic books," in *Proceedings of the 9th International Conference on Advances in Mobile Computing and Multimedia (MoMM)*. New York, NY, USA: ACM, 2011, pp. 218–222.
- [10] V.-G. Blanca, G.-S. Gabriel, and P.-M. Rafael, "Effects of relevant contextual features in the performance of a restaurant recommender system," in *In RecSys11: Workshop on Context Aware Recommender Systems (CARS-2011)*, 2011.
- [11] Y. Koren, "Collaborative filtering with temporal dynamics," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*. New York, NY, USA: ACM, 2009, pp. 447–456.
- [12] N. Sundaresan, "Recommender systems at the long tail," in *Proceedings of the fifth ACM conference on Recommender systems (RecSys)*. New York, NY, USA: ACM, 2011, pp. 1–6.
- [13] R. Sinha and K. Swearingen, "The role of transparency in recommender systems," in *Extended Abstracts on Human factors in Computing Systems (CHI EA)*. New York, NY, USA: ACM, 2002, pp. 830–831.

An Error Detection Strategy for Improving Web Accessibility for Older Adults

Alfred Taylor, Sr., Les Miller, Sree Nilakanta, Jeffry Sander, Saayan Mitra, Anurag Sharda, and Bachar Chama
Iowa State University

Ames, Iowa

ataylor43@gmail.com, lmiller@iastate.edu, nilakant@iastate.edu, jsander@partners.org

Abstract— The ability to use the Internet can provide an important contribution to an older adult's quality of life. Communication via email with family, friends and service providers have become critical factors for improving ones ability to cope with modern society as individual's age. The problem is that as users age, natural physical and cognitive impairments make it more difficult for them to use the required technology. Setting user preferences in browsers has been suggested as a means of dealing with these limitations. However, questions exist as to the effectiveness of older adult's ability to use self-assessment as a means of setting preferences. The present study investigates the use of error detection as a means of improving web access amongst older adults. Specifically, an error detection strategy has been developed and compared to self assessment, written tests, and observation as a means of identifying the impairments of older Internet users.

Keywords—web usability, error detection, older adults

I. INTRODUCTION

The normal aging process can trigger decreases in acuity of vision and cognition as well as physical impairments, which impact Web usability, particularly if Web designs are not user-friendly [8,9]. Web design issues related to fonts, colors, graphics, background images, navigation, and search mechanisms might prevent older adult users from taking full advantage of online health resources. Web designs may also present reading comprehension barriers for the older adult, due to limitations in visual acuity, cognitive abilities, and education levels, all of which may have a consequence on Web usage [6]. Savago, Sloan, and Blat [24] see cognition problems as the largest barrier to computer use by older adults. Cognitive issues place older adults at greater risk for falling for Internet scams [7].

The implication of better health care for older adults is a longer life [12]. It is crucial for them to be able to keep abreast of new developments in health care that can enhance their life [2]. Older adults who have access to the Internet have access to a large number of ways to find information to help them achieve this goal. It also provides an excellent means of interacting with family members, which also has implications for positive health outcomes. Xie [33] has noted that the use of the Internet has changed the

relationship between older adults and their health providers. Many older adults have problems performing daily tasks because of restricted mobility, lack of transportation, inconvenience, and fear of crime [4]. Home computers with an Internet connection can provide access to information and services, and can also be used to manage banking and Internet shopping tasks. This can be of critical importance. Sum et al. [27] have found that Internet usage is an important factor in older adult's ability to deal with loneliness. Kwon and Noh [14] have also noted that using the Internet can help reduce boredom. Hogeboom et al. [10] have shown that Internet usage is important in strengthening older adults' social networking. Madden [16] notes "Social networking among Internet users ages 50 and older almost doubled -- from 22% to 42% during the 2009-2010 period". Uphold [30] found that older adults are the most likely to seek information on the Web.

Sloan et al. [26] and Mazur et al. [17] have noted that the full impact of how older adults use Internet tools is still an open question. Salces et al. [22] provide a detailed discussion of the effect of aging. Interestingly, studies [13,15,16,31] have found that the average age of computer users has continued to increase. Berry [3] suggests that the variation in the older users should be taken into account as methods for older adults are considered.

Xie and Bugg [32] have found that good training for older adults in public libraries can improve effective computer usage. However, training does not help with the user's limitations. One means of dealing with these issues is for website designers and Human Computer Interaction (HCI) professionals to provide services for better interfaces and Websites in order for older adults to effectively use computers and obtain information resources on-line [12]. While such an approach is viable, it restricts use of the Internet to sites that have been designed with such limitations in mind.

To provide a more general solution to the problem, it requires taking the limitations of the users into consideration. Hanson and Crayne [8] make use of user preferences. However, older adults are not as successful as younger users in making use of the preference options provided by the browser [8, 9]. To bridge this gap, we propose the use of an error

detection strategy to determine the level of impairment of user. The proposed error detection strategy is compared against self assessment, written test, and one on one observation

The information on the user's level of expected performance is stored in a user profile and then is used by the server to modify the Web page the user is working with. The use of user profiles is not new, but it has proved to be a useful construct in our tests. Jacko et al. [11] used visual profiles in their work. We ultimately see a user profile as containing information such as font size, cognitive level (reasoning, speed of processing and locus of control), and mobility/motor measures. The present work looks at the development of profile types based on self assessment, written tests, and observation and our error detection strategy and focuses on vision and motor skill issues.

The key question addressed in this research is "Does error detection produce a profile of the older adults' accessibility performance that is comparable to profiles based on self assessment, written tests, or observation?" To test these questions, we constructed a server based software platform that makes use of a user profile to modify Web pages. The platform was used in a user study of 25 older adults to examine our research question.

The next section briefly looks at some related work. Section III looks at the software platform used in the study and Section IV describes the experiment design. Section V presents the results of our study and Section VI provides a discussion of the results. Section VII provides concluding remarks and thoughts about future work.

II. RELATED WORK

Several approaches have been proposed to assist older adults. A number of special purpose devices have been developed to aid users with motor and vision issues [18]. Mice and specialized keyboards are available to aid older adults [4,5] with declining motor skills. Special viewers to magnify the symbols on the screen are available as well. While such devices are very useful, they tend to increase the cost of computer systems and restrict where older adults can access the Internet. Moreover, older adults are less likely to be aware of special hardware [8]. Hanson et al. [9] have looked at voice browsing as a compromise. Pervasive computing [21] has somewhat similar goals, but does not focus on the user's limitations. Sato et al. [23] have built on voice augmentation techniques to help older adults.

The IBM research group at Watson proposed a Web solution approach to Web accessibility for older

adults [9]. They employed a server to reformat Web pages based upon user preference and capability. This Gateway software was built on WebSphere Transcoding Publisher. IBM WebSphere® Transcoding Publisher Version 4.0 for Multi-platforms is server-based software that dynamically translates Web content and applications into multiple markup languages and optimizes it for delivery to mobile devices, such as mobile phones and handheld computers. The software adapts, reformats, and filters content, tailoring it for display on pervasive devices, giving companies better access to mobile employees, business partners and customers.

Another group of IBM's researchers, Nagao et al. [20] continued research in the area of content adaptation through transcoding for accessibility for users with specific needs. Content adaptation is a type of transcoding that considers a user's environment devices, network bandwidth, profiles, and so on [20]. In their implementation, an annotation server annotated and changed the document contents in accordance with profiles.

More recently, Hanson and Crayne [8] have started to stress the use of user defined preferences at the browser level. However, older adults are not as successful as younger users in making use of the preference options provided by the browser [8]. Mobasher et al. [19] explored mining usage data for Web personalization. The rules are used to adapt the content served to a particular user. Collaborative filtering systems, such as Firefly [25], typically take explicit information in the form of user ratings or preferences, and through a correlation engine, return information that is predicted to closely match the users' preferences.

The next section looks at the software platform used in the experiment.

III. SOFTWARE PLATFORM

We start by briefly overviewing the design of the Error Detection System software used to support our study. The purpose of the Error Detection System is to measure efficiency in an unobtrusive and dynamic Internet browsing environment capable of evaluating user performance and providing dynamic modification of Web pages according to individual user profiles.

A. Overview of Platform

The Error Detection System's platform provides the mechanism to collect relevant information (errors) to gain insight into some of the problems that older adults encounter while browsing the Internet. The

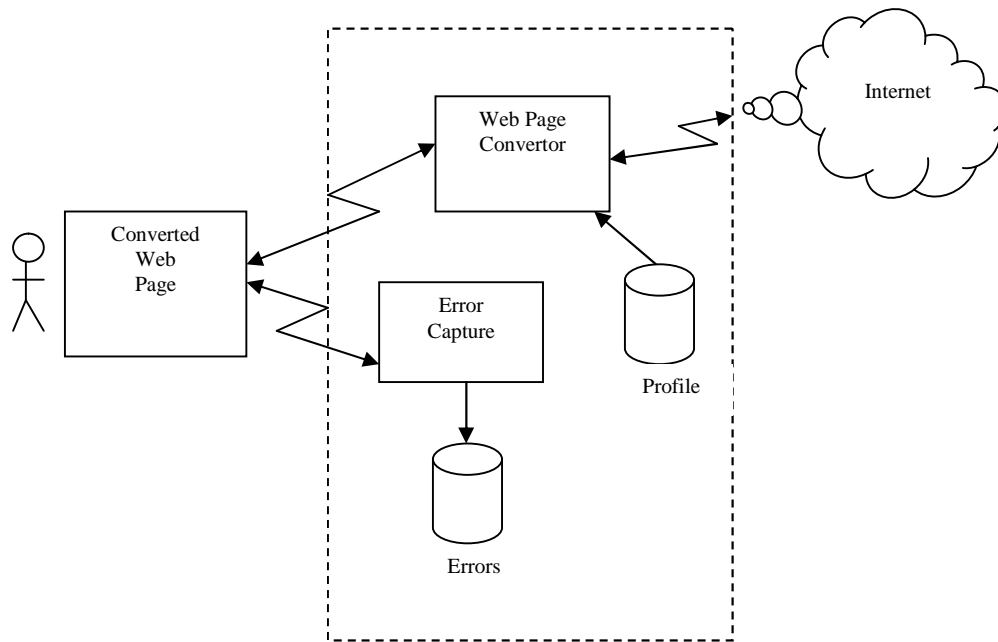


Figure 1. Block diagram of the software platform used by participants.

software uses a user profile for each individual to assist participants surfing the Internet, while tracking their error rates. The result is that users are able to get page modification without having to make manual adjustment with their browser.

The architecture was designed to capture errors related to vision and motor skills. In this study, user performances were compared based on profiles created by self-assessment, written tests, observation, and our error detection software. Details on the design and generation of the four types of profiles are given in Section IV. Here it is sufficient to note that the user profile variables (font size, motor skills) represent the perceived limitations of the owner of the profile.

User Name
Date
Time
Profile Level
Font Size
Motor Skills Score

Figure 2. Profile parameters.

A block diagram of the basic software platform used in the experiment is shown in Figure 1. URL's for the Web pages requested by the user are sent to the *Web Page Convertor* module. The module downloads the requested Web page and modifies it based on the contents of the user's profile (Figure 2). The strategy in converting webpages is to increase

the font size to the value given in the user profile, if necessary. The motor skills and mobility scores are used to enlarge the area of interactive screen features, like buttons and text boxes, using JavaScript. We used the phrase *sensitive area* to represent this enlarged area. When the user clicks inside the sensitive area, the feature is activated (e.g., button is clicked).

When a web page is requested, the page is retrieved and loaded onto the server. The web page is then parsed for font parameters, tags, size and area. If the values are the same or larger than required by the profile the fonts or button size are not changed. Each converted Web page is supplemented with code (JavaScript) to support error detection and collection. The errors made by the user are captured on the webpage and sent to the server level and stored to support analysis. An example of an error would be a user clicking near the sensitive area of a button but not close enough to activate the button.

B. Profiles

Four approaches to constructing the user profile have been used in this work [1,28]. Each profile has the same composition and structure. They only differed in the way values in the fields were generated. More details on the four profiles are given in Section IV. Here we provide a brief overview.

1) *Self Assessment Profile*: Participants were asked to make a self-evaluation of their preferences

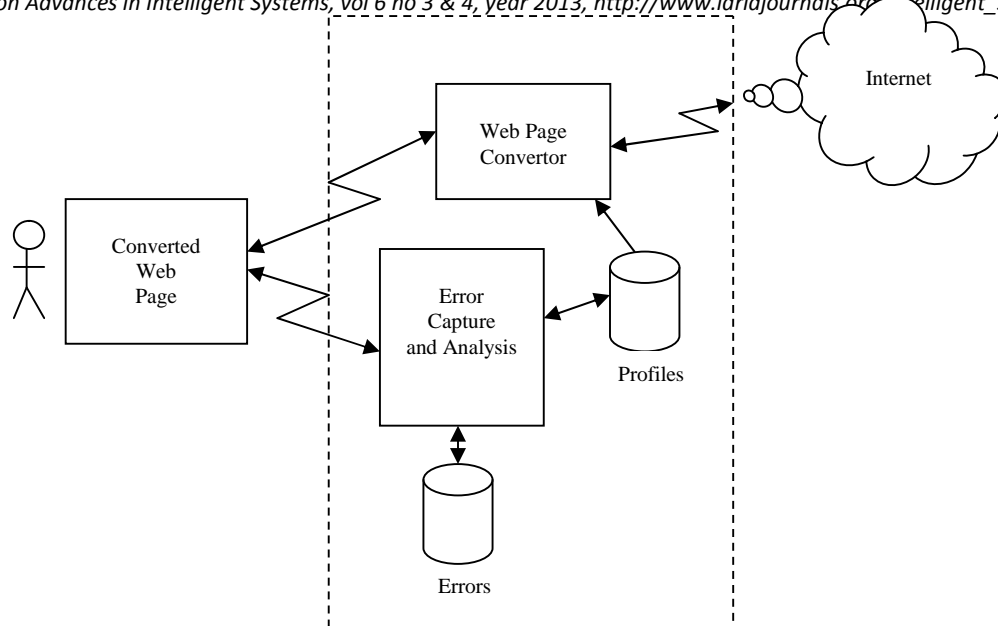


Figure 3. Block diagram of the system platform for supporting development and use of an error based user profile.

for font size and rating their motor skills. The self-evaluation was the participants' perspective of their own ability and their assessment of what they thought was the optimal settings for them to perform effectively. The self-assessment was used to generate a user profile for each user. In the remainder of the manuscript we use the phrase *self profile* to represent this profile.

2) *Written Test Profile*: Each participant was given a written test to check their limitations with respect to vision and motor skills. The exam results were recorded and used to create a user profile. This profile is called the *test profile* in the rest of the manuscript.

3) *Observation Profile*: Participants were observed while they were completing a task set. To ensure consistency, an observation evaluation form was developed with the help of a psychologist (Jennifer Margrett). Moreover, all observations were conducted by the same reviewer to reduce any observer biases. The phrase *observation profile* is used to represent the use of this profile in the rest of the manuscript.

4) *Error Detection Profile*: The errors generated by a user as he/she worked their way through a set of tasks were captured and used to generate a user profile. Figure 3 shows the block diagram of the modified platform. During the tasks, the errors that are captured are analyzed and used to modify the participant's current profile when the number of errors is above a preset threshold. The process continues until the system sees no additional change

in the performance of the user. The number of clicks around a button or link is counted to determine the motor skill score. The font size is set in part based on the user's performance based on giving the users different font sizes to work with. The strategy behind error detection was to provide a transparent tool to measure errors and change the current profile to reduce the user errors. We use the phrase *error profile* in the remainder of the manuscript when referring to the use of this profile.

X-list:

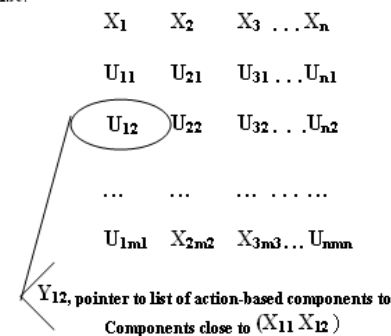


Figure 4. Basic index structure used for Error Detection.

C. Determining Errors

A key aspect of the software instrument is the successful determination of when an error has occurred and how to assign the error type. In the

Participant Information:

Name:

Email Address:

Sex:

☐ Male

☐ Female

Do you have computer experience?

☐ Yes

Submit

Useful links

[Project description](#)

[Participant questionnaire](#)

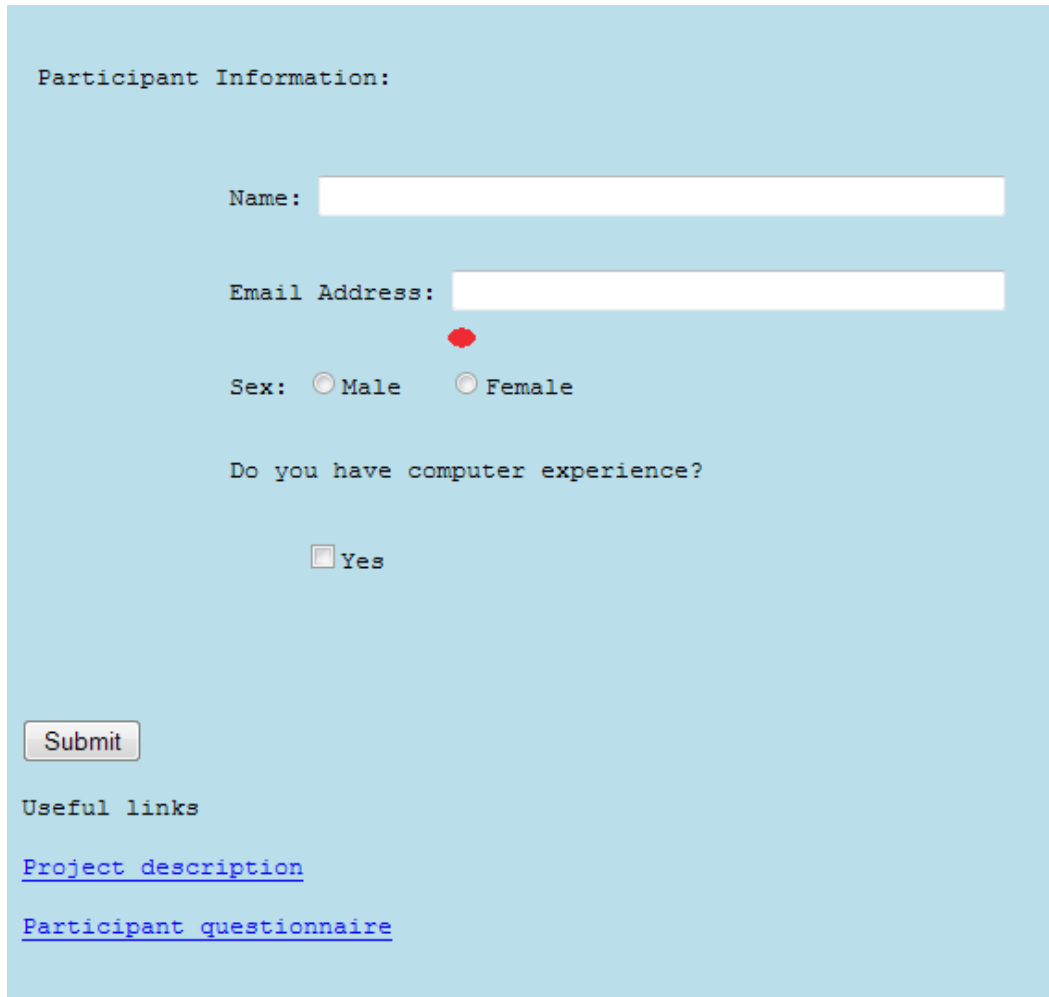
Figure 5. Simple webpage example.

| | | | | | | | | | |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| X-list | 8 | 71 | 163 | 168 | 177 | 209 | 240 | 255 | 538 |
| | | | | | | | | | |
| | 385 | 385 | 195 | 92 | 289 | 491 | 144 | 195 | 92 |
| | 409 | 409 | 210 | 113 | 305 | 506 | 164 | 210 | 113 |
| | 457 | | 289 | | | | 195 | | 144 |
| Y-lists | 475 | | 305 | | | | 210 | | 164 |
| | 491 | | 457 | | | | | | |
| | 506 | | 475 | | | | | | |

Figure 6. Idealized version of the index for the simple webpage from Figure 5.

present work, we have used two error types, namely, motor skills errors and vision errors.

To detect errors, we see the screen real estate as being broken into two disjoint regions, namely, a



Participant Information:

Name:

Email Address:

Sex: ☐ Male ☐ Female

Do you have computer experience?

☐ Yes

Useful links

[Project description](#)

[Participant questionnaire](#)

Figure 7. Mouse clicks shown as oval.

sensitive region and a non-sensitive region. The *sensitive region* of the screen is defined as the portion of the screen where an action is initiated whenever a mouse click occurs within its boundary. For example, if a user clicks on the sensitive area around a web link, the browser action is to transfer the user to the web page indicated by the web link. Similarly, the browser takes actions when a user clicks on a button, a textbox, radio button or any other action-based HTML component.

The *non-sensitive region* of the screen is the region of the screen where a mouse click does not cause an action to occur. The non-sensitive portion of the screen can be made up of empty space or screen components that do not generate actions (such as labels, images or text that are not defined by HTML as web links).

D. Index

To support the detection of mouse click errors, we developed a screen real estate index designed to index the active components on a web page [29]. The structure of the index (shown in Figure 4) makes use of an x-list (the x-values of the set of points on the web page that define the location of the action-based components). For each x_i entry we have a list of u objects, where each u object consists of a y value and a pointer to the list of action-based components that are within a threshold t distance from the (x, y) point indicated by the u entry. Consider the simple webpage shown in Figure 5. An idealized version of the index for the simple webpage from Figure 5 is shown in Figure 6.

To detect errors the index described in the previous subsection is used whenever the user clicks

Participant Information:

Name:

Email Address:

Sex: ☐ Male ☐ Female

Do you have computer experience?

☐ Yes

Useful links

[Project description](#)

[Participant questionnaire](#)

Figure 8. Expanded sensitive areas for textboxes.

in the non-sensitive region of the screen. Suppose a mouse click occurs at location (x^1, y^1) in the non-sensitive region. The x^1 value is used to search the x-list of the index to locate the two x values that bound x^1 (note that x^1 can not be in the x-list or it would not have occurred in the non-sensitive region). Once we have found the two x values (say x_1 and x_2) that bound x^1 , we examine the u-lists for x_1 and x_2 to find the y values that bound y^1 in each list. We can then use the components that are linked to the u-list entries to determine if our mouse click at (x^1, y^1) is sufficiently close to one of the components to label it an error. We define *sufficiently close* to mean that (x^1, y^1) is within a distance t (a system defined threshold) from one or more action-based components. A simple example using the webpage

from Figure 5 and the index from Figure 6 is given in Figure 7.

In Figure 7, the mouse click is represented by the red oval. The location of the mouse click is at (245, 179). The x-value (245) falls between 240 and 255 in the x-list of the index. The y-value (179) falls between 164 and 195 in the y-list for 240 and is less than 195 in the y-list for 255. For the points we have

| Point | Components within $t = 40$ |
|--------------|----------------------------|
| (240, 164) → | 1) Email text box |
| | 2) Female radio button |
| (240, 195) → | 1) Email text box |
| | 2) Female radio button |

Testing the distance from the mouse clicks to the two components, we find that the mouse click is

Table I. Participant Demographics.

| | |
|--------|-------|
| Male | 11 |
| Female | 14 |
| Age | M=77 |
| Range | 62-97 |

closer to the email text box. We assign the error type as a text box error. Since mouse click errors can either be motor skill or vision errors, we first look to extend the sensitive area around the components of the type found to be sufficiently close to the mouse click (incase of ties, the sensitive area around all tied component types are inspected). The sensitive areas are investigated to determine if they can be expanded without causing the sensitive areas of action-based components to overlap. If no overlap is found, the error is considered to be a motor skill error; otherwise, we label it as a vision error. The errors are logged. Figure 8 illustrates what we mean by expanding the sensitive area around a component type. The figure shows the expanded areas around the textboxes as a red rectangle. Since the expanded sensitive areas around the textboxes do not overlap, we assume that the error is a motor skills error and we log the error. Note at this point we are only interested in classifying the error and no actual expansion of the sensitive area during this action. Details on how the user profile will be modified are given in Section IV.

IV. EXPERIMENT DESIGN

To study the effectiveness of the error detection approach, we compared it to the traditional methods: Self-assessment, written test and observation.

A. Participants

Twenty-five participants were recruited for the study. They were comparable in health, and received comparable treatment throughout the study. The background and characteristics of the participants who completed the study were similar to those reported in other studies of usability for older adults. No significant differences in demographic characteristics or baseline performance were observed between the participants who completed the study (N=25). The participants were scheduled individually for each step of the study.

The study was conducted at a retirement community, which provides services for independent living, assisted-living, and nursing care residents. The sample size was 25. There were 24 independent living and one assisted-living participant. The

sample of 25 participants was randomly selected from a pool of volunteers. Table I shows the demographics of the participants. All of the participants met the preconditions of being over the age of 60 years and not having any severe physical impairment such as blindness or could not use the mouse and/or keyboard. They had to be willing to learn and have the ability to sit at a computer for a 30 to 60 minute session. The participants were not paid.

B. Basic Experiment Description

Participants were placed approximately 25 inches from a 20-inch viewable Dell monitor display screen. Screen resolution was set at 1024×768 pixels, with a 32 bit-color setting. The icon and the target folders sizes were 36.8 mm (diagonal distance) based on the findings from Jacko et al. [11]. To perform the experiment, each participant used an IBM Pentium computer. The operating system was Microsoft Windows XP Professional. The computer used a Digital Subscriber Line (DSL) for Internet access. The computer was housed on a computer desk with an accompanying chair in the retirement community. Each participant was instructed to complete a list of tasks. At the initial meeting, each of the participants was given a letter of consent to read that explained the study. They were told that there would be four tests and their activity would be recorded. At the first meeting each user was asked to give a self-assessment of their limitations and their abilities. At the second, a pen and paper test of vision, motor skills and memory was administered and recorded, then their performance using the profile based on the pen and paper test was recorded. At the third meeting, there was an observational assessment of the participant's abilities based on his/her vision, and motor skills. Afterwards, the participants completed a set of tasks and their performance was captured via the server. Finally, the participant was evaluated using the Error Detection approach. We initially set the participant's profile settings according to the test profile settings for the Error Detection study. The

Participant Name: _____

RSU IRB #1
 Approved Date: 06-04-07
 Expiration Date: 27 August 2007
 Initialed by: _____

Participant Self- Assessment Questionnaire

1. Rate your motor (hand movement) skills ability 1 to 5, (5) being excellent.
☐ 1-Poor ☐ 2-Below average ☐ 3-Average ☐ 4-Above average ☐ 5-Excellent
2. Rate your vision level 1 to 5, (5) being excellent.
☐ 1-Poor ☐ 2-Below average ☐ 3-Average ☐ 4-Above average ☐ 5-Excellent

Figure 9. Participant Self-Assessment Questionnaire fragment.

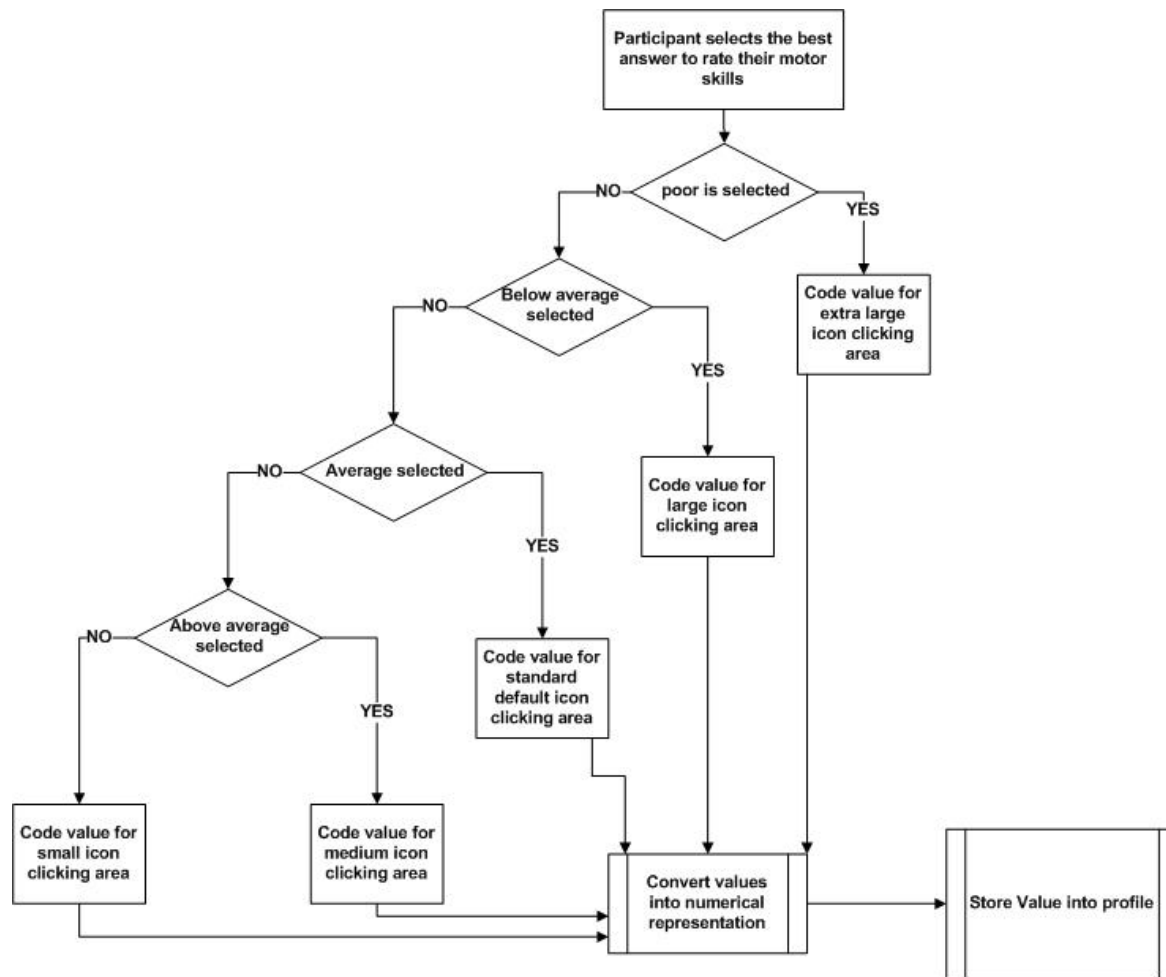


Figure 10. Self-assessment Profile Motor Skill Parameter Process flow.

four experiments were conducted within a twenty-one day period. Each experiment was scheduled for one hour. All of the participants completed the four experiments.

User profiles were created for each participant, based on the self-assessment, test assessment and observational test, respectively. Each participant executed an Internet usability task based upon the self, test and observation profiles. Finally, the participants utilized the Error Detection system to perform a similar task. The Internet usability task was used to capture how many errors the participants made. Each time they missed an icon/button because of low vision, mobility or motor skills it would be captured.

C. Profile Creation

1) *Self-Profile*: Each Participant completed a self-assessment, (Figure 9) of their limitations and computer skills. The responses from the self-assessment were used to define the parameters in the self-profile. The assessment asks the participant how they rated themselves in the context of using the Internet, their vision, motor skills, and their cognitive abilities.

The answers from the self-assessment questionnaire were the basis for the initial coding of the parameters for the self profile. Figure 10 illustrates the process flow for participant answers being translated into coded motor skills parameters for the profile. Figure 11 shows the process flow for creation of vision fonts.

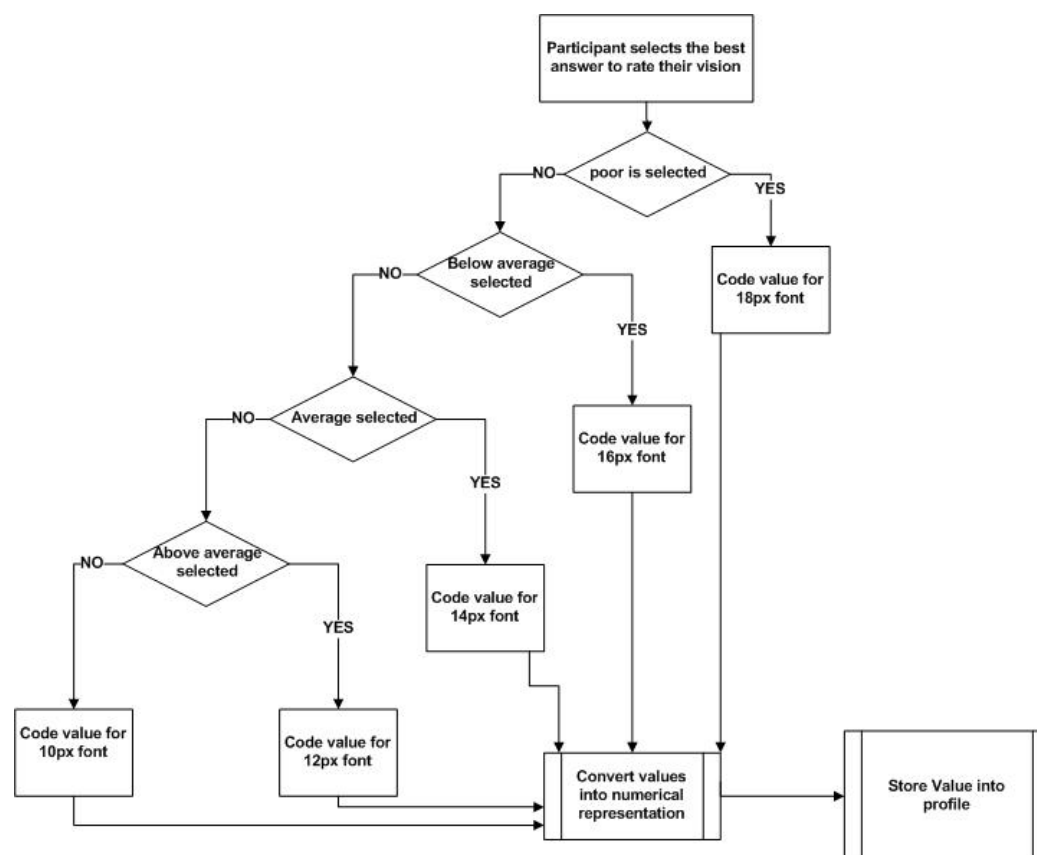


Figure 11. Self-assessment profile vision parameter process flow.

5. Check the word size that you prefer.

☐ Apple ☐ banana ☐ orange ☐ pear ☐ grape

8. Please place a period . in the middle of each box.



9. Please match the symbols with the corresponding number. Place the number in the object.

1. Star 2. Rectangle 3. Square 4. Circle 5. Triangle



Figure 12. Test question fragment used to generate the test profile.

2) **Test Profile:** A paper-based test was used to test the participant's skills to show the capabilities of the participant through the execution of the tasks. Figure 12 shows a fragment of the test used. Figures 13 and 14 show the process flow used to convert the

participant's responses into the test profile values for vision and motor skills, respectively.

3) **Observation Profile:** The observation data was gathered while the participants were administered a set of tasks. The number of instances and actions

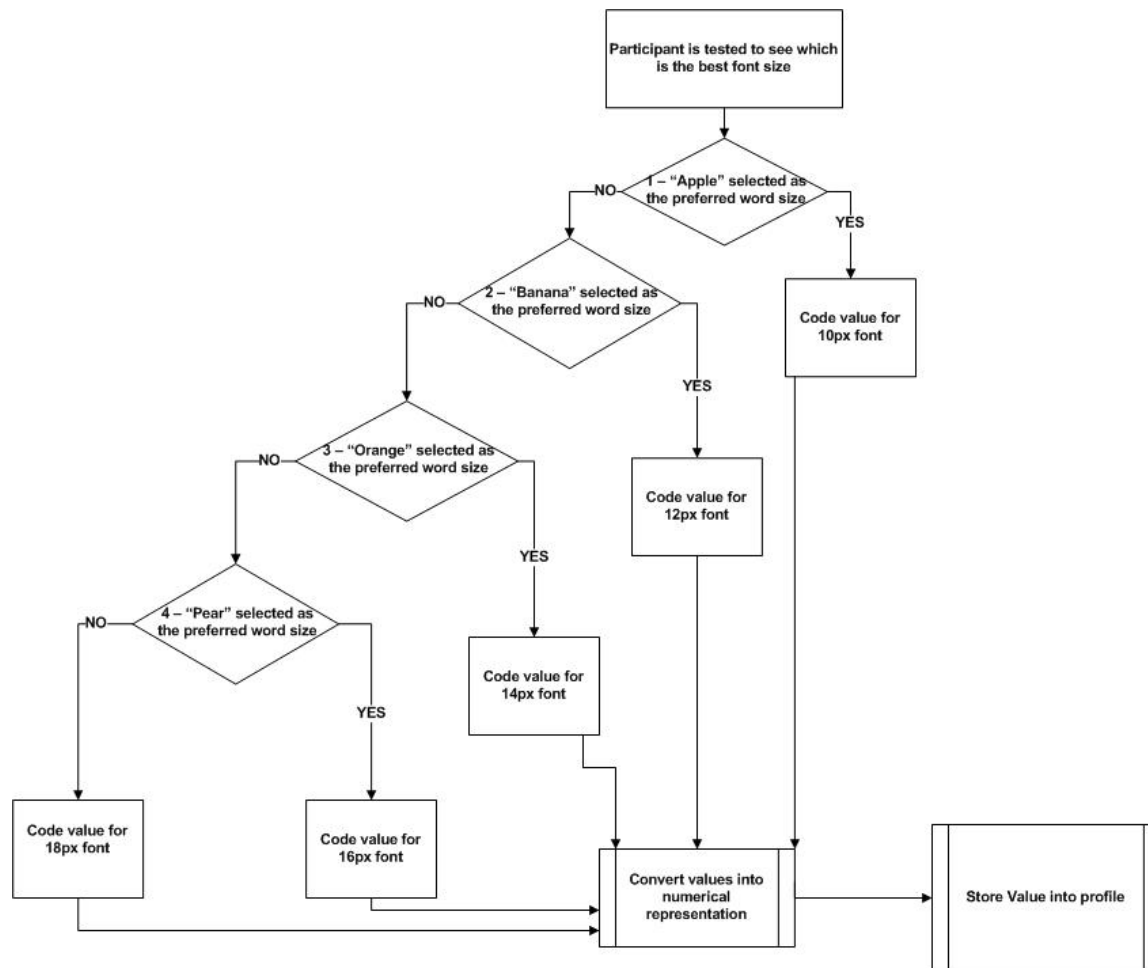


Figure 13. Test profile vision parameter process flow.

while the participants completed the usability task were recorded using an observation form. The results were converted into profile parameter values for font and motor skills. The process flow for converting the values are shown in Figures 15 and 16.

The main objective of the observation of the participant was to collect observable behavior of the participant surfing the web. The usability observation evaluation form was used to collect and record varied characteristics of the interaction of the participant, such as accuracy in moving and clicking the mouse, traversing through a web page, asking questions, talking out loud, and how efficiently they were accomplishing the tasks. The observational behavioral measures were used to evaluate and score specific behaviors that the participant displayed:

- *Screen response*: The ability to respond to prompt and icon presented on Webpage is the second item on the set measuring responsiveness and effect of the Webpage.

- *Type*: The item measures skills such as typing, vision and dexterity.
- *Visually Scanning*: This measure assesses participant's skill in application of material read and analyzes instructional skill. The ability to understand unfamiliar printed words.
- *Non-verbal*: This measure assesses the participant's facial kinesics to capture any computer anxiety or frustration that would be otherwise undetectable.
- *Body Language*: This measure, evaluates the participant's communication using body movements or gestures in the performance of task assigned.
- *Questions*: This measure assesses participant's skill and how well he or she performed the usability task with or without asking questions.
- *Talk aloud*: This measure, gauges the participant's cognitive processing and reasoning skills in applying analysis skills.

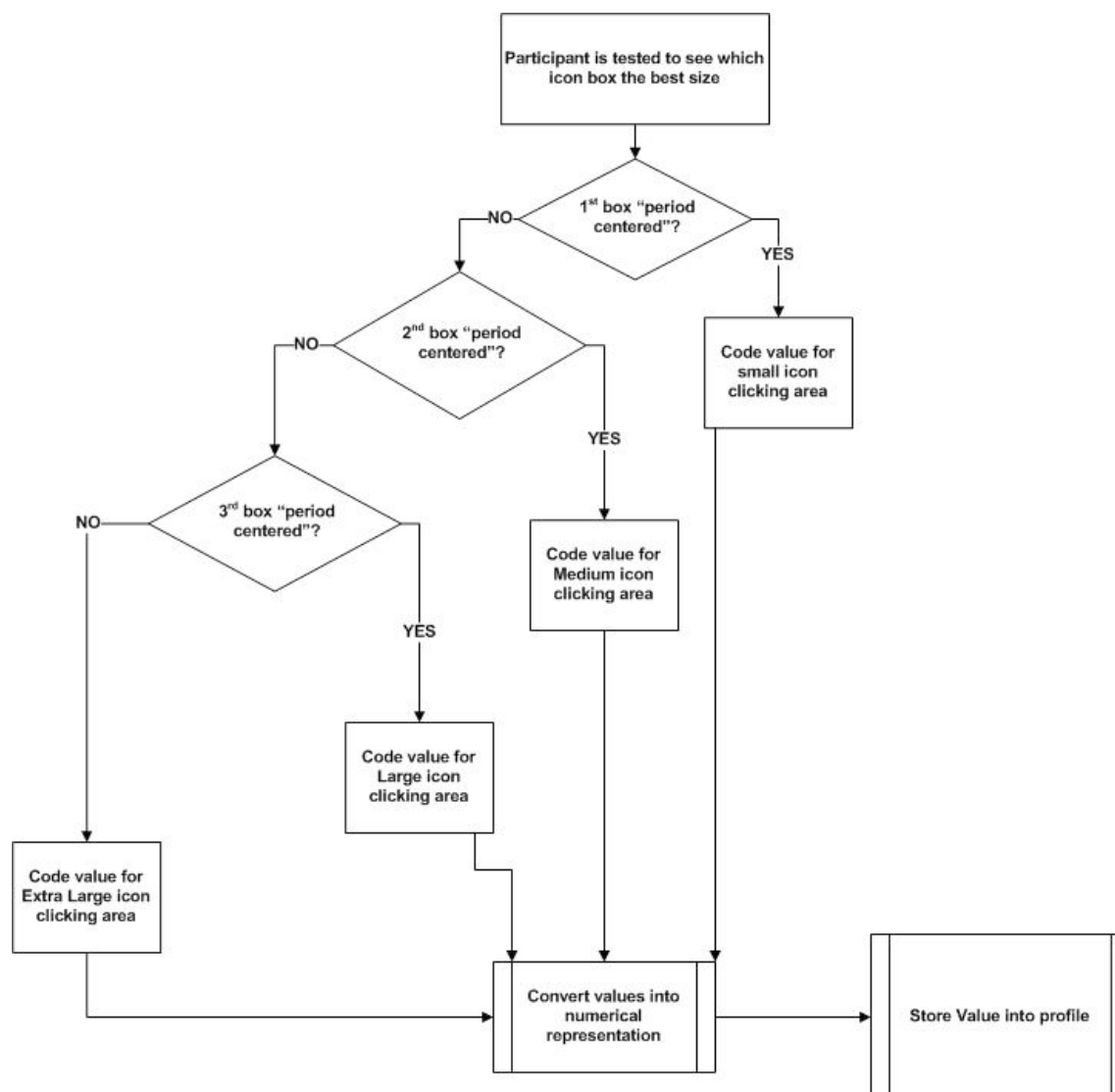


Figure 14. Test profile motor skills parameter process flow.

The results from the Usability Observation Evaluation Form were the basis for the initial coding of the parameters for the Observation profile. Figures 15 and 16 show the process of translating the observation into the font size and the motor skills parameters for the observation profile, respectively.

4) *Error Profile:* Participant data were collected through the system, when participants were working on a set of tasks. The system captured the (vision and motor skill) errors of the user and stored the information in a database maintained on a server. The errors were then used to develop a profile (collection of preferences) of usage for the participant.

The process of surfing and modifying the Web page of the participant was predicated on the reading

a requested Web page and transforming the Web page for the participant based upon parameters that were captured within the error profile of default values. The system keeps track of the errors that are made as the participants worked their way through the task. The system automatically updated the profile based on the errors made. When the number of vision errors from regular screen operations and those obtained during the periods where the system changes font size increases above a preset system threshold, the font size parameter is increased. In a similar fashion, an increase in motor skill errors is used to raise the motor skill parameter value. The process for error profile creation is shown in Figure 17.

The next section looks at the discussion and the results of the study.

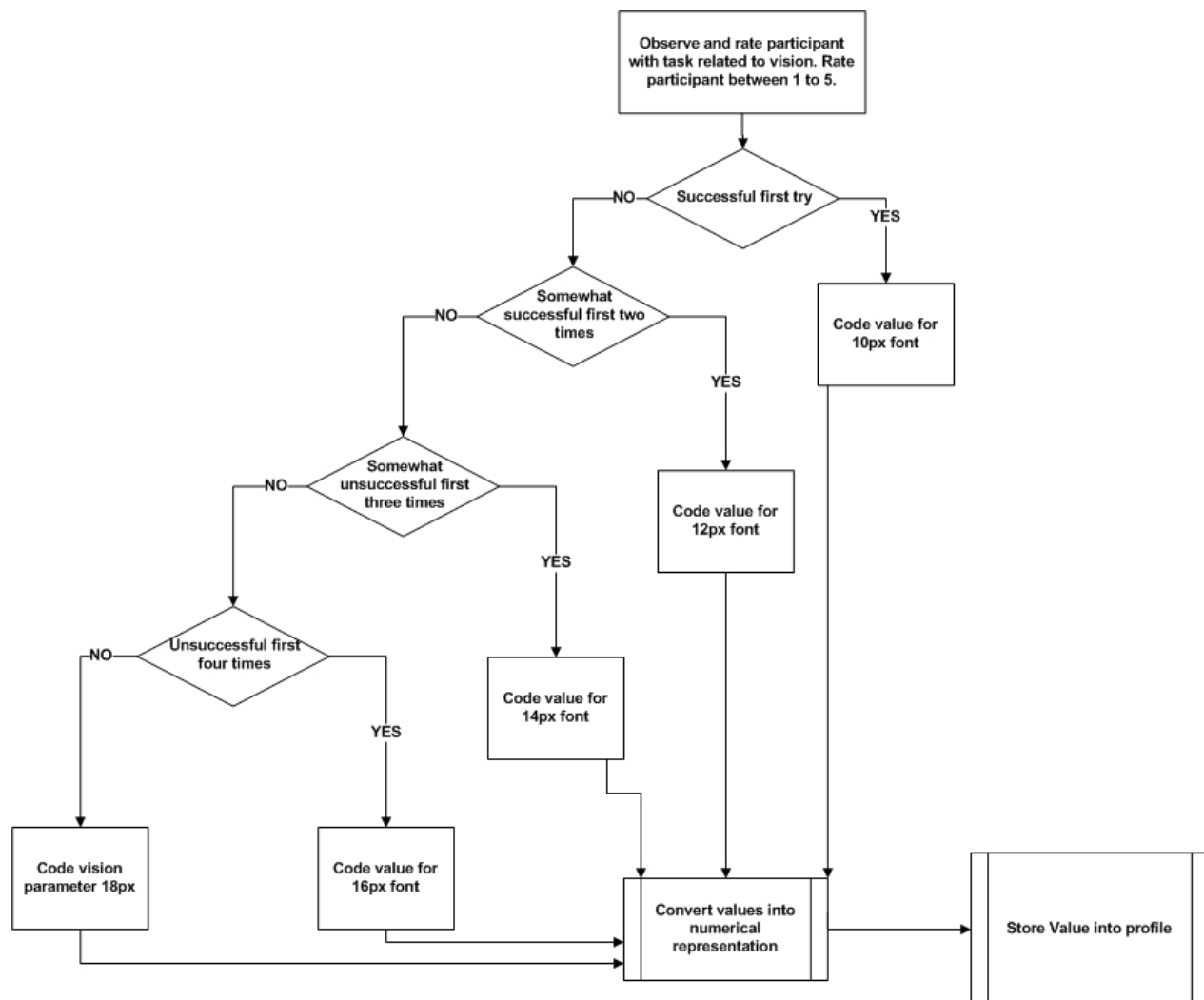


Figure 15. Observation Profile Vision Parameter Process.

V. RESULTS

Performance in the experiment was measured based on the number of errors that participants made while completing the task set. Errors were chosen over time due to our belief that the critical issue for the older adults was successful navigation rather than speed of performance. Table II shows the mean and standard deviation of the errors made for the cases where the Web pages were converted using a profile based on self assessment, written tests, observation, and on error detection.

To consider the key question, “Does error detection produce a profile of the older adults’ accessibility performance that is comparable to profiles based on self assessment, written tests, or

observation?”, we looked at a series of four hypothesis and we used the paired samples t-test to test the individual hypothesis.

Hypothesis 1: Testing provides superior results (with respect to the number of errors a participant makes) to asking older adults for a self assessment of their limitations when using the Internet.

Hypothesis 2: One on one observation provides superior results (with respect to the number of errors a participant makes) to giving older adults written tests to determine their limitations when using the Internet.

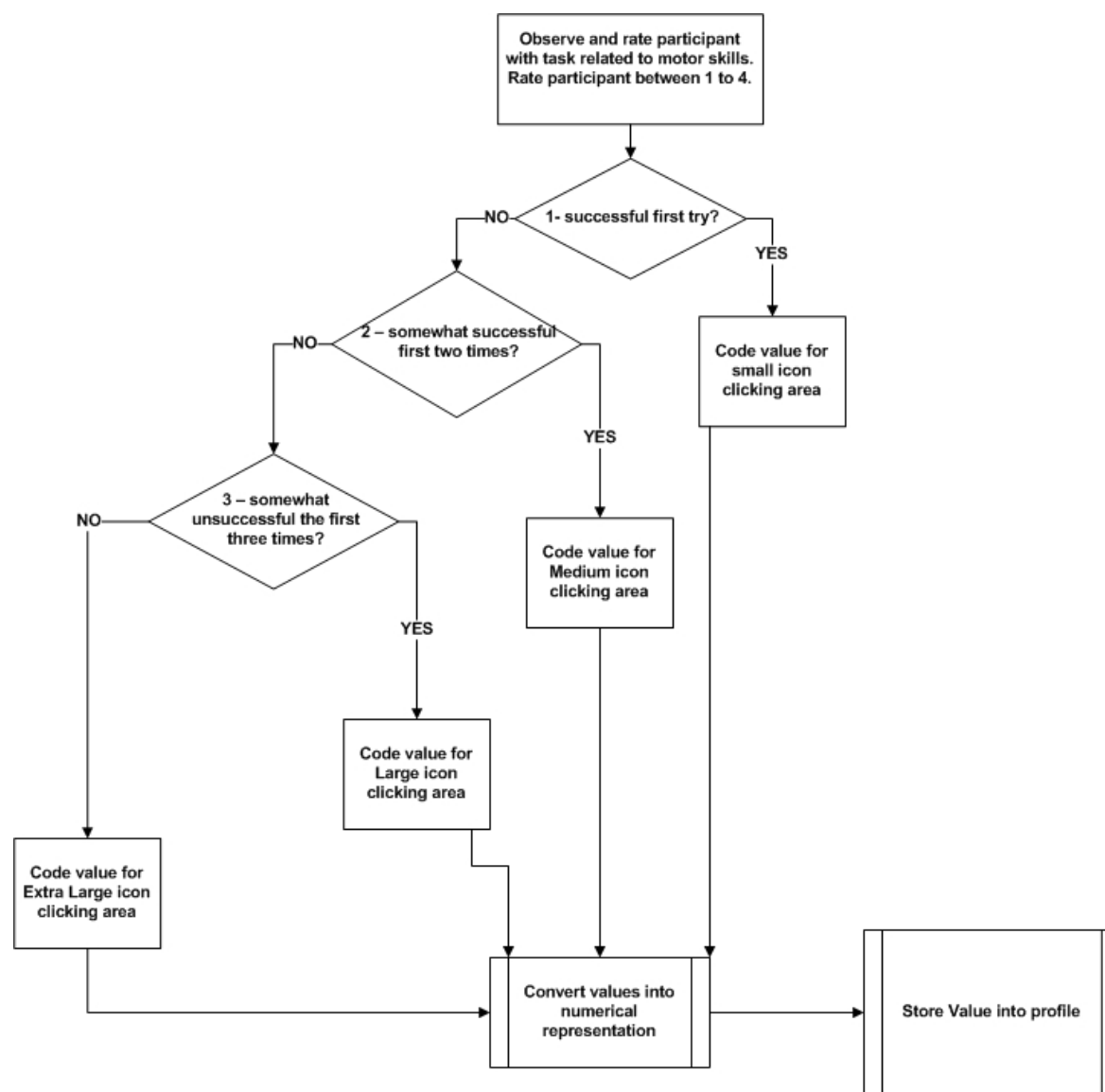


Figure 16. Observation profile motor skills parameter process flow.

Table II. Error means and standard deviations for the 4 approaches tested in the study.

| | Self Assessment | Written Test | Observation | Error Detection |
|------|-----------------|--------------|-------------|-----------------|
| Mean | 57.8 | 11.20 | 7.12 | 6.80 |
| S.D. | 13.952 | 4.003 | 3.621 | 4.010 |

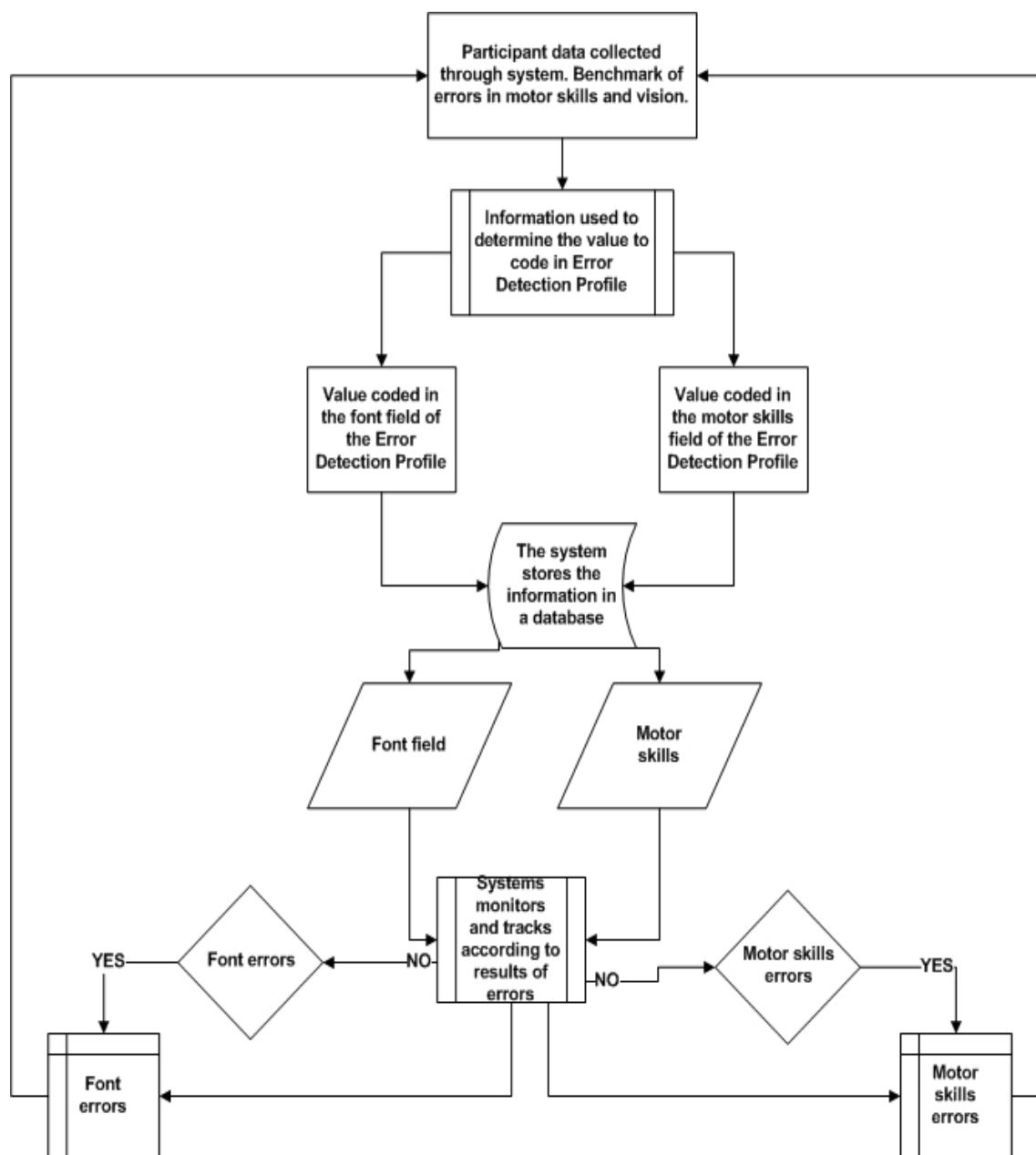


Figure 17. Error detection profile for font size and motor skills parameters process.

Hypothesis 3: One on one observation provides superior results (with respect to the number of errors a participant makes) to asking older adults for a self assessment of their limitations when using the Internet.

Hypothesis 4: Using the error detection strategy described in this paper provides comparable results (with respect to the number of errors a participant

makes) with one on one observation when determining the limitations of older adults when using the Internet. The next section looks at the discussion of the study.

VI. DISCUSSION

Looking at Hypothesis 1, we see from Table III that there is a rather low correlation (0.469) between the error rates of the two methods. The first row in Table IV shows a t value of 18.551 with a significance of 0.0

Table III. Paired Samples Correlation.

| | N | Correlation | Sig. |
|---|----|-------------|-------|
| Errors made using the testing-based profile & a self-assessment-based profile | 25 | 0.469 | 0.018 |
| Errors made using the testing-based profile & a observation-based profile | 25 | 0.627 | 0.001 |
| Errors made using the observation-based profile & a self-assessment-based profile | 25 | 0.385 | 0.057 |
| Errors made using Observation profile & errors using Error detection | 25 | 0.963 | 0.000 |

Table IV. Paired Samples t-test: self assessment, written tests, and observation error detection.

| | Paired Differences | | | t | df | Sig. (2-tailed) |
|--|--------------------|--------|-----------------|--------|----|-----------------|
| | Mean | S. D. | Std. Error Mean | | | |
| Errors made using Self-assessment – Errors made using Testing | 46.600 | 12.560 | 2.512 | 18.551 | 24 | 0.000 |
| Errors made using Testing – Errors made using Observation | 4.080 | 3.366 | 0.673 | 6.061 | 24 | 0.000 |
| Errors made using Self Assessment – Errors made using Observation | 50.680 | 12.996 | 2.599 | 19.498 | 24 | 0.000 |
| Errors made using Observation profile – errors using Error detection | 0.320 | 1.108 | 0.673 | 1.445 | 24 | 0.161 |

indicating that there is a significant difference between the two samples. From these results, we can see that written tests were far superior to self assessment in our study.

The second row of Table IV shows that for Hypothesis 2, we are able to say that observation provides a better estimate of an older adult than what we were able to get from written tests. The significance of the t value (6.061) is again 0.0 showing that there is a significant difference. Table III still shows a low level of correlation between the error rates of the two methods of creating profiles.

Hypothesis 3 compared the error rates of the observation against self assessment. As in the case of Hypothesis 1, self assessment performs very poorly when compared to observation. Again the significance of the t value (19.498) in row 3 of Table IV is 0.0, indicating that there is a significant difference between the two samples. Table III indicates a very low correlation between the error rates of the two methods.

The results of the first three hypotheses indicate that one on one observation is statistically superior to either written exams or self assessment. The problem is that one on one observation is extremely expensive and does not lend itself to periodic retests of older adults.

Hypothesis 4 looks at the comparison of observation to the proposed error detection approach. Table III shows a high correlation (0.963) between the error rates of the two approaches. More important, the result shown in row four of Table IV indicates that the two tailed significance is greater than 0.05 and there was not a significant difference in our study between creating the user profile based on observation or on error detection. The importance of this result comes from the work required to create the profiles. Observation is very labor intensive and is difficult to use with very many users. The use of error detection, on the other hand, places the burden on the computer system. It can be applied to any number of users and is not site specific. Moreover, targeting the accessibility skills of an older adult is not a static target. The physical and cognitive limitations of older adults tend to increase as they age. The dynamic nature of using an error detection strategy allows the profile contents to dynamically change as the user changes.

VII. CONCLUSION

Performance in the experiment was measured based on a user study consisting of 25 older adults was developed and performed to compare the proposed error detection strategy to evaluation strategies based

on self assessment, written tests, and one on one observation. A server based platform was developed for the user study. The platform used a user profile that contained a measurement of the user's impairments for motor skills and vision.

The server converted any Web page that the user requested based on the contents of the user profile. The results of the study were promising. Four hypotheses were tested. The first three compared self-assessment, written tests and one on one observation. The study results indicated that observation was superior with respect to the user error rates. The fourth hypothesis compared one on one observation against the proposed error detection strategy. The study indicated that there was no statistical difference between the means of the results of the observation-based profiles and the results of the error detection-based profiles. This is an interesting result in that doing in depth observations of the potential users is very labor intensive and error detection places the burden on the computer system. Currently, we are looking at the cognitive phase of our project.

ACKNOWLEDGMENT

Al Taylor and Les Miller would like to thank Jennifer Margrett for her help on this project.

REFERENCES

- [1] Taylor Sr., A., L. Miller, S. Nilakanta, J. Sander, S. Mitra, A. Sharda, & B. Charna, "Using Error Detection Strategy for Improving Web Accessibility for Older Adults," *Advances in Computer-Human Interactions (ACHI 2009)*, February 2009, Cancun, Mexico, pp 375-380.
- [2] Becker, S., "A Study of Web Usability for Older Adults Seeking Online Health Resources," *ACM Transactions on Computer-Human Interaction*, vol. 11, no. 4, 2004, pp. 387-406.
- [3] Berry, R. "Older people and the internet: Towards a system map of digital exclusion," London: The International Longevity Centre – UK, June 2011, http://www.ilcuk.org.uk/index.php/publications/publication_details/older_people_and_the_internet.
- [4] Bickmore, T.W., L. Caruso, and K. Clough-Gorr., "Acceptance and usability of a relational agent interface by urban older adults," *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI '05, pp. 1212-15.
- [5] Czaja, S., & Lee, C., "Designing Computer Systems for Older Adults," *The Human Computer Interaction Handbook* 21, 2003, pp. 413-427.
- [6] Czaja, S., & Lee, C., "The Internet and older adults: design challenges and opportunities," *Communication, Technology and Aging: Opportunities and Challenges for the Future*, 2001, pp. 60-80.
- [7] Garg, V., Camp, L. J., Lorenzen-Huber, L. M., and K. Connelly, "Designing Risk Communication for Older Adults," *Symposium on Usable Privacy and Security*

- (SOUPS). 2011, July 20–22, 2011, Pittsburgh, PA USA. Pp. 1-10.
- [8] Hanson, Vicki, John T. Richards, and Chin Chin Lee, "Web Access for Older Adults: Voice Browsing?," *HCI* (5), 2007, pp. 904-913.
- [9] Hanson, Vicki, "The user experience: designs and adaptations," *Proceedings of the international cross-disciplinary workshop on Web accessibility (W4A)*, 2004, New York City, New York, pp. 1 – 11.
- [10] Hogeboom, David, McDermott, Robert, Perrin, Karen, Osman, Hana, and Bethany Bell-Ellison, "Internet Use and Social Networking Among Middle Aged and Older Adults," *Educational Gerontology*, vol. 36, no. 2, 2010, pp. 93-111.
- [11] Jacko, J., Rosa, R., Scott, U., Pappas, C., & M. Dixon, "Visual impairment: The use of visual profiles in evaluations of icon use in Computer-based tasks," *International Journal of Human-Computer Interaction*, vol. 12, no. 1, 2001, pp. 151-164.
- [12] Jonsson, I., M. Zajicek, H. Harris, and C. Nass, "Thank you, I did not see that: in-car speech based information systems for older adults," *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. *CHI '05*, 2005, pp. 1953-56.
- [13] Kahana, Eva and Kahana, Boaz and Lovegreen, Loren and Cronin, Cory and Holger Plaff, "The Proactive Aged: New Players in the Web World," *Proceedings of the ACM WebSci'11*, June 14-17 2011, Koblenz, Germany. p. 1-2.
- [14] Kwon, Wi-Suk and Mijeong Noh, "The influence of prior experience and age on mature consumers' perceptions and intentions of internet apparel shopping," *Journal of Fashion Marketing and Management*, vol. 14, no. 3, 2010, pp. 335 – 349.
- [15] Lee, B. (2012). "Cyber Behaviors among Seniors," pp 233-241, *Encyclopedia of Cyber Behavior*. DOI: 10.4018/978-1-4666-0315-8.ch020.
- [16] Madden, M., "Older Adults and Social Media," *Pew Internet and American Life Project*, <http://pewinternet.org/Reports/2010/Older-Adults-and-Social-Media.aspx> (accessed 31 May 2012).
- [17] Mazur, E., Signorella, M., and M. Hough, "Older Adults and Their Internet Behaviors," pp 608-619. *Encyclopedia of Cyber Behavior*. 2012, DOI: 10.4018/978-1-4666-0315-8.ch052.
- [18] McCullagh, P., Nugent, C., Zheng, H., Burns, W., Davies, R., Black, N., Wright, P., Hawley, M., Eccleston, C., Mawson, S., and G. Mountain, "Knowledge Transfer for a Technology Based Intervention in the Self-Management of Long-term Conditions," *The AAATE Workshop: Assistive Technology – Technology Transfer*, Sheffield, UK, Oct 4-5, 2010.
- [19] Mobasher, B., Dai, H., Luo, T., and M. Nakagawa, "Improving the Effectiveness of Collaborative Filtering on Anonymous Web Usage Data," *Proc. of the IJCAI 2001 Workshop on Intelligent Techniques for Web Personalization(ITWP01)*, 2001, <http://maya.cs.depaul.edu/~mobasher/papers/itwp01.pdf>.
- [20] Nagao, K., Shirai, Y., and K. Squire, "Semantic Annotation and Transcoding: Making Older adults Content More Accessible," *Older adults Engineering*, 1070(986X), pp. 69-81.
- [21] Saba, D., and A. Mukherjee, "Pervasive Computing: A Paradigm for the 21st Century," *IEEE Computer Society*, 0018-9162 2003, pp. 25-30.
- [22] Salces, Fausto J. Sainz, Michael Baskett, David Llewellyn-Jones and David England, "Ambient Interfaces for Elderly People at Home," *Ambient Intelligence in Everyday Life*. Springer. Berlin, 2006 , pp. 256-284.
- [23] Sato, D., Kobayashi, M., Takagi, H., Asakawa, C., and J. Tanaka, "How voice augmentation supports elderly web users," *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, Oct 24-26, 2011. Dundee, Scotland. pp. 155-162 doi: 10.1145/2049536.2049565.
- [24] Sayago, S, Sloan, D. and J. Blat, "Everyday use of computer-mediated communication tools and its evolution over time: An ethnographical study with older people," *Interacting with Computers*, vol. 23, no. 5, pp. 543-554. doi: 10.1016/j.intcom.2011.06.001
- [25] Shardanand, U., and P. Maes, "Social information filtering: algorithms for automating word of mouth," *CHI*, 1995, pp. 210-217.
- [26] Sloan, D., Atkinson, M., Machin, C., and Y. Li, "The potential of adaptive interfaces as an accessibility aid for older web users," *Proceedings of 2010 International Cross-Disciplinary Conference on Web Accessibility (W4A)* Raleigh, US, April 2010, Article 35.
- [27] Sum, Shima, R. Mark Mathews, Ian Hughes and Andrew Campbell, "Internet Use and Loneliness in Older Adults," *CyberPsychology and Behavior*, vol. 11, no. 2, 2008, pp. 208-211.
- [28] Taylor Sr., Alfred, Les Miller, and Sree Nilakanta, "Evaluating Older Adults to Improve Web Accessibility," *ISCA 24th International Conference on Computers and Their Applications*, New Orleans, La. 2009, pp. 13-18.
- [29] Taylor Sr., Alfred, Karthik Ramalingam, and Les Miller. "Using a Screen Real Estate Index to Support Adaptive User Interfaces," *Twenty-seventh International Conference on Computers and Their Applications*, Las Vegas, NV, March 12-14, 2012, pp. 89-94.
- [30] Uphold, C. R., "Transitional Care for Older Adults: The Need for New Approaches to Support Family Caregivers," *J. Gerontol Geriatric Res* vol. 1, no.2, pp. 107, 2012, <http://dx.doi.org/10.4172/jggr.1000e107>.
- [31] Wagner, Nicole, Khaled Hassanein, and Milena Head, "Computer Use by Older Adults: a Multidisciplinary Review," *Computers in Human Behavior*, vol. 26, 2010, pp. 870-882.
- [32] Xie, Bo and Julie M. Bugg, "Public library computer training for older adults to access high-quality Internet health information," *Library & Information Science Research*, vol. 31, no. 3. 2009, Pages 155-162.
- [33] Bo Xie, "Older Adults' Health Information Wants in the Internet Age: Implications for Patient-Provider Relationships," *Journal of Health Communication*. vol. 14, no. 6, 2009. pp 510-524.



www.iariajournals.org

International Journal On Advances in Intelligent Systems

✦ ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, ENERGY, COLLA, IMMM, INTELLI, SMART, DATA ANALYTICS

✦ issn: 1942-2679

International Journal On Advances in Internet Technology

✦ ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING, MOBILITY, WEB

✦ issn: 1942-2652

International Journal On Advances in Life Sciences

✦ eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO, SOTICS, GLOBAL HEALTH

✦ issn: 1942-2660

International Journal On Advances in Networks and Services

✦ ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION, VEHICULAR, INNOV

✦ issn: 1942-2644

International Journal On Advances in Security

✦ ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS

✦ issn: 1942-2636

International Journal On Advances in Software

✦ ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, IMMM, MOBILITY, VEHICULAR, DATA ANALYTICS

✦ issn: 1942-2628

International Journal On Advances in Systems and Measurements

✦ ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL, INFOCOMP

✦ issn: 1942-261x

International Journal On Advances in Telecommunications

✦ AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA, COCORA, PESARO, INNOV

✦ issn: 1942-2601