UBICOMM 2014

The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies

ISBN: 978-1-61208-353-7

August 24 - 28, 2014

Rome, Italy

UBICOMM 2014 Editors

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain

Christoph Steup, FIN - OvGU, Germany

Sönke Knoch, German Research Center for Artificial Intelligence (DFKI GmbH), Germany

UBICOMM 2014

Forward

The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2014), held on August 24 - 28, 2014 - Rome, Italy, was a multi-track event covering a large spectrum of topics related to developments that operate in the intersection of mobile and ubiquitous technologies on the one hand, and educational settings in open, distance and corporate learning on the other, including learning theories, applications, and systems.

The rapid advances in ubiquitous technologies make fruition of more than 35 years of research in distributed computing systems, and more than two decades of mobile computing. The ubiquity vision is becoming a reality. Hardware and software components evolved to deliver functionality under failure-prone environments with limited resources. The advent of web services and the progress on wearable devices, ambient components, user-generated content, mobile communications, and new business models generated new applications and services. The conference made a bridge between issues with software and hardware challenges through mobile communications.

The goal of UBICOMM 2014 was to bring together researchers from the academia and practitioners from the industry in order to address fundamentals of ubiquitous systems and the new applications related to them. The conference provided a forum where researchers were able to present recent research results and new research problems and directions related to them.

Advances in web services technologies along with their integration into mobility, online and new business models provide a technical infrastructure that enables the progress of mobile services and applications. These include dynamic and on-demand service, context-aware services, and mobile web services. While driving new business models and new online services, particular techniques must be developed for web service composition, web service-driven system design methodology, creation of web services, and on-demand web services.

As mobile and ubiquitous computing becomes a reality, more formal and informal learning will take pace out of the confines of the traditional classroom. Two trends converge to make this possible; increasingly powerful cell phones and PDAs, and improved access to wireless broadband. At the same time, due to the increasing complexity, modern learners will need tools that operate in an intuitive manner and are flexibly integrated in the surrounding learning environment.

Educational services will become more customized and personalized, and more frequently subjected to changes. Learning and teaching are now becoming less tied to physical locations,

co-located members of a group, and co-presence in time. Learning and teaching increasingly take place in fluid combinations of virtual and "real" contexts, and fluid combinations of presence in time, space and participation in community. To the learner full access and abundance in communicative opportunities and information retrieval represents new challenges and affordances. Consequently, the educational challenges are numerous in the intersection of technology development, curriculum development, content development and educational infrastructure.

We take here the opportunity to warmly thank all the members of the UBICOMM 2014 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to UBICOMM 2014. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the UBICOMM 2014 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success. We gratefully appreciate to the technical program committee co-chairs that contributed to identify the appropriate groups to submit contributions.

We hope the UBICOMM 2014 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in ubiquitous systems and related applications.

We hope Rome provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

UBICOMM 2014 Chairs:

UBICOMM Advisory Committee
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Zary Segal, UMBC, USA
Yoshiaki Taniguchi, Kindai University, Japan
Ruay-Shiung Chang, National Dong Hwa University, Taiwan
Ann Gordon-Ross, University of Florida, USA
Dominique Genoud, Business Information Systems Institute/HES-SO Valais, Switzerland
Andreas Merentitis, AGT International, Germany
Timothy Arndt, Cleveland State University, USA
Tewfiq El Maliki, Geneva University of Applied Sciences, Switzerland

Yasihisa Takizawa, Kansai University, Japan
Jens Haupert, German Research Center for Artificial Intelligence (DFKI), Germany

UBICOMM Industry/Research Chairs
Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Carlo Mastroianni, CNR, Italy
Michele Ruta, Politecnico di Bari, Italy
Jose Manuel Cantera Fonseca, Telefonica Investigacion y Desarrollo, Spain
Yulin Ding, Defence Science & Technology Organization Edinburgh, Australia
Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Qiong Liu, FX Palo Alto Laboratory, USA
Hamed Ketabdar, Deutsche Telekom Laboratories / TU Berlin, Germany
Inas Khayal, Masdar Institute of Science and Technology - Abu Dhabi, United Arab Emirates
Donnie H. Kim, Intel, USA
Cornel Klein, Siemens AG/Corporate Research and Technologies - Münich, Germany
Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA
Ian Oliver, Nokia, Finland
Serena Pastore, INAF- Astronomical Observatory of Padova, Italy
Jyrki T.J. Penttinen, Finesstel Ltd, Finland
Jorge Pereira, European Comission, Belgium
Miroslav Velev, Aries Design Automation, USA
Yu Zheng, Microsoft, USA
Christoph Steup, FIN - OvGU, Germany

UBICOMM Publicity Chairs
Roland Dutzler, University of Technology Graz, Austria
Raul Igual, University of Zaragoza, Spain
Andre Dietrich, Otto-von-Guericke-University Magdeburg, Germany
Rebekah Hunter, University of Ulster, UK
Francesco Fiamberti, University of Milano-Bicocca, Italy
Sönke Knoch, German Research Center for Artificial Intelligence (DFKI GmbH), Germany
Adriana Wilde, University of Southampton, UK

UBICOMM 2014

Committee

UBICOMM Advisory Committee

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Zary Segal, UMBC, USA
Yoshiaki Taniguchi, Kindai University, Japan
Ruay-Shiung Chang, National Dong Hwa University, Taiwan
Ann Gordon-Ross, University of Florida, USA
Dominique Genoud, Business Information Systems Institute/HES-SO Valais, Switzerland
Andreas Merentitis, AGT International, Germany
Timothy Arndt, Cleveland State University, USA
Tewfiq El Maliki, Geneva University of Applied Sciences, Switzerland
Yasihisa Takizawa, Kansai University, Japan
Jens Haupert, German Research Center for Artificial Intelligence (DFKI), Germany

UBICOMM Industry/Research Chairs

Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Carlo Mastroianni, CNR, Italy
Michele Ruta, Politecnico di Bari, Italy
Jose Manuel Cantera Fonseca, Telefonica Investigacion y Desarrollo, Spain
Yulin Ding, Defence Science & Technology Organization Edinburgh, Australia
Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Qiong Liu, FX Palo Alto Laboratory, USA
Hamed Ketabdar, Deutsche Telekom Laboratories / TU Berlin, Germany
Inas Khayal, Masdar Institute of Science and Technology - Abu Dhabi, United Arab Emirates
Donnie H. Kim, Intel, USA
Cornel Klein, Siemens AG/Corporate Research and Technologies - Münich, Germany
Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA
Ian Oliver, Nokia, Finland
Serena Pastore, INAF- Astronomical Observatory of Padova, Italy
Jyrki T.J. Penttinen, Finesstel Ltd, Finland
Jorge Pereira, European Comission, Belgium
Miroslav Velev, Aries Design Automation, USA
Yu Zheng, Microsoft, USA
Christoph Steup, FIN - OvGU, Germany

UBICOMM Publicity Chairs

Roland Dutzler, University of Technology Graz, Austria
Raul Igual, University of Zaragoza, Spain

Andre Dietrich, Otto-von-Guericke-University Magdeburg, Germany
Rebekah Hunter, University of Ulster, UK
Francesco Fiamberti, University of Milano-Bicocca, Italy
Sönke Knoch, German Research Center for Artificial Intelligence (DFKI GmbH), Germany
Adriana Wilde, University of Southampton, UK

UBICOMM 2014 Technical Progam Committee

Afrand Agah, West Chester University of Pennsylvania, USA
Rui Aguiar, Universidade de Aveiro, Porutgal
Tara Ali-Yahiya, Paris Sud 11 University, France
Mercedes Amor, Universidad de Málaga, Spain
Timothy Arndt, Cleveland State University, USA
Mehran Asadi, Lincoln University, U.S.A.
Zubair Baig, Edith Cowan University, Australia
Sergey Balandin, FRUCT, Finland
Matthias Baldauf, FTW Telecommunications Research Center Vienna, Austria
Michel Banâtre, IRISA - Rennes, France
Felipe Becker Nunes, Federal University of Rio Grande do Sul (UFRGS), Brazil
Aurelio Bermúdez Marin, Universidad de Castilla-La Mancha, Spain
Daniel Bimschas, University of Lübeck, Germany
Carlo Alberto Boano, University of Lübeck, Germany
Bruno Bogaz Zarpelão, State University of Londrina (UEL), Brazil
Jihen Bokri, ENSI (National School of Computer Science), Tunisia
Sergey Boldyrev, Nokia, Finland
Diletta Romana Cacciagrano, University of Camerino, Italy
Jose Manuel Cantera Fonseca, Telefonica Investigacion y Desarrollo, Spain
Juan-Vicente Capella-Hernández, Universidad Politécnica de Valencia, Spain
Rafael Casado, Universidad de Castilla-La Mancha, Spain
Everton Cavalcante, Federal University of Rio Grande do Norte, Brazil
José Cecílio, University of Coimbra, Portugal
Bongsug (Kevin) Chae, Kansas State University, USA
Konstantinos Chatzikokolakis, National and Kapodistrian University of Athens, Greece
Sung-Bae Cho, Yonsei University - Seoul, Korea
Mhammed Chraibi, Al Akhawayn University - Ifrane, Morocco
Michael Collins, Dublin Institute of Technology, Dublin, Ireland
Andre Constantino da Silva, IFSP, Brazil
Kyller Costa Gorgônio, Universidade Federal de Campina Grande, Brazil
Pablo Curiel, DeustoTech - Deusto Institute of Technology, Spain
Gennaro Della Vecchia, ICAR-CNR - Naples, Italy
Steven A. Demurjian, The University of Connecticut, USA
Gianluca Dini, University of Pisa, Italy
Yulin Ding, Defence Science & Technology Organization Edinburgh, Australia
Roland Dodd, Central Queensland University, Australia
Charalampos Doukas, University of the Aegean, Greece
Jörg Dümmler, Technische Universität Chemnitz, Germany
Tewfiq El Maliki, University of Applied Sciences of Geneva, Switzerland
Alireza Esfahani, Instituto de Telecomunicações - Pólo de Aveiro, Portugal

Josu Etxaniz, University of the Basque Country, Spain
Andras Farago, The University of Texas at Dallas - Richardson, USA
Sana Fathalla, Sophia Antipolis / CNRS, France
Ling Feng, Tsinghua University - Bejing, China
Gianluigi Ferrari, University of Parma, Italy
Renato Ferrero, Politecnico di Torino, Italy
George Fiotakis, University of Patras, Greece
Rita Francese, Università degli Studi di Salerno, Italy
Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Franco Frattolillo, University of Sannio, Italy
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Crescenzio Gallo, University of Foggia, Italy
Junbin Gao, Charles Sturt University - Bathurst, Australia
Shang Gao, Zhongnan University of Economics and Law, China
Marisol García Valls, Universidad Carlos III de Madrid, Spain
Dominique Genoud, HES-SO Valais Wallis, Switzerland
Chris Gniady, University of Arizona, USA
Paulo R. L. Gondim, University of Brasília, Brazil
Francisco Javier Gonzalez Cañete, University of Málaga, Spain
Ann Gordon-Ross, University of Florida, USA
George A. Gravvanis, Democritus University of Thrace, Greece
Dominic Greenwood, Whitestein Technologies - Zürich, Switzerland
Markus Gross, ETH Zurich, Switzerland
Fikret Gurgen, Isik University - Istanbul, Turkey
Norihiro Hagita, ATR Intelligent Robotics and Communication Labs, Kyoto, Japan
Jason O. Hallstrom, Clemson University, USA
Jens Haupert, German Research Center for Artificial Intelligence (DFKI), Germany
Arthur Herzog, Technische Universität Darmstadt, Germany
Hiroaki Higaki, Tokyo Denki University, Japan
Sun-Yuan Hsieh, National Cheng Kung University, Taiwan
Shaohan Hu, UIUC, USA
Xiaodi Huang, Charles Sturt University - Albury, Australia
Javier Alexander Hurtado, University of Cauca, Colombia
Raul Igual, University of Zaragoza, Spain
Marko Jaakola, VTT Technical Research Centre of Finland, Finland
Tauseef Jamal, University Lusofona - Lisbon, Portugal
Jongpil Jeong, Sungkyunkwan University, South Korea
Jun-Cheol Jeon, Kumoh National Institute of Technology, Korea
Faouzi Kamoun, Zayed University, UAE
Fazal Wahab Karam, Gandhara Institute of Science and Technology, Pakistan
Rehana Kausar, Queen Mary, University of London, UK
Nobuo Kawaguchi, Nagoya University, Japan
Hamed Ketabdar, Deutsche Telekom Laboratories / TU Berlin, Germany
Subayal Khan, VTT, Finland
Inas Khayal, Masdar Institute of Science and Technology - Abu Dhabi, United Arab Emirates
Donnie H. Kim, Intel, USA
Cornel Klein, Siemens AG/Corporate Research and Technologies - Münich, Germany
Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA

Sönke Knoch, German Research Center for Artificial Intelligence (DFKI), Germany
Eitaro Kohno, Hiroshima City University, Japan
Jakub Konka, University of Strathclyde, UK
Shin'ichi Konomi, University of Tokyo, Japan
Dmitry Korzun, Petrozavodsk State University / Aalto University, Russia / Finland
Natalie Kryvinski, University of Vienna, Austria
Jeffrey Tzu Kwan Valino Koh, National University of Singapore, Singapore
Frédéric Le Mouël, INRIA/INSA Lyon, France
Nicolas Le Sommer, Université de Bretagne Sud - Vannes, France
Juong-Sik Lee, Nokia Research Center, USA
Valderi R. Q. Leithardt, Federal University of Rio Grande do Sul, Brazil
Pierre Leone, University of Geneva, Switzerland
Jianguo Li, Motorola Mobility, USA
Yiming Li, National Chiao Tung University, Taiwan
Jian Liang, Cork Institute of Technology, Ireland
Kai-Wen Lien, Chienkuo Institute University - Changhua, Taiwan
Bo Liu, University of Technology - Sydney, Australia
Damon Shing-Min Liu, National Chung Cheng University, Taiwan
Qiong Liu, FX Palo Alto Laboratory, USA
David Lizcano Casas, Open University of Madrid (UDIMA), Spain
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Jaziel Souza Lobo, Instituto Federal de Sergipe, Brazil
Juan Carlos López, University of Castilla-La Mancha, Spain
Jeferson Luis Rodrigues Souza, University of Lisbon, Portugal
Paul Lukowicz, German Research Center for Artificial Intelligence (DFKI), Germany
Elsa María Macías López, University of Las Palmas de Gran Canaria, Spain
Victor Emmanuilovich Malyshkin, Technical University of Novosibirsk, Russia
Gianfranco Manes, University of Florence, Italy
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Teddy Mantoro, University of Technology Malaysia, Malaysia
Sergio Martín Gutiérrez, UNED-Spanish University for Distance Education, Spain
Oscar Martinez Bonastre, University Polytechnic of Valencia, Spain
Carlo Mastroianni, ICAR-CNR - Rende, Italy
Roseclea Duarte Medina, Universidade Federal De Santa Maria (UFSM), Brazil
Natarajan Meghanathan, Jackson State University, U.S.A.
Andreas Merentitis, AGT Group (R&D) GmbH, Germany
Kathryn Merrick, University of New South Wales & Australian Defence Force Academy, Australia
Elisabeth Métais, CNAM/CEDRIC, France
Markus Meyer, Technische Hochschule Ingolstadt, Germany
Daniela Micucci, University of Milano - Bicocca, Italy
Hugo Miranda, Universidade de Lisboa, Portugal
Moeiz Miraoui, Gafsa University, Tunisia
Claudio Monteiro, Science and Technology of Tocantins, Brazil
Costas Mourlas, University of Athens, Greece
Tatsuo Nakajima, Waseda University, Japan
Wolfgang Narzt, Johannes Kepler University - Linz, Austria
Rui Neves Madeira, New University of Lisbon, Portugal
David T. Nguyen, College of William and Mary, USA

Kazunori Takashio, Keio University, Japan
Yoshiaki Taniguchi, Kindai University, Japan
Adrian Dan Tarniceriu, Ecole Polytechnique Federale de Lausanne, Switzerland
Markus Taumberger, VTT Technical Research Centre of Finland, Finland
Saïd Tazi, LAAS-CNRS, Université de Toulouse / Université Toulouse1, France
Manos Tentzeris, Georgia Institute of Technology, USA
Tsutomu Terada, Kobe University, Japan
Stephanie Teufel, University of Fribourg, Switzerland
Parimala Thulasiraman, University of Manitoba, Canada
Lei Tian, University of Nebraska-Lincoln, USA
Marco Tiloca, University of Pisa, Italy
Chih-Cheng Tseng, National Ilan University, Taiwan
Jean Vareille, Université de Bretagne Occidentale - Brest, France
Dominique Vaufreydaz, INRIA Rhône-Alpes, France
Miroslav Velev, Aries Design Automation, USA
Massimo Villari, Universita' di Messina, Italy
Wei Wei, Xi'an University of Technology, China
Chao-Tung Yang, Tunghai University, Taiwan
Xiao Yu, Aalto University, Finland
Zhiwen Yu, Northwestern Polytechnical University, China
Mehmet Erkan Yüksel, Istanbul University Turkey
Zhifeng Yun, University of Houston, USA
Hao Lan Zhang, Zhejiang University, China
Gang Zhao, National University of Singapore, Singapore
Yu Zheng, Microsoft, USA
Nataša Živić, University of Siegen, Germany

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Handling Positioning Errors in Location-based Services

Philipp Marcus and Claudia Linnhoff-Popien
Institute of Computer Science
Ludwig-Maximilians-Universität München, Germany
Emails: {philipp.marcus, linnhoff}@ifi.lmu.de

***Abstract*—Location–based services (LBS) shall typically only be provided within an authorized zone. This is enforced by location–based access control (LBAC) and affected by occurring positioning errors. Recent research has brought up different approaches for LBAC strategies. However, up to now it is unclear which strategy should be chosen for a given LBS and positioning system under realistic boundary conditions. In detail, the false authorization decision may cause severe additional costs when operating the underlying LBS. Hence, this paper presents a methodology to analyze the expected costs of LBAC strategies under the occurrence of positioning errors. The correlation to the practical costs occurring under realistic conditions is evaluated in an extensive case study. It is shown that in certain situations risk–based authorization is easily mislead by imprecise position error estimates and thus games away its theoretical superiority. In such situations, ignoring estimated errors may even yield lower expected costs of operating the LBS. The presented methodology contributes to finding the most suitable authorization strategy when deploying a LBS. This finally helps to reduce costs occurring from false authorization decisions when operating the LBS.**

Keywords*—*Location–based Access Control; Risk–aware Authorization; Positioning Errors

I. INTRODUCTION

Mobile devices with their integrated positioning capabilities enabled LBS, which nowadays are of substantial importance for most users and service providers [1]. A subclass of LBS are zone–based. Here, a user is granted the authorization to use a LBS if he resides within a predefined authorized zone. For example, imagine a museum that provides an audio guide for mobile devices. The guide's explanations for an exhibition room shall only be audible if the user payed the room's entrance fee and is inside. In order to enforce such authorization semantics, LBAC systems have been developed. Those systems employ the user's current location measurement to decide about the permission to use a given LBS. Such LBS require precise location measurements, which are unfortunately inherently affected by a varying degree of uncertainty, for example due to changing environmental influences or imprecise sensors. Thus, the most precise formalization of the real user location is done by adhering an according probability density function (pdf) as the position estimate, which is finally derived from measurements.

An appropriate LBAC strategy is crucial, as false authorization decisions typically cause costs for false negatives and false positives. Early approaches to LBAC focused on extending the expressiveness of existing access control policies with spatial information. In those approaches, possibly occurring errors are ignored and only the most likely geographical point is used as location estimate. Here, authorization decisions are derived by checking whether the punctual location estimate is contained within the polygon of the authorized zone. Those approaches ignore possibly occurring positioning uncertainty and costs and are thus called risk–ignoring for the rest of this paper. The second category comprises threshold–based approaches to LBAC. Here, the position estimate pdf is employed to derive the probability that the user resides within the authorized zone. A threshold is predefined by the policy designer as the minimum required probability in order to be authorized for using the LBS. Here, costs are not considered and often, the derivation of an appropriate threshold is left unspecified. Even risk–based LBAC strategies were developed. Here, authorization decisions are finally derived based on cost functions and the probability that the user resides within the authorized zone. In detail, such LBAC systems only grant the authorization if the expected costs of a false positive undershoot the expected cost of a false negative. This method is theoretically optimal [2].

However, up to now it is unclear which LBAC strategy should be chosen in practice. Furthermore, it has not been studied how the superiority of the risk–based approach depends on the cost functions for false authorization decisions and the accuracy of the positioning system. In detail, there is urgent need to clarify the effect of statistically imperfect position estimates on the superiority of the risk–based approach. A methodology for choosing that LBAC strategy with the lowest expected cost of false authorization decisions when operating the LBS is non–existent. Nevertheless, such a methodology is urgently needed in order to avoid costly wrong decisions when choosing the authorization strategy for a LBS.

In order to solve this problem, this paper presents a novel approach to analyze the expected costs of false authorization decisions in the forefront of a LBS's deployment. In detail, a methodology for computing the expected costs when operating the risk–ignoring, threshold–based and risk–based LBAC strategy is proposed. Given the error characteristics of the underlying positioning system, this facilitates the computation of expected savings when operating the theoretically optimal risk–based strategy compared to the risk–ignoring strategy. This allows to illustrate the LBS's sensitivity to statistically imprecise position estimates of the positioning system. The need for such analysis is demonstrated for an indoor–LBS in a typical office environment with WiFi fingerprinting as location provider. The theoretical optimality of the risk–based approach is shown to be highly dependent on the LBS's parameters and the quality of the reported position estimates. The rest of the paper is structured as follows: Section II gives a detailed view on related work. Next, Section III presents the theoretical approach to analyze the superiority of the risk–based LBAC strategy. Section IV presents a case study to illustrate the urgent necessity of a detailed analysis before choosing a LBAC strategy. Finally, Section V concludes the paper.

 ISBN: 978-1-61208-353-7

II. Related Work

Location information has been widely used for spatial authorization systems and provisioning of LBS in particular. Often, these methods are called LBAC or spatial access control. One important subset of LBAC systems employs the estimated location in order to determine if the user resides within a prescribed authorized zone. If true, the access right is granted. Important approaches provide sophisticated spatial extensions to role–based access control (RBAC) [3]–[5]. However, even recently published work, for example from Abdunabi et al. uses the reported user location without any consideration of measurement uncertainty [6]. Unfortunately, these approaches do not show the effectiveness of this strategy when applied to realistic and error–prone location providers. It is left unclear, if this strategy is suitable for a given authorized zone and a concrete location provider. Ardagna et al. proposed an approach which employs a confidence value for the probability that the user resides within a predefined authorized zone [7]. If this value overshoots a predefined threshold, the access right is granted. Thresholds are derived empirically based on estimates about the positioning system's sensitivity to changing weather and environmental conditions. Also, the number of sensors is mentioned as an important factor when defining a threshold. However, no concrete methodology to provide any justification of derived thresholds is given. Shin et al. define an authorization policy, which also grants access if the user resides within an authorized zone with a confidence value larger than a predefined threshold [8]. Here, the uncertainty of a position fix is modeled as a probability distribution. The confidence value is derived by integrating the probability distribution over the authorized zone. The thresholds are derived for each access rule individually depending on whether the authorized zone demands high security or is an area of low sensitivity. Again, only very abstract and vague statements about deriving a suitable threshold are mentioned. Krautsevich et al. consider costs when making authorization decisions based on the values of discrete attributes with uncertain values [2]. A threshold–based authorization strategy is shown to be cost–optimal in their scenario for a certain threshold based on cost functions. However, the approach does not show the influence of the uncertainty estimates' quality on the cost–effectiveness of their strategy. Marcus et al. proposed a risk–aware approach for trajectory–based authorization using probabilistic trajectories derived from an adapted particle filter in combination with WiFi fingerprinting [9]. Here, expected costs and the corresponding risk are minimized by adhering assigned cost functions of false authorization decisions and the probability that the user's trajectory satisfies the authorization condition.

Error estimators of positioning systems are needed to operate threshold–based and risk–based LBAC strategies. Basically, given an estimated location μ, an error estimate is a probability distribution $\mathrm{P}(\mu|x)$ describing the likelihood to observe an estimated location of μ when standing at position x in the real–world. Hightower et al. [10] use a commercial infrared badge system and an ultrasound time–of–flight badge system. The infrared error estimates are a static bivariate Gaussian with a covariance matrix $\Sigma = \left(\begin{smallmatrix} 2.3\,m & 0 \\ 0 & 2.3\,m \end{smallmatrix}\right)$. The values are derived from the vendor specification of the system's range. The error estimates of the ultrasound system are retrieved from a lookup table built from previously recorded measurement errors within the test lab. Zandbergen et al. observed GPS errors with a root mean square error of 9-11 m for modern Smartphones, which highly increase in urban areas [11]. Zandbergen et al. also found that the positioning errors of GPS are not perfectly approximated by Gaussian distributions and hence, outliers need to be expected [12]. The error distribution of GPS is found to be approximate to a Rayleigh distribution. Marcus et al. proposed an error estimator for SMARTPOS, an indoor positioning system based on WiFi fingerprinting [13]. Here, the errors were shown to be approximately normally distributed with a mean of 1.2 m.

III. Location–based Authorization Under Positioning Errors

This section first discusses the characteristics of positioning systems and subsequently describes the methodology to theoretically derive the expected costs for each LBAC strategy. The expected savings of uncertainty–aware strategies are compared to the risk–ignoring strategy.

A. Positioning Systems and Error Estimators

The key technology for LBS are positioning systems, which determine the user's location either terminal- or infrastructure based. In outdoor scenarios, GPS emerged as the most important positioning technique, while in indoor scenarios WiFi fingerprinting showed very promising results [12] [13]. In the following, the returned position measurements are called position fixes and noted as μ with $\mu \in \mathbb{R}^2$. In all cases, position fixes are subject to physical perturbations due to interference, reflections, multipath propagation, humidity, imprecise sensors, and so on [1]. Consequently, all position fixes are inherently affected by an error of varying degree as discussed in Section II. The user's ground truth position around the returned position fix μ can be modeled as a probability density function (pdf), which is derived by error estimators.

In all cases, such error estimates are derived from singularities of the underlying measurement by an according error estimator. For example, in case of WiFi fingerprinting, the distribution of the k nearest neighbors around μ was shown to be a good indicator for the occurring error [13]. Given a position measurement with an estimated position of μ, the error estimator finally derives a scale parameter Σ which defines a pdf around μ in order to describe the ground truth location. In case of WiFi Fingerprinting, the scale parameter represents the covariance matrix of the underlying bivariate normal pdf. For the rest of this paper, position fixes μ are reported with a scale parameter Σ of an appropriate error estimate pdf and finally written as (μ, Σ). The larger the estimated scale parameter Σ, the more uncertainty exists with the position fix (μ, Σ). In the following, the accuracy of positioning systems is described by a distribution F_{err} of reported scale parameters Σ. Practical experiments have shown that inverse Gaussian distributions give a very good fit for F_{err} in case of WiFi fingerprinting. However, the distribution of F_{err} is finally needed to analyze the suitability of single LBAC strategies for a concrete scenario and needs to be known for the employed positioning system.

B. Risk–ignoring, threshold– and risk–based LBAC strategies

The task of a LBAC strategy is to derive an authorization decision $\texttt{auth} \in \{\texttt{true}, \texttt{false}\}$ based on a position fix

(μ, Σ). The most basic LBAC strategy is the risk–ignoring authorization strategy as employed in [3]–[6]. Here, only the estimated position μ is considered, when deriving the authorization decision. In detail, this strategy performs a simple point in polygon test to determine, if the estimated position μ is contained within the authorized zone $\mathcal{Z}$:

$$\texttt{auth} \Leftrightarrow \mu \in \mathcal{Z} \tag{1}$$

The main advantage of such systems is the low computational overhead and the efficiency of point in polygon tests. In detail, no error estimate needs to be derived and no complex numerical operations need to be performed.

In order to consider the occurring uncertainty of a position fix (μ, Σ), the threshold–based LBAC strategy derives its authorization decision based on the probability $p_{\mathcal{Z}}$ that the user resides within the authorized zone $\mathcal{Z}$, [2,7,8]. This probability is derived from the estimated position μ and the error estimate Σ and needs to overshoot the threshold:

$$\texttt{auth} \Leftrightarrow p_{\mathcal{Z}}(\mu, \Sigma) > \textit{threshold} \tag{2}$$

A drawback of this strategy is the dependence on reliable error estimators and the computational overhead of computing $p_{\mathcal{Z}}(\mu, \Sigma)$. Furthermore, it is not clear, which threshold makes sense for a given LBS.

A more sophisticated strategy is the risk–based strategy [2] [9]. Here, the expected costs of granting or denying the authorization request are compared. In particular, the authorization request is only granted, if the expected cost c_{fp} of a false positive undershoot the expected cost c_{fn} of a false negative. The expected cost can also be interpreted as the risk of each outcome:

$$\texttt{auth} \Leftrightarrow (1 - p_{\mathcal{Z}}(\mu, \Sigma)) \cdot c_{fp} < p_{\mathcal{Z}}(\mu, \Sigma) \cdot c_{fn} \tag{3}$$

This has the same computational complexity as the threshold–based strategy. Its main advantage is its decision–theoretical optimality given statistically perfect error estimates Σ, which will be discussed in detail later in Section IV.

The risk–based strategy is a real generalization of the threshold–based strategy. Obviously, according to (3), the risk–based strategy depends on the static costs c_{fp} and c_{fn} and the ratio $\frac{c_{fp}}{c_{fn}}$ in detail. For each ratio $\frac{c_{fp}}{c_{fn}}$, there exists exactly one value of *threshold* such that a threshold–based strategy with *threshold* behaves exactly like a risk–based strategy with $\frac{c_{fp}}{c_{fn}}$. This easily follows from resolving (3) to $p_{\mathcal{Z}}$, which finally represents the corresponding value of *threshold*:

$$\textit{threshold} = \frac{\frac{c_{fp}}{c_{fn}}}{1 + \frac{c_{fp}}{c_{fn}}} \tag{4}$$

This correspondence is depicted in Figure 1. Clearly, the higher the cost of a false positive compared to the cost of a false negative, the higher the corresponding value of *threshold*, which converges to 1. The knowledge of this correspondence has two positive effects. On the one hand, existing LBAC policies based on the threshold–based strategy can be assigned comprehensible thresholds given the cost functions of the underlying service or resource to be granted. This correspondence finally allows to derive such values of *threshold* that the threshold–based strategy also yields risk–optimal decisions given the specific cost functions for the LBS to be deployed.

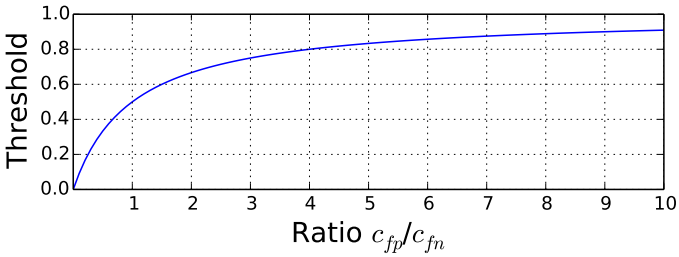


Fig. 1. Correspondence of a risk–based LBAC strategy with a given value of $\frac{c_{fp}}{c_{fn}}$ to a threshold–based LBAC strategy with a predefined value of *threshold*.

On the other hand, analysis of a risk–based strategy deployed for a LBS also allows to assess a threshold–based strategy with a corresponding threshold. However, in real situations, both strategies are subject to the quality of the underlying error estimator. Statistically imperfect error estimators may under- or overestimate the real error, which is a severe weak point compared to the risk–ignoring strategy. In order to finally derive the most suitable LBAC strategy for a given LBS, first, the theoretically expected costs for each LBAC strategy are derived in the next section.

C. The Expected Costs of LBAC strategies

In this section, a methodology is derived to compare the LBAC strategies w.r.t. the expected costs of false authorization decisions. These expected costs are highly dependent on c_{fp} and c_{fn} and the uncertainty of underlying position fixes. In order to decide about the authorization of a requesting user to use a given LBS, the underlying LBAC strategy is provided a position fix (μ, Σ) to check the authorization conditions according to the aforementioned methods. Given this position fix, the statistical distribution of the user's ground truth position x around μ is denoted as the probability density function $F_{\mu,\Sigma}(x)$. Note, the distribution of error estimation scale parameters Σ is denoted as F_{err} in the following.

In the next step, a methodology is presented, which allows to assess the expected costs $E(costs_i)$ for each LBAC strategy $i \in \{\textit{risk–ignoring}, \textit{risk–based}, \textit{threshold–based}\}$. This is achieved by employing a function $costs_i(\mu, \Sigma)$ denoting the expected costs arising from a possibly false decision when authorizing a user with position fix (μ, Σ) with LBAC strategy i. Therefore, for each LBAC strategy i, the expectation of occurring costs is derived w.r.t. the set $\mathcal{R} \subseteq \mathbb{R}^2$ of possible estimated locations $\mu \in \mathcal{R}$ and the distribution of occurring error estimates $F_{err}(\Sigma)$:

$$E(costs_i) = \frac{1}{|\mathcal{R}|} \int_{\mu \in \mathcal{R}} \int_0^{\infty} costs_i(\mu, \Sigma) \cdot F_{err}(\Sigma)\, \mathrm{d}\Sigma\, \mathrm{d}\mu \tag{5}$$

The rest of this section focuses on deriving the function $costs_i(\mu, \Sigma)$ for each LBAC strategy i in order to finally derive its expected overall costs $E(costs_i)$ according to (5).

In order to derive $E(costs_i)$ for each LBAC strategy i, the function $costs_i(\mu, \Sigma)$ needs to be specified first. Here, a prerequisite is the computation of the probability $p_{\mathcal{Z}}(\mu, \Sigma)$ that a user with position fix (μ, Σ) resides within the authorized zone $\mathcal{Z}$:

$$p_{\mathcal{Z}}(\mu, \Sigma) = \int_{x \in \mathcal{Z}} F_{\mu,\Sigma}(x)\ \mathrm{d}\, x \tag{6}$$

In the following, p is used as an abbreviation for $p_{\mathcal{Z}}(\mu, \Sigma)$ if no ambiguities exist. Finally, this allows to derive the expected costs for each LBAC strategy given a position fix (μ, Σ).

1) Deriving Expected Costs of Risk–ignoring Approaches: Given a position fix (μ, Σ), the risk–ignoring approach simply checks if $\mu \in \mathcal{Z}$. Nevertheless, the user's ground truth position x is distributed according to the position fix' pdf $F_{\mu,\Sigma}(x)$ and consequently lies outside of $\mathcal{Z}$ with probability of $(1-p)$. Thus, if the authorization request with the estimated location μ is denied, the probability of a false negative is p. Contrary, if the authorization is granted, the decision is a false positive with probability $(1-p)$. Assume both cases to cause costs c_{fn} and c_{fp}, respectively. This finally allows to derive the expected costs:

$$costs_{risk\text{–}ignoring}(\mu, \Sigma) = \begin{cases} (1-p) \cdot c_{fp}, & \text{iff } \mu \in \mathcal{Z} \\ p \cdot c_{fn}, & \text{else} \end{cases} \quad (7)$$

Note, that given a position fix μ, the risk–ignoring strategy unfortunately may even choose that authorization decision with higher expected cost.

2) Expected Costs of Threshold–based Approaches: Given a position fix (μ, Σ), the threshold–based strategy derives its authorization decision based on checking whether probability p that the user resides within $\mathcal{Z}$ exceeds a predefined threshold. The probability is derived according to (6). Finally, this allows to compute the expected costs for the threshold–based strategy for each possible position fix (μ, Σ):

$$costs_{threshold\text{–}based}(\mu, \Sigma) = \begin{cases} (1-p) \cdot c_{fp}, & \text{iff } p \geq threshold \\ p \cdot c_{fn}, & \text{else} \end{cases} \quad (8)$$

Again, despite the expected cost of an authorization decision, only the satisfaction of the threshold is considered.

3) Expected Costs of Risk–based Approaches: The risk–based LBAC strategy first computes the expected costs of either granting or denying an issued authorization request. Clearly, this results from multiplying the cost of c_{fp} and c_{fn} with their individual probability of occurrence $(1-p)$ and p. Finally, that authorization decision is taken, which promises lower expected costs. Formally, the expected costs for granting or denying the authorization request compute as:

$$costs_{risk\text{–}based}(\mu, \Sigma) = \min\left((1-p) \cdot c_{fp},\, p \cdot c_{fn}\right) \quad (9)$$

Clearly, the risk–based approach always derives that authorization decision with the minimal expected costs. Thus, whenever the risk–ignoring or threshold–based strategy choose that authorization decision with lower expected costs by chance, the risk–based strategy will consequently also choose that decision. This implies, that the expected costs of the risk–based strategy theoretically are a lower bound for the expected cost of any other LBAC strategy.

D. Analyzing the Expected Costs of LBAC strategies

As seen above, the expected costs of the risk–based LBAC strategy are a theoretical lower ound for the expected costs of the risk–ignoring and threshold–based strategy. The percentaged savings $E(S)$ w.r.t. any LBAC strategy $j \in \{risk\text{–}ignoring, threshold\text{–}based\}$ compute as:

$$E(S) = \frac{E(costs_j) - E(costs_{risk\text{–}based})}{E(costs_j)} \quad (10)$$

The theoretically expected savings $E(S)$ from fewer false authorization decisions are laying the foundation for deciding about the overall most cost–effective authorization strategy. Given the boundary conditions $\mathcal{Z}$, F_{err} and $\mathcal{R}$ of a LBS, the expected savings $E(S)$ strongly depend on the ratio $\frac{c_{fp}}{c_{fn}}$ of the LBS's costs for false authorization decisions. Hence, the theoretically expected savings $E(S)$ for a ratio of $\frac{c_{fp}}{c_{fn}}$ will finally play a major role when choosing the practically most cost–effective LBAC strategy. This will be explained later in Section IV in detail.

The dependence of $E(S)$ on $\frac{c_{fp}}{c_{fn}}$ is exemplary depicted in Figure 2 for five theoretical examples with increasingly more inaccurate positioning systems and an authorized zone of 5×5 m. Note, μ is the mean of F_{err} in this figure. The curves are derived using (5). Clearly, the expected savings $E(S)$ show

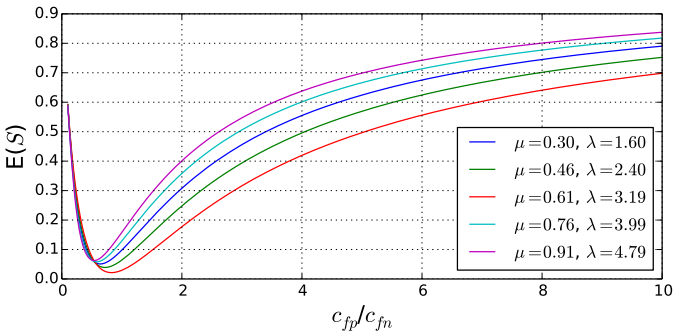


Fig. 2. Theoretically computed expected savings $E(S)$ for five different distributions of inverse Gaussian distributions of F_{err}.

a minimum for each distribution of errors F_{err}. The valleys around the minima of the graph $E(S)$ spread with increasing accuracy of the positioning system. This coincides with the intuition, that the expected savings from operating the risk–based strategy are higher if the positioning system is more inaccurate.

The dependence of the expected savings on the ratio of costs $\frac{c_{fp}}{c_{fn}}$ is a direct consequence of the dependence of $costs_{risk\text{–}based}(\mu, \Sigma)$ on $\frac{c_{fp}}{c_{fn}}$. Given a fixed value of Σ, savings can only arise for such estimated locations μ where the risk–based approach takes a different authorization decision than the risk–ignoring approach. Clearly, the larger the set of such estimated locations μ, the larger the expected savings. This correlation is depicted in Figure 3 in 1D for a authorized zone of 5 m, and Gaussian error estimates with a fixed value of $\Sigma = 3.5$ m. The difference of cost functions for the risk–based and risk–ignoring LBAC strategy is marked green. Here, cost functions $c_{fp} = c_{fn} = 1$ were chosen, i.e., the risk–based LBAC strategy only authorizes a request with position fix (μ, Σ) if $p > 0.5$ according to (4). As the risk–ignoring strategy authorizes all $\mu \in \mathcal{Z}$, the risk–based approach only derives a more cost–effective decision for such μ with $p(\mu, 3.5\text{ m}) < 0.5$.

If the ratio of costs was increased, the shape of cost functions in Figure 3 changes accordingly. In detail, the threshold required by the risk–based LBAC strategy rises according to (4), which increases the set of $\mu \in \mathcal{Z}$ where the risk–based approach denies the authorization and thus has lower expected costs. However, if the threshold implied

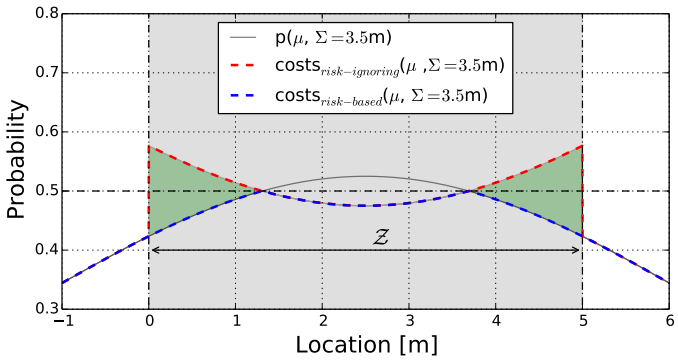


Fig. 3. A 1D example of an authorized zone $\mathcal{Z}$ showing the expected costs for the risk–ignoring and risk–based strategy for a fixed value Σ and cost functions $c_{fp} = c_{fn} = 1$. The green area marks difference of costs.

by the ratio of costs finally overshoots the maximum of p within $\mathcal{Z}$, all authorization requests will be denied by the risk–based strategy. Equally, if the ratio is decreased, the threshold required by the risk–based strategy decreases according to (4). If the required threshold undershoots the value of p on the boundary of $\mathcal{Z}$, the set of location estimates authorized by the risk–based strategy also includes $\mu \notin \mathcal{Z}$. If the ratio of costs converges to 0, the set of authorized μ and the expected savings converge to infinity. Both authorization strategies yield the same expected costs if the threshold implied by the ratio of costs according to (4) corresponds to the value of p on the boundaries of $\mathcal{Z}$. In that case, the risk–based and the risk–ignoring approach show identical behavior. If the distribution of F_{err} shows a high probability for such Σ which cause an identical or nearly identical behavior of the risk–based and risk–ignoring strategy, the expected savings $E(S)$ will finally have a lower minimum. Intuitively, the lower the value of this minimum and the wider the valley around it, the more sensitive is the risk–based strategy to statistically imprecise error estimates Σ. In detail, the risk–based strategy might be mislead by wrong error estimates and game away the theoretically small benefit. Thus, the next section evaluates the weakness of the risk–based strategy when deployed with realistic error estimators.

IV. Use Case: Deploying a Zone–based LBS in an Office Environment

The expected savings of the risk–based strategy are now exemplary evaluated in a use case in a typical office environment. Here, WiFi fingerprinting is used as the underlying positioning system. A radiomap of 206 fingerprints was recorded within an area of 1400 m^2 as depicted in Figure 4. An overall set of 1500 test fingerprints was recorded, each labeled with the room where it was recorded. All areas outside the labeled rooms shown in Figure 4 were assigned the label *outside*. Positioning is performed according to prior work, where a kNN approach with $k = 4$ is used [13]. The position estimate is derived as the weighted mean of the nearest neighbors. Two error estimators, a Laplace and Gaussian error estimator, were used in order to compare the impact of the statistical quality of returned error estimates on the expected savings $E(S)$. The Gaussian error estimator returns bivariate normal distributions and is defined according to prior work [13]. In detail, the aforementioned scale parameter Σ corresponds to

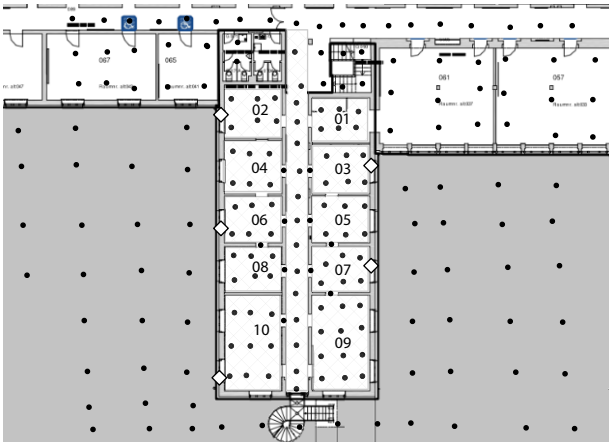


Fig. 4. The recorded radiomap of fingerprints (points) used for WiFi fingerprinting, installed access points (diamonds) and labeled offices.

the covariance matrix of the returned Gaussian and is defined as $\Sigma = \left(\begin{smallmatrix} \sigma & 0 \\ 0 & \sigma \end{smallmatrix}\right)$. Here, σ is derived as the weighted average of the kNNs' distances to the position fix. In contrast, the Laplace error estimator returns rotational symmetric bivariate Laplace distributions with a mean of μ and a diversity b of $\Sigma = b$. Again, the scale parameter is derived as the weighted average of the kNNs' distances to the position fix μ.

The distribution of estimated scale parameters Σ is shown in Figure 5 and mainly follows an inverse Gaussian with parameters $\mu = 0.8$ and $\lambda = 9.5$. In the evaluation, the scale parameter is estimated for a derived position estimate μ and used twice as the variance for a Gaussian and accordingly as the diversity for the Laplace distribution. A set of authorized

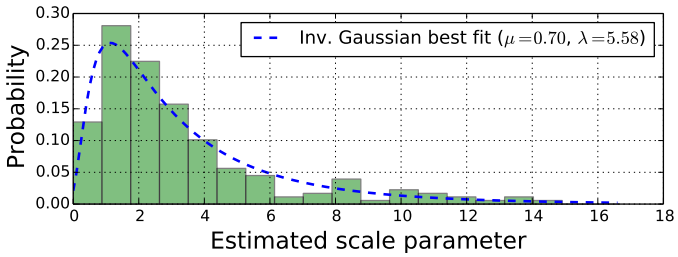


Fig. 5. The distribution F_{err} of scale parameters estimated on the recorded WiFi fingerprinting testset approximately follows an inverse Gaussian.

zones was defined as the labeled rooms shown in Figure 4. In order to compare the effects of the error estimators, the recorded testset was applied to each of the authorized zones, once using the Laplace error estimator and once using the Gaussian error estimator. In order to identify the impact of the authorized zones' size, a second run was performed, where the authorized zones consisted of all possible unions of labeled neighbored rooms from Figure 4. The results are depicted in Figure 6. All runs approximate the theoretically derived shape with a single minimum. For both runs, with single or aggregated rooms, the percentaged expected savings $E(S)$ obtained by employing the Laplace error estimator clearly overshoot the value of $E(S)$ obtained when applying a Gaussian error estimator. However, the theoretical optimality of the risk–based strategy is not given in all cases here. All runs except for the one with a Laplace error estimator and non–aggregated

 ISBN: 978-1-61208-353-7

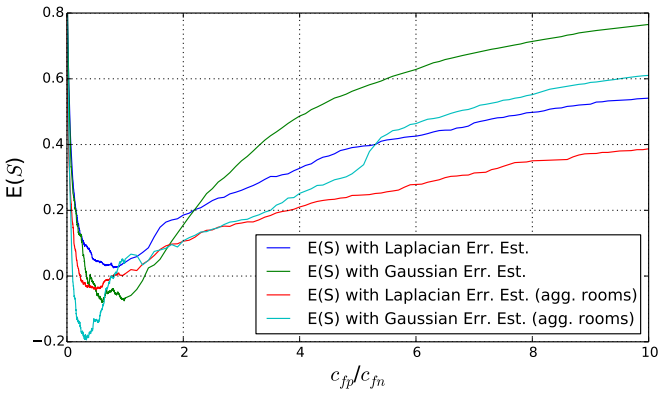


Fig. 6. In our experiment, the Gaussian error estimator clearly compromises the theoretical optimality of the risk–based approach under realistic conditions.

rooms have a negative minimum. For both, the Laplace and the Gaussian error estimator, the expected savings for the run with aggregated rooms are slightly lower than for the single, non–aggregated rooms. This stems from the intuitive fact, that the application of the risk-based strategy is more promising if the rooms are small compared to the estimated errors. Consequently, for aggregated rooms, the minimum for the theoretically expected savings is lower than for non–aggregated rooms and hence, the small superiority is gamed away more easily by imprecise error estimates. Finally, the evaluation results show several important implications. First, the optimality of the risk–based strategy strongly depends on the statistical correctness of the underlying error estimator. The extent of its statistical error compared to the authorized zone also shows a negative effect on the expected savings of the risk–based LBAC strategy. Hence, the cost–optimal LBAC strategy can only be determined by recording a set of test data around the authorized zone of the LBS in the forefront of its deployment. This test data needs to be evaluated with a suitable error estimator for the underlying positioning system. If the expected savings $E(S)$ are negative or near 0 for the LBS's ratio of costs $\frac{c_{fp}}{c_{fn}}$, the application of the risk–based strategy is most likely inferior to the risk–ignoring strategy. However, when such evaluations are performed in order to decide about the most suitable LBAC strategy, a large number of test data is necessary in order to obtain statistically sound results.

V. Conclusion and Future Work

This paper examined the problem of choosing an appropriate location–based authorization strategy, for example needed for LBS, under the occurrence of positioning errors. First, expected costs of operation were theoretically derived for three distinct authorization strategies. The risk–ignoring, threshold–based and risk–aware strategy. It was shown that the superiority of the risk–aware to the risk–ignoring strategy strongly depends on the ratio of costs of false positive and false negative decisions and is minimal for a specific ratio of these costs. In a practical evaluation, the risk–aware policy was shown to be easily mislead to suboptimal decisions when operated with statistically imperfect error estimators. This clearly shows that in practice the widely accepted theoretical superiority of risk–based authorization strongly depends on the ratio of costs and the quality of the error estimator. Furthermore, it is shown that the superiority of the risk–based approach needs to be empirically asserted if the LBS's ratio of costs is near the theoretical minimum of expected savings. Clearly, when deploying a LBS, choosing the right authorization strategy is crucial in order to minimize the expected costs arising from false authorization decisions. Regardless of the importance of the correct choice, this question has not been studied under realistic assumptions up to now. Thus, the results presented in this paper show that the theoretically optimal strategy is not the most effective strategy in all cases under realistic boundary conditions. A methodology to assess the theoretically expected savings of the risk–based and threshold–based strategy is presented. This finally allows to analyze if the risk–aware strategy shows only little improvement given a specific LBS and finally indicates if its application needs to be empirically justified. Finally, the presented approach helps to deploy LBS more cost–efficiently and thus supports their acceptance and efficiency. Future work is seen in developing quality of service metrics for LBS based on the expected costs of their operation. In detail, the effects of imprecise position estimates on LBS require further research.

References

[1] A. Küpper, *Location-Based Services: Fundamentals and Operation*. Wiley, 2005.

[2] L. Krautsevich, A. Lazouski, F. Martinelli, and A. Yautsiukhin, "Cost-effective enforcement of Access and Usage Control Policies under Uncertainties," *Systems Journal, IEEE*, vol. 7, no. 2, pp. 223–235, 2013.

[3] F. Hansen and V. Oleshchuk, "SRBAC: A Spatial Role-based Access Control Model for Mobile Systems," in *Proceedings of the 7th Nordic Workshop on Secure IT Systems (NORDSEC'03)*. Citeseer, 2003, pp. 129–141.

[4] S. M. Chandran and J. B. Joshi, "LoT-RBAC: A Location and Time-based RBAC Model," in *Web Information Systems Engineering–WISE 2005*. Springer, 2005, pp. 361–375.

[5] I. Ray and M. Toahchoodee, "A Spatio-temporal Role-based Access Control Model," in *Data and Applications Security XXI*. Springer, 2007, pp. 211–226.

[6] R. Abdunabi, I. Ray, and R. B. France, "Specification and Analysis of Access Control Policies for Mobile Applications." in *SACMAT*, 2013, pp. 173–184.

[7] C. Ardagna, M. Cremonini, E. Damiani, S. di Vimercati, and P. Samarati, "Supporting Location-Based Conditions in Access Control Policies," in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '06. ACM, 2006, pp. 212–222.

[8] H. Shin and V. Atluri, "Spatiotemporal Access Control Enforcement under Uncertain Location Estimates," in *Data and Applications Security XXIII*. Springer, 2009, pp. 159–174.

[9] P. Marcus, M. Kessel, and C. Linnhoff-Popien, "Enabling Trajectory Constraints for Usage Control Policies with Backtracking Particle Filters," in *MOBILITY 2013, The Third International Conference on Mobile Services, Resources, and Users*, 2013, pp. 52–58.

[10] J. Hightower and G. Borriello, "Particle Filters for Location Estimation in Ubiquitous Computing: A Case Study," in *UbiComp 2004: Ubiquitous Computing*. Springer, 2004, pp. 88–106.

[11] P. Zandbergen, "Accuracy of iPhone Locations: A Comparison of Assisted GPS, WiFi and Cellular Positioning," *Transactions in GIS*, vol. 13, pp. 5–25, 2009.

[12] P. A. Zandbergen, "Positional Accuracy of Spatial Data: Non-normal Distributions and a Critique of the National Standard for Spatial Data Accuracy," *Transactions in GIS*, vol. 12, no. 1, pp. 103–130, 2008.

[13] P. Marcus, M. Kessel, and M. Werner, "Dynamic Nearest Neighbors and Online Error Estimation for SMARTPOS," *International Journal On Advances in Internet Technology*, vol. 6, no. 1 and 2, pp. 1–11, 2013.

 ISBN: 978-1-61208-353-7

Event Driven Adaptive Security in Internet of Things

Waqas Aman and Einar Snekkenes

Norwegian Information Security Laboratory (NISLab)
Gjøvik university College, Norway
Email: {waqas.aman, einar.snekkenes}@hig.no

***Abstract*—With Internet of Things (IoT), new and improved personal, commercial and social opportunities can be explored and availed. However, with this extended network, the corresponding threat landscape will become more complex and much harder to control as vulnerabilities inherited by individual *things* will be multiplied. Conventional security controls, such as firewalls, intrusion detection systems (IDS) etc., may show some level of resistance to this self-organizing network but, as standalone mechanisms, are not sufficient to analyze the threat in a particular context. They fail to provide the essential context of a threat and yields false positives-negatives which can trigger pointless re-configurations, service unavailability and end user discomfort. Such unwanted events can be very catastrophic, for instance, in an IoT enabled eHealth services. We need to have an autonomous adaptive risk management solution for IoT, which can analyze an adverse situation in a distinct context and manage the risk involved intelligently so that the end user, service and security preferences are well-preserved. This paper details an event driven adaptive security model for IoT to approach the objective specified and explicates how it can be utilized in an eHealth scenario to protect against a threat faced at runtime.**

***Keywords*—Adaptive Security; Internet of Things; Event Correlation; eHealth; Ontology.**

I. INTRODUCTION

According to an analysis conducted by the International Data Corporation (IDC), the IoT expected install base will consist of approximately 212 billion things among which 30.1 billion will be autonomous [1]. Indeed, IoT has the potential to create new huge opportunities for personal, business and social services. However, the research this far is still inconclusive on various topics, such as standardization, networking, QoS, etc., among which security and privacy are the most challenging [2].

Things carry inherited vulnerabilities and corresponding threats. Physical exposure, user lack of knowledge, unattended management, remote implementation, communicating wirelessly, low resources, etc., are the common weaknesses which are mostly exploited when devices at the edge of the network are attacked. Bringing them to the IoT will make the threat faced more complex and hard to control. Traditional controls, such as IDS, Antiviruses, etc., as standalone measures may provide protection to some level but are limited in providing a clear context of a situation. As a result, false positives and negatives are triggered and create service disruptions, unnecessary changes and sometimes panic [3]. For instance, an IDS trigger a critical alarm that someone is trying a port scan looking for an open File Transfer Protocol (FTP) port and suggest to close that immediately. This might take the administrator to a total panic situation, and he might close the port on the file server without the fact that it is adequately protected by a strong password. Thus, a simple lack of contextual information might yield to service disruption and panic.

An effective way to approach this problem will be to collect the appropriate network and system information (status or any changes), analyze them in a context and decide an action accordingly. This approach is called adaptive security or adaptive risk management. It is the process of understanding, analyzing and reacting to an adverse situation in a particular context [4] and can be seen in a number of proposals, such as, [5][6]. Common problems with these models are, either they focus on only one security service, such as authentication, or provide a generic architecture without detailing the methods used within each architectural component. Also, existing approaches are either focused on threat analysis or adaptation individually. We realize an absence of a model with specific methods to address and connect both analysis and adaptation as a holistic solution to the problem. Hence, we approach these issues as a set of two questions, i.e., *how to monitor and collect security changes in a real time and analyzed them in a specific context?* And, *how can the analyzed information be used to adapt security settings such that user and service preferences are preserved?*

In this paper, we address the first question by utilizing Open Source Security Information Management (OSSIM) [7], which provides a platform to filter and normalize primitive events collected from *things* in the monitored scope. Correlation directives are specified to model adverse situations in which security events are correlated and analyzed in a particular context. The adaptation question is addressed by utilizing a proposed Adaptation Ontology which leverages on the risk information from the event correlation and adapt security settings accordingly. Using the ontology an optimum mitigation action is selected from an action pool in a manner such that its utility, in terms of usability, QoS and security reliability, is maximum among the possible actions as per user requirements.

The main contribution of this paper is our autonomic security adaptation ontology. OSSIM does not provide such capability and relies on manual reconfigurations which may not address user and service requirements. Also, OSSIM is focused on the traditional computing environment including

 ISBN: 978-1-61208-353-7

servers, desktops and corresponding applications where event processing is relatively a common task. This paper extends event driven security to the IoT where environment becomes more complex due to things diversity and mobility for which traditional protocols and tools seem to be inefficient to approach event processing. Hence, the concept of the paper itself can be considered as contribution.

The rest of the paper is structured as follows: In Section II, work related to event monitoring, correlation and adaptation is presented. The proposed Event Driven Adaptive Security model is detailed in Section III. In Section IV, an eHealth case study will be presented to show how the model can be utilized to protect against a threat at runtime. Finally, the paper will be concluded in Section V along with an overview of our near future plan.

II. Related Work

The related work is categorized into three major areas of relevance, i.e., event monitoring, event correlation and security adaptation in order to get a clear understanding of the specific methods used.

A. Event Monitoring

The objective of monitoring is to collect primitive events from various sources in the environment, filter out the unwanted, categorize them into interested areas of investigation, such as authentication, routing, confidentiality, etc., and normalize them to a common language specification for further analysis. In most of the event driven architecture (EDA), this phase is considered to be a typical task yet, requires knowledge of the target system event specification.

1) Event Collection: The two common approaches are agent-based and agent-less collection. An agent is a small additional program that is installed on the monitored source in order to collect and send events or log files remotely [8]. Agents can be customized to accomplish more specific objectives. The agent-less approach does not require any additional component to be installed. Instead, it utilizes built-in protocols and services, such as System Log (Syslog), Windows Management Instrumentation (WMI), SNMP, etc., to store, access and communicate information at different levels of a monitored system in a standardized manner [9].

One has to address the attributes of flexibility, lightweight, platform in-dependency and management when either of these approaches is adopted. With agent-based, the first three properties can be somehow achieved using expert skills, open source tools and libraries; however, it will be quite a challenge to manage agents across a complex network. The management and control issues can be complex when it comes to a network like IoT. Agent-less approach faces the problem of detail customization thus lacks flexibility and might require additional tools for detail diagnosis [8].

Many commercial and open source event analysis tools, such as [7], use mixed strategies to overcome the flexibility and cross-platform issues. However, most of them use third party apps, for instance, [10][11], where updating and controlling is still a matter of discussion. In [8], the author presented an order-based approach which can provide all the mentioned properties by defining a monitoring scope and using system utilities. However, the method applies only to distributed computing environment where diagnosis utilities are supposed to be already in use. The approach apparently shows lacking when considered in the IoT environment where the monitored objects are more likely to be low-end and resource-less sensors.

2) Event Filtering : The objective of event filtering is to discard the redundant or unwanted events [12]. It defines the targeted event scope to be investigated. Filtering is normally achieved using regular expression where a pattern is matched against the collected events. Non matched events are dropped as redundant events. Two important issues that need to be addressed here are: what events are redundant and how to assure minimal information loss during the process? [13].

The authors in [14] explain that event redundancy scope can be defined using two approaches. Temporal filtration can be used to filter out events generated repeatedly over time with the same information. On the other hand, spatial filtration can provide a mechanism to remove similar event reported by a different system within a given time frame, t. They also propose casual filtration where events collected from different sources are removed based on the fact that they may have different syntax but conveys the same semantics.

Threshold values or time frames can be maintained in temporal or spatial filtration techniques to guarantee minimal information loss. Such flags and offsets will ensure that the information contained in the event will not change potentially and will also take into considerations, e.g., compression rates [13][15].

3) Event Classification: Event classification seems to be based on primitive knowledge about events. Every event generated and stored by a source has a unique set of attributes which can be used to classify an event, for instance, see event structures [16][17]. These attributes designate the event source/destination, timestamps, type, user IDs and the event severity level whose ranges changes as per the source event model and specification.

B. Event Correlation

Correlation is the heart of EDA. It aims to investigate a complex relationship among events and assist to provide enough contextual information to analyze errors, bugs and security threats . Broadly, correlation methods can be classified into two categories, Deterministic and Anomaly-based, either of which can observe events in spatial, temporal or both of the domains [18]. Both the approaches have their associated advantages and disadvantages. Thus, qualifying which of them is a better approach can be determined by evaluating them in a specific application domain [19].

1) Deterministic Approach: In deterministic approach, a predetermined knowledge is utilized to observe and evaluate a given situation. A knowledge base is maintained with application specific information, which is accessed whenever a

 ISBN: 978-1-61208-353-7

particular event pattern is matched. So as a fact, a more expert knowledge can analyze a given security threat, problem or situation more precisely. The knowledge itself and the control to it can be characterized in a number of ways as discussed underneath:

Rule-base Correlation: Rule-based event correlation or threat analysis is the most common way to implement deterministic approaches. Most IDS and security event monitoring tools, for instance [7][20][21], uses a rule based correlation to analyze a threat faced. The knowledge is represented in the form of a predefined rule set which dictates defined alarms and alerts when a specific condition during analysis is met.

State Machine Automata based Correlation: Finite State Machine (FSM) is used to study the behavior and state of underlying systems. In the context of event correlation, various defined states for a system behavior (normal and abnormal) are designed and stored as knowledge base as FSM tuples [22]. A runtime diagnosis engine observes user, application and device behavior and foresees the next system state. Alerts are generated as a flagged state is or about to be triggered. Some of the event correlation models proposed on FSM are [23][24].

The Codebook/Correlation Matrix Techniques: The codebook approach utilizes a symptom-problem relationship. Different suspected events (symptoms) are mapped to their associated abnormal behaviors (problems) and are stored as a knowledge base in a binary matrix, called correlation matrix or a codebook. Events generated are matched against this matrix to identify associated threats or problems. Event correlation models based on codebook techniques can be found in [18].

2) Anomaly-based Approach: Computing and networking environments are very dynamic and the attack vector changes frequently. Some events may not provide certain information and are thus subjected to probabilistic correlation and processing to resolve the uncertainty problem [19]. Unlike predetermined situations in deterministic methods, anomaly-based event correlation aims to identify anomalies without any prior knowledge and can be used to analyze unknown threats. However, they inherit the problems of generating false positive alarms.

Statistical Correlation: As mentioned earlier, events can be filtered, categorized and correlated in both time and space domains to extract rich contextual statistics. For instance, grouping the number of repeated login failure attempts events can provide credible statistics on whether the attempt is a legitimate or that somebody is trying to break-in using a guessing, dictionary or brute force method. High level events, such as alarm/alerts, generated by various security controls, such as IDS, can be used to perform statistical correlation. Statistical information can also be drawn from diverse events having similar attributes/parameters, such as event source, destination, timestamps, etc. Mostly used in anomaly based IDS, these attributes are used as random variables which are later utilized in statistical inferences [25][26].

Probabilistic Modeling: Bayesian networks tend to model relationship among interested random variables. Events can be mapped to random variables. Bayesian model can be illustrated as directed acyclic graphs where nodes represent events of interest and the connecting edges represent the relationships or inter-dependency between them. The probability of a node (situation or event) is inferred by utilizing conditional probability assigned to each node (event) in a given network (scenario) [27]. In most cases Bayesian modeling is coupled with other models techniques, such as Hidden Markov Model and Kalman filters, to investigate complex events in depth [18].

C. Security Adaptation

Assuming that during the analysis an adverse situation or a risk has been discovered, what choices do we have to adapt the security in accordance? How can we utilize the information or context of the analyzed risk to adapt our security? Following is a list of approaches that can be used to answer these questions.

1) Security Policies: Policies remained one of the earliest methods to dictate an action against a given situation. They are a set of rules specifying how a particular situation should be tackled. Edwards et al. in [28] pointed that security policies can be divided into three groups, fixed (e.g., kernel level implementation), customizable (e.g., firewall, router ACLs, etc.) and dynamic, based on the flexibility they offer. Dynamic policies can be detailed on individual user or service level thus providing more flexible adaptation. Some related work include [29][30].

Utility and Probabilistic Models: Utility expresses the measure of efficacy or profit of a choice for a given user or service. In event driven adaptive security, adaptive decisions can be expressed in utilities on the basis of user acceptance, accuracy, power usage, etc. for a given analyzed risk (event). For instance, Alia and Lacosta in [31] used various QoS and security properties corresponding to a required security service to manipulate the utility of an autonomic adaptive response using a non-probabilistic (utility) predictor function. Probabilistic models of utility, such as, [32][33], provides a fair understanding of how security and trust adaptation can be modeled with utilities.

Besides utility theory, probabilistic models such as Bayesian Networks have also been used in a variety of adaptive applications. Bayesian models can be used to select a suitable algorithm from available list [34]. They can also be advantageous in rules discovery [35] to resolve a conflict where an analyzed risk (high level event) two different rules under a given policy [36]. Game theoretic models have also been proposed where intrusion and defense are modeled as games to adapt and defend system security [37][38][39].

Ontologies: Ontologies are used to capture and structure the knowledge about entities, instances and their relationship within an organization. They can be used both for design and runtime purposes [40]. In [41], the author describes an ontology where the knowledge required for security adaptation such as risk, security services and metrics, etc., are related to be assessed at runtime. Denker et. al [42], the authors used security ontologies for annotating functional aspects of electronic resources. However, these ontologies did not discuss

 ISBN: 978-1-61208-353-7

how user requirements and preferences should be valued during the adaptation.

III. The Model

The model presented, Event Driven Adaptive Security (EDAS), addresses the notion of security adaptation in IoT as an EDA in feedback loop manner. We believe that the basic element of change available within the network is the event generated by various application and devices recorded into log files. They provide a primitive context about *who*, *when*, *where* and *what* of a change and contain vital information, such as timestamps, sources, destinations, user activity, severity levels, etc., necessary to reason about the risk situation associated with an event.

EDAS uses Open Source Security Information Management (OSSIM)[7] which provides a platform for writing scripts, called *plugins*, to filter and normalize primitive security events collected from the monitored sources. Correlation in OSSIM is supported with XML rules through which specific situations, in both temporal and spatial view, can be modeled to correlate and investigated events for potential security risks. The model utilizes a runtime adaptation ontology to adapt a best mitigation action from the available actions based on the stored user and service preferences and risk information produced by the correlation engine. A reference model is shown in Figure 1. It includes three major components Monitor, Analyzer and Adaptor. The input, method(s) utilized by individual component along with the details of the output they produced are explained below:

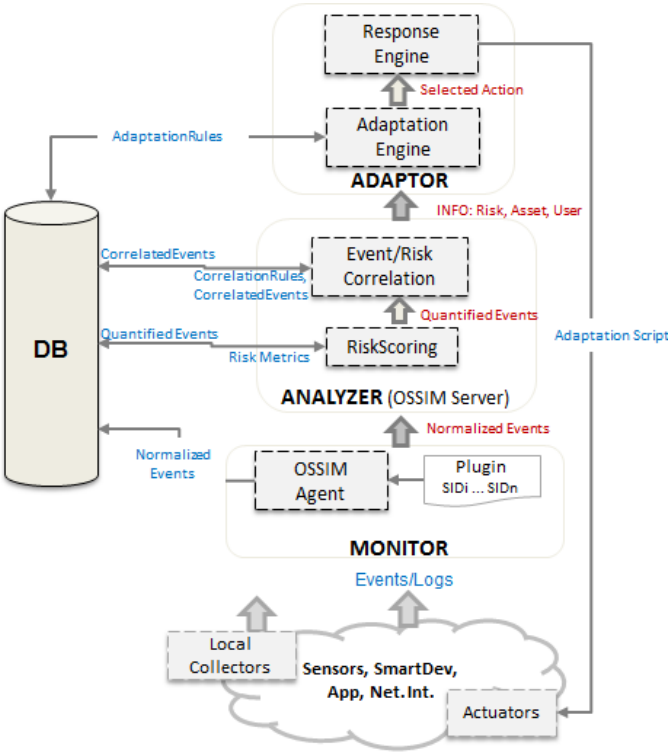


Fig. 1: Event Driven Adaptive Security-Reference Model

A. Monitor

The monitor, OSSIM Agent, collects various events (logs) from diverse things in the IoT, filters the unwanted events and normalizes them to a common language for correlation (analysis).

1) Event Collection: Events generated by monitored *things*, e.g., devices, applications, security tools, are collected remotely by the Monitor enabled with OSSIM Agent. Both, agent and agent-less, methods are used to collect methods. OSSIM uses a variety of methods for remote collection including Syslog and SNMP. These two protocols are only used when a device or application supports them otherwise; an agent is installed on the monitored object. OSSIM does recommend some agents, such as Snare [11] and OSSEC [10], which translate events onto the Syslog stream. However, these agents are not supported by devices at the edge of the network enabling IoT, for instance, smart devices and wireless or body sensors. Thus, we opt for an agent based on MQ Telemetry Transport (MQTT). MQTT is a lightweight M2M messaging transport protocol specifically designed for IoT with platform independence support [43]. The MQTT client hooks onto the event API of the device to collect security events generated and will transport them to the monitor component, the OSSIM Agent, where they are stored in a specific log file.

2) Event Filtration: Security events are extracted using a script, called *Plugin*, designed for individual event source. Writing the script requires some knowledge of the source and the events it is generating. Plugin, identified by a unique ID and other necessary parameters, is a configuration file that dictates from which queue events should be read and which of them needs to be filtered out. OSSIM utilizes a white-listing mechanism where only interested events are sent for further processing. A regular expression specifies these interested events. A match with the expressions is given a unique security ID (SID) which is further used in event correlation. An example plugin configuration is given in Figure 2 showing a specific SID corresponding to a login success event. A different SID can be defined for other events, for instance, a login failure event.

3) Event Normalization: Normalization is performed due to the fact that different *things* in the IoT will generate events in different formats. It is, therefore, necessary to transform them into a single common format for correlation and analysis. It is done during SIDs extraction and aims to extract vital attributes of an event transforming them into a common format for correlation. Attributes vary from event to event depending upon the primitive context they carry. In the above example, date and event source IP is normalized into a normalized common format and src_ip respectively.

B. Analyzer

1) Risk Scoring: Before the normalized events are correlated, they are assigned risk score. OSSIM uses three metrics used for the event (SID) risk quantification [44].

- Asset Value: Specifies the importance of event source or destination within the monitored scope. Ranges from 0-5.

```
[DEFAULT]
plugin_id=1008

[config]
type=detector
enable=yes
source=log
location=/var/log/mydevice.log

[my-device-login-success]
#Apr 2 12:45:12 192.168.5.18 my device:192.168.30.18
login success
event_type=event
regexp="(?P<date>\w{3}\s+\d{1,2}\s\d\d:\d\d:\d\d)\s+(?P
<sensor>\S+)\s+my\device\[\d{1,2}\]:+(?P<src>\d{1,3}\.\d
{1,3}\.\d{1,3}\.\d{1,3})\s+login\success"
date={normalize_date($date)}
sensor= $sensor
plugin_sid=1
src_ip={$src}
…
```

Fig. 2: Example Plugin

- Priority: Specifies the impact of the event. Ranges from 0-5.
- Reliability: Determines the probability or confidence of the fact the event will corresponds to a compromise. Thus, gives a weight to it false positivity. Reliability ranges from 0-10.

For each event, X, risk is quantified as:

$Risk(X) = (Priority * AssetValue * Reliability)/25$

The division of 25 is made to keep the risk values in the range of 0-10 which reflects the risk level of each event. These values are stored in the DB against each SID and are assigned as they arrive in the Risk Scoring engine. They can be changed as required manually. However, priority and reliability values can take different values automatically during event correlation as per the rules.

2) Event Correlation: The correlation engine investigates normalized events coming from the Monitor. It is done using correlation directives stored in XML. They are triggered when a specific SID is encountered, and thus a new event is generated with a new reliability value. The engine increases and decreases this value with respective to defined attributes within the directive rules. Hence, risk is dynamically assessed when SIDs are correlated over time. An SSH login failure example taken (simplified) from OSSIM wiki [45] is given in Figure 3.

```
<directive id="500000" name="SSH Brute Force Attack Against DST_IP" priority="4">
  <rule type="detector" name="SSH Authentication failure" reliability="0"
      occurrence="1" from="ANY" to="ANY" port_from="ANY" port_to="ANY"
      plugin_id="4003" plugin_sid="1,2,3,4,5,6,9,10,12,13,14,15,16,20">
      <rules>
         <rule type="detector" name="SSH Successful Authentication (After 1 failed)"
                        reliability="1" occurrence="1"
                        from="1:SRC_IP" to="1:DST_IP"
                        port_from="ANY" time_out="15" port_to="ANY"
                        plugin_id="4003" plugin_sid="7,8"/>
         <rule type="detector" name="SSH Authentication failure (10 times)"
                        reliability="2" occurrence="10" from="1:SRC_IP"
                        to="1:DST_IP"
                        port_from="ANY" time_out="40" port_to="ANY"
                        plugin_id="4003"
                        plugin_sid="1,2,3,4,5,6,9,10,12,13,14,15,16,20"
                        sticky="true"/>
      </rules>

   </rule>
</directive>
```

Fig. 3: Correlation Directive & Rules

It can be seen that rules can be defined up to n-levels of correlation depending upon the requirements. As the level is increased, more precise information is used, such as the time out, occurrence, source and destination, to validate the reliability and context of an event. In the mentioned example, reliability is increased which increases the risk level correspondingly. Similarly, using a rule, reliability during correlation can also be decreased if a login success event (SID) is encountered within the acceptable threshold range of the *occurrence* variable. Also, logical operators can be utilized when certain conditions are to be assured during the correlation.

Event correlation produces high level events which either goes for in-depth correlation or are flagged as alarms to be managed. Alarms are correlated events with risk level above risk acceptance threshold. Information carried by an alarm includes source and destination IDs, the user involved, risk level, threat details and the correlation directive responsible for generating it. This information is utilized during the adaptation process where the confronted risk is mitigated.

C. Adaptation

In order to utilize the available knowledge precisely and adapt security settings in an optimized manner, we propose an Adaptation Ontology. To be traversed at runtime, the ontology considers all the entities and their relationships necessary for optimal security adaptation. We will be utilizing this entire EDAS model in the IoT enabled eHealth scenario where a patient is remotely managed over the traditional internet or cellular network. To do so, we establish three different contexts in the proposed ontology as shown in Figure 4.

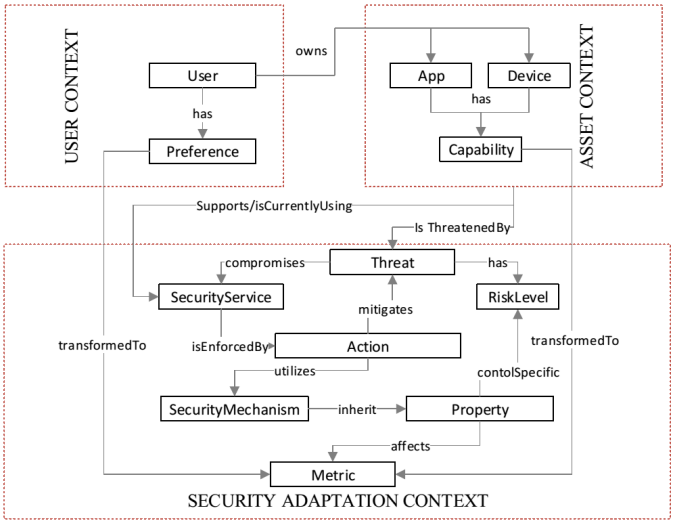


Fig. 4: Security Adaptation Ontology

- *User Context* corresponds to the patient and medical staff preferences which have to be considered before the adaptation
- Each user owns or utilizes a set of application, such as the eHealth app, Skype for patient-doctor communication,

etc. and devices, such as body sensors, smart device or desktop/Laptop, in the scope IoT-eHealth infrastructure. The corresponding information for instance, type, asset value, etc., along with their capabilities is contained in the *Asset Context*.

- The entities and associated settings required for optimized security adaption is grouped under the *Security Adaptation Context*.

An optimal mitigation action is selected from the actions pool following the procedure shown in Figure 5. The Response engine articulate a message based on the details of the action provided by the adaptation engine. Using MQTT transport, the message is sent to an actuator (MQTT Client) installed on the monitored *thing*. The actuator is hooked the specific component API, for instance a login API, and passes the message as variables to be reconfigured.

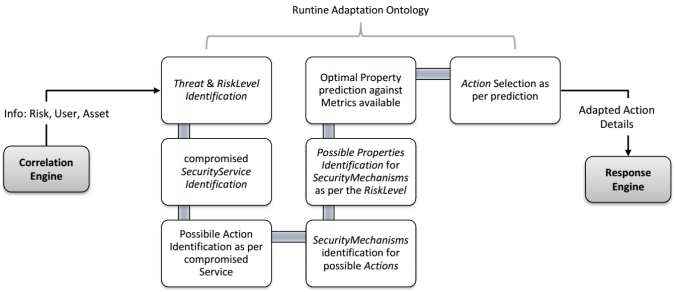


Fig. 5: Security Adaptation Process

A predictor function chooses the action with maximum utility. Subjective weights are assigned to affected metrics against each property, which correspond the overall utility of the property (to be used in the adapted action) for a specific user. Metrics reflect parameters, such as usability, reliability, service cost, etc., which can be negatively or positively influenced by a security property selection. For the time being, metrics are grouped into three categories, User, QoS and Security, to capture influences concerning user preferences, overall QoS and security reliability. However, we are still exploring metrics and measures, such as described in [46], to make our adaptation process more focused and convincing for user and service requirements besides dealing with security issues. A description of individual entities along with example instances is listed in Table I whereas, relations among them are detailed in Table II.

IV. eHEALTH CASE STUDY

IoT can substantially increase service quality and reduce cost, if enabled in the eHealth paradigm where patient vital signs are remotely diagnosed and managed via internet or cellular network. A number of projects, such as [47][48], aim to investigate different aspects of IoT-eHealth to make it more reliable and convenient. This section describes an IoT-eHealth home scenario in which a patient residing at home, Lynda, is equipped with various body sensors. Her vital signs are monitored through these sensors and are transmitted over a Wifi or cellular network to remote hospital site for further diagnosis. She frequently uses her smart phone, part of this infrastructure, installed with an eHealth app to keep track of health status as well as for billing payments besides personal use. We intend to explicate how our model fits into this scenario to defend against a security threat faced.

Home Scenario–Authentication: Lynda wants her credentials saved in the eHealth app to be protected. The app installed on her smart phone is protected with a password that is used to protect her credit card credentials, billing information and local Patient Health Information (PHI).

Adverse Situation: An insider having access to Lynda's smart phone with the intention of stealing her credit card information is trying to login into the eHealth app by guessing different passwords repeatedly.

Preferences: Lynda prefers medium level password instead of a complex one. She does not want her account to be locked out as she has to check her diabetes level frequently.

A generalized message sequence of the whole adaptation process as per the scenario is given in Figure 6. The defense against the situation is detailed as follow:

Model Go-Through – The Runtime Defense:

Event Collection & Monitoring: Smart phone login failure events will be collected by the MQTT client and will be sent to the Monitor. Plugin, e.g., `pluginID=20`, specified for the smart phone will read these events on the OSSIM Agent. The *login failure on eHealth App* SID, with `SID=3`, will extract and normalize the important attributes such as timestamps, user, source, and will add other attributes, such as the number of attempts made.

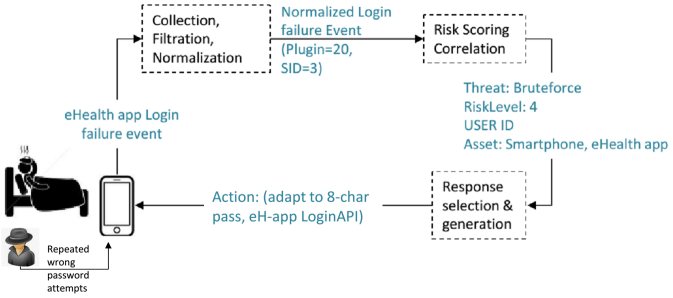


Fig. 6: Attack-Defense Case Study Message Diagram

Risk Quantification: Considering the risk acceptance level for repeated login failure is 4 let the smart phone be a critical asset, so `Asset Value=5`. To give space to for the accidental wrong attempts, let the `Reliability=0` for the first encounter and suppose the importance of the event is considerable so, `Priority=5`.

Event Correlation: The correlation directive shown in Figure 7 specifies 3 levels of correlation. The first wrong attempt is considered as normal so Reliability is not increased. For the next 5 wrong attempts, Reliability is increased to 2 and the engine waits for 10 seconds as a time out. Risk, as per the equation stated earlier, at this stage becomes 2. Similarly,

 ISBN: 978-1-61208-353-7

TABLE I. Ontology Entities

Context	Entity	Description	Example Instances
User	User	The registered user	Patient, Medical Staff, IT staff
	Preference	User preferences that affects or are affected by the adaptation decision	App/device usage knowledge, Current Health Status, Location, Environmental Context, etc.
Asset	App	Any soft components used in the IoT-eHealth infrastructure	eHealth app, communication software such as Skype, email, Security tools, etc.
	Device	Any hard components used to send receive and store User information	Body Sensors, Smart phones, Tablets, Laptops, Desktops
	Capability	The resources offered by individual **Asset**	Battery life time, CPU power, Memory, Supported Protocols etc.
Security Adaptation	SecurityService	The **security services** supported/Currently used by each **Asset**	e.g., Authentication, Encryption and Integrity modules
	RiskLevel	Event/Alarm Risk Level (analyzed by the event correlation/analysis engine) which threatens a **SecurityService** and **Asset**	Range(0-10)
	Threat	Threat information dictated by Correlation Directive	Brute Force, DoS, etc.
	Action	A list of adaptation actions (options) associated with a given **SecurityService**. Actions enforces a specific **SecurityService** in order to control a **Threat** faced	Changing Password, Locking a user for a specific time, changing encryption methods, Adapting a secure authentication protocol, etc.
	SecurityMechanism	Methods/algorithms associated with a given **Action** which are utilized in order to enforce a **SecurityService** challenged by a **Threat**	WEP, WP2, DES, AES, Captcha, SHA1, Disabling User Account etc.
	Property	Available attributes of a specific **SecurityMechanism** which can be adjusted for adaptation	AES (key length), Password (length, character type), captcha (image, audio), Account Locking time (seconds, minutes)
	Metric	Factors affecting security adaptation. Derived from user **Preferences**, device **capabilities** and the overall security against a given **Property** in terms of expected utilities.	Usability, PowerCost, ExecutionTime, ServiceLevelCost, Reliability, etc.

after 6 wrong repeated attempts Reliability is increased to 3 and so does the associated risk level. Finally, an alarm will be generated a risk of level 4 is raised after consecutive 20 attempts when Reliability is increased to 4. Risk is assessed dynamically and instances of the same events are correlated over a period of time as context becomes more evident.

```
<directive id="100" name="Password Brute Force against DST_IP" priority="5">
    <rule type="detector" name="eHealth APP Login failure" reliability="0"
       occurrence="1" from="ANY" to="ANY" port_from="ANY" port_to="ANY"
          plugin_id="20" sid="3">
          <rules>
               <rule type="detector" name="eHealth App Successful Login (After 1 failed)"
                  reliability="2" occurrence="1"
                  from="ANY" to="DST_IP"
                  port_from="ANY" time_out="10" port_to="ANY"
                  plugin_id="20" plugin_sid="3"/>
              <rule type="detector" name="eHealth App Login failure (6 times)"
                  reliability="3" occurrence="6" from="ANY"
                  to="DST_IP"
                  port_from="ANY" time_out="40" port_to="ANY"
                  plugin_id="20"
                  sid="3" />
             <rule type="detector" name="eHealth App Login failure (20 times)"
                  reliability="4" occurrence="20" from="ANY"
                  to="DST_IP"
                  port_from="ANY" time_out="60" port_to="ANY"
                  plugin_id="20"
                  sid="3" />
          </rules>

    </rule>
</directive>
```

Fig. 7: Correlation Directive & Rules for Repeated Login Failures

Security Adaptation: Proceeding logically with the procedure shown in Figure 5. An optimal mitigation action can be selected as:

- *Threat & Risk Level:* Password Brute Force
- *Compromised Security Service:* Authentication
- *Possible Actions:* Suppose, Password Change, Account Lockout & Enforcing Captcha
- *Security Mechanisms:* As per each action, Password Change (keyLength), Enforcing Captcha (Captcha), Account Lockout (Time Restriction)
- *Security Properties Metrics & Utilities:* As a hypothesis, consider Table III showing the affected metrics by individual properties with associated utilities (ranging from 1-10). The properties listed are considered to mitigate risk level 4 or above for password brute force attempts on the smart phone. Furthermore, it is assumed that the utilities are assigned as per service and user preferences.

TABLE III. Properties, Metrics & Utilities

Metric	PROPERTIES					
	KeyLength		Captcha		Time Restriction	
	8-char,	10-char.	Audio	Visual	15 min.	30 min.
Usability	8	5	6	7	6	3
QoS	8	7	5	5	6	6
Reliability	7	8	4	4	7	8
Total Utility	23	20	15	16	19	17

The predictor function will identify that the optimal action to circumvent this threat is to change the password on the smart phone eHealth app to an 8-characters. If it is already in use, it will go back and select the second best option. The selected action along with the user, concerned API and asset details will be given to the Response engine which will send a

 ISBN: 978-1-61208-353-7

TABLE II. Ontology Relations

Context	Relation	Classes Involved	Example Relations
User	has	User, Preference	Patient has a Preference of having easy to remember credentials Patient prefers service over security while being outside home Doctor prefers strict confidentiality while being outside hospital
User, Asset	owns	User, Asset	Patient owns a tablet to read his vital signs Patient owns (wears) ECG sensor Doctor owns a desktop machine to communicate with Patient over Skype
Asset	has	App, Device, Capability	Patient tablet has DualCore processor installed eHealth app installed on patient tablet has a medium level password ECG sensor does not support DES 128 bit algorithm Smart phone has 1 hour of talk time left
Asset, Security Adaptation	Supports, Currently Using	Asset, SecurityService	ECG Sensor supports/currentlyUsing Confidentiality, Authentication
	IsThreatenedBy	Asset, Threat	eHealth app is threatened by a password brute force attack In home Wifi network is threatened by DeAuth flooding
	compromises	Threat, SecurityService	Password Brute force compromises eHealth app Authentication WifiDeAuth flooding compromises network integrity
	has	Threat, RiskLevel	Password Brute force on eHealth app has a HIGH Risk Level
	isEnforcedBy	SecurityService, Action	eHealth App is authentication is enforced by a medium strength password Wifi Network authentication is enforced by WPA policy
	mitigates	Action, Threat	Changing user password mitigates a password brute force threat Restricting user login attempts to t-seconds mitigates a password brute force
	utilizes	Action, SecurityMechanism	A password change action utilizes the password length & complexity Restricting user login attempts utilizes the time limit Increase encryption level action utilizes AES
Security Adaptation	Inherit	SecurityMechanism, Property	Password length inherit the property of 6, 8 or 10 characters Password complexity inherit the property of character type
	controlSpecific	Property, RiskLevel	A password with 6 digit key length controls LOW level brute force attempts A password with 10 digit key length controls HIGH level brute force attempts
	affects	Property, Metric	10 character password affects (decreases) usability and (increase) security reliability 3G network affects (increases) Service Quality and (decreases) device battery
User, Security Adaptation **Asset, Security Adaptation**	transformedTo	Preference, Metric	User preferences are transformed to Usability User location is transformed to QoS, Security \& Privacy attributes
		Capability, Metric	Supported protocols (can be) transformed to QoS and Security metrices

message containing the instructions as appropriate variables to the MQTT client residing on the smart phone as an actuator. The actuator will identify the API mentioned and will pass the message variable. The API will implement the changes and will ask the user/adversary to enter a new 8-character password based on the older one.

V. Conclusion & Future Work

Existing detective and preventive controls as individual components seems to be inefficient in providing the required context to investigate security threats. We presented an event driven adaptive security model, EDAS, which leverages the capabilities of existing event models of diverse things in IoT and OSSIM correlation to adapt security settings by keeping the user and service utility at maximum. Primitive knowledge about security changes is collected and is analyzed in a definitive and established security context. The runtime adaptation ontology provides a structured knowledge of all the elements necessary to select appropriate mitigation action as user and service preferences. MQTT as a transport mechanism for the collection and actuation processes makes the model more extendable, platform independent and cost effective.

In the near future, we intend to develop a prototype for EDAS to test its processes as a real world IoT-eHealth artifact. Preliminary plans are to investigate the overall reliability, service response timings and building universal collectors and actuators for devices at the network edge, such as body sensors and personal smart devices. The prototype will be validated with confidentiality, availability, integrity and mobility scenarios as they are deemed to be the most critical aspects in remote patient management systems.

Acknowledgements

The work presented in this article is a part of the ASSET (Adaptive Security in Smart IoT in eHealth) project. ASSET (2012-2015) is sponsored by the Research Council of Norway under the grant agreement no: 213131/O70.

References

[1] "The internet of things is poised to change everything, says international data corporation," Press Release, October 2013, last access date: 31 May 2014. [Online]. Available: http://www.idc.com/getdoc.jsp?containerId=prUS24366813

 ISBN: 978-1-61208-353-7

[2] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[3] D. Shackleford, "Real-time adaptive security," SANS, Tech. Rep., December 2008, last Accessed on 4 April 2014. [Online]. Available: http://www.sans.org/reading_room/analysts_program/adaptiveSec_Dec08.pdf

[4] M. Nicolett and K. M. Kavanagh, "Magic quadrant for security information and event management," *Gartner RAS Core Reasearch Note (May 2009)*, 2011.

[5] RSA, "Rsa adaptive authentication. a comprehensive authentication and risk management platform," 2013, accessed on: 31 May 2014. [Online]. Available: http://www.emc.com/collateral/data-sheet/h11429-rsa-adaptive-authentication-ds.pdf

[6] H. Abie, "Adaptive security and trust management for autonomic message-oriented middleware," in *Mobile Adhoc and Sensor Systems, 2009. MASS'09. IEEE 6th International Conference on*. IEEE, 2009, pp. 810–817.

[7] "Ossim: the open source siem." last access date: 31 May 2014. [Online]. Available: http://www.alienvault.com/open-threat-exchange/projects

[8] L. Kufel, "Security event monitoring in a distributed systems environment," *IEEE Security Privacy*, vol. 11, no. 1, pp. 36–43, Jan. 2013.

[9] P. Bellavista, A. Corradi, and C. Stefanelli, "Java for on-line distributed monitoring of heterogeneous systems and services," *The Computer Journal*, vol. 45, pp. 595–607, 2002.

[10] "OSSEC: open source SECurity," last access date: 31 May 2014. [Online]. Available: http://www.ossec.net/

[11] "InterSect alliance - snare agents," last access date: 31 May 2014. [Online]. Available: http://www.intersectalliance.com/snareagents/index.html

[12] O. Etzion and P. Niblett, *Event Processing in Action*, 1st ed. Greenwich, CT, USA: Manning Publications Co., 2010.

[13] Z. Zheng, Z. Lan, B.-H. Park, and A. Geist, "System log pre-processing to improve failure prediction," in *IEEE/IFIP International Conference on Dependable Systems Networks, 2009. DSN '09*, Jun. 2009, pp. 572–577.

[14] A. Oliner and J. Stearley, "What supercomputers say: A study of five system logs," in *37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, 2007. DSN '07*, Jun. 2007, pp. 575–584.

[15] M. F. Buckley and D. P. Siewiorek, "A comparative analysis of event tupling schemes," in *Fault Tolerant Computing, 1996., Proceedings of Annual Symposium on*. IEEE, 1996, pp. 294–303.

[16] "System logger: Syslog linux man page," last access date: 31 May 2014. [Online]. Available: http://linux.die.net/man/3/syslog

[17] "Event schema elements (windows)," last access date: 31 May 2014. [Online]. Available: http://msdn.microsoft.com/en-us/library/windows/desktop/aa384367(v=vs.85).aspx

[18] G. Jiang and G. Cybenko, "Temporal and spatial distributed event correlation for network security," in *American Control Conference, 2004. Proceedings of the 2004*, vol. 2, Jun. 2004, pp. 996–1001 vol.2.

[19] J. P. Martin-Flatin, G. Jakobson, and L. Lewis, "Event correlation in integrated management: Lessons learned and outlook," *Journal of Network and Systems Management*, vol. 15, no. 4, pp. 481–502, Dec. 2007.

[20] "Snort : Open source IDS/IPS," last access date: 31 May 2014. [Online]. Available: http://www.snort.org

[21] "The bro network security monitor," last access date: 31 May 2014. [Online]. Available: https://www.bro.org

[22] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to automata theory, languages, and computation*. Addison-Wesley, 2001.

[23] M. Sifalakis, M. Fry, and D. Hutchison, "Event detection and correlation for network environments," *IEEE Journal on Selected Areas in Communications*, vol. 28, no. 1, pp. 60–69, Jan. 2010.

[24] J. Tan, X. Pan, S. Kavulya, R. Gandhi, and P. Narasimhan, "SALSA: analyzing logs as StAte machines." *WASL*, vol. 8, pp. 6–6, 2008.

[25] P. Garca-Teodoro, J. Daz-Verdejo, G. Maci-Fernndez, and E. Vzquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & Security*, vol. 28, no. 12, pp. 18–28, Feb. 2009.

[26] N. Ye, S. Member, S. M. Emran, Q. Chen, and S. Vilbert, "Multivariate statistical analysis of audit trails for host-based intrusion detection," *IEEE Transactions on Computers*, vol. 51, pp. 810–820, 2002.

[27] B. Gaudin, P. Nixon, K. Bines, F. Busacca, and N. Casey, "Model bootstrapping for auto-diagnosis of enterprise systems," in *International Conference on Computational Intelligence and Software Engineering, 2009. CiSE 2009*, Dec. 2009, pp. 1–4.

[28] W. K. Edwards, E. S. Poole, and J. Stoll, "Security automation considered harmful?" in *Proceedings of the 2007 Workshop on New Security Paradigms*, ser. NSPW '07. New York, NY, USA: ACM, 2008, pp. 33–42.

[29] D. Kulkarni and A. Tripathi, "Context-aware role-based access control in pervasive computing systems," in *Proceedings of the 13th ACM symposium on Access control models and technologies*. ACM, 2008, pp. 113–122.

[30] T. E. Maliki and J.-M. Seigneur, "A security adaptation reference monitor (SARM) for highly dynamic wireless environments," in *Proceedings of the 2010 Fourth International Conference on Emerging Security Information, Systems and Technologies*, ser. SECURWARE '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 63–68.

[31] M. Alia and M. Lacoste, "A QoS and security adaptation model for autonomic pervasive systems," in *Computer Software and Applications, 2008. COMPSAC '08. 32nd Annual IEEE International*, Jul. 2008, pp. 943–948.

[32] D. Quercia and S. Hailes, "MATE: mobility and adaptation with trust and expected-utility," *International Journal of Internet Technology and Secured Transactions*, vol. 1, no. 1, 2007.

[33] S.-W. Cheng, D. Garlan, and B. Schmerl, "Architecture-based self-adaptation in the presence of multiple objectives," in *Proceedings of the 2006 international workshop on Self-adaptation and self-managing systems*. ACM, 2006, pp. 2–8.

[34] H. Guo, "A bayesian approach for automatic algorithm selection," in *IJCAI 2003 Workshop on AI and Autonomic Computing, Mexico*. Citeseer, 2003, pp. 1–5.

[35] R. Sterritt, "Autonomic networks: engineering the self-healing property," *Engineering Applications of Artificial Intelligence*, vol. 17, no. 7, pp. 727–739, 2004.

[36] E. Lupu and M. Sloman, "Conflict analysis for management policies," in *Integrated Network Management V*. Springer, 1997, pp. 430–443.

[37] J. Stiborek, M. Grill, M. Rehak, K. Bartos, and J. Jusko, "Game theoretical adaptation model for intrusion detection system," in *Advances on Practical Applications of Agents and Multi-Agent Systems*. Springer, 2012, pp. 201–210.

[38] C. B. Simmons, S. G. Shiva, H. S. Bedi, and V. Shandilya, "ADAPT: a game inspired attack-defense and performance metric taxonomy," in *Security and Privacy Protection in Information Processing Systems*. Springer, 2013, pp. 344–365.

[39] W. Jiang, B.-x. Fang, H.-l. Zhang, Z.-h. Tian, and X.-f. Song, "Optimal network security strengthening using attack-defense game model," in *Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on*. IEEE, 2009, pp. 475–480.

[40] A. Evesti, E. Ovaska, and R. Savola, "From security modelling to run-time security monitoring," *Security in Model-Driven Architecture*, pp. 33–41, 2009.

[41] A. Evesti and E. Ovaska, "Ontology-based security adaptation at run-time," in *Self-Adaptive and Self-Organizing Systems (SASO), 2010 4th IEEE International Conference on*, Sept 2010, pp. 204–212.

[42] G. Denker, L. Kagal, and T. Finin, "Security in the semantic web using owl," *Information Security Technical Report*, vol. 10, no. 1, pp. 51–58, 2005.

[43] "Mq telemetry transport, mqtt," last access date: 31 May 2014. [Online]. Available: http://mqtt.org/

[44] "Ossim risk calculation," last access date: 31 May 2014. [Online]. Available: https://www.alienvault.com/wiki/doku.php?id=user_manual:dashboards:risk:risk_metrics#risk_calculation

[45] "Ossim-writing correlation directives," last access date: 31 May 2014. [Online]. Available: https://www.alienvault.com/wiki/doku.php?id=user_manual:intelligence:writing_correlation_directives

[46] R. M. Savola and H. Abie, "On-line and off-line security measurement framework for mobile ad hoc networks." *Journal of Networks*, vol. 4, no. 7, 2009.

[47] "Asset - adaptive security for smart internet of things in ehealth," last access date: 31 May 2014. [Online]. Available: http://asset.nr.no/asset/index.php/ASSET_-_Adaptive_Security_for_Smart_Internet_of_Things_in_eHealth

[48] "Butler ubiquitous, secure internet-of-things with location and context-awareness," last access date: 31 May 2014. [Online]. Available: http://www.iot-butler.eu/

 ISBN: 978-1-61208-353-7

A Certificate-based Context Aware Access Control Model For Smart Mobile Devices In Ubiquitous Computing Environments

Davut Cavdar, Ahmet Yortanlı, Pekin Erhan Eren, Altan Koçyiğit
Middle East Technical University, Informatics Institute,
Ankara, Turkey
Emails: {dcavdar, e129848, ereren, kocyigit }@metu.edu

***Abstract*—In this paper, a context-aware access control model is provided to be used by people with mobile devices in ubiquitous computing environments. The model utilizes a certificate-based approach and aims to create an infrastructure for regulating access requests through mobile devices to resources and services in a local environment. The model also allows users from different domains access to local resources and services within the scope of agreements between domains. In addition to conceptual design of the model, a working prototype implementation is developed and successful application of the model is demonstrated. In the prototype implementation, an application running on a real smart mobile phone is developed for generating access requests; a gateway device is utilized for context management and access control in a local ubiquitous environment with real physical sensors. Sample use cases are applied on the prototype in order to demonstrate the applicability and feasibility of the model.**

Keywords-certificate; access control; smart mobile device; context awareness

I. INTRODUCTION

Following rapid developments in the technology, smart mobile devices have become smaller than personal computers. Also, device diversity has increased to cope with the needs of humans for daily life activities. Besides standard computers, mobile devices, sensors, actuators have started to be used by humans in daily life. Unlike standard devices, these ubiquitous computing devices have effective interaction capabilities with both humans and other electronic devices. The actual contribution of ubiquitous computing is that these small and smart computers are densely distributed in the environment; they work and interact in the background invisibly and without disturbing people [12].

In today's world, humans are involved in many interactions with ubiquitous devices. Like any other system, access requests to such devices should be controlled and regulated. Especially, it is important to prevent unauthorized access to resources in ubiquitous environments.

Mainly two types of authentication methodology are offered: one is static authentication, the other is dynamic authentication. Password-based authentication is the most popular authentication method for static authentication. Smart card, biologic, Universal Serial Bus (USB) token and certificate based authentications are examples of dynamic authentication. For ubiquitous computing environments, static authentication is not suitable because access evaluation results can change according to user, location, time etc. contexts. In this paper, a certificate-based context aware access control model for smart mobile devices is provided.

This paper consists of five sections. Section 2 includes background information about certificate-based access control models. Section 3 explains the proposed model, including the main components and the activity flow. A prototype implementation of the model is introduced in Section 4. Finally, Section 5 provides the conclusion.

II. RELATED WORK

Different systems and solutions use certificates in order to regulate or check authenticity during access to resources. Web sites, mail systems, mobile applications are main areas of this usage. A user authentication method using smart cards is offered as a certificate-based authentication [10]. In this method, user certificate and other private information are stored in a smart card and the system performs authentication process based on the combination of smart card information and related context. The method focuses on authentication transactions; however, access control mechanisms and process are not discussed.

Another offered solution to access control uses certificates for access control for inter-domain environments is presented by Thompson et al. [9]. In this solution, users send their certificates storing their roles in order to reach resources However, major deficiency of this solution is that it is not designed for context aware environments and context usage.

An access control method is offered for healthcare systems by Koufi and Vassilacopoulos [6]. This method is designed for context aware environments. The system validates user roles stored in user certificates and evaluates certificate data with context information. However, this model does not support giving access to other domain users for reaching resources.

An extension model of Role Based Access Control (RBAC) is suggested for access control by Chadwick et al. [3]. The model uses X.509-based certificates that store user roles and definition for accessing resources. Access rules are defined as XML-based policy rules and they are stored in Lightweight Directory Access Protocol (LDAP) [3]. The

 ISBN: 978-1-61208-353-7

model also controls certification cancellation status using certificate revocation lists. However, the model is not suitable for context aware environments; this can be considered as the main shortcoming.

For Grid environments, certificate-based access control model namely "Sygn" is also offered by Ludwig et al. [4]. It provides decentralized permission storage and management for dynamically changing resources. Although it creates on-demand permissions without central permission systems, it is hard to regulate permissions with, in the certificates, in terms of inter-domain and security approaches.

Our proposed model combines three main properties of ubiquitous computing environments. The model provides a context aware access control and smart mobile device usage and also the model works in inter-domain environments in order to allow users access to resources of other domains.

III. PROPOSED MODEL

A. *Main characteristics of the proposed Model*

There are three main components of this model, (i) first is the user processes running on a mobile device, (ii) second is the main gateway, performing duties such as certificate control, applying rules, etc., and (iii) third is local resources or services such as reaching sensors values or printer usage. The general structure of the proposed model is shown graphically in Figure 1.

The proposed model uses certificates for authorized access to resources. After connecting to the local service wirelessly, the user sends his/her certificate to the gateway by using his/her smart mobile devices. After that, the gateway gathers required information and performs actions accordingly, and finally, produces a result. This result is received by the user via his/her mobile device again.

Another important characteristic of the proposed system is that it provides a mobile usage environment to the system users. In ubiquitous environments, computers are hidden and resources/services are widely distributed. Also, people are in transition to more mobility in terms of life styles and technology trends. Therefore, people have frequent interactions with embedded computers or resources when they are mobile. In the proposed model, home or other domain users can explore local resources/services when they are in a different location and they can send access requests to gateways.

Context-awareness is another important feature of the proposed model. In order to respond to received requests correctly according to access policy rules, the system gathers context information from the environment. Since the proposed model is a context-aware system, it senses contextual information like location, mobile user id, time, resource type and it performs required actions according to this gathered contextual information.

The proposed model can perform authorization and access control requests conducted by not only different domain users, but also by other domain users. To do that, domains provide an access policy rules agreement for their own users when they are using different domains' resources/services.

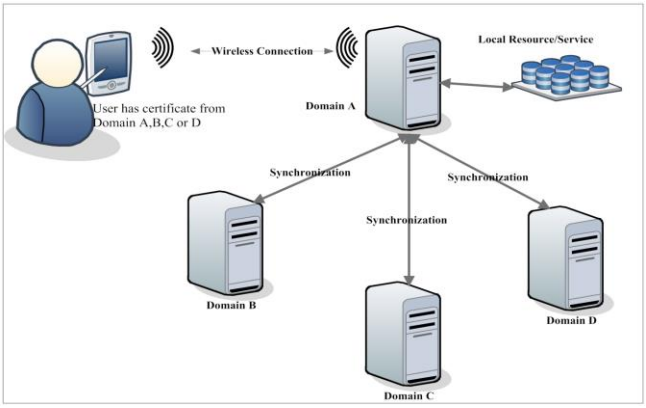


Figure 1 The general structure of the proposed model.

According to the agreement between domains, each domain sets access rules for both home and other domains' users' requests. After these agreements, the system checks other domain certificate lists in order to update access lists of other domains at each pre-defined synchronization time intervals.

There are two cases for users' domain status. When the user requests an access to a local resource/service, if s/he is home domain user, the system checks access policy rules, acquires context information, evaluates request and produces a response accordingly. If s/he belongs to a different domain, the system first checks an agreement between domains, if it is available then it checks access rules for that user, collects contextual information and finally performs an evaluation and creates a response

B. *Advantages of using certificates*

The systems need to check the identity of users for each access control by communicating with the home domain of the user. Or, at least, each domain should provide an authentication mechanism before the access decision for requests is made in order to validate user from the home domain. Such a validation mechanism requires communication between the servers of the domains and this consumes tangible amount of time and network bandwidth.

By using certificates, users carry their own authentication credentials with them so that communication between domains to validate users can be minimized. This certificate-based approach provides a faster access control compared to the approach based on authentication via the home domain. Moreover, by using certificates, users not only carry their own identity, but also carry their domain's identity with them. As a result, inter-domain service level agreement rules to access the resources can be defined in domain identity level and access control evaluations can be done based on domain identity when a user tries to access a resource with his/her certificate. That is, instead of storing users' identity in rules, storing certificate provider's identity in rules is enough in order to evaluate individual user requests.

 ISBN: 978-1-61208-353-7

C. Components of the proposed model

There are three main components of the proposed model, (i) first is user processes running on a mobile device, (ii) second is the main gateway, performing duties like certificate control, applying rules, etc., and (iii) third is local resources or services.

The gateway is the main unit of the model that accomplishes critical tasks and behaves like a bridge between the user (client) and the requested resources or services. Sub-components of the gateway are; Certificate Service, Context Engine, Decision Engine, Database Service and Management Panel.

The Certificate Service is mainly responsible for checking certificates sent by the user during access request process. After certificate information reaches the gateway, Certificate Service parses it, and checks identifier sections of the certificate. Also, the Certificate Service performs synchronization between domains. It checks domains' active certificate lists and if any change (add, delete, update) has occurred in these lists, Certificate Service updates required lists between home and other domains.

The Context Engine has mainly two duties in the proposed model. Its first task is context acquisition. Because model offers a context aware environment, context information such as location, time, group etc. about the requested resource and the user should be collected. The Context Engine collects required context information and sends them to the Decision Engine when the user demands access to the resource. Secondly, the Context Engine is responsible for managing context rules for the model. When the Decision Engine requests related rules for the defined user and resource, the Context Engine finds correct rules that will be applied for the request and sends them to the Decision Engine.

The Decision Engine is the core component of the proposed model. The user sends requests as an envelope to the Decision Engine. After receiving requests it opens envelope and defines user certificate data and resource IDs. Then, the Decision Engine demands required context information and related rules for the user from the Context Engine and sends certificate data to the Certificate Service in order to check certificate accuracy and also validity. It reaches a decision after collecting all these required data and rules.

The Database Service provides a communication infrastructure for all system modules and database. When a system module needs information stored in the database such as user group, resource ID, policy rules, it uses the Database Service to get access to the related database.

The Management Panel allows system administrators to manage system parameters by using its interface. Administrators can perform management tasks such as add or delete rules, user groups, etc., by using this panel.

D. Activity flow of the proposed model

When a user requests to reach a local resource or service via his/her smart mobile device, first s/he downloads or saves his/her certificate provided by his/her host domain into his/her mobile device, then s/he establishes a wireless connection with the resource gateway. By using the application running on the mobile device, the user selects his/her certificate and requested resource type. This type may only be one such as printer usage or more than one such as sensors providing more than one resource type like temperature, light, etc. The mobile application generates a message envelope including "Certificate Data" and "Resource Type" and sends this envelope to the gateway.

After retrieving the request envelope, the Decision Engine opens it and parses the data. Certificate data is sent to the Certificate Service for validation process. The Certificate Service first parses certificate data for default validity check, also it controls synchronized active certificate lists of other domains for certificate validity and sends the result to the Decision Engine. After that, if the certificate passes the validity check, the Decision Engine requests related contexts and rules from the Context Engine. According to the user and resource type, the Context Engine finds related rules from the database and contexts and this information are delivered to the Decision Engine. During this process, the local resource sends required data to the gateway. This data type may vary according to the designed application. If it is a file access control system, data may be up-to-date version of files, if it is a printer access control system, data may be printer current status, if it is a sensor data access control system, data may be values read by sensors.

Finally, the Decision Engine collects required information from related modules and makes an evaluation. According to the results of the evaluation, the resource/service access is allowed or denied by the system. This activity flow of the proposed model is illustrated in Figure 2 in detail.

E. Synchronization of domains' certificate lists

In the proposed model, Certificate Service performs a synchronization task in a pre-defined time period in order to make all agreed domains' active certification lists up-to-date. Each domain should be aware of certification cancellations in order to prevent access of unauthorized users into local resources. The method is based on broadcasting Certificate Cancellation List (CCL), Domains get other domains' Active Certificate List (ACL) and then each domain broadcasts its CCL in some pre-defined time intervals via its Certificate Service.

Using CCL-based synchronization method seems to be more suitable for the proposed system. However, it also has some problems; if synchronization time period is too long, this may cause unauthorized user access. An administrator of one domain may cancel one certificate; however, it may still seem to be active in the home domain due to the fact that there is time to the synchronization process.

 ISBN: 978-1-61208-353-7

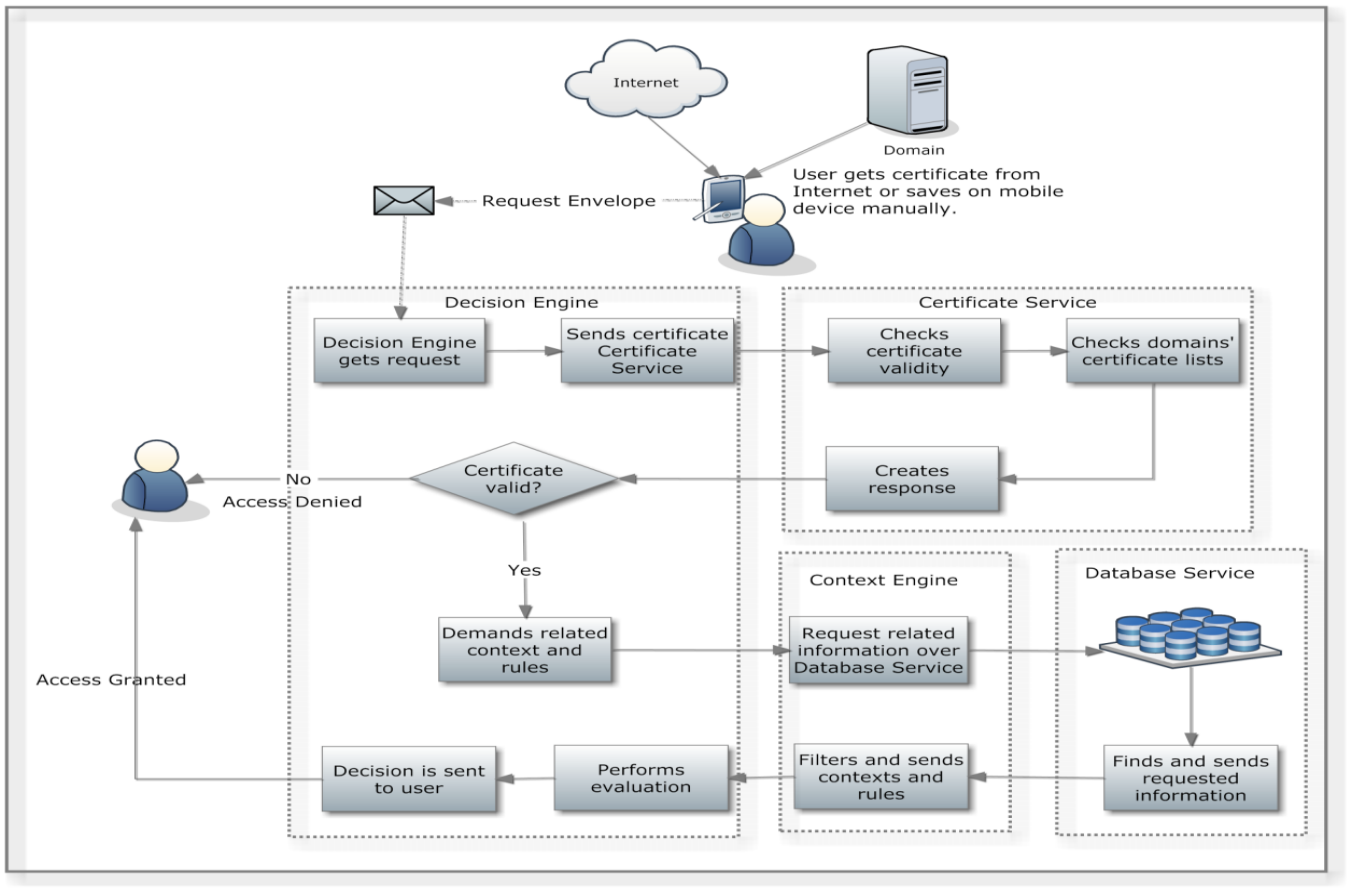


Figure 2 Activity Flow of Proposed Model

Synchronization time intervals should be set as minimum as possible according to the network communication traffic load. Another problem of the method is that when a request occurs from a different domain user, Certificate Service checks user domain's CCL, this may also cause network traffic. To overcome this problem, CCL lists of domain can be stored regularly in the home domain. These specifications and functions can be adapted according to the system environment and network conditions.

F. Management of Access Rules (AR)

The proposed model requires Access Rules (AR) in order to make decisions toward requests. The Context Engine is responsible for management, retrieving and sending required rules to requester component of the proposed model.

System Administrator defines ARs and adds them into the database. When a user requests an access, the Decision Engine asks for the related rules from the Context Engine, and then the Context Engine gets ARs using the Database Service. After getting ARs, it controls requested rules and sends them back to the Decision Engine.

ARs are transferred between the Decision Engine and the Context Engine in Extensible Markup Language (XML) format. XML is a platform and programming language-independent notation format; therefore, this usage provides flexibility for future module addition and deletion or structure change.

When a rule is requested by the Decision Engine, the Context Engine receives user's certificate/provider and resource id and then, it queries related rules with these identifiers. The Database Service finds and sends all related rules to the Context Engine. Sent rules are checked by the Context Engine and they are converted into XML structure. After that, XML based rules are sent to the Decision Engine for evaluation process. An example of Access Rule (AR):

```
<Access Rule>
<subject type = " certificate_provider "> METU
</subject>
<resource type = "service"> II/printers </ resource >
<context type = "time "> week_days </ context >
<decision > allow </ decision >
</Access Rule>
```

G. Permission evaluation method

Context Engine sends most suitable one rule to Decision Engine in order to avoid conflicts.

 ISBN: 978-1-61208-353-7

1. If there is a "deny" rule among queried rules, it has superiority over other "allow" rules for the same user and same contexts.
2. If there is no rule for requested access, Decision Engine sends "deny" response to the user.
3. If rule has time "deny" definition for current time context, Decision Engine sends "deny" response to the user.
4. If rule has time "allow" definition for current time context, however it has location "deny" definition for current location context, Decision Engine sends "deny" response to the user again.
5. If rule has "allow" definition for both time and location and if these parameters are "true" for current time and location context, Decision Engine sends "allow" response to the user and performs required operations.

IV. PROTOTYPE IMPLEMENTATION

In order to show the feasibility of the proposed model and demonstrate its applicability, a working system is developed based on the proposed model. The prototype mainly consists of an Android-based mobile application that works on a mobile device, gateway software that works on a personal computer and temperature/light sensors that work on an electronic board (microcontroller). The working logic and interactions of software modules are described in Section 3 in detail.

In the prototype, the Android-based mobile application represents mobile domain of the proposed model, J2EE-based software installed computer represents the gateway of the proposed model and temperature/light sensors represent the resource/service domain of the proposed system.

The application converts access requests into Simple Object Access Control (SOAP) envelope messages. SOAP is an XML-based messaging structure for communicating Web Services-based on Web Services Description Language (WSDL) [17].

The software in the gateway consists of five modules and these are Certificate Service, Context Engine, Decision Engine, Database Service and Data Receiver. Except the Data Receiver, other modules communicate with each other, and also with the mobile device using web services based on WSDL structure.

A. Sensor data retrieving

The microcontroller, together with the sensors mounted is connected to the gateway. The gateway detects it as a serial connection and gives it a serial port number such as "COMx". This port number is defined in the Data Receiver module of the gateway software. The Data Receiver module starts to listen to this port and after data flow starts from sensors, it detects this data and shows them in the console. After detecting, it writes these values into an external file. These values are parsed and interpreted as two different sensor values when an authorized user wants to get these data.

B. Activity flow of the prototype

Sensor data retrieving continues regularly as long as sensors work and sense the environment, therefore, this process is independent from user activity flow. Temperature sensor produces real environment temperature and light sensor measures light and gives a value between 0 and 1024 as a result of light level of environment.

The application starts with login page. The credentials on the login page are not used for authorization; they are only for program usage and can be obtained from domain administrators. After logging in successfully, the user is forwarded to the main screen of the application. In this screen, the user performs certificate transactions using two buttons. The first button "Select Certificate" forwards the user to his/her host domain to get a certificate.

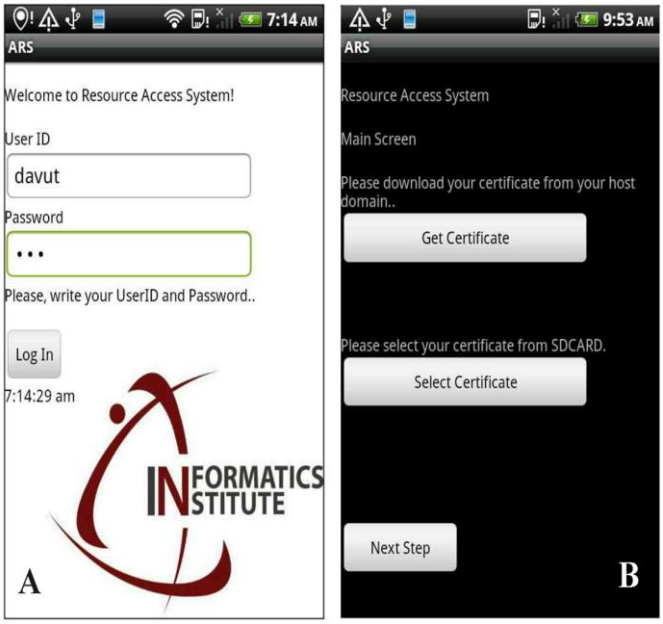


Figure 3 Login (A) and Certificate Selection (B) Interfaces of Mobile Application

The second button "Select Certificate" allows the user to select his/her certificate from the storage of the mobile device. After selecting a certificate, the application shows the certificate content for 3 seconds. If it is not possible, it gives an "unable to read file" warning. The certificate content disappears and application creates a text notification about selected certificate and its path on the SD Card. These processes are illustrated in Figure 2 (B).

The user proceeds to the final step by clicking "Next Step" and in the final interface: the application shows available resource types. According to the gateway that is connected to, the application shows which resource and resource types are available.

In the prototype implementation, available resources are sensors in the Wireless Lab of the METU Informatics Institute. This information is shown on the screen and the user selects temperature or light from drop down menu as the resource type. This selection is illustrated in Figure 3(A). With the resource type selection, the application brings together certificate data and resource type and creates SOAP access request envelope. This envelope is sent to the gateway using SOAP mechanism via the wireless connection.

 ISBN: 978-1-61208-353-7

The Decision Engine of the gateway software first receives the envelope and opens it. After opening the envelope, data is parsed, and certificate data and resource type are separated. Certificate data is sent to the Certificate Service for validity check, if certificate can pass this control, then the Decision Engine demands required context data and rules from the Context Engine. After all required data are collected, the Decision Engine performs an evaluation and makes a decision. If the user is authorized to reach temperature or light sensor data, the gateway sends related data directly to the mobile device instead of sending "allow" information only. If the user is not authorized after the evaluation process, the gateway sends "deny" response to the user's mobile device.

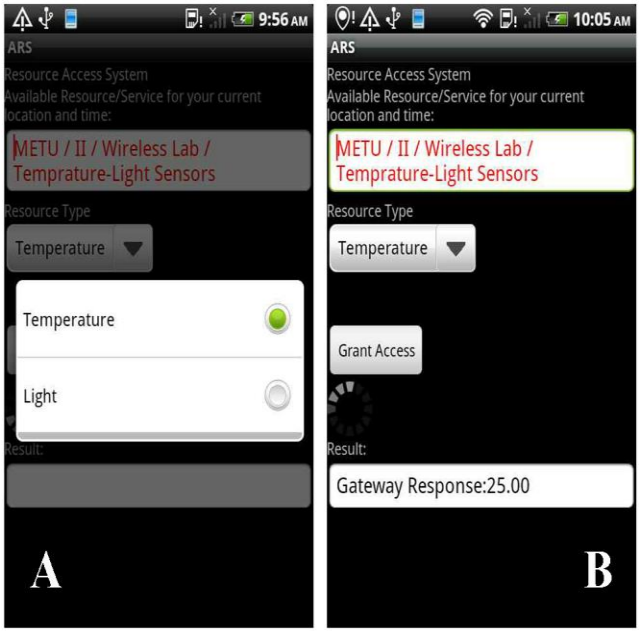


Figure 4 Resource Selection (A) and Temperature Sensor Data Response (B) Interfaces of Mobile Application

The demonstration of temperature sensor data is illustrated in Figure 3(B); also, light sensor data and access denial demonstration are illustrated in Figure 4.

C. Use cases of prototype

Different cases about trying to reach resource sensor data according to related rules and context will be analyzed. Sensors are located in the Informatics Institute (II) at Middle East Technical University (METU) and users of METU or member of user groups of METU_II and METU_CENG (Department of Computer Engineering) have different rules and privileges. Also, it is assumed that METU and Bogazici University (BOUN) have inter-domain resource usage agreement between them and users of BOUN have access to reach sensor data according to defined rules. CCL list synchronization is performed every 10 seconds. Ten different Access Rules (AR) are defined, as indicated in Table 1.

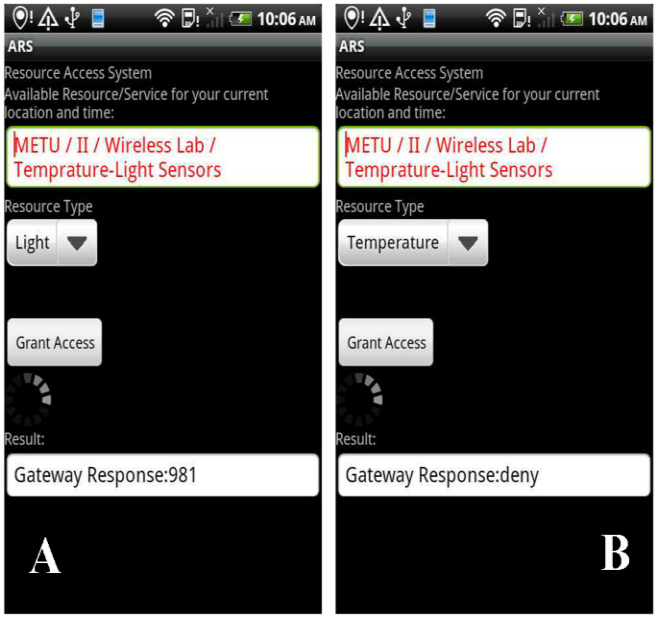


Figure 5 Light Sensor Data (A) and Access Denial (B) Interfaces of Mobile Application

TABLE 1: ACCESS RULES FOR USAGE CASES

Context	User or User Group	Resource	Response
Everytime	davut	all	allow
Everytime	serhat	light	allow
Weekend	METU_II	temp	allow
Monday	BOUN	temp	deny
Evening	serhat	temp	allow
Weekend	BOUN	all	deny
FallTerm	BOUN	light	allow
Evening	METU_CENG	light	deny
FallTerm	METU_CENG	all	allow
FallTerm	BOUN	temp	allow

D. Sample Cases

Case 1: METU domain user "serhat" wants to reach light sensor data with the following time context.
Time of request: 22.08.2011-20:30:00 (Evening, Monday)
Result: The system allows user "serhat" to access light data, because related user has two following Access Rules and first rule allows "serhat" to access light sensor data all time.

Context	User	Resource	Response
Everytime	serhat	light	allow
Evening	serhat	temp	allow

Case 2: METU_II user "ahmet" wants to reach temperature sensor data with the following time context.
Time of request: 22.08.2011-19:30:00 (Evening, Weekday)
Result: The system does not allow user "ahmet" to access light data and returns "deny" response to user , because

related user has one following Access Rule, however, related rule indicates requests with the "Weekend" time context. The time of request is not in the "Weekend" range; therefore, Context Engine does not send any rule to Decision Engine and access is not granted to the user.

Context	User Group	Resource	Response
Weekend	METU_II	temp	allow

These two and some other different usage cases are applied on the prototype implementation and they work correctly according to related rules and contexts. If the user is allowed access, the mobile application presents temperature sensor data as in Figure 3 (B), light sensor data as in Figure 4 (A) and if user access is denied, the mobile application gives "deny" response as in Figure 4 (B).

V. CONCLUSION

In this study, a certificate-based context-aware access control model using smart mobile devices for ubiquitous computing environments was presented. Using smart mobile devices for access requests and reaching resources is the major contribution of this study. In the ubiquitous computing environments, resources are distributed in the environment and in order to reach resources and use services effectively, mobile devices need to be used.

The proposed model combines three main properties of ubiquitous computing environments. The model provides a context-aware access control and smart mobile device usage and also provides inter-domain synchronization process for active certificates lists.

REFERENCES

[1] G. D. Abowd and E. D. Mynatt. "Charting Past, Present, and Future Research in Ubiquitous Computing". ACM Transactions on Computer-Human Interaction,vol. 7, no. 1, 2000, pp. 29–58,

[2] D. G. Abowd and A. K. Dey. "Towards a better understanding of context and context-awareness" 1st international symposium on Handheld and Ubiquitous Computing, London, 1999, pp. 304-307

[3] D. W. Chadwick, A. Otenko, and E. Ball. "Role-Based Access Control With X.509 Attribute Certificates" Ieee Internet Computing March- April 2003, pp. 62-69

[4] S. Ludwig, J. Pierson, and L. Brunie, "Sygn: A certificate based access control in grid environments." Tech. Report RR-LIRIS-2005-011, Laboratoire d'InfoRmatique en Images et Systmes d'information (LIRIS), 2005.

[5] T. Kindberg, J. Barton, J. Morgan, G. Becker, D. Caswell, and P. Debaty. "People, Places, Things:Web Presence for the Real World" ACM MONET (Mobile Networks & Applications Journal), 2002, pp. 365-376

[6] V. Koufi and G. Vassilacopoulos, "Context-Aware Access Control for Pervasive Access to Process-Based Healthcare Systems," eHealth Beyond the Horizon IOS Press, 2008, pp. 679-684

[7] Y. Lee, C. Min, Y. Ju, S. Pushp, and J. Song, "A Mobile Context Monitoring Platform for Pervasive Computing Environments," in 5th IEEE International Conference on Digital Ecosystems and Technologies (IEEE DEST 2011),, Daejeon, 2011, pp. 345-348

[8] M. Satyanarayanan, "Pervasive Computing:Vision and Challenges," IEEE Personal Communications August, 2001, pp. 10-17

[9] M. Thompson, W. Johnston, S. Mudumbai, G. Hoo, K. Jackson, and A. Essiari, "Certificate-based access control for widely distributed resources," in 8th conference on USENIX Security Symposium, CA, USA,1999, pp. 17-23.

[10] C.-D. Wang, L.-C. Feng, and Q. Wang, "Zero-knowledge-based user" in International Conference on Multimedia and Ubiquitous Engineering, Washington, DC, 2007, pp. 874 – 879.

[11] R. Want, "An introduction to ubiquitous computing, Ubiquitous Computing Fundamentals" J. Krumm, Ed. Redmond,Washington, U.S.A:CRC Press, ch.1, 2009, pp. 2-27

[12] M. Weiser, "The Computer for the 21st Century" Scientific American, vol. 265 no. 3, 1991 , pp. 94-104.

[13] M. Weiser and J. S. Brown, "The Coming Age Of Calm Technology," Copernicus, New York, NY, USA, 1997, pp. 75-85.

[14] B. Song, I.Yu, and J.Son, D. Baik, "An effective access control mechanism in home network environment based on SPKI certificates," Information Theory and Information Security (ICITIS), 2010 IEEE International Conference, 2010, pp. 592,595

[15] J. J. Haas, Y.-C. Hu, and K. P. Laberteaux, "Design and analysis of a lightweight certificate revocation mechanism for VANET" The sixth ACM international workshop on VehiculAr InterNETworking, New York, 2009, pp. 89-98.

[16] H. Kun, Y. Jing, D. Xiaoming, and W. Lu, "Distributed Access Control Model over Multi-trust Domain," Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on , vol. 2, 2012, pp. 595-598,

[17] Y. Kortesniemi and M. Särelä, "Survey of certificate usage in distributed access control" Computers & Security, vol. 44, July 2014, pp. 16-32

 ISBN: 978-1-61208-353-7

Machine to Machine Trusted Behaviors

Defining and Implementing Effective and Efficient Trust Mechanisms

Margaret L Loper and Jeffrey D. McCreary
Georgia Tech Research Institute
Atlanta, GA USA
{Margaret.Loper, JD.McCreary}@gtri.gatech.edu

***Abstract*—In the coming decades, we will live in a world surrounded by tens of billions of devices that will interoperate and collaborate in an effort to deliver personalized and autonomic services. Our reliance on these machine-to-machine systems to make decisions on our behalf has profound implications, and makes mechanisms for expressing and reasoning about trust essential. The Georgia Tech Research Institute recently started a strategic initiative on the Internet of Things focusing on trust. We are developing a trust framework for the machine-to-machine domain that classifies leadership functions into three dimensions. We are also developing a live, virtual, constructive platform for the design and validation of trust technologies for fully connected, ubiquitous systems. This work is in an exploratory stage, and our approach and future plans are described in the paper.**

Keywords-Internet of Things; Machine-to-Machine Systems; Trusted Behaviors.

I. INTRODUCTION

The International Telecommunication Union (ITU) predicts that there will be as many as 25 billion devices online within the next decade, outnumbering connected people 6-to-1 [1]. This will lead to a pervasive presence around us of objects and things (e.g., radio-frequency identification tags, sensors, actuators, cameras and mobile phones), which will have some ability to communicate and cooperate to achieve common goals. This paradigm of objects and things ubiquitously surrounding us is called the Internet of Things (IoT). The ITU defines IoT as a "global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on, existing and evolving, interoperable information and communication technologies" [2]. The IoT covers different modes of communication, including: between people and things, and between things (Machine-to-Machine or M2M). The former assumes human intervention, and the latter none (or very limited).

A primary aim of IoT is to deliver personalized or even autonomic services by collecting information from and offering control over devices that are embedded in our everyday lives. The reliance of IoT on simple, cheap, networked processors has implications for security; the potentially invasive nature of the information gathered has implications for privacy; and our reliance on machine-to-machine systems to make decisions on our behalf makes mechanisms for expressing and reasoning about trust essential.

While security, privacy and trust are all critical research areas for IoT, our research is focused on trust. The need for trust has long been recognized, as stated recently by Moulds in [3], the "… pivotal role in … decision making means it is essential that we are able to trust what these devices are saying and control what they do. We need to be sure that we are talking to the right thing, that it is operating correctly, that we can believe the things it tells us, that it will do what we tell it to, and that no-one else can interfere along the way."

This work provides an initial concept for trust in the M2M domain. We have completed a seedling phase of this work, which included defining the approach, testbed and use cases, with the detailed work beginning mid-summer. From our exploratory work, our main contributions to trust will be requirements for three dimensions of a trust framework, incorporating leadership functions in these dimensions as would be needed in complex M2M environments, a live virtual constructive research platform for design and evaluation of trust frameworks, and a future focus on cognitive adaptive trust, so that machines learn and recognize situations in which trust should be varied.

The remainder of this paper is organized as follows. In Section II we will define trust and its importance. Section III will briefly describe three dimensions of trust and outline our early work in developing a trust framework. Using intelligent streetlights as a platform for conducting our trust research will be described in Section IV. The remaining two sections will discuss our conclusions and future work.

II. WHY TRUST?

Trust is the belief in the competence of an entity to act dependably, securely and reliably within a specified context [4]. In M2M systems, trust is commonly accomplished using information security technologies, including cryptography, digital signatures, and electronic certificates. This approach establishes and evaluates a trust chain between devices, but it does not tell us anything about the quality of the information being exchanged among machines.

Trust is a broader notion than information security; it includes subjective criteria and experience. Trust is a human belief that someone or something is reliable, good, honest, effective, etc. Trust includes concepts, such as

- Perception – awareness of something through the senses;

 ISBN: 978-1-61208-353-7

- Memory - past history and experience; and
- Context – trust may exist in one situation, but less or not at all in another.

A key challenge is whether the human-to-human concept of trust can be extended to machine-to-machine communication. To make that extrapolation, we must define a way for machines to express and reason about trust. Expressing trust involves defining a rich language for M2M communication, including ontologies to capture the context of the environment. Reasoning about trust must consider the trust chain established among machines, as well as whether the machine is designed for the context in which the trust is required, whether it can accomplish the intended function with the desired results, and whether it has demonstrated a history of reliable performance in the intended function.

Reasoning about trust will vary over time, as machines dynamically join and leave networks. Therefore, the technical theme of our work is to develop a cognitive adaptive trust framework, focusing on core issues of M2M trust in open, decentralized systems with dynamic configuration of networks of objects. The cognitive adaptive aspects of this work are an important long-term goal, but will not be the initial focus of the work.

III. DIMENSIONS OF TRUST

There are several strategies in the literature that define trust as dimensions. Ahn et al. [5] described the concept of multi-dimensional trust by different agent characteristics, such as quality, reliability and availability. For Matei et al. [6], trust refers to the trustworthiness of a sensor, whether it has been compromised, the quality of data from the sensor, and the network connection. To address behavior uncertainty in agent communities, Pinyol and Sabater-Mir [7] define three levels of trust based on human society: security, institutional and social. Lastly, Leisterm and Schultz [8] identify technical, computational, and behavioral trust, but focus primarily on a behavioral trust indicator.

Our M2M trust framework will focus on three dimensions. These dimensions will work together to create a trusted environment in which machines can independently make decisions on behalf of humans. Our approach to defining trust dimensions is loosely based on the work described in [8] but includes aspects of leadership trust as defined by Covey [9]. This work also has some relationship to Saied et al.'s work [10] in that it considers trust in a heterogeneous IoT architecture involving nodes with different resource capabilities. The dimensions in our framework are described below.

- Technical Trust: establishing and evaluating a trust chain between devices using information security technologies. One way to describe this dimension is integrity - accuracy of algorithms, freedom from virus/malware, machine is operational, and no malfunctions or failures.
- Computational Trust: trustworthy devices that assemble data into actionable information. This dimension covers two qualities: intent and capability. Intent is whether the machine is designed for the context in which the trust is required, and whether it can be tasked with function by other machines. Capability is whether the machine(s) can accomplish the intended function with the desired results, and based on its design, is it suitable for the requesting machine's mission.
- Behavioral Trust: perception of the trustworthiness of information and devices for optimizing the mission performance. In other words, whether the machines demonstrate a history of reliable performance in the intended function.

To illustrate these dimensions, consider the operation of intelligent streetlights (iSL). Intelligent streetlights refer to public street lighting that adapts its behavior based on interactions with pedestrians, cyclists, cars and other environmental conditions. Streetlights can be made intelligent using a variety of sensors to ingest observable data, and networking technology that enables them to behave as a collaborative system. Intelligent streetlights can provide many services, but this example will focus on a simple example of adaptive lighting, where streetlights communicate with their neighbors to create dynamic lighting that follows the presence of pedestrians, bicycles and cars.

If the mission of the intelligent streetlight system is to provide lighting that adjusts based on the presence of humans, all the streetlights in a geographic area must communicate and collaborate to accomplish this mission. Trust in the streetlight system can be broken out across the dimensions as follows:

- Behavioral trust: does the intelligent streetlight system demonstrate a history of reliable performance providing adaptive lighting?
- Computational trust: do the lights turn on in the appropriate area, do they provide adequate light coverage based on speed of the vehicle or pedestrian, and can they predict when someone will reach the next streetlight? Is the light capable of detecting the presence of a vehicle or pedestrian, can it detect the speed they are traveling, can it detect when another light is not working appropriately, and is it capable of changing brightness?
- Technical trust: can the lights be turned on/off, are sensors operating properly, and is there power to operate the system?

This example is intentionally simple to convey the basic ideas of trust dimensions. If the intelligent streetlight system has multiple types of sensors (beyond motion and light) and is tasked to accomplish a variety of missions (e.g., adaptive lighting, rerouting traffic, identifying emergency situations, notifying people about emergency events or evacuations, etc.), then assessing trust along these dimensions becomes more critical.

IV. INTELLIGENT STREETLIGHTS AS A PLATFORM

In order to conduct our research, we need a problem domain with several key attributes:

 ISBN: 978-1-61208-353-7

- A variety of sensors, devices and machines that allow us to look at machine-to-machine communications;
- The ability for people to interact with the sensors and devices that allow us to look at people-to-machine communications; and
- A problem that can scale to very large numbers of machines and people in order to understand security, privacy and trust as the number of connected systems grows to the hundreds of thousands.

To design and evaluate the M2M trust framework, we will use intelligent streetlights (as described in the previous section) as a demonstration platform. Some initial use cases for evaluating trust include:

- When a pedestrian, cyclist or car is detected, it will communicate this to neighboring streetlights, which will brighten so that people are always surrounded by light.
- When a medical emergency occurs in a crowded area, streetlights can provide communication and location services to medical personnel and responders.
- When an emergency situation occurs in a geographic area, streetlights can notify pedestrians to capture information and evacuate for their personal safety.

The first step in our research will be to develop a simulation of the intelligent streetlight network in order to design and evaluate different algorithms and strategies for security, privacy and trust in fully connected systems. The simulation will be capable of representing large numbers of sensors and machines in order to look at scalability issues related to trust. The second step will be to develop an intelligent streetlight lab on the Georgia Tech (GT) campus. We are targeting a location that provides a variety of behaviors - people walking, sitting in the green space, biking, as well as car traffic. Future expansion of this system could reach further into campus, as well as downtown areas surrounding campus.

Our focus on a simulated and live environment to design and evaluate trust motivates the need for a research platform that can support Live, Virtual and Constructive (LVC) systems. The LVC categorization comes from the distributed simulation community, and refers to the way in which humans interact with simulations. Live involves real people operating real systems for simulated reasons; virtual involves real people operating simulated systems; and constructive involves simulated people (or no people) operating simulated systems [11]. We believe an LVC research platform is key to understanding the interactions and behavior between the physical and virtual world.

The iSL testbed concept is shown in Figure 1. Establishing an outdoor lab will enable us to validate the simulation with actual behavior of the system. This will be important as we begin work on scalability of trust.

The Georgia Tech Research Institute is currently working in different aspects of trust as well as cognitive reasoning, which will be leveraged to support this research. Our expertise in machine learning, modeling and simulation, systems engineering, networking and communications, autonomy, and sensors, will be required to develop the live, virtual and constructive platform to design and test cognitive adaptive trust.

V. CONCLUSIONS

Both government and commercial users/providers are trending towards significantly increasing reliance on fully automated complex M2M interactions. Our current work in unmanned systems, cyber, and complex spectrum operations require improved "trust" to achieve their full potential across acquisition/business and operational communities.

To fully realize the desired end state, we must understand the limits of what M2M missions are acceptable; how to visualize and understand trust; and acceptable mission design, execution and degradation parameters. It is also important to explore and validate the role and scope of M2M decision-making or human-in/on-the loop. Ultimately, generating trust in different dimensions will allow decision-makers to confidently invest in and employ M2M, and understand M2M self-optimization.

The work presented in this paper provides an initial concept for trust in the M2M domain. Our main contributions to trust will be well-defined requirements along three dimensions. Understanding the relationship of trust functions to leadership roles will be needed in complex M2M environments. We will also develop a live virtual constructive research platform for design and evaluation of trust frameworks. This LVC environment will connect the physical and virtual worlds, thereby enabling us to define and implement efficient trust mechanisms beyond our demonstration platform. A future focus will be on cognitive adaptive trust, so that machines learn and recognize situations in which trust should be varied.

VI. FUTURE WORK

After GTRI demonstrates M2M trust at technical, computational and behavioral levels in simple constructs, the goal is to demonstrate scalability, as well as performance and effectiveness in increasingly complex systems and scenarios. One desired future goal is to demonstrate fully translating and implementing human intent into M2M cognitive adaptive, "creative" execution.

ACKNOWLEDGMENT

We would like to thank the Georgia Tech Research Institute chief scientists for funding this work.

REFERENCES

[1] International Telecommunication Union, "The State of Broadband: Achieving Digital Inclusion for All," Broadband Commission for Digital Development technical report, September 2012.

[2] International Telecommunication Union, Recommendation ITU-T Y.2060 "Overview of the Internet of Things," June 15, 2012.

[3] R. Moulds, "The internet of things and the role of trust in a connected world," The Guardian, January 23, 2014. Available from: http://www.theguardian.com/media-network/media-network-blog/2014/jan/23/internet-things-trust-connected-world. [retrieved: June 2014]

 ISBN: 978-1-61208-353-7

[4] T. Grandison and M. Sloman, "A survey of trust in internet applications," IEEE Communications Surveys and Tutorials, vol. 3, issue 4, 2000, pp. 2-16.

[5] J. Ahn, D. DeAngelis and S. Barber, "Attitude driven team formation using multi-dimensional trust," Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT '07), Nov. 2007, pp. 229 –235.

[6] I. Matei, J. Baras and T. Jiang, "A composite trust model and its application to collaborative distributed information fusion," Proceedings of the 12th International Conference on Information Fusion (FUSION 2009), July 2009, pp. 1950 –1957.

[7] I. Pinyol and L. Sabater-Mir, "Computational trust and reputation models for open multi-agent systems: a review," Artificial Intelligence Review, July 2011, pp. 1–25.

[8] W. Leisterm and T. Schultz, "Ideas for a Trust Indicator in the Internet of Things," Proceedings of the First International Conference on Smart Systems, Devices and Technologies (SMART 2012), IARIA, May 2012, pp. 31-34.

[9] S. Covey, The Speed of Trust, Free Press, 2008.

[10] Y. Saied, A. Olivereau, D. Zeghlache, and M. Laurent, "Trust management system design for the Internet of Things: A context-aware and multi-service approach," Computers & Security, vol. 39 part B, Nov 2013, pp.351-365.

[11] A. Henninger, D. Cutts, M. Loper, R. Lutz, R. Richbourg, R. Saunders, and S. Swensen, "Live Virtual Constructive Architecture Roadmap (LVCAR) Final Report", M&S CO Project No. 06OC-TR-001, Sept 2008. Available from: http://www.msco.mil/LVC.html. [retrieved: June 2014].

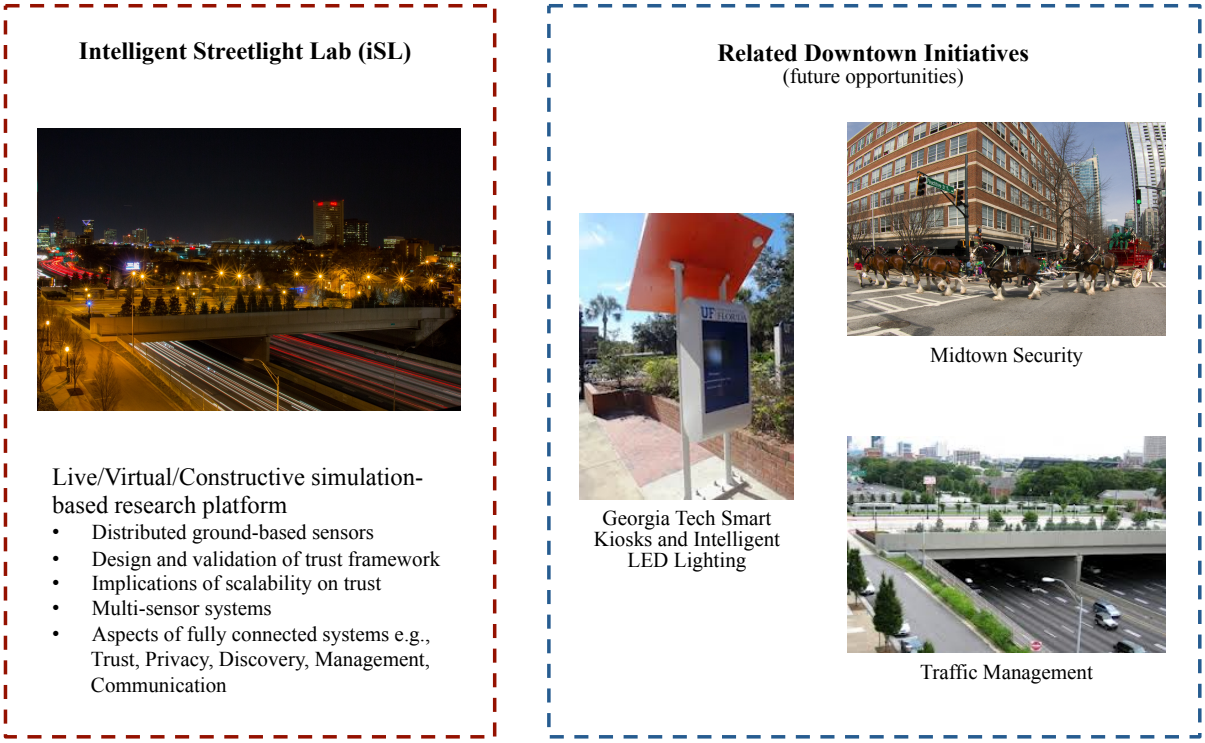


Figure 1. Intelligent Streetlight Laboratory

The Pervasive Information System Adaptation: Android Device Context

Fatma Achour, Anis Jedidi, Faiez Gargouri

MIRACL
Multimedia, Information Systems and Advanced Computing Laboratory
Sfax University, Tunisia

Email: fatma.achour@gmail.com, anis.jedidi@isimsf.rnu.tn, faiez.gargouri@isimsf.rnu.tn

***Abstract*—The conceptual step in the pervasive adaptation system is mostly aimed at the success of such a system. It is for this reason that the development of a conceptual adaptation system helps developers to implement their adaptation systems. Accordingly, several categories of contextual information can be presented in a pervasive information system; as such, the network context, the location context, the service context, the application context, the device context, and the person context. Again, a detailed description of each of these contextual information categories allows achieving a better adaptation of the applications to the contexts of use. However, the device context shows the focal point of any information system. In fact, the adaptation of the pervasive information systems, when using contexts, rests on the existing physical constraints in this context. In this paper, we present the architecture of adapting the applications to the pervasive system based on the semantic web services. Specifically, we are interested in the adaptation of the device context. The device contextual data are collected by using an Android program. The adaptation rules are also created using the Jena toolkit.**

Keywords–Adaptation; Model; Android Device; Jena rules; RDF; OWL; OWL-S.

I. INTRODUCTION

The information systems evolution is correlated with telecommunication as well as connectivity hardware and software development. These types of information systems are designed to create a transparent and an inter-operable environment to ensure better information shared between the various types of information systems with heterogeneous information resources.

Thanks to the technological developments and the new technologies integration in all applications of everyday life, as a matter of fact, connectivity enhanced accessibility to the resources. This progress has enormously given the user a free interaction to access the different resources anywhere, anyhow and at any time: the systems have consequently become pervasive and ubiquitous.

The pervasive or the ubiquitous systems are actually designed to make information available anywhere and at any time. These systems, however, must be used in different contexts according to the user's environment and profile as well as the used terminal. One of the major problems of the pervasive system is the adaptation of the applications to the user's situation [1]. Several research works [1][2], indeed, are dedicated to seek out a solution to this problem and to develop an adaptation framework. Nevertheless, the research works in the pervasive system conceptual adaptation domain are too limited. Every researcher in this field seeks to develop a system allowing the implementation of adaptation without taking into consideration the design phase of this adaptation. In this paper, we suggest a complete, a generic and a scalable architecture to conceptually adapt the application to the user's situation. In this, we mainly focus on the Android devices adaptation in the user's situation based on the semantic web services creation. To do this, we propose a generic model to design the device context in the pervasive system. We also integrate the proposed model in the semantic web service description or structure. Eventually, we define a set of rules that can be applied to this description.

In this paper, we present in the first part a state of the Art on the pervasive computing. In the first section, we present the pervasive system adaptation. In the second section, we describe the web service semantic description structure OWL-S Ontology. In the second part, we present our proposed framework. The first section of this part is dedicated to present the proposed description of the device context. The following section deals with the use of this description to adapt the applications in the pervasive system.

II. STATE OF THE ART

Context modeling is a vital aspect in pervasive computing. Because context-aware applications must be adapted to the changing situations, they need a detailed model of the user's activities and entities in the surroundings that let them share the user's perceptions of the real world. One of the basic steps in the development of context-aware applications is, therefore, to provide a formalized representation and standardized access mechanisms to the context information.

In this paper, we present the existing works to adapt the application to the use's context in the pervasive system and we show the existing structure to integrate the contextual information in the semantic web service description.

A. The pervasive system adaptation

Several research works are all for the adaptation of the application to the pervasive system. Since we are inclined to the conceptual adaptation, we present the conception phase (the second phase in the software engineering lifecycle) of two existing adaptation works: SECAS (Simple Environment For Context-Aware Systems) [1] and COCA (A Collaborative Context-Aware Service Platform for Pervasive Computing) [2].

 ISBN: 978-1-61208-353-7

The context models that use the markup scheme approaches are commonly used for the profile data representation. This type of model is used by several research works. Among these works, we can cite the COCA [2] and the SECAS [1] platforms. The COCA platform proposes a semantically rich model for collaboration, representation and context management [2]. It uses a contextual model representation based on a hybrid approach using ontology and relational databases (HCoM/EHRAM). EHRAM is a conceptual context representation meta-model and HCoM is a hybrid model that uses the components of EHRAM in ontology and relational schema. The ontology part represents the semantic aspect of the context data and the relational schema represents the context data itself. The EHRAM model includes: person context, device context, physical environment context, network context, activity context, service context, and location context.

The limitations of this work are inherent in the separation of the various categories of contextual information. For example, there are no semantic relationships between the two contexts: the personal context and the location context. However, we present the relation"locatedIn" between these two contexts (the context personal context is located in the location context).

SECAS attaches a great importance to the context management without showing how to modify the behavior of the application to adapt the context [3]. The application use context is defined as a five-dimensional vector: terminal, communication, user, location, and environment. To store the context parameters, it uses an XML representation based on the CSCP model [4].

In SECAS work, we note the absence of a design general model to conceive the information used to adapt the application in the pervasive system. Also, the adaptation is based on the Petri Nets representation which describes the services of the application and their dependencies. Therefore, the adaptation is functional but is not semantic.

B. The OWL-S Ontology

In the second part of this state of the art, we present the existing works so as to integrate the contextual information in the semantic web service description. In the first place, we present the used structure to describe the semantic web service: OWL-S. In the second place, we present the existing OWL-S extension to integrate the contextual information in order to provide an adapted semantic web service.

1) The OWL-S Ontology Presentation: OWL-S is a Web Services ontology that specifies a conceptual framework to describe the semantic web services. OWL-S is also a language based on the DARPA work of its DAML program and takes the result of DAML-S (DARPA Agent Markup Language Service). It was incorporated into W3C in 2004 within the interest group of semantic web services at the OWL recommendation [5].

The initial purpose of OWL-S is to implement the semantic web services. OWL-S is based on OWL to define the abstract categories of entities and events in terms of classes and properties. OWL-S uses this ontology language description to define a particular ontology for the web services. This ontology is used to describe the web service properties as well as its services available to the public. The OWL-S structure regroups a set of ontology. Each one provides a functionality to describe the web service semantically. The ontology main classes described by OWL-S [6] are defined by the following figure (see Figure 1).

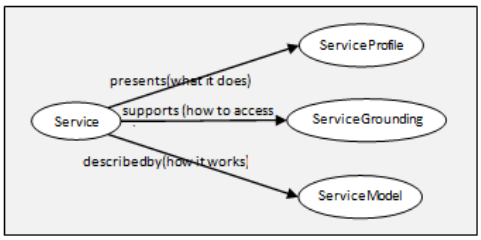


Figure 1. The principal OWL-S classes.

The necessity to use the OWL-S ontology is justified by the creation of a semantic web service that has a dynamic description. This dynamic is provided by the addition of contextual descriptions to the OWL-S structure. The description depends on the use of the context of a pervasive system.

2) The existed OWL-S extension: Several research works take the advantage of the existing OWL-S structure to describe the different contexts. In this paper, we present two research works of Qiu et al. [7][8] and Ben Mokhtar [9]. In [7][8], the authors propose an adaptation system based on the service composition approach. To do this, Qiu et al. [8] offer three context categories : the user's context, the web service context and the environmental context.

The user's context ("U-Context") specifies the context information about the user. In this context, the authors defined two types of contextual information: the user's static context (profile, interest, and preferences) and the user's dynamic context (location, current activity and task trying to achieve). The web service context ("W-Context") includes the not functional contextual information (price, execution time, confidence degree). The environmental context ("E-Context"): this category collects the context information about the user's environment (time, date, etc.).

Each context category is represented by the OWL ontology and is integrated in the existing OWL-S extension ontology to introduce the OWL-SC (OWL-S for context) [8]. The latter is intended to describe a general contextual information (see Figure 2) based on the users' description.

The proposed structure focuses only on the user's context description. However, it presents a vision for the integration or the addition of more information to the OWL-S structure. Ben Mokhtar et al. [9] research works propose a system to adapt the web services to a pervasive environment [9]. The context definition includes the description of four types of contextual information: the context sensors, services, devices and users.

In addition, the contextual adaptation in this work is based on the service representation and the user's task representation. In the service representation, the authors describe the services using OWL-S extended with context information. This information is divided into a high level context attributes,

 ISBN: 978-1-61208-353-7

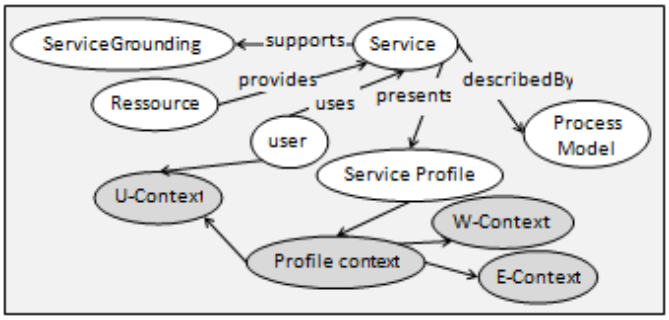


Figure 2. The OWL-SC Ontology.

preconditions and contextual effects. However, via the user's task representation, the user's task representation is performed while extending the OWL-S service model ontology (see Figure 3). To do this, Ben Mokhtar et al. [9] propose to integrate the quality conditions service descriptions and the context conditions required by the user's task in the OWL-S structure.

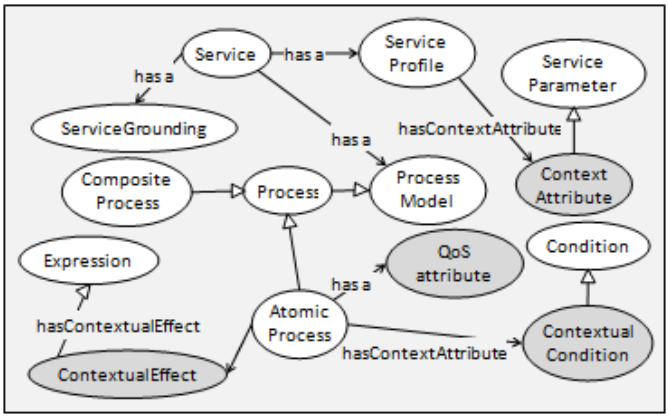


Figure 3. The OWL-S ontology extension for the pervasive system.

III. THE PROPOSED FRAMEWORK

In the proposed framework, we are interested in the semantic web service to adapt the application to the pervasive system. The target audiences of our proposed framework are the developers in the pervasive system. Therefore, we proposed two description levels (see Figure 4):

- **The generic level:** At this level, we create a semantic web service. In this semantic web service description, we create OWL-S ontology. In this ontology, we integrate the highest level of the pervasive system description. This level is created to collect the context description corresponding to the user's needs and to apply the user's rules to the collected information.
- **The specific level:** At this level, we create a set of six semantic web services corresponding to the pervasive information categories: network, device, user, location, service and application. In each semantic web service, we integrate the context description. These semantic web services are created to instantiate and to describe the user's situation.

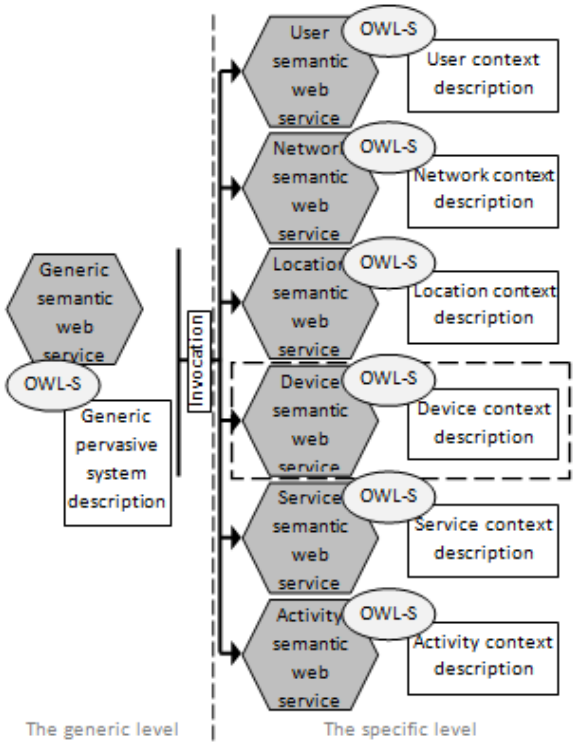


Figure 4. The purposed architecture levels.

The goal of the information separation in the specific level is to facilitate the conceptual adaptation by the use of the most appropriate contextual information and to reduce the search time by the specific semantic web service invocation used in the adaptation phase. In this paper, the device context is of a growing interest.

A. The proposed description framework: the generic and device contexts

The pervasive information system entails six information categories (device, network, activity, service, location and user) [10]. In our proposal, we distinguish two levels of description. The first level presents the generic level and the second level presents the specific level (see Figure 4). In the generic level, we create a generic semantic web service. In the OWL-S description structure of the generic web service, we integrate the generic OWL description. In the specific level, we create a semantic web service for each pervasive information system categories. Also, for each category, we create a semantic web service. In the specific semantic web service description, we integrate the information system category OWL description in the OWL-S structure.

1) The proposed description of the generic level: The generic semantic web service generally describes the pervasive environment. It has all the necessary information about each web service context shown in the second level. This information is described in an extended OWL-S ontology (see Figure 5). The extended OWL-S ontology includes the basic information examining a pervasive system.

The pervasive context is presented by the "PervasiveContext" OWL class. The activities in a pervasive system are

presented by the "A-Context" OWL class. They exist in a device. The latter is symbolized by the "D-Context" OWL class. Each device ("D-Context") offers services in a pervasive system. The services are presented by the "S-Context" OWL class. The latter regroups the characteristics of the services provided by the pervasive system. All the devices existing in the pervasive works are modeled through the "N-Context" OWL class.

The two classes "D-Context" and "U-Context" represent the entire agent that exists in a pervasive system. For this reason, we position the two classes as a sub-class of the» Agent" OWL class.

Each of the classes presented in the proposed OWL-S structure will be transformed into ontology. The latter regroups the classes and the attributes shown by the OWL semantic relation "owl:onProperty". The ontological structures are used to detail the contexts defined in the pervasive system.

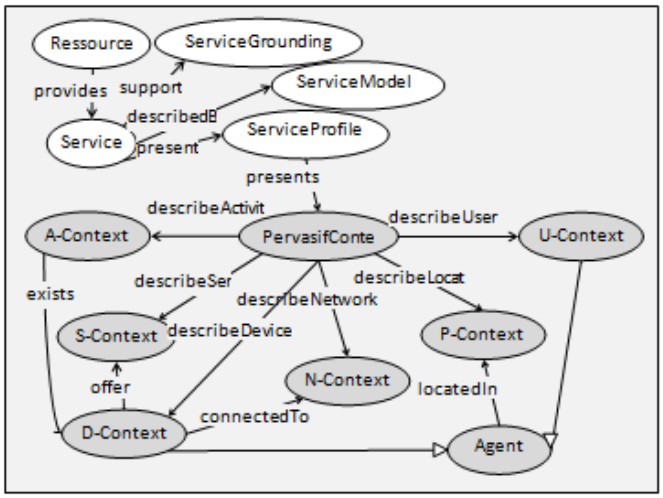


Figure 5. The proposed structure for the first level of description.

2) The proposed description of the device context: The device side is very significant in the pervasive information system since the pervasive system is accessible anywhere, anyhow and to anything. Indeed, such a system can be executed according to the existing hardware.

The OWL ontology is created to describe the device context and the activity profiles in the pervasive information system. Each device has a configuration, rules and preferences. In the mobile system, a new communication method has emerged to satisfy the user's needs. Such a method paves the way for the propagation of intelligent systems through the invention of smart phones, such as "blackberrys", "iPhone", and touch pads, such as the "iPad". In order to ensure adaptation, the pervasive system must capture a material characteristic to ensure the answer to the query in accordance with the hardware configuration. To ensure the generality of the proposed device context model we define a "key" and "values" properties for each class. These properties can catch any value depending on the user's contextual situation.

This ontology is inserted in the OWL-S structure. The original purpose of OWL-S is to implement the semantic web services. OWL-S is based on OWL to define the abstract categories of entities and events in terms of classes and properties. OWL-S uses this ontology language description to define a particular ontology for the web services. This ontology is used to describe the web service properties as well as its services available to the public. The OWL-S structure regroups a set of ontology. Each one provides a functionality to describe the web service semantically. The ontology main classes described by OWL-S are defined by the following figure (see Figure 6).

The necessity to use the OWL-S ontology is justified by the creation of a semantic web service that has a dynamic description. This dynamic is provided by the addition of contextual descriptions to the OWL-S structure. The description depends on the use of the context of a pervasive system. To integrate the device context ontology, we use the two classes "service" and "service profile".

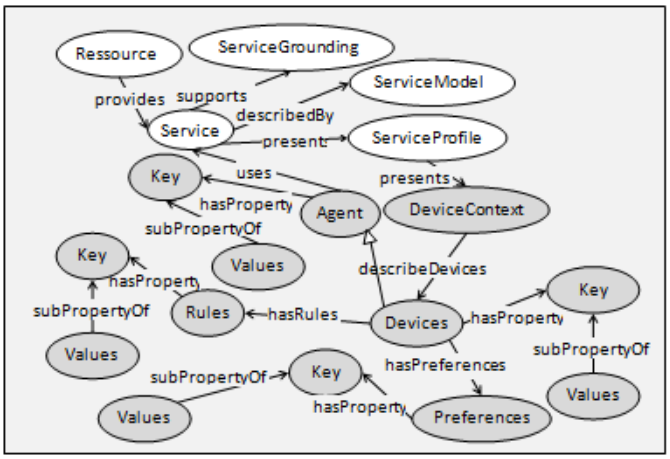


Figure 6. The proposed structure for the device context.

3) The data properties used in the device context: In this section, we present a list of properties and sub-properties (see. Table 1) used to describe the Android device context in the pervasive system. These properties are used as an instance of the presented "key" and "value" properties (see Figure 6). This list is detected automatically using Android programs and submitted to the device semantic web service using SOAP. This list is used to create an RDF instance to the device context in conformity with the presented OWL-S structure (see Figure 6). This service is proposed by the device semantic web service. Since we propose two generic properties in the created model in our proposed framework, we present a Java Server Page to add other properties to the created RDF device instance and to the created OWL-S extension. This Java Server Page is designed for the developer in the pervasive system representing the target audience of our proposed framework.

B. The proposed adaptation framework

In the proposed adaptation framework, we present two different works. The first work is concerned with the proposal of the adaptation phase and communication between the generic and the device context. The second work is about applying the rules to the created models which offer a complete conceptual adaptation system.

TABLE I. THE DEVICE CONTEXT PROPERTIES.

Classes	Sub-properties	Properties
Agent	AgentCharact	AgentType AgentName
Devices	BatteryCharact	Voltage Temperature Technology Status Scale Presence Plugged Level Health
	DBluetoothChar	DBluetoothAdress DBluetoothName
	DeviceCharact	DeviceType ScreenBrightness ScreenDimension DeviceName CPUSpeed
	LBluetoothChar	LBluetoothAdress LBluetoothName
	Volume	MaxVolume ActualVolume VolumeMode
	MemoryCharact	TotInteMem AvailInteMem AvailExteMem TotExteMem TotalRAMMem AvailRAMMem
	CameraCharact	CameraNumber CameraSize CameraEncodFmt
Preferences	Language	LanguageCountry UsedLanguage
Rules	SupScreenMode	ScreenModeValue SupResolution
		ResolutionValue
	SupPictureFormat	PictFormatVal
	SupAudioFormat	AudioFormatVal
	SupVideoFormat	VideoFormatVal

1) The generic and the device semantic web services communication: The proposed architecture is made up of two fundamental parts (see Figure 7). The first one consists in developing a generic semantic web service and the second one regroups a set of specific web services corresponding to the pervasive information system categories [10]. We develop a set of semantic web services using an OWL-S structure. In the generic semantic web service case, we integrate the semantic relation between the different specific semantic web services. For each one of the pervasive information system categories (person, device, network, service, application and location), we create a specific semantic web service. Also, we integrate their classes into the OWL-S description which corresponds to their semantic web service.

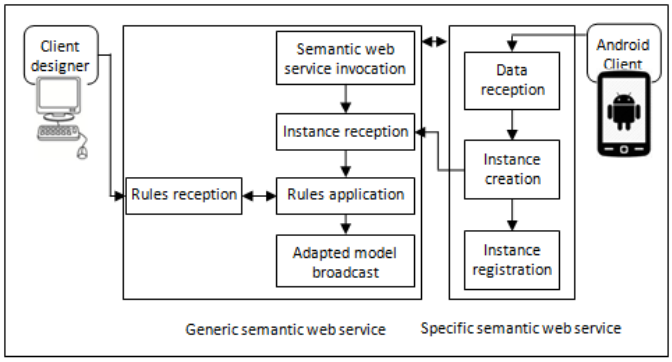


Figure 7. The purposed Framework Architecture.

The value of the generic web service creation resides in ensuring the specific web services communication by regrouping their created instances and applying the user's rules to the latter. The specific web services are created in an attempt to receive the Android device characteristics and to create an RDF instance in accordance with the OWL-S description.

2) The semantic web service invocation: The classical web service architecture is composed of three elements. The first element represents the user. The second element stands for the provider. The last element is the registry. The interaction between the three elements is ensured by SOAP. The purpose of the semantic web service is to integrate the ontology description in the OWL-S web service description. In this section, we present the interaction between the Android client, the generic semantic web service client and the specific semantic web service (see Figure 8).

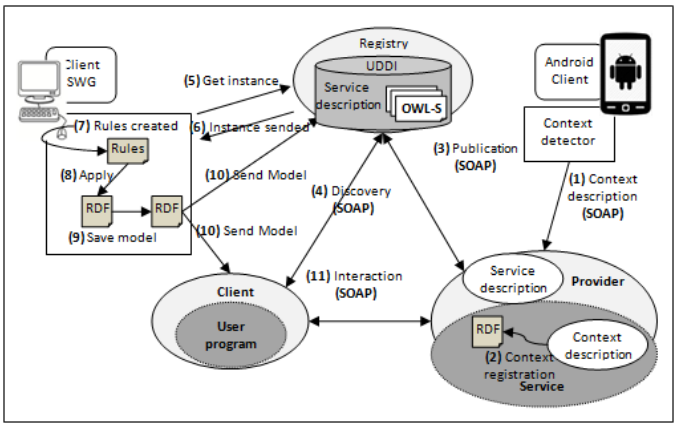


Figure 8. The proposed semantic web service architecture.

(1) To participate in the adaptation framework, the Android devices must send their characteristics to the specific semantic web service.

(2) Based on the received information from the Android device, the provider in the specific semantic web service creates an RDF instance conforming to the OWL description inserted in OWL-S description.

(3) The provider publishes the service and the created instance in the registry using OWL-S description.

(4) The client discovers the service and the created instance.

(5) The generic web service client requests the created instance from the registry.

 ISBN: 978-1-61208-353-7

(6) The registry sends the requested instance to the generic web service client.

(7) The generic web service client creates his rules using a jsp page.

(8) The generic web service client applies the created rules to the created instance.

(9) The generic web service client saves the adapted model.

(10) The adapted model is sent to the registry and to the specific web service client.

(11) The interaction between the client and the provider is started again.

In the next section, we will present the pervasive information system adaptation framework to the device context.

3) The Jena rules: To validate the proposed instance, we propose to execute two Jena rules in the created RDF instance (see Figure 9). The result of the rules execution is a model used by the developers in the pervasive system domain to ensure the adaptation.

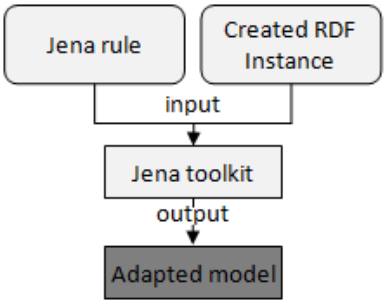


Figure 9. The Jena input/output.

We decide to use Jena [11] because it defines a java package that can manipulate the files in any of the standard RDF storage formats. Additionally, Jena can store and read RDF data in a relational database. Jena offers a statement-centric (based on the subject-predicate-object structure) support to manipulate the RDF and OWL data, and comes with a built-in RDF query language, SPARQL. Jena provides a programmatic environment for RDF, RDFS, OWL and SPARQL and includes a rule-based inference engine. In the next section, we will present an example of two Jena rules.

Rule 1: If the battery level is lower than 50% and the screen brighteness equal to 100%. The value of this latter is changed to 30%.

The Rule description:

$$((Battery.Level(50) \wedge DeviceCharacteristics.ScreenBrightness(100)) \Longrightarrow (DeviceCharacteristics.ScreenBrightness(50) \wedge Volume.ActualVolume(5.0)))$$

The Jena code:

```
@prefix rdf:
http://www.w3.org/1999/02/22-rdf-syntax-ns#
@prefix perSys:
http://example.org/PervasiveSystem#
@prefix xs:
http://www.w3.org/2001/XMLSchema#
[PreferredVolume:
(?d ?rdf:type ?t1),
(?c ?rdf:type ?t2),
(?c perSys:Level ?a)
lessThan(?a, 50)
(?c perSys:ScreenBrightness "100%")
(?c perSys:ActualVolume "15.0")
->
(?c perSys:ScreenBrightness "30%")
(?c perSys:ActualVolume "5.0")
```

The execution result of the second rule presents the RDF model as described below. The RDF model regroups the device characteristics after the execution rule where the screen brightness equals to 100% and the battery level equals to 50%. Also, the execution result of this rule is to change the screen brightness value into 30%.

```
<perSys:Device> <rdf:Description
rdf:about=
"http://PervasiveSystem
#DeviceCharacteristics">
<perSys:CPUSpeed
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#float"
>1.0</perSys:CPUSpeed>
<perSys:ScreenBrightness rdf:datatype=
"http://www.w3.org/2001/XMLSchema#float"
 >30</perSys:ScreenBrightness>
...
<perSys:DeviceName
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#string"
>Unknown sdk</perSys:DeviceName>
</rdf:Description>
</perSys:Device>
...
<perSys:Device>
<rdf:Description
rdf:about=
"http://PervasiveSystem
#BatteryCharacteristics"> <perSys:Voltage
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#int"
>0</perSys:Voltage>
...
<perSys:Presence rdf:datatype=
"http://www.w3.org/2001/XMLSchema#string"
>true</perSys:Presence>
<perSys:Plugged
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#int"
>1</perSys:Plugged>
 <perSys:Level
rdf:datatype=
```

```
"http://www.w3.org/2001/XMLSchema#int"
>50</perSys:Level>
<perSys:Health
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#int"
>2</perSys:Health>
</rdf:Description>
</perSys:Device>
<rdf:Description
rdf:about="http://PervasiveSystem#Volume">
<perSys:MaxVolume
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#string"
>15.0</perSys:MaxVolume>
<perSys:ActualVolume rdf:datatype="http:
//www.w3.org/2001/XMLSchema#string"
>5.0</perSys:ActualVolume>
<perSys:VolumeMode
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#string"
>Normal mode</perSys:VolumeMode>
</rdf:Description>
</perSys:Device>
```

The device context does not present the totality of information included in the pervasive system. In fact, the pervasive system presented a collection of six categories of contextual information. We presented in previous work our design of each category in this paper we present a rule applied to two contextual information categories: the device context and the location context.

Rule 2: The devices located in "Tunisia, ISIM Gabes, Amphi 1", the preferred Volume Mode must be changed from "Normal Mode" to "Silent Mode".

The Rule description:

$$((Adress.CountryName(\text{"}Tunisia\text{"}) \wedge Adress.Region(\text{"}Gabes\text{"}) \wedge Adress.FeatureName(\text{"}ISIM\text{"}) \wedge Adress.SpecificLocation(\text{"}Amphi1\text{"}) \wedge Volume.VolumeMode(\text{"}Normalmode\text{"})) \Longrightarrow Volume.PreferedMode(\text{"}Silentmode\text{"}))$$

Jena code:

```
@prefix rdf:
http://www.w3.org/1999/02/22-rdf-syntaxns#
@prefix perSys:
http://PervasiveSystem#
@prefix xs:
http://www.w3.org/2001/XMLSchema#
[PreferredLocation:
(?d ?rdf:type ?t1),
(?c ?rdf:type ?t2),
(?c perSys:CountryName "Tunisia"),
(?c perSys:Region "Gabes"),
(?c perSys:FeatureName "ISIM"),
(?c perSys: SpecificLocation "Amphi1")
(?h ?rdf:type ?t3),
(?h perSys:DeviceName ?r)
->
(perSys:DesignedDevice ?rdf:type ?t3)]

@prefix rdf:
http://www.w3.org/1999/02/22-rdf-syntaxns#
@prefix perSys:
http://PervasiveSystem#
@prefix xs:
http://www.w3.org/2001/XMLSchema#
[PreferredVolumeMode:
(?v ?rdf:type ?t3),
(?w ?rdf:type ?t4),
(?w perSys:DeviceName ?r)
(?y ?rdf:type ?t5),
(?y perSys:VolumeMode "Normal mode")
->
(?y perSys:PreferedMode "Silent mode") ]
```

The execution of the second rule aims to define two Jena rules: "PreferredLocation" and PreferredVolumeMode". The first one permits to know the devices name located in "Amphi1" and the second one permits to define the desired volume mode ("Silent Mode").

```
<perSys:Location>
<rdf:Description
rdf:about="http://PervasiveSystem#Adress">
<perSys:FeatureName>ISIM
</perSys:FeatureName>
<perSys:Region>Gabes
</perSys:Region>
<perSys:CountryName>Tunisia
</perSys:CountryName>
<perSys:SpecificLocation>Amphi1
</perSys: SpecificLocation>
<perSys:CountryCode>TN
</perSys:CountryCode>
</rdf:Description>

...
<rdf:Description
rdf:about="http://PervasiveSystem
#DesignedDevice"> Samsung GT-S5360
</rdf:Description>
</perSys:Location>
<perSys:Device>
<rdf:Description
rdf:about=
"http://PervasiveSystem
#DeviceCharacteristics">
<perSys:CPUSpeed
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#float"
>1.0</perSys:CPUSpeed>
<perSys:ScreenBrightness rdf:datatype=
"http://www.w3.org/2001/XMLSchema#float"
```

```
>30</perSys:ScreenBrightness>
...
<perSys:DeviceName
rdf:datatype=
"http://www.w3.org/2001/XMLSchema#string"
> Samsung GT-S5360</perSys:DeviceName>
</rdf:Description>
<rdf:Description
rdf:about=
"http:PervasiveSystem#Volume">
<perSys:PreferedMode> Silent mode
</perSys:PreferedMode>
<perSys:MaxVolume>15.0
</perSys:MaxVolume>
<perSys:ActualVolume>3.0
</perSys:ActualVolume>
<perSys:VolumeMode>Normal mode
</perSys:VolumeMode>
</rdf:Description>

...
</perSys:Device>
```

IV. CONCLUSION

The most detailed description or modeling provides an accurate level of adaptation. In this paper, we present a pervasive information system adaptation using an Android device. In order to do this, we created a model including all the contextual information necessary to adapt the application to the device context. In fact, we proposed an extensible model to describe the device context through the OWL-S extension. Also, we defined an important number of contextual information.

Moreover, we tried to define a JSP interface to add more data to the suggested extensible model. In this paper, we propose a complete approach to provide a framework to adapt applications to the device context. In the future work, we seek to complete the proposal of the pervasive information system adaptation framework gathering all contextual information categories (user, device, network, application, location and service).

REFERENCES

[1] T. Charri, "Adaptating pervasive applications in multi-contextual environments," Computer Science, National Institute of Applied Sciences, LIRIS, September 2007.

[2] D. E. Dedefa, "Context modeling and collaborative context-aware services for pervasive computing," Computer Science, National Institute of Applied Sciences, INSA de Lyon, December 2007.

[3] T. Chaari, F. Laforest, and A. Celentano, "Adaptation in Context-Aware Pervasive Information Systems: The SECAS Project," Int. Journal on Pervasive Computing and Communications(IJPCC), vol. 3, no. 4, Dec. 2007, pp. 400–425, Available: http://liris.cnrs.fr/publis/?id=2996 Retrieved: June, 2014.

[4] T. H. Sven Buchholz and G. Hübsch, "Comprehensive structured context profiles (cscp): Design and experiences," in Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, ser. PERCOMW '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 43–, Available: http://dl.acm.org/citation.cfm?id=977405.978623 Retrieved: June, 2014.

[5] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara, "Owl-s: Semantic markup for web services." Internet [http://www.w3.org/Submission/2004/SUBM- OWL-S-20041122/], 2004, Available: http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/ Retrieved: June, 2014.

[6] M. D, M. Burstein, M. D, McIlraith, M. Paolucci, K. Sycara, M. D, M. Burstein, M. D, M. S, M. Paolucci, K. Sycara, M. D, E. S, and N. Srinivasan, "Bringing semantics to web services with owl-s," World Wide Web Journal, vol 10, no. 3, pp. 243-277, no. CMU-RI-TR-, August 2007.

[7] L. Qiu, Z. Shi, and F. Lin, "Context optimization of ai planning for services composition," in ICEBE. IEEE Computer Society, 2006, pp. 610–617.

[8] L. Qiu, L. Chang, F. Lin, and Z. Shi, "Context optimization of ai planning for semantic web services composition," Service Oriented Computing and Applications, vol. 1, no. 2, 2007, pp. 117–128.

[9] S. B. Mokhtar, D. Fournier, N. Georgantas, V. Issarny, S. B. Mokhtar, D. Fournier, and N. Georgantas, "Context-aware service composition in pervasive computing environments," in RISE, vol. 3943. Springer, 2005, pp. 129–144.

[10] F. Achour, A. Jedidi, and F. Gargouri, "A two-leveled semantic web service description of the pervasive information system." in MobiWIS, ser. Lecture Notes in Computer Science, F. Daniel, G. A. Papadopoulos, and P. Thiran, Eds., vol. 8093. Springer, 2013, pp. 35–48.

[11] B. McBride, "Jena: a semantic web toolkit," IEEE Internet Computing, vol. 6, no. 6, 2002, pp. 55–59.

 ISBN: 978-1-61208-353-7

Formal Modeling For Pervasive Design of Human-Computer Interfaces

Ines Riahi, Faouzi Moussa
National School of Computer Sciences, Cristal Laboratory
University of Manouba, Tunisia
Emails: {ines.riahi@yahoo.fr, faouzimoussa@gmail.com}

***Abstract*—The advent of mobile interfaces induces an evolution on the Human Computer Interaction (HCI) field. We observe the emergence of several mobile devices and sensors that gave birth to the ubiquitous environments. In our research, we focus on: (i) how to adapt the interface to its environment, specifically in its context of use and (ii) what relationship has the context with the users' task. This paper will propose a formal approach for specifying user interfaces adapted to the context of use. We will focus on the strength of formal approach to context and user's modeling and how to infer users' requirements through the model of the task for critical domains. Our approach will be illustrated by a case study on the monitoring of diabetic patients.**

Keywords-pervasive user interfaces; ubiquitous computing; formal modeling; critical domains.

I. INTRODUCTION

Ubiquitous environment is a physical space in which technology is seamlessly integrated in order to assist the user in performing tasks to reach its goals more conveniently.

Ubiquitous environments are often considered highly dynamic environments and the contextual information can change at runtime. User interfaces should provide the right information for the right person at the right time [1]. In order to cope with such a complexity, new methods need to be developed. The research field of HCI has introduced many techniques for interaction design that are partially suitable for ubiquitous environment. However, it has an enormous rate of environmental information and the user's task becomes difficult to identify.

The specification of user interface adapted to the context of use presents several problems. The consideration of the user's task represents an important criterion in an environment where the context has a direct impact on the user's task. The users' interface must change according to the context and task at a specific moment.

Modeling ubiquitous environment and user's task poses several issues. In fact, pervasive environments are extremely dynamic and hold a vast amount of information. In critical domains, such as health, nuclear and transport systems, modeling pervasive system must be rigorous with minimal percentage of error risk. If the user's interface shows wrong information, it will have a disastrous impact on the user's task. So, the use of formal approach in such domains is required to guarantee a valid interface since the modeling stage. To tackle these problems, we study, in the second section, different related works in the literature of context-aware systems. In the third section, we study the state of the art of modeling context and user's task based on Petri Nets (PN). Then, we introduce, in the fourth section, our formal approach for specification of adapted user interface to the context of use. We will focus on the advantage of the formal model by representing the relation between the context's model and the user's task and illustrating how to deduce the users' requirements from the task models. This approach will be illustrated, in the fifth section by a critical case study on the monitoring of diabetic patients in a smart hospital.

II. RELATED WORK

Researchers in the context adaptation area have not introduced a generic and pragmatic definition of the notion of context. Following the study of the main definitions, we have concluded that the majority agrees on the definition proposed by Dey. For our research work, we will consider the definitions of Dey [2] and Calvary [3], which define the context as the triplet of <user, platform, environment>. These definitions help to clarify the notion of context in Human-Computer Systems (HCS).

Over the years, a large number of context-aware systems have been developed for different domains. Based on different context models, these applications are able to gather, manage, evaluate and disseminate context information [4]. Among these approaches, we mention the SOCAM (Service-Oriented Context-Aware Middleware) architecture that was especially proposed to convert the physical spaces; thus, contextual information can be converted into a semantic space and can be shared between the context-aware services [5]. Moreover, the key component in the CoBrA (Context Broker Architecture) architecture is responsible for managing and processing the contextual information while maintaining the contextual model [6]. The SECAS (Simple Environment for Context-aware Systems [7]) architecture is based on three components: context management, adaptation layer and the application core. Context management is composed by the context provider, the context interpreter, the broker and the context repository. The adaptation layer considers three type of adaptation: content, behavior and user interface adaptation. The Context Toolkit, proposed by Dey [8], provides a toolkit for the development of context-aware applications. It has a layered architecture that permits the separation of context acquisition, representation and the adaptation process. This architecture is based on: (i) Widget: a software component that provides applications with access to context information from their operating environment; (ii) Interpreter: used to interpret low-level context information and convert it into

 ISBN: 978-1-61208-353-7

higher level information; (iii) Server: a connection between the applications and the widgets.

Recent researches take into account context-awareness and complex dynamic system in which context variables change over time, such as GECAF (Generic and Extensible Context Aware Framework) [9], which uses a generic framework that supports all elements found in existing systems. Also, the conceptual architecture for Adaptive Human-Computer Interface of a PT Operation Platform (AHCI of PT platform [10]) based on context-awareness improves usability, simplifies the operation process, reduces operation complexity, provides needed information timely and properly and supports user needs diversification.

Having analyzed related work, it then becomes necessary to define comparison criteria to work out the advantages and disadvantages of each architecture. The main criteria in our research are: (i) the model of context: the use of any inappropriate model for context representation could lead to incorrect interpretation of contextual information. This could compromise the entire functioning of the architecture and provide the user with inadequate adaptations. Using a simplistic model could result in conflicts in the interpretation and the description of the current context of use. (ii) User interface validation: in critical domains, the generation of the user interface must be validate. The interface will guide the user to accomplish his task. Any errors in the interface can lead to a critical situation. Notwithstanding its obvious importance, this factor is not seriously treated in the majority of research work studied. For the first criterion, the SOCAM and CoBra architecture use ontologies for modeling context. SECAS utilize XML for context description. Furthermore, the Context Toolkit applies the Key/Value model for the specification of context. In our research, we focus on graphical modeling approach.

III. CONTEXT AND USER'S TASK MODELING BASED ON PETRI NETS

Several graphical modeling approaches to context-aware systems have been proposed, such as Unified Modeling Language (UML) [11], Object Role Modeling (ORM) [12] and Petri Nets (PN) [13]. In this section, we will focus on PN. PN, proposed by Carl Adam Petri in 1962, is a mathematically-based formalism dedicated to the modeling of parallelism and synchronization in discrete systems [13]. Recently, many context-aware systems modeling approaches based on PN have been proposed. They have been recognized as promising for the representation of context [14].

Context modeling approaches using PN differ depending on the purpose. Some authors are mainly interested in modeling the behavior of context-aware application; others try to solve the problem of time and resources in applications. There are several extensions of PN, such as: (i) Synchronized PN: Reignier introduces an approach to the representation of the context and the behavior of the application [15]. (ii) Colored PN (CPN): Silva proposes to combine 3D modeling tools with CPN for modeling 3D environments. In this model, the place is used to indicate the current state of the user and the transition is used to deduce the movement of the user and the behavior of the components [16].

The approach of modeling context must verify several requirements of pervasive environments, such as partial validation, formal language and formal verification. The last two requirements are essential in our research:

- Partial validation: It is preferable to be able to partially validate contextual information because of the complexity of contextual interrelationships, which may be responsible for any modeling error.
- Formal language: The chosen modeling method should have formal semantics. Formal methods comprise formal specification using mathematics to specify the desired properties of the pervasive system.
- Formal verification: the model must be verified through rules or mathematical property. This can be helpful in proving the accuracy of pervasive systems; this ensures the validation of user interfaces before they are implemented.

The PN-based methods have all the required characteristics. Indeed, the use of PN for modeling paves the way for formal verification and validation of the interfaces. These criteria are very important in our research work. The modeling of pervasive systems in a critical domain requires a rigorous validation of the interface in order to present the best solution to face an urgent situation.

This saves considerable time in the development cycle, particularly during the validation phase. Moreover, PN have a formal definition; they are highly capable of expressing aspects such as parallelism, timing, concurrency, etc. They possess many techniques for an automatic verification of properties. They provide, in addition, an unconstrained graphic representation.

We use "small granularity" PN ensuring the accuracy of our model and the partial validation. In fact, the context model is decomposed in small granularity and can be done through a simple validation based on partiality.

Modeling the user's task has a tough impact on the design of the user interface. In recent years, there have been different approaches to the specification of the task and how it relates to the area of application. Several notations have been proposed (ConcurTaskTrees CTT [17], Collaborative Task Modeling Language CTML [18], and PN [19]). The tasks are organized hierarchically to represent the task's decomposition, which is executed to meet a particular purpose. The process of task decomposition is ceased once the atomic task 'action' is obtained.

The PN are continuously expanding and they are a suitable tool for modeling the HCS. Initially, they were only used to describe tasks that were to be computerized. But later, especially with the emergence of High Level PN, they were used to model the HCI.

In the next section, we present our formal approach for specification of pervasive user interfaces.

 ISBN: 978-1-61208-353-7

IV. APPROACH FOR FORMAL SPECIFICATION AND GENERATION OF USER INERFACES ADAPTED TO THE CONTEXT

The overall objective of our research is to generate in real-time a user interface adapted to the current context of use in critical situations. The specifications of the HCS must consider the context modeling. As we mentioned in the third section, the captured context data will be modeled using PN. As the context is defined by the triplet <user, platform, environment>, each element will have its own PN: (i) The User's PN describes the different users of the application; (ii) The Platform's PN presents the different platforms that host our application; (iii) The Environment's PN describes the different information of the environment (i.e., geographical location, time, etc.).

Since each component of the context is modeled by its own PN, the marking of all these networks determines, at any given time, the current state of the context. Indeed, the marked places in the three networks determine the values of the triplet <user, platform, environment> and characterize the current context. Furthermore, according to the context in which the user operates, the user's task may vary. Indeed, each user task is specific to a given context. Thus, a set of pairs (context, task) will compose, among others, the model of the HCS. The user's task will also be modeled using PN. Each task will be decomposed into elementary tasks to be modeled using elementary structure of PN.

A. *Elementary structure of PN*

The modeling of an elementary structure is illustrated by Figure 1.

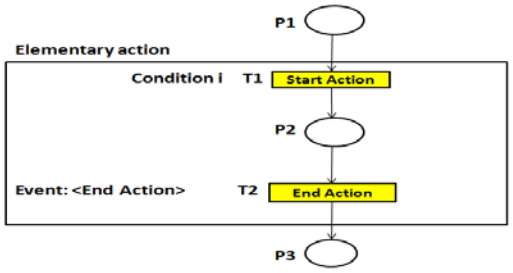


Figure 1. Structure of an elementary action

The validation of the condition i (transition T1) models the fact that the user will start the execution of the action relative to that condition. After the event, the "end action" (transition T2) expresses the fact that the user action was performed and ended. The place P2 represents a waiting state for the end of the action's execution, while the places P1 and P3 model the state of the user before and after the execution of his action. For example, P1 models the user's mental intention in order to act. The place P3 expresses his state at the end of the action's execution.

All the user's actions and components context behavior (elementary or composed) are arranged according to typical compositions: sequential, parallel, alternative, choice, iterative or of-closure. We present below the principle of parallel and alternative composition:

- The parallel composition expresses the possibility of simultaneous execution. The parallelism is ensured thanks to an input synchronization place. This place activates at the same time all the places of initialization of the parallel actions to be executed. Note that the effective parallelism can only be done if the actions to be executed do not use the same resources. Otherwise, a partial or complete sequencing would be necessary. Obviously, the number of places Pn must be equal to the number of parallel actions Ai. Thus, to ensure the parallel composition of actions, it is necessary to synchronize the places of entry and those of exit of those actions (Figure 2).

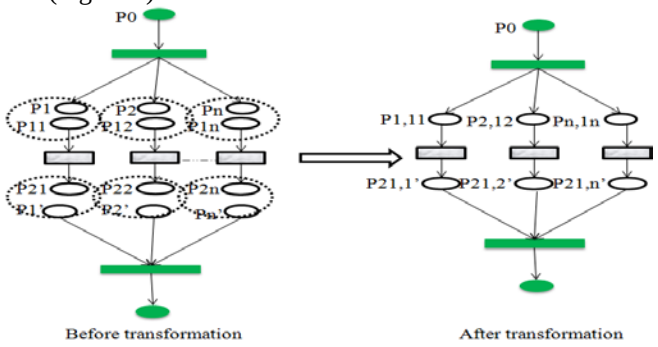


Figure 2. Parallel composition

- The alternative composition of n actions reflected a performance always exclusive of these actions. To avoid an actual conflict, conditions are associated with transitions to unambiguously determine which action should be executed. The alternative composition of n networks is realized by composing them sequentially with an ALT structure and merging all the end places of these networks. ALT structure allows the validation of a single condition at a time. ALT structure comprises a set of transitions equal to the number of networks to be composed alternately. These transitions are from the same input place P0. They allow, through the conditions associated with them, without ambiguity to initialize a single PN from the n modeled, which guarantees the absence of actual conflict (Figure 3). More details are presented in [20].

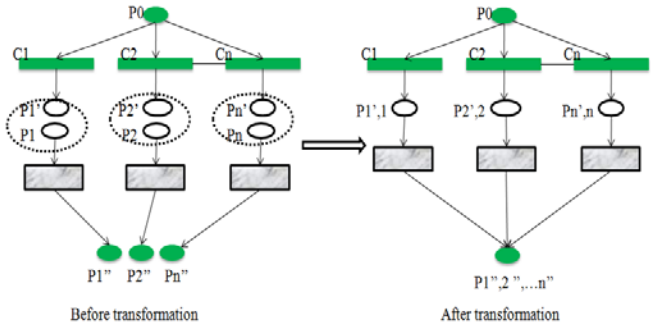


Figure 3. Alternative composition

The elementary structures represent our Meta-model. All these structures are defined manually and stored in a database.

B. Proposed approach

At a given moment, the marking of (i) the three PN of the context <user, platform, environment> and (ii) the PN of the user task, give the state of the ubiquitous HCS. The values of these markings are previously identified analyzed and stored in a database. So, this database will contain pairs of (context, task). At any time, if the values of the PN marking, describing the current context, are already included in this database, then this will be considered as an expected and well known situation and the user task will be identified, otherwise it will be considered as an unexpected situation. Managing unexpected situation will be the subject of our future research.

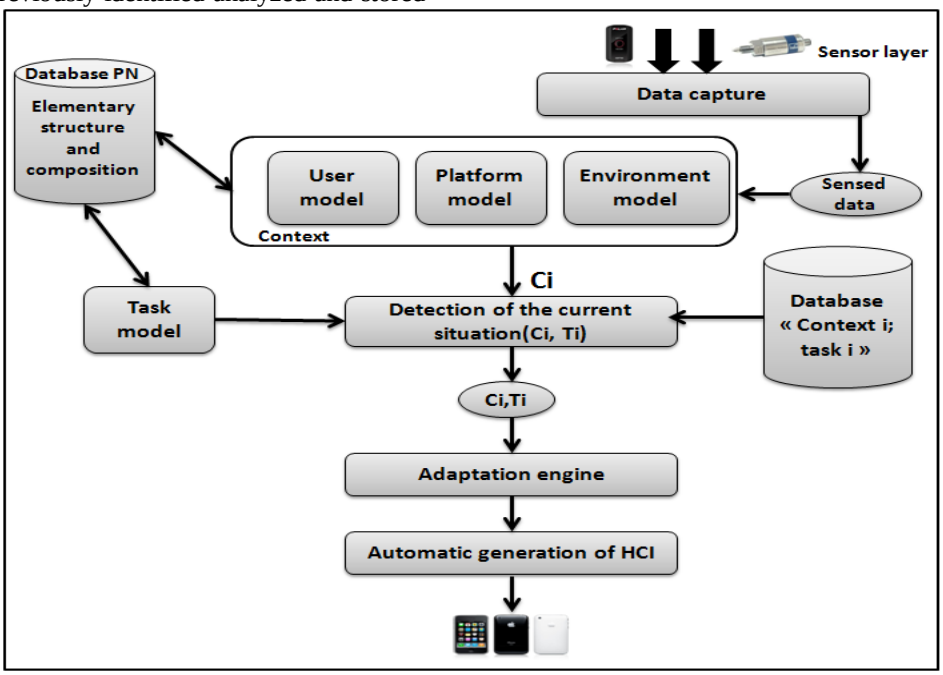


Figure 4. Proposed approach

Figure 4 illustrates our approach stating that once the data is collected from the sensor layer, it will be modeled and decomposed using PN in a user model, a platform model and an environment model. The user's task will be modeled using PN. This modeling is realized using the database of PN which contains the elementary structures and the compositions. All the (context, task) couples will be identified and stored in the "database Context i; Task i". "Context, task" database (Figure 5) is composed by two tables: context and task, connected by the association context-task. Figure 6 describes very summarily the logic diagram of database of PN. The PN is composed by elementary structures. This diagram will lead to the building of the associated database schema.

At a specific moment, the marking of the PN modeling the context will determine its current state Ci. In order to know the proper task Ti, it is required to browse the database of "context, task".

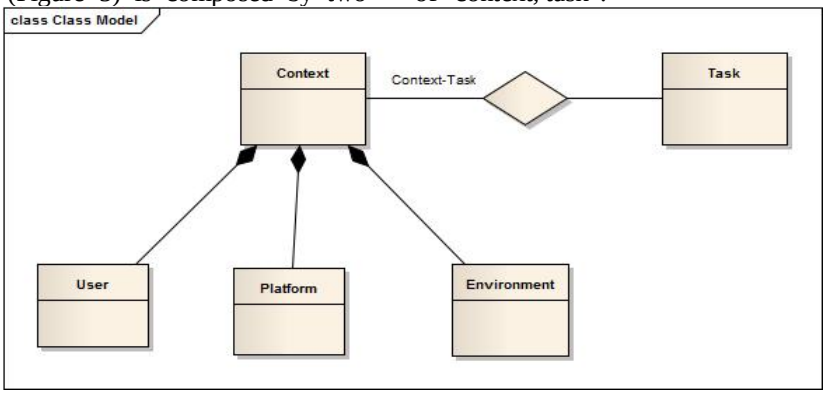


Figure 5. Logic schema of "Context-Task"

 ISBN: 978-1-61208-353-7

Whenever, the detection of the current context is made, the couple of values are transmitted to the adaptation engine. If its value is null, the adaptation engine will launch the script to deal with unexpected situation. Otherwise, it will trigger scripts to adapt to a "Known" situation. Finally, the adaptation engine will activate the automatic generation of the user interface adapted to the current context.

To manage an unexpected situation, the user will be given prior studied information and will intervene manually on the system. Actually, this is a very challenging problem that we will tackle it in our future work.

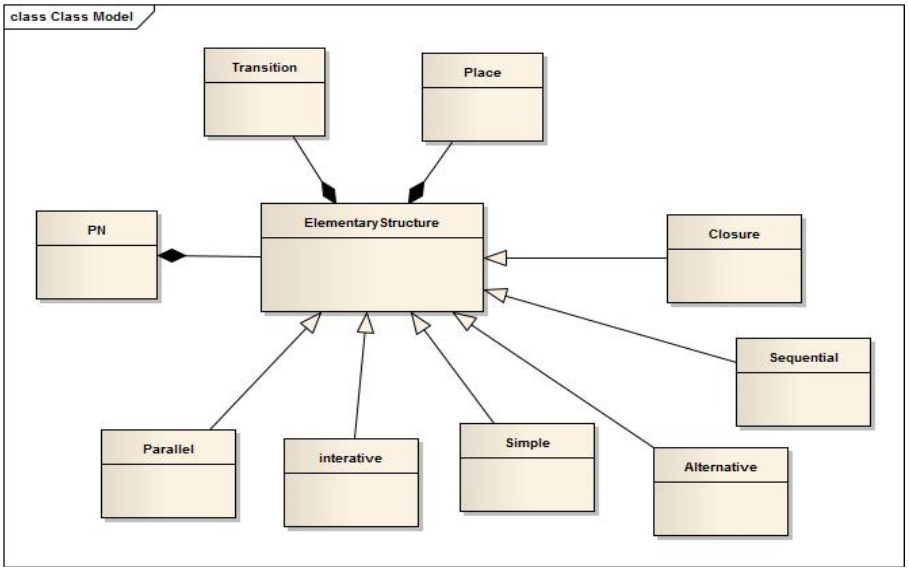


Figure 6. Logic schema of PN database

Our approach presents several advantages: first, it includes the five layers of a context-aware system, namely, context acquisition, interpretation, storage, diffusion and application layer. It separates the acquisition and modeling of context from its use. Each component of our architecture fulfills a particular task. Second, due to the complexity of context data, we choose to decompose this data into three models and to consider the user's task at the stage of modeling context. The originality of our approach lies in the proposal of the couple (context task). Indeed, the HCS became context aware and according to the context in which the user operates, the user's task may vary. In fact, each user's task is specific to a given context. That is why a set of pairs (context, task) were defined to compose the model of the pervasive HCS. Our model describes the behavior of contextual information. It decomposes the context's components into small granularity to ensure the validity of the model. To do this, we use a set of "well-organized" elementary PN structures.

Pervasive application presents a high level of dynamism so that many actions must be done in parallel. In our opinion, the aspect of parallelism in PN is very important especially in a critical domain. Certain situation requires the intervention of two or more users at the same time to meet a particular circumstance. The use of formal method to describe the behavior of a context-aware system allows us to deduct the properties of the system and the users' requirements in order to generate the appropriate interface at a given moment.

Context-aware approaches for user interface generation still have serious difficulties to dynamically and automatically adapt and generate interfaces, meeting users' requirements. These approaches are not formal and do not cover the validation of the user interface. We try to fill these gaps by proposing elements of solutions for:

Automatic deduction of user requirements: the database "Context i, Task i" is responsible to identify the appropriate task meeting the values of context.

Automatic adaptation of user interface to the context: The context changes are managed automatically by the "detection of the current situation" component. It can be considered as a representation of the smart environment.

Validation of the user interface: Thanks to the PN modeling approach, the modeled system and the generated interfaces verify the main PN properties as reachability, boundedness, liveness, etc. This guaranties the validity of the generated interfaces

Automatic generation of graphical interfaces: The adaptation engine component assumes the automatic generation of the user interface by identifying the most suitable widget to meet the user needs [21].

Comparing to the proposed approaches, seen in the second section, our architecture is centered on the context modeling and the generated user interface. The choice of the used approach for modeling context is very important. The information of context has a direct impact on the generated interface especially in the critical domains which justifies the use of formal methods in our approach.

 ISBN: 978-1-61208-353-7

The question that arises at this stage is: "how can we deduce the users' requirements from the context and the task model?"

C. Deduction of user's requirements

In a ubiquitous environment, the context model will trigger the appropriate task model. Indeed, the task depends on the context, and it is not a fixed model. Furthermore, according to the values of the context at a given moment, we can deduce the appropriate user's task. The proposed PN for context and user's task modeling is an Interpreted Petri Net IPN (Figure 7) defined by the set: < P, T, E, OB, Pre, Post, μ, Precond, Action, Info-Transition > where:

- P= set of places = {P1, P2,..., Pn},
- T = set of transitions = {T1, T2,..., Tm},
- E = set of events including the event "always present" <e>,
- OB = set of graphical objects of the interface,
- Pre: P x T → N defines the weight of the bow joining a place pi of P to a transition tk of T,
- Post: P x T → N defines the weight of the bow joining a transition tk of T to a place pi of P,
- μ: T → E associates to each transition the appropriate triggering event,
- Precond: T → Boolean Expression defines the necessary passing condition for each transition,
- Action: T → A defines the eventual and appropriate action procedure associated to each transition,
- Info-Transition: T → OB associates to each transition, the appropriate interface objects [13].

This type of networks introduces the notions of event, condition and the notion of action. Indeed, a passing condition (Cj), a trigger event (Evj) and a potential action (Aj) are associated with each transition Tj of an IPN.

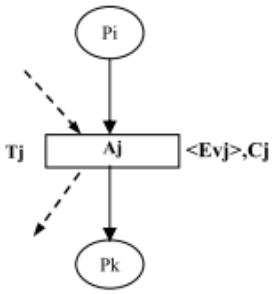


Figure 7. Interpreted Petri nets

Once the behavior of the user in its context of use is modeled, users' requirements can be deducted.

For a better management of the situation, the user needs, instantly, a lot of information. This information will be transmitted to the user by different interface components (messages, values, graphics, etc.). Since these objects are related to the context of the ubiquitous environment, i.e., the contextual parameters, we must identify the appropriate set of informational parameters for each state.

Moreover, in order to perform the tasks, the user needs to adjust some parameters in order to correct an abnormal situation or abnormal information, and/or to operate in particular situations or in collaborative tasks. For that, the interface will present a set of control components through which the user can monitor the situation; the set of these control and informational parameters constitutes the user requirements. Having presented our approach, we demonstrate its feasibility using a case study in the monitoring of a diabetic patient.

V. CASE STUDY: MONITORING OF A DIABETIC PATIENT

As a first experiment of our approach, we conducted a case study of a medical system for monitoring of a diabetic patient. This example is designed to monitor at real-time the evolution status of diabetic patients in a smart hospital. This monitoring is made possible by biological sensors implanted under the skin of the patient, which periodically control the patient's glucose levels. The ubiquitous system must continuously verify the changing state of each patient, which provides guidance on any medical interventions, or otherwise may deem any intervention to be unnecessary.

One of the problems that can arise from such a case study is to know how to notify the medical team (doctor / nurse) for an urgent and immediate intervention, and how should we proceed to carry on. This intervention should take into account the status of the patient and the location of the medical team, nurse or doctor.

The ubiquitous system will therefore generate real-time user interfaces adapted to their preferences, profiles, activities and geographical location. It will guide the user to best accomplish his task, while taking into account the various constraints of the context. In such system, the intervention of the medical team must be immediate, as the risk of loss of human life can be high. User interfaces should be validated and must present relevant and reliable information. Errors at the interfaces can cause the deaths of patients.

As a first step of our approach, we must model the information of the context. We consider the context as the triplet <user, platform, environment>. Each component of this triplet is modeled by an independent PN. Those components are:

- User's PN (Figure 8.A): it aims to identify the profile of the user (doctor or nurse). The marking of the network at a given time defines the type of the connected user. The doctor can be specialist, resident or internal.
- Platform's PN: Figure 8.B describes the different platforms that can be used by users. User interface can be hosted on various platforms. For our case study, we consider that a user can connect using a tablet, a PC or a mobile phone.
- Environment's PN: it describes the different values of our environment. For our example, after the opening of the session, various sensors intercept in parallel, the glucose level (GL) of the patient, the geographical coordinates of the user and the time. Concerning the geographical data, a user can be in the hospital, in the cabinet, at home or outside (i.e.,

in a restaurant or on the road, etc.). Concerning the time, it can have three different values: morning, afternoon or evening. Tokens present in different places, will describe the state of the environment by specifying the value of time, geographical data and glucose level (Figure 8.C).The deduction of the environment's properties must be done at the same time. This later is possible through the parallel composition of PN.

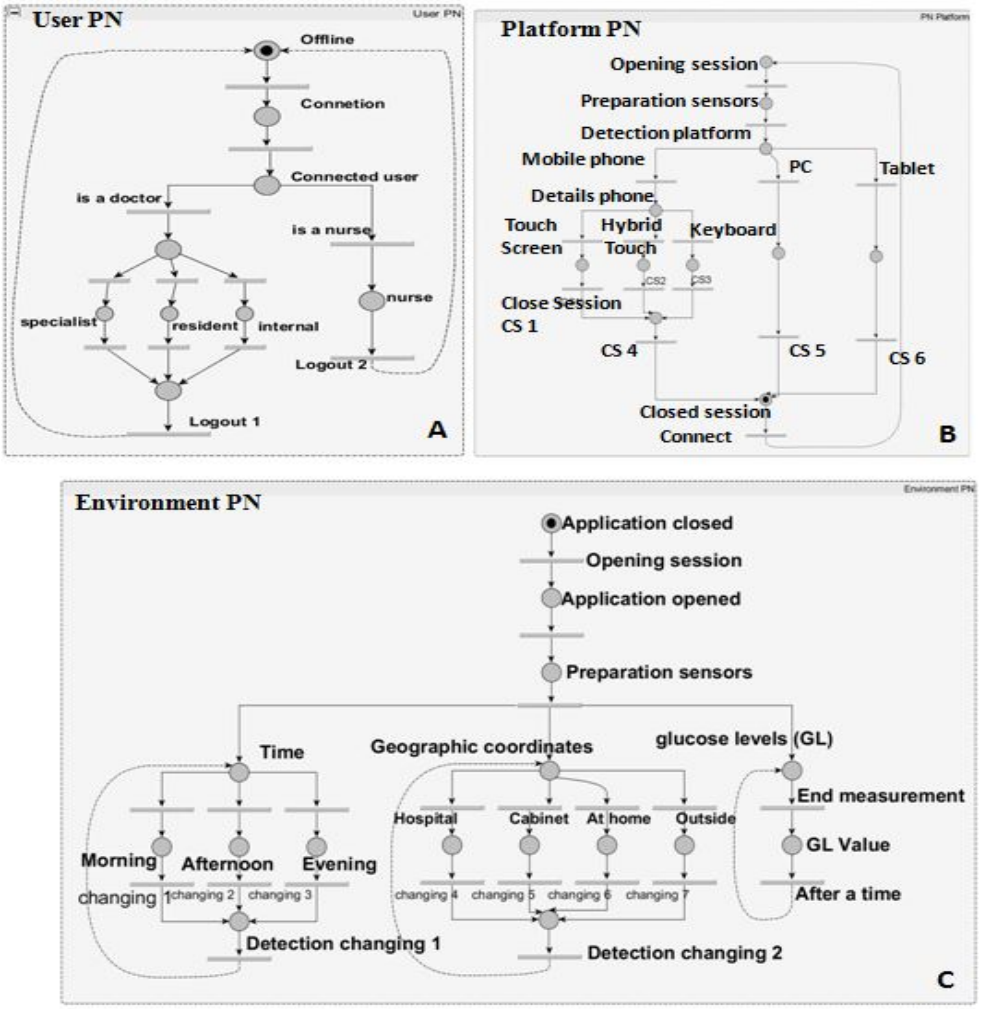


Figure 8. Context model

The environment's PN must be watchful to any changes that may happen to the environment. This action is possible by transitions "changing detection 1 and 2", which will monitor the possible changes in the environment. If any change occurs, then our sensors measuring will catch the new data. After modeling the different components of the context, we model the task.

For our example, we consider a critical situation where several actions must be done at the same time. Here is the scenario: Let us suppose that the patient is hyperglycemic (i.e., the Glucose Level >=4mmol) and the relevant doctor is not in the hospital. This situation is very critical and has to be treated by several actors. The doctor and the nurse must be aware of this critical situation at the same time, so they must perform actions in parallel. Let us notice that the system must select the most relevant replacing doctor according to his geo-location, his availability and his profile. The doctor receives a notification for an urgent and immediate intervention. In this case, the nurse cannot face alone such a situation, which requires a doctor's intervention. These two users must perform their actions at the same time. The interfaces will guide the doctor and the nurse each one according to his profile, by presenting appropriate information about the patient. The patient, suffering from a very serious condition of hyperglycemia, is in a coma. The doctor must inject insulin in him and check patient's status and measure its glucose level:

- If the rate of GL <=4 mmol, the doctor must give food containing sugar, wait 15 minutes and repeat the measurement of glucose;
- If the rate of GL > 4 mmol, the doctor must repeat the insulin injection;
- If the status is normal, then the patient's condition should be monitored for any possible change.

 ISBN: 978-1-61208-353-7

At the same time, the nurse must give plenty of water to reduce the patient's glucose. Then, she must make a urine levy. This situation is very critical and must have an immediate intervention to avoid the risk of patient's death. The actions of doctor and nurse, illustrated in Figure 9, must be done in parallel. PNs are very efficient to do this modeling.

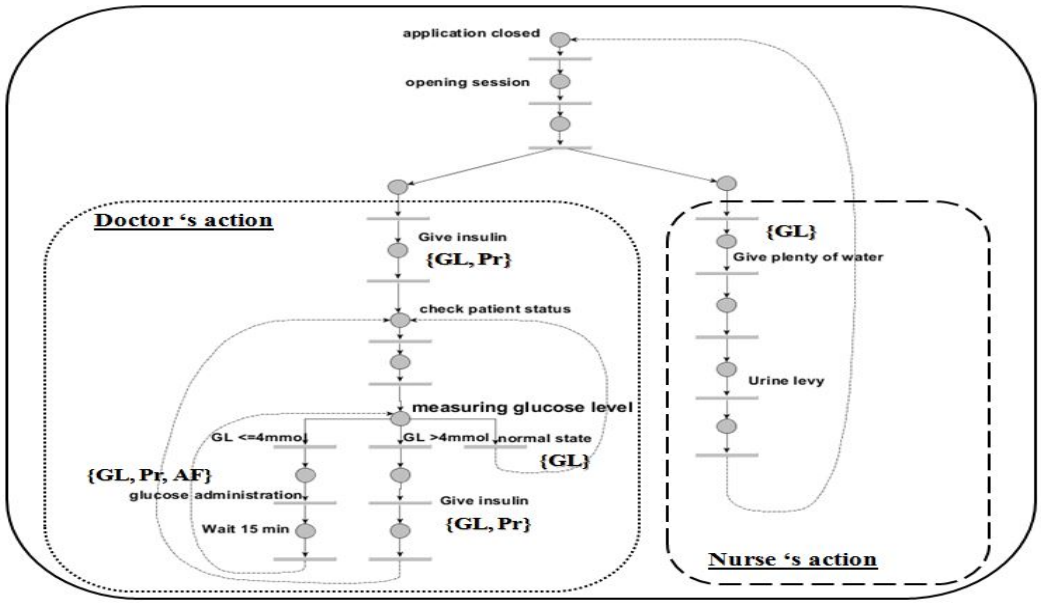


Figure 9. Task model: critical intervention on hyperglycemic patient

The values of the tokens in the context's PN trigger the appropriate Task. Concerning the deduction of user's requirements, we consider the PN transition "Begin Action". To these transitions, we associate the adequate parameter(s) of interface, either informational or control, which refer to the user's requirements. Thus, at the point of "measuring glucose level", the user disposes of the relevant user requirements in order to adequately perform their action, i.e., the Glucose Level (GL). This value comes from the context model, more precisely from the environment's PN. For instance, the informational parameter GL informs the user of the Patient's usual level of glucose. At the state "glucose administration", the user disposes of the glucose rate and other information related to the prescription of medicines (Pr). At the point of "give insulin", the user will dispose of an authorized food menu for the patient (AF) to select from. At the point of "Give plenty of water"; the user disposes of patient's glucose rate.

The compositions of elementary structure offer the possibility to the user to perform multiple tasks simultaneously. All tasks can be modeled according to the compositions shown in Section 4. The principle of the elementary compositions applied in the context modeling is also applicable to the composition of several tasks.

Once the informational and the control parameters have been identified, we can deduce the necessary components of the User interface: (i) We associate an Informational graphical object to each informational parameter; (ii) We associate a Control graphical object to each control parameter. These parameters are very important because they will lead to the graphics interface components. This interface will guide each kind of user throughout his intervention. In critical situations, user interface will play crucial role especially since the physician is not the patient's treating doctor. So he does not know the patient information. These graphic components will adapt depending on the doctor's context and profile. If he is a specialist, then no need to provide all the information and if he is an internal then the interface must provide the maximum of information to reduce the risk of errors.

For the implementation of our approach, we use a SOA (Service Oriented Architecture) [22]. First, we transform our PN model into a PNML representation (Petri Net Markup Language [23]). PNML is PN XML-based standard. Its main scope is to facilitate information exchange between PN models. Each PN is considered as a labeled graph. It contains a set of objects: places, transitions and arcs. Each object has a unique identifier and a set of labels (annotations and attributes) representing the name of a place, the inscription of a transition, etc. [23]. Algorithms are written in Java to parse the PNML file of context in order to detect the location of the tokens and consequently the current context. Once the triplet of context identified through the xml code, we browse the Mysql database "context i, Task i" in order to identify the appropriate task for the current situation. We use JavaScript Object Notation (JSON) web service [24] to query the database and to extract the name of the appropriate task. The monitoring of diabetic patient is developed using Android.

Our methodology applies to all types of applications. In fact, pervasive HCS comprises a triplet: system, task and context. Each component will be modeled using PN. The

 ISBN: 978-1-61208-353-7

designer has to analyze the system and deduce all relevant context information. Each component of the context such as user, environment and platform as well as its behavior, must be known in advance. All context values are stored in a database in which we can identify the appropriate user's task. The sensor layer will guide us on the value of the context.

VI. CONCLUSION AND FUTURE WORK

This paper presented an approach for context and task modeling based on Petri nets. We model the pervasive Human-Computer System using a composing process of elementary PN in order to verify the relevant properties of the system before the generation of the interfaces. In this paper, we have tried to demonstrate the context influence over the user's task modeling in a critical domain and the strength of PN in critical system's modeling. We also explained how to deduce the users' requirements through the task model. This formal approach is illustrated by a case study on the monitoring of diabetic patients in a smart hospital.

In the near future, we plan to address the following issues: (i) we will explain the running of each component of our approach namely the implementation of the adaptation engine and the automatic generation of user interface; (ii) we will explain how acquisition context layer will supply the model of context. We are now working on human-centred evaluations. We create an experimental platform implementing a realistic scenario that volunteer doctors and nurses will use in their work.

REFERENCES

[1] M. Wurdel, S. Propp, and P. Forbrig,"HCI-Task Models and Smart Environments", Human-Computer Interaction Symposium IFIP International Federation for Information Processing, vol. 272, 2008, pp. 21-32.

[2] A.K. Dey, D. Salber, M. Futakawa, and G. D. Abowd,"An Architecture To Support Context-Aware Applications", GVU Technical Reports, 1999.

[3] G. Calvary, A. Demeure, J. Coutaz, and O. Dâassi, "Adaptation des Interfaces Homme-Machine à leur contexte d'usage", Revue d'intelligence artificielle, vol. 18, no. 4, 2004, pp. 577-606.

[4] M. Knappmeyer, S. L. Kiani, E. S. Reetz, N. Baker, and R. Tonjes, "Survey of Context Provisioning Middleware", IEEE Communications Surveys & Tutorials, 2013, vol. 15, no. 3, pp. 1492–1519.

[5] T. Gu, H. K. Pung, and D. Q. Zhang, "A service-oriented middleware for building context-aware services", Network Computer Application, 2005, vol. 28, no. 1, pp. 1-18.

[6] H. Chen, T. Finin, and A. Joshi, "An Intelligent Broker Architecture for Context-Aware Systems", Adjunct Proc. of Ubicomp, 2003, pp. 183-184.

[7] T. Chaari, F. Laforest, and A. Celentano, "Adaptation in Context-Aware Pervasive Information Systems: The SECAS Project", Int. Journal on Pervasive Computing and Communications (IJPCC), vol. 3, no. 4, 2007, pp. 400-425.

[8] A. K. Dey, G. D. Abowd, and D. Salber, "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications", Hum.-Comput. Interact, vol. 16, no. 2, 2001, pp. 97-166.

[9] A. Angham, A. Sabagh, and A. Al-Yasiri, "GECAF: a framework for developing context-aware pervasive systems", Computer Science - Research and Development, Springer, 2013, pp. 1-17.

[10] Q. Xue, X. Han, M. Li and M. Liu, "A Conceptual Architecture for Adaptive Human-Computer Interface of a PT Operation Platform Based on Context-Awareness", Discrete Dynamics in Nature and Society Journal, 2014, pp.1-7.

[11] J. Bauer, "Identification and Modeling of Contexts for Different Information Scenarios in Air Traffic", In: Workshop on Advanced Context Modelling, Reasoning and Management, UbiComp - The Sixth International Conference on Ubiquitous Computing, Nottingham/England , 2004.

[12] K. Henricksen, J. Indulska, and A. Rakotonirainy, "Generating Context Management Infrastructure from High-Level Context Models", In: Industrial Track Proceedings of the 4th International Conference on Mobile Data Management (MDM), 2003, pp. 1–6.

[13] M. Moalla, "Reseaux de Petri interpértés et Grafcet", TSI de l'AFCET, vol. 4, 1985.

[14] S. Han and H. Y. Youn, "Petri net-based context modeling for context-aware systems", Artificial intelligence review, vol. 37, 2011, pp. 43-67.

[15] P. Reignier and O. Brdiczka, "Context-aware environments: from specification to implementation", Exp Syst, vol. 24, no. 5, 2007, pp. 305–320.

[16] J. L. Silva, J. C. Campos, and M. D. Harrison, "An infrastructure for experience centered agile prototyping of ambient intelligence", EICS '09: Proceedings of the 1st ACM SIGCHI symposium on engineering interactive computing systems, 2009, pp. 79–84.

[17] M. Giulio, P. Fabio, and S. Carmen, "Ctte: Support for developing and analyzing task models for interactive system design", IEEE Trans. Softw.Eng, vol. 28, no. 8, 2002, pp. 797–813.

[18] M. Wurdel, S. Propp, and P. Forbrig, "Hci-task models and smart environments", in: P. Forbrig, F. Patern, A. Pejtersen (Eds.), Human-Computer. Interaction Symposium of IFIP International Federation for Information Processing, Springer US, vol. 272, 2008, pp. 21–32.

[19] I. Riahi, M. Riahi, and F. Moussa, "Xml in formal specification, verification and generation of mobile hci", in: J. Jacko (Ed.), Human-Computer Interaction. Towards Mobile and Intelligent Interaction Environments, vol. 6763 of Lecture Notes in Computer Science, Springer BerlinHeidelberg, 2011, pp. 92–100.

[20] I. Riahi, F. Moussa, and M. Riahi, "Petri Nets context modeling for the pervasive Human-Computer Interfaces", Eighth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT'13), Lecture Notes in Computer Science, vol. 8175, 2013, pp. 316-329.

[21] F. Moussa, I. Ismail, and M. Jarraya, "Towards a Runtime Evolutionary Model of User-Adapted Interaction in a Ubiquitous Environment: The RADEM Formal Model", Cognition, Technology & Work, Springer, 2014.

[22] E. Newcomer and G. Lomow, "Understanding SOA with Web Services (Independent Technology Guides", Addison-Wesley Professional, 2004.

[23] L. M. Hillah, E. Kindler, F. Kordon, L. Petrucci, and N. Trèves, "A primer on the Petri Net Markup Language and ISO/IEC 15909-2", Petri Net Newsletter 2009.

[24] C. J. Ihrig, "JavaScript Object Notation", Pro Node.js for Developers, 2013, pp. 263-270.

 ISBN: 978-1-61208-353-7

Mobile Staff Planning Support for Team Leaders in an Industrial Production Scenario

Sönke Knoch, Melanie Reiplinger

German Research Center for Artificial Intelligence
Research Department Intelligent User Interfaces
Saarbrücken, Germany
Email: [first name].[last name]@dfki.de

Rouven Vierfuß

Miele
imperial-Werke oHG
Bünde, Germany
Email: rouven.vierfuss@imperial.de

***Abstract*—Team leaders in industrial production scenarios face the problem of short-termed reactions to unforeseeable events in a factory. Adjustments in staff planning require many calculation and coordination efforts that consume valuable time. A mobile application is presented that assists team leaders in this challenging task. The selection and identification of relevant information, its case-specific processing and visualization as well as the persistence of worker data in a semantic network form the cornerstones of this work in progress.**

Keywords–Decision support; Staff planning; Mobility; Domain model.

I. Introduction

The developments towards Industry 4.0 [1] let the gap between real and virtual worlds melt and foster the utilization of the potentials of the Internet of Things [2]. The future industrial production aims at highly individualized products and an increasing flexibility of production processes. These developments involve a challenge for staff planning engineers who have to answer this flexibility in production with an adaptive staff planning strategy. Rising quantity of data and complexity makes this planning process even harder.

The coordination between planning engineers—in this scenario called team leaders—to find additional qualified workers within the factory consumes time and is a difficult task. To assist teamleaders in this process, the concept of a planning support system is suggested. In a defined scenario, a mobile application will support team leaders by visualizing the current worker-to-production-line allocation and interconnecting the team leaders to coordinate personnel placement in an efficient way. Therefore, information from several distributed information sources is prepared and presented on a mobile device. Relevant information dependent on the user's role and specific context of use is shown. Human ressource allocation can be edited and optimized directly using the user interface.

In Section II, an overview of related work is given and the distinction to the suggested approach is drawn. Section III describes the industrial scenario that is taken as a basis for the considerations made in the following sections. The IT-infrastructure necessary to run the system is presented in Section IV. In Section V, the domain model that gathers and stores worker profiles and supports the system in its task is developed. Section VI shows mockups of the mobile application. Finally, Section VII contains a discussion of the suggested approach and gives an outlook on future work.

II. Related Work

Current staff planning software is implemented as stand alone solution or integrated in enterprise resource planning (ERP) or manufacturing execution systems (MES). Time tracking is commonly part of the MES whereas time management can be part of both, MES and ERP [3, 199-212]. Most of these staff planning or workforce management systems focus on the scheduling and optimal resource allocation task. User interfaces present timetables with a view over a planned period in a very functional way. Some of these software providers offer mobile applications that allow access to the staff planning systems. These systems lack adaptive context- and role-specific information processing. The collaboration and coordination aspects tackled in this work are neglected. Additionally, the installation, application and maintenance costs, especially related to MES and ERP systems, are too high for many medium-sized and small factories.

In research, the project ENgAge4Pro [4] focuses on age-appropriate staff planning. Physical attributes, such as body weight and height, are regarded therefore. The research project EPIK [5] focuses on the optimal allocation of resources to enhance efficiency. A mobile application was developed that supports the worker with context-specific information. The research project KapaflexCy [6] covers short-term production scheduling. A mobile application allows to send employment requests to workers. After receiving these requests, the workers coordinate the takeover of the employment themselves. From a hierarchical perspective, this is a bottom-up approach.

Unlike the project ENgAge4Pro, the main concern of the staff planning support system is not ergonomics. In contrast to the EPIK project, the system suggested is developed to support team leaders in their planning task. Therefore, more attention is paid to the appropriate presentation of relevant information on the device. Optimized resource allocation will be included in form of an additional function (not the object of research). Compared to the KapaflexCy project, the team leader application developed here implements the allocation of employment in a top-down way. Nevertheless, worker-related information can be used to provide feedback for their work.

III. Scenario

The production of kitchen appliances in Germany is facing several challenges. Companies require the ability to produce their products under optimum cost, flexibility due to rising variants and a competitive market. Therefore, it is necessary

 ISBN: 978-1-61208-353-7

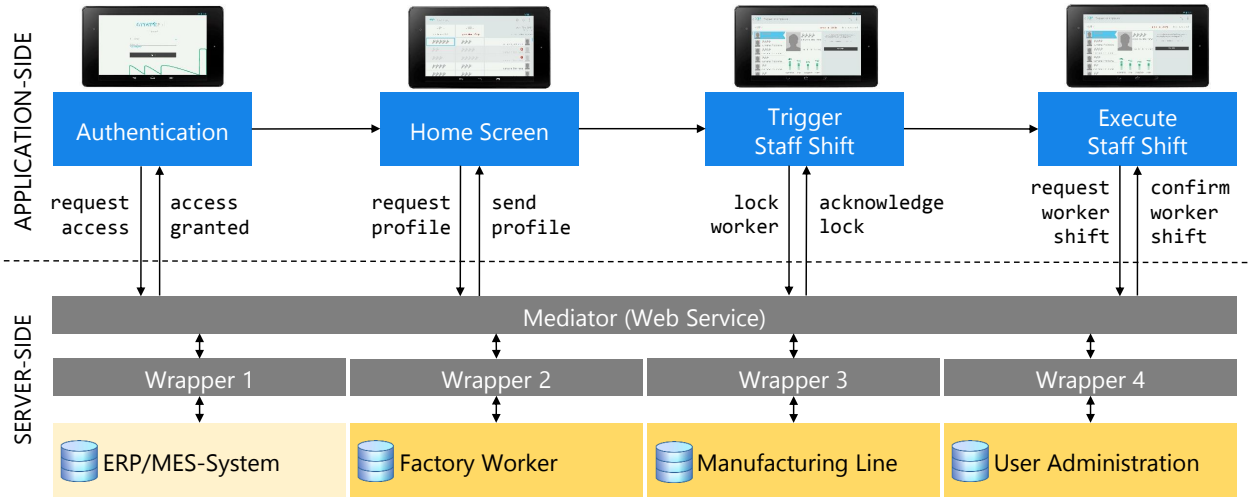


Figure 1. Staff switching process and involved distributed data sources.

that the manufacturing industry makes efficient use of resources and energy in order to keep the high-cost country Germany a competitive production location. Manufacturing of steam ovens in Imperial/Miele plant floors follows the "Miele Value Creation System". Multiple U-shaped production lines for diverse product classes allow for highly flexible handling of varying production programs. Each steam oven is assembled by a single worker in a one-piece-flow setting, which entails high responsibility and a complex work content for all employees.

Once per week, the plant's foremen and team leaders plan the production on the shop floor level, assigning resources and capacities to production orders. Detailed planning is done on a daily basis, considering the production program and availability of workers. In case of unexpected staff shortage or modified production volumes at short notice, the team leaders re-assign available workers to production orders and assembly stations, also across assembly lines. The process of staff planning is demand-oriented and flexible, and quickly becomes complex and time consuming while trying to meet the demands of multi-variant production scenarios with varying production programs, small lot sizes on multiple lines and customer-individual products. Furthermore, team leaders want to foster a broad skill set in all employees by organizing a rotating assignment of workers to varying tasks while, at the same time, the high quality standards of Miele need to be guaranteed by intense training on each particular product class.

A great deal of experience is needed in order to make the right decisions, and an adaptive assistant system that transparently combines data from production orders, human resource management and the plant floor in order to support decision-finding could significantly facilitate and speed up the daily staff planning process.

IV. IT-Infrastructure

To tackle the challenges described in Section III, we developed an application that supports the team leader in adjusting the daily routing when specific events occur. This kind of ad hoc planning involves heterogeneous information sources that need to be prepared and consulted in order to support reasonable decision making. The integration of these sources is discussed in Section IV-B. The integrated information needs to be preprocessed for the specific context and user role. The development of a suitable user interface for the industrial planning scenario is discussed in Section IV-A.

A. Application-side

Figure 1 shows the process that appears when a worker is scheduled. The *Authentication* forms the entry point to the application. The application loads role-specific profiles for each user. As each team leader manages different manufacturing lines, the respective lines are loaded and currently allocated workers are presented on the *Home Screen*. If the system registers any deviations from the planned schedule, it visualizes warnings and encourages the team leader to act in an ad hoc manner. One common reaction to hold the planned quantities at the end of the day is the search for suitable workers. For that reason, the application suggests suitable workers for a respective production line and product. If the team leader selects one of these workers the *staff shift is triggered*. To avoid multiple allocation of the same human resource, the selected worker is locked for 2 minutes. During this time interval, the team leader is able to execute the staff shift or search for another worker. If it was decided to schedule the selected worker he or she *executes the staff shift*. The team leader who supervises the respective worker is informed and can deny or confirm the worker shift. In the latter case, the shift is persisted and written to the database.

To compute the qualification of a worker (w) for a respective manufacturing line (l) and product (p), three parameters were defined:

$experience(w, l, p)$
: Total time by a worker.

$quality(w, l, p)$
: Defective pieces per shift by a worker.

$productivity(w, l, p)$
: Pieces per shift by a worker.

 ISBN: 978-1-61208-353-7

These parameters are visualized in the detail view of the front-end described in Section VI and allow the team leader a vague estimation of the worker's skills. The ranking function $rank(w, l, p)$ forms a weighted aggregation of these parameters and represents the skill level of each worker on a scale between 0 and 5. These weights are dynamic and adapted to the specific use case.

B. Server-side

The information that is necessary to support the planning process of a team leader as described in Section IV-A is located at four different points that are sketched at the bottom of Figure 1. In the present case, the information system can be described as a multi-computer, partitioned, distributed, shared nothing system. Thus, a suitable strategy to integrate the information has to be selected. To preserve the autonomy of the sources, a virtual integration strategy was selected which leaves the data at the sources. This kind of on-demand integration in a decentralized manner enables us to keep the system design easy to extend and to transfer data only when needed from solely relevant sources. A mediator-based approach was chosen to realize the virtual integration system. The *Mediator* provides an interface implemented as a Web service and communicates with the application. It lies in the responsibility of the mediator to provide a structural and semantic data integration. *Wrappers* are implemented for each information source to overcome the heterogeneity on the data level and to enable the data flow between mediator and sources.

Focusing on the four data sources at the bottom of Figure 1, the *ERP/MES-System* forms the only system that is already existing and available in a common factory. An ERP or MES system in the considered scenario provides—in collaboration or separately—access to time tracking and management data. The domain model for the *Factory Worker* encodes the worker's skill matrix and working history. It is described in detail in Section V. The domain model of the *Manufacturing Line* describes the process steps and involved manufacturing equipment and allows estimations about quantities that can be achieved when a specific worker is scheduled on that line. In combination with processing times it can be used to calculate optimal resource allocations with algorithms from the operations research field. The factory worker model constitutes new input data that allows new forms of optimization based on the user model. The model of the manufacturing line is not part of this work and is developed separately. The *User Administration* is responsible for authentication and lock requests.

V. Domain Model of a Factory Worker

The information about workers' skills and experience that is needed by the mobile application in order to generate useful recommendations is taken from a worker model. The worker model represents semantic relationships between the horizontal and vertical roles of an employee in the factory (i.e., his tasks, work content, and his position within the staff hierarchy) on the one hand, and his skills (e.g. work experience) and individual requirements on the other hand. Examples for the latter are an employee's handedness, language skills, allergies against specific materials, or an inability to lift heavy weights or to distinguish colors. A small excerpt of the worker model

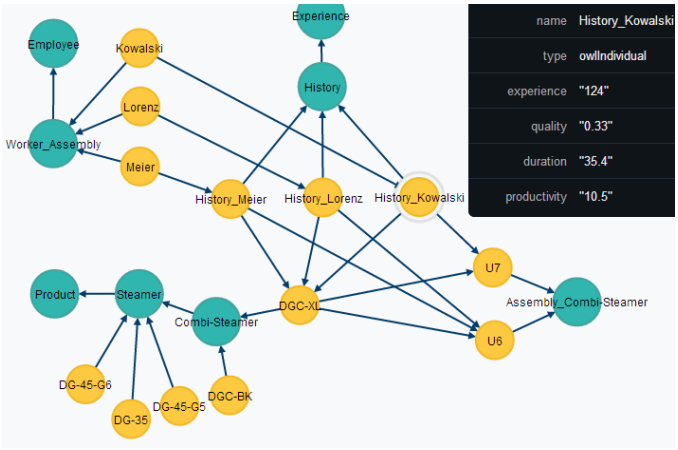


Figure 2. Snippet of the worker model.

is visualized in Figure 2. Here, the qualification of workers is encoded within the *History*-nodes of a semantic network between products, assembly lines, and employees. The application can, for instance, query the model for workers that are experienced assemblers of product p at assembly line l, and rank them using weighted aggregation as described in IV-A. The depicted example graph shows how information about the worker history is encoded. The node *History_Kowalski* represents the qualification-related data for worker *Kowalski* wrt. the product *DGC-XL* when assembling at line *U7*. The graph connects a model of the factory floor with a topology of manufactured products and the staff's profiles.

The worker model is implemented using the open-source, Java-based graph database Neo4j [7]. Querying of the model from within the application is realized by accessing Neo4j's RESTful interface. The model's contents are derived from a domain ontology (OWL) by automatically mapping the ontology's concepts and relations into the Neo4j graph database format (T-Box). Once created, the graph can be populated and updated at runtime with dynamically-changing data like a worker's history, retrieved by logging assembly operations.

There are several reasons for choosing a graph database above conventional storage formats like, e.g., SQL databases. Regarding performance for path operations (like in the above example) in highly-connected data, relational databases can become overburdened by queries of increasing complexity due to joins and index lookups; whereas in graph databases, which use index-free adjacency in traversing from node to node, query latency is relatively independent of the database size and the number of connections ([8], chapter 2). Note that denormalization for relational databases is not an alternative here, since the data model is not tailored exclusively to the needs of the staff planning app, but is instead meant to provide a flexible, multi-purpose source of semantic information, with diverse applications reading from and writing to the model. This implies that a) we cannot anticipate what relations will be queried most frequently at runtime, and b) we need to prepare for easy model update (e.g., based on sensor data), in order to keep the graph a realtime model of the staff data. Consequently, there is no use in optimizinig read access for specific relations at the cost of slower write access.

 ISBN: 978-1-61208-353-7

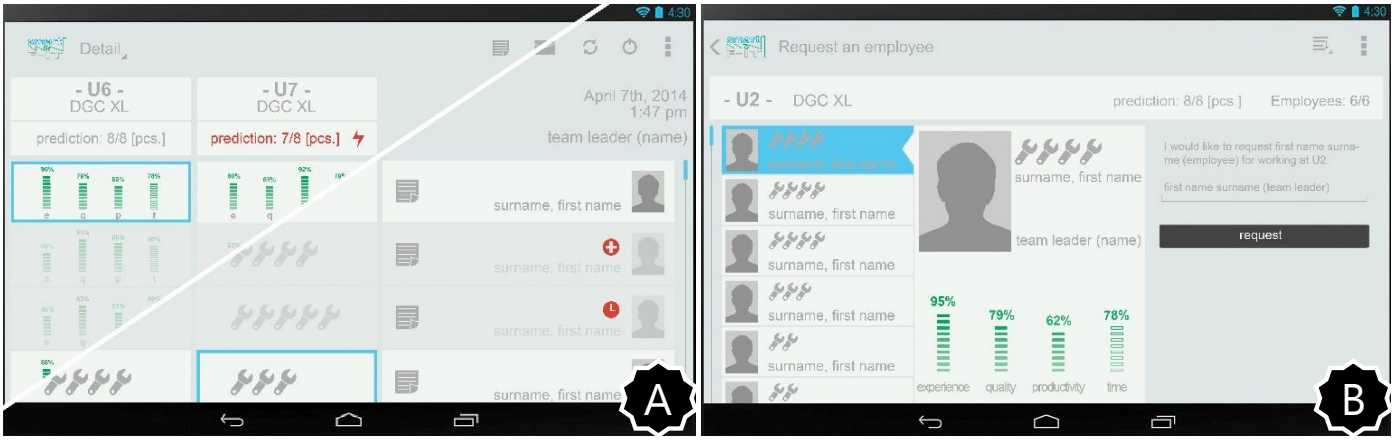


Figure 3. Mockups of the mobile application.

The most crucial benefit of graph databases in the context of Industry 4.0 is that they allow for an explicit, intuitive, and easily-expandable modeling of the complex semantic dependencies that exist in modern factories, and that the semantics of such graphs can easily be understood even by users unfamiliar with conventional modeling languages like UML.

VI. User Interface Design

A 7 inch tablet was identified to be most suitable for the daily deployment in a factory environment. It is small enough to fit in a team leader's pocket and offers enough space on the screen to present relevant information. In order to start the development of the screens, mockups of the mobile application were designed. The result is shown in Figure 3.

After a successful sign in, the home screen is presented (*screen A*). The screen visualizes the manufacturing lines and allocated workers the respective team leader is supervising. Events such as sick notes or delays are visualized by specific symbols. The absence of a worker can result in a deviation of the planned quantities for the day, which is visualized on top of the manufacturing line columns. To get an impression of the worker's fitness for the manufacturing process on a line, a 5 point ranking is visualized in form of wrenches.

If the team leader wants to break down this aggregation of worker skills, he or she can switch to a detailed view. The three qualification parameters *experience*, *quality*, and *productivity* (cf. Section IV-A) are visualized in form of bar graphs. As a fourth parameter, *time* indicates when the recommended dwell time limit on this line is reached. To trigger or execute a staff shift, the team leader changes to *screen B* by touching the matrix position he or she wants to edit on screen A. Screen B presents a list of suitable workers according to the weighted worker ranking. If the team leader selects a worker, his or her impact on the planned quantity numbers is visualized on top of the screen.

VII. Discussion and Outlook

In this work in progress, the concept of a mobile application for team leaders, to support them in making ad hoc planning decisions, was proposed. Recommendations of suitable staff shifts are based on a dynamic worker model that was implemented using a graph database. A Web service makes relevant information available that is presented on a mobile user interface.

In the next step, a prototype of the application is realized. The linking of relevant information sources forms the first challenge to allow a field test in the factory. Feedback of the team leaders will show if the desired improvements were achieved and will flow into enhancements of the prototype. Additionally, a useful data exchange between information sources, such as the factory worker model and the manufacturing line model, might result in new optimization strategies for an optimal worker-to-manufacturing-line allocation. On the other side, the information from the factory worker model might be used to provide workers motivating feedback on their daily work.

Acknowledgment

We would like to thank Nadja Rutsch for the design of the graphical user interface mockups presented in Figure 3. This research was funded in part by the German Federal Ministry of Education and Research (BMBF) under grant number 01IS13015 (project SmartF-IT). The responsibility for this publication lies with the authors.

References

[1] BMBF, "Project of the Future: Industry 4.0," 2014, URL: http://www.bmbf.de/en/19955.php [accessed: 2014-04-10].

[2] K. Ashton, "That 'internet of things' thing," 2009, URL: http://www.rfidjournal.com/articles/view?4986 [accessed: 2014-04-10].

[3] J. Kletti, Manufacturing Execution Systems (MES). Berlin; London: Springer, 2007.

[4] RWTH Aachen University IAW, "ENgAge4Pro Project," 2014, URL: http://www.iaw.rwth-aachen.de/index.php?article_id=20&projid=89&proalias=ENgAge4Pro [accessed: 2014-04-10].

[5] Fraunhofer IAO, "EPIK Project," 2014, URL: http://www.epik-projekt.de/ [accessed: 2014-04-10].

[6] ——, "KapaflexCy Project," 2014, URL: http://www.kapaflexcy.de/ [accessed: 2014-04-10].

[7] Neo Technology, Inc., "Neo4j," http://www.neo4j.org/, [Accessed 23-June-2014].

[8] I. Robinson, J. Webber, and E. Eifrem, Graph Databases. Beijing: O'Reilly, 2013. [Online]. Available: http://my.safaribooksonline.com/9781449356262

 ISBN: 978-1-61208-353-7

Unified and Conceptual Context Analysis in Ubiquitous Environments

Ali Jaffal, Manuele Kirsch-Pinheiro, Bénédicte Le Grand
Centre de Recherche en Informatique, Université Paris 1 Panthéon – Sorbonne, Paris, France
Ali.Jaffal@malix.univ-paris1.fr, Manuele.kirsch-Pinheiro@univ-paris1.fr, Benedicte.Le-Grand@univ-paris1.fr

Abstract— This article presents an original approach for the analysis of context information in ubiquitous environments. Large volumes of heterogeneous data are now collected, such as location, temperature, etc. This "environmental" context may be enriched by data related to users, e.g., their activities or applications. We propose a unified analysis and correlation of all these dimensions of context in order to measure their impact on user activities. Formal Concept Analysis and association rules are used to discover non-trivial relationships between context elements and activities, which, otherwise, could seem independent. Our goal is to make an optimal use of available data in order to understand user behavior and eventually make recommendations. In this paper, we describe our general methodology for context analysis and we illustrate it on an experiment conducted on real data collected by a capture system. Thanks to this methodology, it is possible to identify correlation between context elements and user applications, making possible to recommend such applications for user in similar situations.

Keywords—Context Analysis; Recommendation; Formal Concept Analysis; Ubiquitous Computing; Context-aware Systems.

I. INTRODUCTION

Context management consists in collecting, grouping and exploiting context information on behalf of the user. Several challenges currently affect context management, as for example, the analysis of a large data volume to analyze [1][2]. Indeed, the challenge is no longer collecting data, but to explore it efficiently, which depends on the impact of context on user behaviors and actions. This is particularly important for context-aware systems, whose goal is to adapt their behavior to the user's context. Several questions arise from this scenario: "How can relevant context information be identified?", "In which context is a specific action performed?", "What is the impact of a given context on user's actions, and what recommendation can be proposed?".

In this work, we propose a methodology for analyzing the impact of context information on the user actions. We focus in particular on user behavior when using mobile devices, such as tablets or smartphones. Our goal is to provide a way to identify context elements that influence user activities, to understand the relation between them, and to construct a knowledge base connecting actions and context situations.

The proposed methodology is based on Formal Concept Analysis (FCA) [3][4][5] to cluster context and actions based on their relationships. In conjunction with this analysis, we propose to extract association rules in order to recommend particular actions or applications to the user. Association rules are complementary to FCA as they allow quantifying and materializing the causal links between actions and context elements, but also among actions or context elements themselves. Thus, FCA gathers user actions according to the context under which they have been used in the past, and association rules allow making recommendations of future actions to users in a given context.

This paper is organized as follows: Section 2 discusses related work on context management and describes the basic principles of FCA and association rules. Section 3 describes our original methodology to perform our conceptual analysis of context in ubiquitous systems. In Section 4 we present the results from a real field experiment illustrating the methodology. Finally, Section 5 concludes this paper and presents the perspectives of our work.

II. RELATED WORK

A. Context in Ubiquitous Computing

Context-awareness stands for the ability of a system to adapt its behavior (operations, services or content) to the current context without explicit user intervention [6][7]. Context-aware systems thus aim at increasing their own usability and effectiveness by taking into account environmental context [6].

Context is a widely concept, as pointed out by Bazire and Brézillon [8] and Coutaz et al. [9]. The most largely accepted definition considers context as *any information that can be used to characterize the situation of an entity (a person, place, or object) that is considered relevant to the interaction between a user and an application* [7]. The relevance of context information is central in this definition and determines its possible use in context-aware systems. According to Greenberg [10], several elements may contribute to the notion of context, and their relevance highly depends on specific situations. Every context-aware system has to determine context elements that will be observed according to its own goals. It is indeed impossible to enumerate in advance a full set of context elements that will apply to any system. This represents an important drawback for these systems, since the relevance of a context element indicates whether this information can be used for adaptation purposes.

Although important, relevance of context information did not receive enough attention on the literature, whose focus is on context management and adaptation. Several context management proposals can be found [2][7][11][12][13], most of them considering context elements and their relevance as predefined a priori during design time. The final purpose of these works is most often adaptation [9][14], namely, adaptation of software components [12], adaptation of supplied content or services [14], adaptation of service composition [13], adaptation by recommending a content or an action according to the user's context [15]. Whatever the purpose of

 ISBN: 978-1-61208-353-7

this adaptation, it is up to the context management infrastructure to offer all the necessary resources for handling and interpreting context information on context-aware systems.

B. Context and recommendation

Recommendation systems are personalization mechanisms that can help users find out interesting information or services [15]. Context-based recommendation systems [16][17] try to recommend content or services to users, based on context information observed during previous uses of the system.

Indeed, context information can be seen as a major criterion for recommendation systems [16]. Nevertheless, the notion of context adopted by traditional recommendation systems is often limited. For instance, Pignotti et al. [16] consider as context only time, user's location and history of previously invoked services. Other works, such as [17][18][19], propose recommendation mechanisms that are not limited to particular context elements. Najar et al. [17] present a prediction mechanism that intends to anticipate user's needs, recommending services according to previously observed context elements, organized in clusters. Similarly, Mayrhofer [18] uses recommendation techniques for anticipating context information and to predict the next likely situation of the user, while Sigg et al. [19] suggest to recommend context information in order to fulfill context description with missing elements based on similar previously observed contexts.

Most context-based recommendation systems use statistical methods and similarity measures for data analysis. Typical data analysis techniques adopted on these works include Bayesian Networks [20] and Markov Chains [17][19] models. Although obtained results are interesting, these methods suffer from some drawbacks. First, classification methods often ignore overlapping classes, preventing context elements to belong to multiple classes simultaneously, even if a context element can be observed in different situations. Besides, classes identified by classification methods are not necessarily understood by final users, which may lead to inappropriate recommendations. Finally, these methods usually require large sets of context data, which are not always available.

In this paper, we focus on context relevance, addressing these issues with a methodology for context analysis based on FCA and on association rules. On the one hand, FCA is a data analysis method that is able to group data at different levels of granularity and to organize them in a coherent set of overlapping classes. On the other hand, association rules allow discovering and quantifying relevant relations among observed values. Although well-known in traditional recommendation systems, FCA and association rules extraction algorithms are not fairly applied to context data. To the best of our knowledge, only a few works [21][22] have tried to apply these approaches to context data. Vanrompay et al. [21] use lattices to group common context data into communities of users, while Ramakrishnan et al. [22] combine Bayesian Networks and association rules to discover frequent correlation between context elements (without recommendation purposes). Indeed, applying these approaches to context data presents some challenges, notably related to data collection and formatting, due to the dynamic and heterogeneous nature of context data. None of these works [21][22] deals with such challenges, contrarily to our methodology. Before presenting it, the next sections introduce underlying analysis methods.

C. Formal Concept Analysis

FCA [3][4][5] allows performing a conceptual clustering, which helps discovering and structuring knowledge. FCA relies on the lattice theory, which defines a lattice as follows:

Definition 1: let $\leq$ be an ***order relation*** of a set E. $\leq$ defines a total order on E if all its elements may be compared by $\leq$: $\forall$ x, y $\in E^2$, x $\neq$ y => (x $\leq$ y $\vee$ y $\leq$ x). An order which is not total is partial.

Definition 2: a ***lattice*** is a partially ordered set (E, $\leq$) where each pair of elements has an upper and a lower bound. A lattice is complete *iff* any part S $\subseteq$ E has an upper bound (top) and a lower bound (bottom).

From a binary relation between a set of objects and a set of attributes, a ***Galois lattice*** (or concept lattice) builds a hierarchy of clusters called **formal concepts** [5]. These concepts are built from a table called **formal context**, which expresses the binary relation between objects and attributes.

Definition 3: a ***binary relation*** between sets *M* and *N* is a set of *(m, n)* pairs where $m \in M$ and $n \in N$. $(m, n) \in R$, also noted *mRn*, means that the element *m* is in relation with the element *n*.

Definition 4: a ***formal context*** is a triple $K = (G, M, I)$, where *G* and *M* are respectively the set of objects and the set of attributes and $I \subseteq G \times M$ is a binary relation between *G* and *M*. $(o, a) \in I$ means that *a* is an attribute of object *o*.

Derivation operations $(.)^I$ are defined for $O \subseteq G$ and $A \subseteq M$: $O^I = \{a \in M \mid \forall\ o \in O: o\ I\ a\}$ and $A^I = \{o \in G \mid \forall\ a \in A: o\ I\ a\}$. O^I is the set of attributes that are common to all objects of O and A^I is the set of objects that have all attributes of *A*.

Definition 5: a ***formal concept*** of context *(G, M, I)* is a pair *(O, A)*, where $O \subseteq G$ and $A \subseteq M$, $O = A^I$ and $A = O^I$. The set *O* is called the **extent** of concept *(O, A)* and *A* is its **intent**.

Definition 6: the set of all formal concepts and the partial order relation between them constitutes a lattice called ***Galois lattice*** of context *K*.

A Galois lattice [5] clusters objects into clusters (i.e., formal concepts) according to their common attributes. A lattice also specifies generalization or specialization relationships among these concepts. Indeed, some of them cluster objects with many common attributes (specific concepts) whereas some contain objects that share very few attributes (generic concepts). The most generic concept (upper bound) contains all objects in its extent, and the most specific one (lower bound) has all attributes in its intent.

In the methodology presented in Section III, a lattice clusters context elements according to user actions and reciprocally. The relationships between context elements and user actions are indeed made explicit in concepts. This allows identifying actions that occur in similar contexts. It also shows correlations among context elements, which is useful in case of missing data. Moreover, in order to quantify causal relations among them, we combine FCA with association rules, described in the following section.

D. Association Rules

Association rules extraction aims at discovering significant relationships between attributes extracted from databases [23]. Compared to other recommendation techniques, association

 ISBN: 978-1-61208-353-7

rules do not require computing a similarity measure. This is particularly interesting when the context elements and actions are not necessarily comparable, which is our case, since we do not make assumption about context elements.

An association rule is defined as an implication between two itemsets: $R: X \Rightarrow Y$, with $X \subseteq I$, $Y \subseteq I$ and $X \cap Y = \emptyset$. The rule R is said to be based on the frequent itemset $X \cup Y$ and the itemsets X and Y are called, respectively, *premise* and *conclusion* of R.

To check the validity of an association rule R, two measures are commonly used:

– *Support*: the support of rule R, denoted support(R) is equal to the frequency of simultaneous occurrences of itemsets X and Y *i.e., support(X U Y)*.

– *Confidence*: it expresses the conditional probability that a transaction contains Y, given that it contains X. The confidence of rule R, denoted *confidence(R)* is measured by the ratio *support (XUY) / support(X)*.

The extraction of association rules consists in determining the set of valid rules, i.e., whose support and confidence are at least equal, respectively, to a minimum support threshold *minsup* and a minimum confidence threshold *minconf* set by the user. This problem is decomposed in [24]: (i) *extraction* of all frequent itemsets with support greater than or equal to *minsup*; and (ii) *generation* of valid association rules (i.e., with confidence greater than or equal to *minconf*) based on the frequent itemsets extracted previously.

We have discussed in this section some related works on context-aware computing, pointing some of its challenges. The relevance and volume of context data that should be analyzed are examples of these challenges, especially for context-based recommendation. Next section presents our methodology that tries to overcome these challenges.

III. METHODOLOGY

We propose a methodology for unified and conceptual analysis of context, based on FCA and association rules. In this methodology, the impact of context elements is studied in two ways: by clustering context elements with FCA, identifying relationships among them, e.g., to detect redundant data or to complete missing data (due to measurement problems); and by extracting association rules to make explicit and to quantify the strength of relations among context elements themselves and between these and user actions. Our methodology is then divided into three steps: 1) collection and formatting of context elements from the user environment; 2) application of FCA and computation of lattices to structure collected context elements; and 3) extraction of association rules for evaluating the impact of context elements on the user actions, for recommendation purposes. It is worth noting that we focus on discovering user's usual behavior in order to analyze how context elements influence it and then to propose him/her applications (or actions) that he/she is more likely to execute in this context. Different from traditional recommendation systems, we are not interested in influencing the user's choices, but in detecting and reproducing them, similarly to a prediction mechanism.

A. Data Collection and Formatting

1) Collection of Context Elements and User Activities:

The starting point of our methodology is a set of raw data collected by sensors or recorded in log files. This step consists in gathering data related to user activities (i.e., applications executed on a mobile device) and environment (e.g., temporal information, location, network connection, etc.).

Storing contextual data is necessary for computation of lattice and association rules, as those are based on previously observed data. This data collection should, of course, respect privacy legislation, in particular in terms of explicit user agreement and anonymization. No assumption is made about the context elements we collect. Our approach considers, in a unified way, any context element. Potential interdependencies (or redundancies) among them will be identified during the analysis described in the following sections.

2) Data Formatting:

Collected raw data must be formatted in order to be processed by FCA. The input data is organized as a set of user activities (our objects) and a set of Boolean attributes, corresponding to observed values of context elements. During this phase, temporal data (i.e., timestamps) are transformed into time intervals, and location information (e.g., GPS coordinates) into geographical zones. At the end of this pre-treatment phase, each user activity can be associated to obtained values. Next step aims at extracting implicit relationships among contextual and activity data.

B. Extraction of Relationships among Data with FCA

As explained in Section II, FCA is a mathematical method that clusters data into concepts in lattices. We use FCA to organize contextual information into overlapping classes at different levels of granularity. Unlike other analysis methods, FCA can find a natural data structure, combining the user actions to the context elements observed during previous uses. This structure enables the connection of contextual data with user actions and allows building a knowledge base (*e.g.*, actions 1 and 2 are always executed in a similar context, which may suggest some proximity between these actions). Besides, the obtained concept lattice translates the hierarchical relationships between formal concepts, and can be used for classification and prediction purposes.

1) Formal Context Specification:

This step consists in identifying the data elements that will become the objects and attributes in a formal context. Formal context corresponds to a table, similar to Table I, that combines the activities performed by a user (objects A_i) and the corresponding context elements (attributes C_j). Thus, (A_i, C_j) indicates if the activity A_i has been performed in the presence of the context element C_j. For instance, activity A_2, in Table I, has been performed in contexts C_2 and C_3.

TABLE I. MATRIX REPRESENTING A FORMAL CONTEXT

Attributes / Objects	C1	C2	C3
A1	1	0	1
A2	0	1	1
A3	1	0	0

 ISBN: 978-1-61208-353-7

2) Construction of Galois Lattices and sub-lattice:

The formal context specified above is used to build a Galois lattice, clustering user activities and observed context elements, as illustrated in Figure 2.

When the lattice is not too big, its graphical representation can be interpreted visually to identify relationships between objects and attributes. However, the lattice grows fast when the number of objects and attributes increases. It is thus necessary to divide it into sub-lattices, by dividing context elements into subsets and computing the corresponding sub-lattices.

The cross-interpretation of sub-lattices allows identifying semantic links between context elements. Each sub-lattice (focused on a given context element) brings knowledge about activities conducted in this context and about correlations with other context elements. This information is extremely useful to complete missing data due to problems during data capture.

However, the analysis based on FCA has some limitations. First, the choice made when splitting context elements into sub-lattices may hide some relationships, which become more difficult to see depending on the way attributes have been separated. The second limit is that causal links between context elements and user activities are not quantified in lattices concepts. In other words, when multiple context elements are observed for a given activity, the lattice does not tell whether a context element is more "important" than another. The extraction of association rules, conducted in the following step, answers this question and can be used to propose recommendations.

C. Association Rules Extraction for Recommendation

Since activities are often used together with other activities, those can be considered as context elements as well. Indeed, there is often a semantic link between consecutive activities. This link is particularly interesting for recommendation purposes and thus for to the extraction of association rules. We have, therefore, considered for this work that any activity conducted within an interval of 30 minutes before a given instant belongs to the user context at this instant. These results form an enriched formal context, which is used as input data for the identification of association rules.

Before extracting association rules, we define two variables used to filter input data. Indeed, activities and context elements that are too frequent may hide interesting association rules among less frequent ones.

Activity frequency = n° of observations in which an activity appears / total n° of distinct values of (extended) context elements

Context element frequency = n° of observations in which this context element is associated to an activity / total n° of distinct observed activities

We have used the well-known Apriori [24] algorithm for the construction of frequent itemsets and the extraction of association rules. This algorithm operates in two phases: it first identifies the frequent itemsets that have a minimal support, then analyzes them to determine association rules whose confidence index is superior to a given threshold.

We consider that association rules with sufficient confidence value can be used for recommendation purposes. A recommendation such as $C_1,C_2,C_3,A_4,A_5 \Rightarrow A_6$ means that if a user is in the context (C_1,C_2,C_3) and has recently made actions A_4 and A_5, then action A_6 is recommended to him/her (as it is very likely that he/she will perform it). The recommendations based on association rules allow anticipating the next action of a given user based on previous usage.

We have applied our methodology to a case study, presented in the following section.

IV. CASE STUDY

In order to demonstrate our methodology, we applied it in a real data set, collected by a capture system on a user's tablet. The capture system is an Android application running in the background, which observes at regular time intervals the applications used and their execution context, without interfering with them. We have experimented this capture system with a single user during 69 days. The collected data has been stored into a SQLite database. For this first experiment, we have decided to consider just one user in order to better evaluate with him the results obtained with FCA and association rules.

A. Data Collection and Formatting

1) Collection of Context Elements and User Activities:

The capture system collects information about: *(i) applications* launched by the user on her/his device (*e.g.*, Facebook, Maps, Dropbox, etc.). We have identified 47 distinct applications of different categories (news, games, social networks, etc.); *(ii) geographical locations* visited by the user, which correspond to observed GPS coordinates; *(iii)* information on internal and external *memory* states of the device; *(iv) networks* to which the user has been connected over time. The capture system periodically observes these context elements, associating each observation to a *timestamp*. According to privacy legislation, the user involved in this experiment has been informed of the data collection process and has a full access to collected data, since these are locally stored in his personal device. The user also keeps full control of the collecting application, actively launching it.

2) Data Formatting

From the raw data collected in the previous phase, we have created the objects (corresponding to applications) and attributes (corresponding to observed context elements) needed to define formal contexts. These measured values cannot be taken into account as they are, since lattices require Boolean attributes. This is not a problem for network data, as for example the *Network_1* attribute is set to 1 if the user is connected to that network and to 0 otherwise. The possible values of (internal and external) memories are 1GB, 2GB and 3GB. We create, therefore, 6 attributes (*MemInt1*, *MemInt2*, *Memint3*, *MemExt1*, *Memext2*, *MemExt3*).

In the following, we detail the processing of temporal and geographical data, which are more complex.

3) Temporal data processing

We have transformed the measured timestamps into 6 time intervals per day: morning (*t_Morning*: 6h-12h), noon (*t_noon*:12h-13h), afternoon (*t_Afternoon*: 13h-20h), evening (*t_Evening*: 20h-00h), night (*t_Night*: 00h-06h). These intervals can, of course, be different depending on users and on their behavior. Each time range corresponds to a Boolean attribute, associated to applications used on the device.

 ISBN: 978-1-61208-353-7

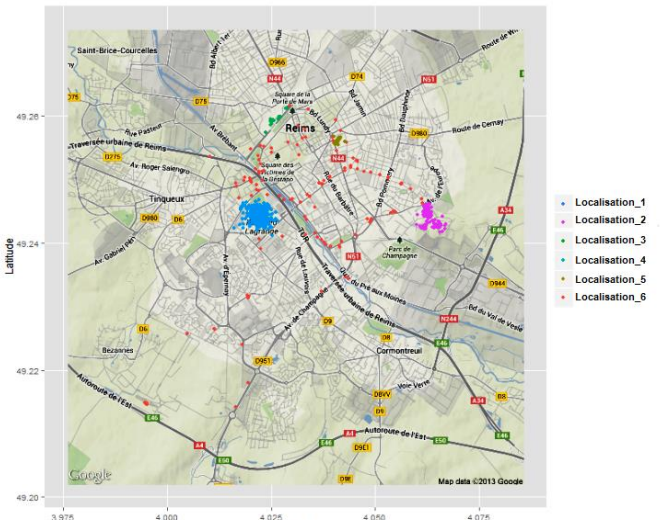


Figure 1. Zoom on the use of the tablet in a city.

4) Geographical data processing

We have processed collected geographical data in order to map them to relevant zones. We have identified these zones using R software, instead of dividing the longitude and latitude data into regular rectangular zones that would not be meaningful. The strength of this approach is that the number of zones is not fixed in advance and these zones do not need to have the same surface. Figure 1 shows the locations observed in our experiment. We may notice that most locations belong to a zone, whereas others places are visited much less frequently by the user. The mapping between each point and applications used at this location is then achieved later according to the identified geographical zones.

We identify all points in the dataset, which belong to a dense zone. Points that are not associated to a dense zone are then studied, in order to see if new dense zones appear. At the end of this process, points that are not associated to any zone are considered as movement locations (on the path between two zones), corresponding to a new zone. We have finally labeled each zone (*Location_1*, *Location_2*, etc.).

B. Formal Concept Analysis

Several tools exist for building Galois lattices, such as *Lattice Miner* [25] and *Conexp* [26]. We have used *Conexp* with the formal context described in previous section and built the associated lattice, illustrated in Figure 2.

As explained in Section III, the direct interpretation of the whole lattice may be difficult. We have therefore divided the original formal context and built sub-lattices. We have built the 3 sub-lattices corresponding to attributes related to location, networks and time periods respectively. Figure 3 shows the sub-lattice built from location attributes. It contains concepts corresponding to groups of applications used in various locations. This sub-lattice is much more readable than the whole lattice of Figure 2 and can be interpreted visually.

In order to know the applications that are used in a given geographical zone (*e.g.*, *Location_2*), we only have to find the corresponding concept in the formal context and identify the applications in its extent, as well as all inheriting applications (below that node). For instance, the applications used at the

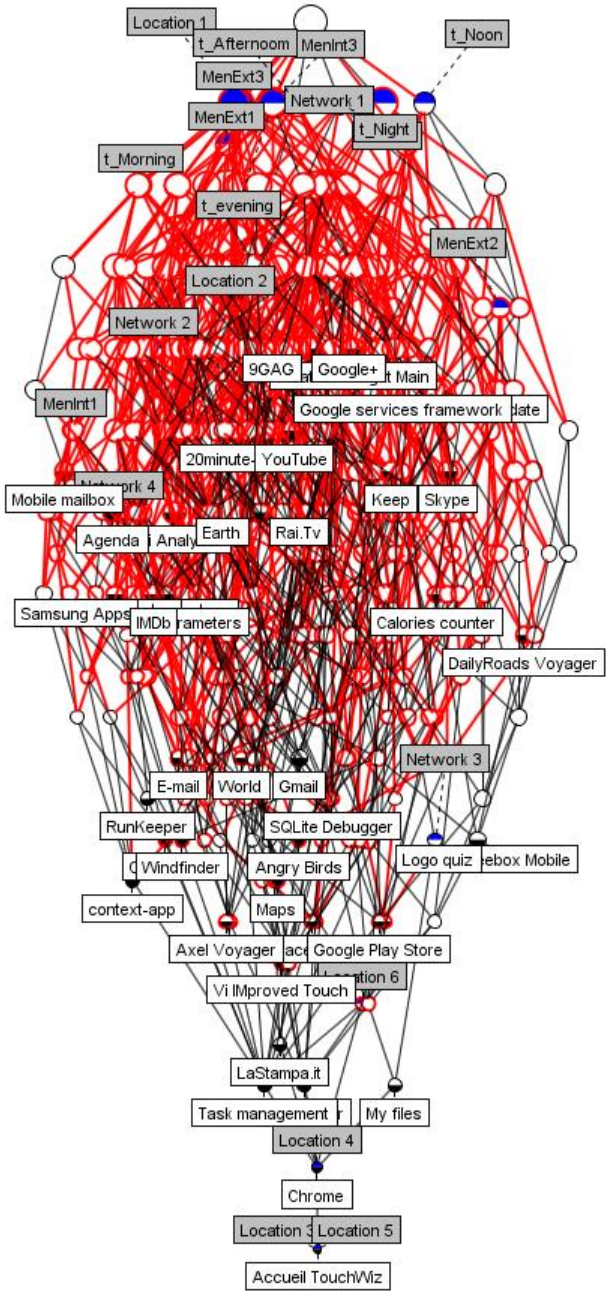


Figure 2. Global concept lattice (with all context elements).

Location_2 are *ConnectBot*, *camera*, *Drive*, *E-mail*, *Calendar*, *Chrome*, *TouchWiz*, *etc*. Applications that are common to two locations appear in a new node, which inherits from original nodes. The applications that have been used both at *Location_2 and at Location_3 are Calendar*, *Chrome and TouchWiz*. The lower a concept is in the lattice, the more specific it is, i.e., it contains more attributes in its intent. We proceed similarly to build the sub-lattice related to connection networks (Figure 4).

It should be noted that dysfunctions of the capture system may result in missing information. Some applications could thus not be associated to any context attribute. Therefore, they only appear in the upper bound of the lattice, which contains no attribute in its intent (cf. Figure 3 and Figure 4). For example, no location could be associated to the applications: *Alarm*,

 ISBN: 978-1-61208-353-7

Google+, Skype, Youtube... Likewise, no access network has been captured for the calculator application.

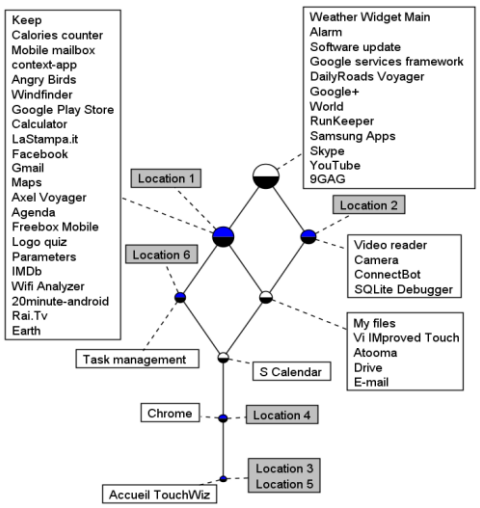


Figure 3. Sub-lattice corresponding to the locations context.

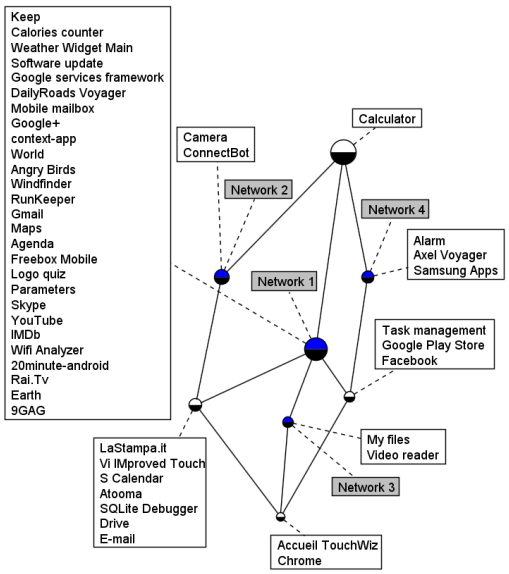


Figure 4. Sub-lattice corresponding to the network context.

We notice a strong relationship between *Location_1* and *Network_1* contexts, through the numerous intersections between these context attributes in terms of related applications (*Maps, Gmail, Agenda, etc.*), as shown on Figures 3 and 4. This relationship has been validated by the user, who confirmed that indeed *Network_1* is physically located in *Location_1*. Intersections like this one allow recovering missing information, i.e., information that could not be captured. We can, for instance, infer the access networks associated to the *calculator* application (on the top of the lattice in Figure 4), since it is often used in *Location_1*, which can be associated with the *Network_1* context.

C. Association Rules Extraction for Recommendation

During this step we have used an extended formal context as an input, in which we have considered recent applications as part of the user context, in addition to temporal, geographical and network connection attributes. We have also discarded very frequent applications, which do not bring relevant information, such as system applications. The frequency diagram of remaining applications is presented in Figure 5; applications with high frequencies are used in a significant proportion of contexts. This is the case for *Chrome*, with a frequency equal to 0.65.

Figure 6 shows the frequencies of context elements. When the frequency is high, the corresponding context element is frequently associated to applications. For example, Location_1 (which is the user's home) has a frequency of 0.65, which means that many applications are used from there.

We have applied the Apriori algorithm [24] to the extended and filtered applications and context elements. Apriori first computes the set of frequent itemsets together with their support measure, such as:

E1= [E-mail, Gmail, Rai.Tv, 20minute-android, t_Evening, Network 1, Google+], supp:11.62%

E1 is a frequent itemset (both context elements and applications) with a support equal to 11.62 %. E1 is a set of context elements for the user.

We have obtained about a hundred frequent itemsets, such as the ones shown in Table II. We have only kept the itemsets with a support superior to 10% for the extraction of association rules. The choice of a low support value allows considering a large fraction of frequent itemsets (further filtering is made later, as explained in the following). Among all generated rules, user has rejected only a small set (about 23%), most of

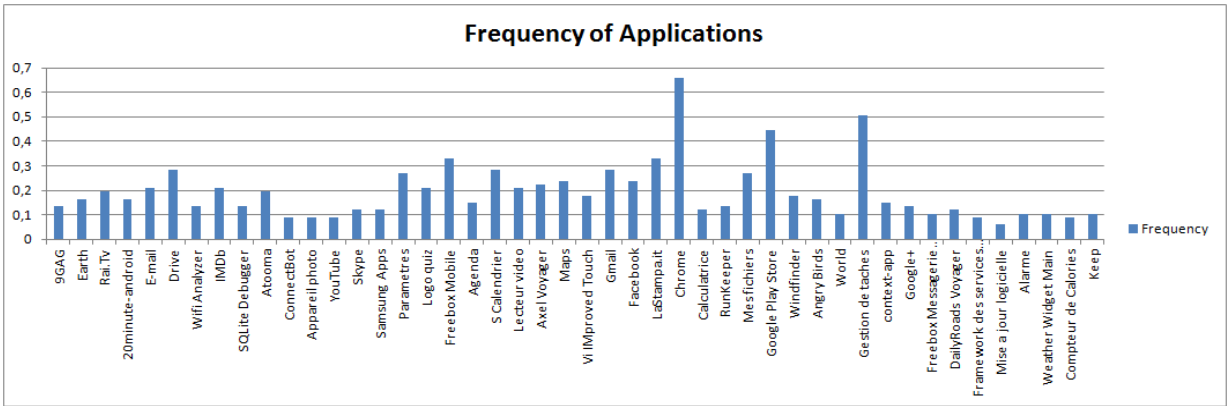


Figure 5. Applications frequency diagram

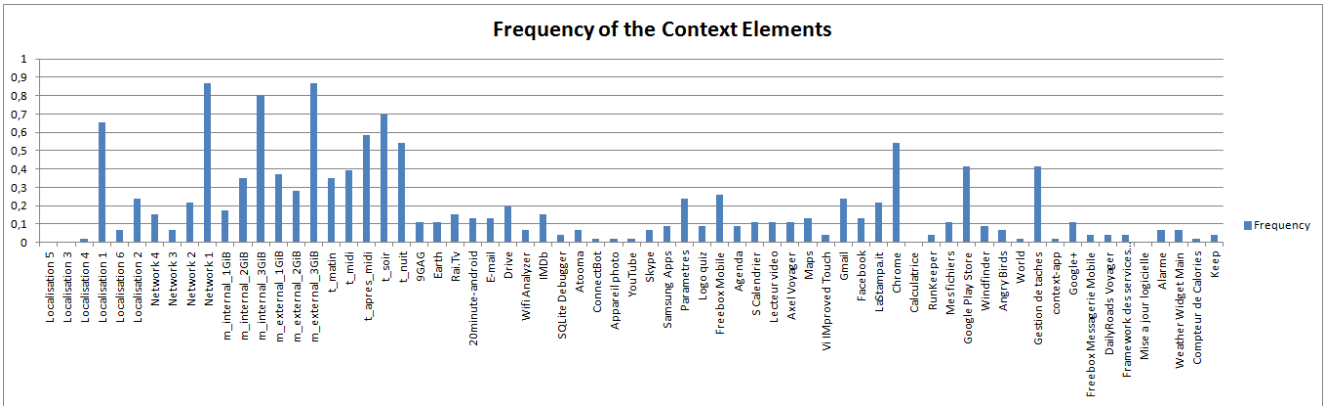


Figure 6. Context elements frequency diagram.

them with lower confidence values. When considering confidence above 75%, rejection decreases to 18,5%, which represents a promising result for us.

TABLE II. EXAMPLE OF FREQUENT ITEMSETS.

FI id	Frequent itemset	Supp
EIF69	[t_Evening, Chrome, Facebook]	13.95%
EIF95	[t_Evening, Network1, Maps, 9GAG]	11.62%
EIF100	[Network1, LaStampa.it]	25.58%
EIF101	[t_Evening, Localisation1, Network1, LaStampa.it]	20.93%
EIF106	[t_Evening, Network1, Gmail]	23.25%
EIF112	[t_Evening, Network1, Chrome]	34.88%
EIF117	[t_Night, Chrome]	37.20%
EIF118	[Network1, Chrome]	46.51%

TABLE III. EXAMPLES OF RECOMMENDATIONS

	Recommendations (association rules)	Conf
R1	[E-mail, Gmail, t_Evening, Network1] ⇒ [Google+]	83.33%
R43	[Network1] ⇒ [Google+]	13.51%
R57	[E-mail, t_Evening, Network1] ⇒ [Google+]	83.33%
R71	[t_Night, Localisation1, Chrome] ⇒ [Facebook]	35.71%
R82	[Network4] ⇒ [Facebook]	83.33%
R86	[t_Evening, Network4, Chrome] ⇒ [Samsung Apps]	83.33%
R95	[t_Evening, Network1, Maps] ⇒ [9GAG]	100.0%
R101	[t_Evening, Localisation1, Network1] ⇒ [LaStampa.it]	45.0%
R109	[t_Evening, Localisation1] ⇒ [Gmail]	40.90%
R123	[t_Afternoon, Localisation1] ⇒ [Chrome]	61.11%

From these itemsets, we have extracted all association rules. Then, we have eliminated (filtered) all rules whose conclusion is a context element, as our goal is to recommend applications. We have also filtered all the rules that contain incompatible contexts in their premise, e.g., with two different locations at the same time, e.g., *[t_Afternoon, t_Evening, Network1] => [Chrome]*. This rule will never be used as the user will never be simultaneously in the afternoon and in the evening. Table III shows a sample of recommendations for our case study (with different confidence values).

We have obtained in total 144 recommendations, 38 of which rely on association rules with a confidence equal to 1: these recommendations correspond indeed to the behavior identified by the user himself. Moreover, 103 recommendations have a confidence greater or equal to 50%.

V. CONCLUSION AND FUTURE WORK

In the approach presented in this paper, we have used FCA for the management of context and association rules for making recommendations. We have described existing context management approaches and shown their limitations. We have presented our methodology for context management, based on the analysis of formal contexts, the construction of Galois lattices and the extraction of association rules, in order to study the relationships between user actions and contextual information and to be able to give recommendations to users. We have described FCA in mathematical terms for explaining our method and the different underlying notions. Based on this theory, we have proposed a methodology for context analysis consisting of 3 steps. A data filtering and formatting are performed first in order to extract a formal context and thereafter build a lattice. The itemsets of this lattice are then interpreted with association rules to make appropriate recommendations and facilitate decision making. However, if the global lattice is too large, a decomposition into sub-lattices allows performing a visual analysis and making both an individual interpretation of each sub-lattice and a cross interpretation.

We have applied our methodology to a case study based on real data: we have used the data obtained by a capture system installed on the tablet of a user. The results have provided important information on the context of application usage, as well as relations between the different context elements. We have thus deduced information about the applications on all dimensions (contexts). We could also make appropriate recommendations with association rules. We could also complete the missing data due to occasional dysfunctions of the capture system.

The approach and solution proposed in this article open many perspectives for future work. The first one consists in devising mechanisms to automate cross-interpretations and associated recommendations. We will apply them to the concepts generated by the Galois lattice and all links between

 ISBN: 978-1-61208-353-7

these concepts, so as to automatically deduce an interpretation and recommendations with association rules.

We have used so far a limited number of context elements (geographical location, time, network connection and device memory). In the future, we will study the relevance of the other types of context elements to extend our approach. The same applies to users. Indeed, we are currently extending our case study to several users with different profiles (age, professional activity, etc.). We will also extend our experiment to include both automatic and non-automatic data collection, in order to identify other context elements that could be observed.

As future work, we will try to build the relations between actions and contexts themselves, and the relations between users' profiles. Thereby we seek to model a user profile (age, sex, student/employee, needs), according to available information, used applications, and moments, and then add information about the applications according to the different categories (Games, News, Entertainment, Economics, Education, Finance, Books, Weather, Sports, Travel), then define recommendations (Marketing, Increase sales, increase users' satisfaction, increase the audience). With this additional information we can provide new predictions and recommend new actions.

We will also apply our methodology to other case studies with more users, and we will evaluate the recommendations from them. In the longer term, we wish to manage context in real time, which will raise scalability issues due to the volume of data to be analyzed and to temporal constraints. A possible track consists in using distributed approaches for the construction and the update of the Galois lattice.

REFERENCES

[1] C. Bettini et al., "A survey of context modelling and reasoning techniques". Pervasive and Mobile Computing, vol. 6, no. 2, Elsevier Science Publishers, April 2010, pp. 161–180.

[2] J.P. Arcangeli et al.,"INCOME - Multi-scale Context Management for the Internet of Things", AmI '12 : International Joint Conference on Ambient Intelligence, Pisa : Italy, 2012, pp. 338-347.

[3] R. Wille, "Formal Concept Analysis as Mathematical Theory of Concepts and Concept Hierarchies". Formal Concept Analysis, B.Ganter et al., eds., pp. 1-33, Springer-Verlag, 2005.

[4] U. Priss, "Formal Concept Analysis in Information Science". In: Blaise, C. (ed.) Annual Review of Information Science and Technology, ASIST, vol. 40, 2006, pp. 521-543.

[5] A. Guénoche and V. Mechelen, "Galois approach to the induction of concepts". In I. Van Mechelen, J. Hampton, R.S. Michalski, & P. Theuns (Eds.), Categories and concepts: Theoretical and inductive data analysis, London: Academic Press, pp. 287-308, 1993.

[6] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems". International Journal of Ad Hoc and Ubiquitous Computing, vol. 2, no.4, 2007, pp. 263-277.

[7] A.K. Dey, "Understanding and Using Context". Personal and Ubiquitous Computing, vol. 5, no. 1, 2001, pp. 4-7.

[8] M. Bazire and P. Brézillon, "Understanding Context Before Using It". In: Dey, A.K., Kokinov, B., Leake, D.B., Turner, R. (eds.) CONTEXT 2005. LNCS (LNAI), vol. 3554, 2005, pp. 29–40.

[9] J. Coutaz, J. L. Crowley, S. Dobson, and D. Garlan, "Context is key". Communcations of the ACM, vol. 48, no. 3, 2005, pp. 49-53.

[10] S. Greenberg, "Context as a dynamic construct". Human-Computing Interaction, vol. 16, no. 2-4, 2001, pp. 257-268.

[11] D. Conan, R. Rouvoy, and L. Seinturier, "Scalable Processing of Context Information with COSMOS", Proc. 7th IFIP International Conference on Distributed Applications and Interoperable Systems, Springer-Verlag, Lecture Notes in Computer Science Volume 4531, Paphos, Cyprus, June 2007, pp. 210-224.

[12] M.Wagner, R. Reichle, and K.Geihs, "Context as a service - Requirements, design and middleware support". Int. Conf. on Pervasive Computing and Communications Workshops, 2011, pp. 220-225

[13] N. Taylor et al., "Pervasive Computing in Daidalos", Pervasive Computing, vol. 10, no.1, 2011, pp. 74 -81.

[14] J. Gensel, M. Villanova-Oliver, and M. Kirsch-Pinheiro, "Modèles de contexte pour l'adaptation à l'utilisateur dans des Systèmes d'Information Web collaboratifs" ("Context models for user adaptation on collaborative Web Information Systems"). 8èmes J. Francophones d'Extraction et Gestion des Connaissances (EGC'08), Atelier sur la Modélisation Utilisateur et Personalisation d'Interfaces Web (in French), 2008, pp. 5-15.

[15] Q. Wen and J. He, "Personalized Recommendation Services Based on Service-Oriented Architecture". In: IEEE Asia-Pacific Conference on Services Computing, 2006, pp. 356–361.

[16] E. Pignotti, P. Edwards, and G.A. Grimnes, "Context-Aware Personalised Service Delivery". European Conference on Artificial Intelligence, ECAI, 2004, pp. 1077-1078.

[17] S. Najar., M. Kirsch-Pinheiro, Y. Vanrompay, L.A. Steffenel, and C. Souveyet, "Intention Prediction Mechanism In An Intentional Pervasive Information System". Intelligent Technologies and Techniques for Pervasive Computing, 2013, pp. 251-275.

[18] R. Mayrhofer, "An Architecture for Context Prediction". PhD thesis, Johannes Kepler University, 2004.

[19] S. Sigg, S. Haseloff, and K. David, "An Alignment Approach for Context Prediction Tasks in UbiComp Environments". Pervasive Computing, vol.9, no. 4, 2010, pp. 90-97.

[20] A.K. Ramakrishnan, D. Preuveneers, and Y. Berbers, "A Loosely Coupled and Distributed Bayesian Framework for Multi-context Recognition in Dynamic Ubiquitous Environments", 10th International Conference on Ubiquitous Intelligence & Computing and 10th International Conference on Autonomic & Trusted Computing (UIC/ATC 2013), IEEE,2013, pp. 270-277.

[21] Y. Vanrompay, M. Kirsch Pinheiro, N. Ben Mustapha, and M.-A. Aufaure, "Context-Based Grouping and Recommendation in MANETs", In K. Kolomvatsos, C. Anagnostopoulos, and S. Hadjiefthymiades (Eds.), Intelligent Technologies and Techniques for Pervasive Computing, IGI Global, pp. 157-178, 2013.

[22] A.K. Ramakrishnan, D. Preuveneers, and Y. Berbers, "Enabling Self-learning in Dynamic and Open IoT Environments", The 5th International Conference on Ambient Systems, Networks and Technologies (ANT-2014), the 4th International Conference on Sustainable Energy Information Technology (SEIT-2014), Procedia Computer Science, vol. 32, 2014, pp. 207-214.

[23] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases". In Proceedings of the ACM SIGMOD Intl. Conference on Management of data, Washington, USA. June 1993, pp. 207–216.

[24] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules". Proceedings of the 20th International Conference on Very Large Databases, June 1994, pp. 478–499.

[25] B. Lahcen and L. Kwuida, "Lattice Miner: A Tool for Concept Lattice Construction and Exploration". In: 8th Int. Conf. on Formal Concept Analysis, Agadir, Marocco http://lattice-miner.sourceforge.net/ [retrieved: June 2014], 2010.

[26] S.A. Yevtushenko, "System of data analysis "Concept Explorer"". In: Proceedings of the 7th national conference on Artificial Intelligence, Russia, http://conexp.sourceforge.net/ [retrieved: June 2014], pp. 127–134 (in Russian), 2000.

 ISBN: 978-1-61208-353-7

Standardized Scalable Relocatable Context-Aware Middleware for Mobile aPplications (SCAMMP)

Fatima Abdallah

Faculty of Sciences
Lebanese University
Beirut, Lebanon
Email: f.3abdallah@gmail.com

Hassan Sbeity and Ahmad Fadlallah

Faculty of Computer Studies
Arab Open University
Beirut, Lebanon
Email: {hsbeity,afadlallah}@aou.edu.lb

***Abstract*—The penetration of handheld devices (especially smartphones) is predicted to be over one billion in the next five years. These devices are increasingly equipped with new sensors offering a great potential for developing context-aware mobile applications that can enhance user experience. Unfortunately, the data provided by these sensors are of low-level (raw data) and diverse, ranging from physical to virtual. This makes embedding contextual information into mobile applications a difficult task. Presenting these raw sensors' data in a unified format, augmenting them into useful high-level context information and offering them through a well formalized standard middleware service can make this task easier. In this work, we present the architecture of a middleware platform that provides an open standard interface offering high-level information to the application layer. This platform maintains user context information in a finite state machine through state engines. State engines that represent user states can be added and removed any time and hence the openness of the platform originates. The platform uses layered approach and is composed of two relocatable layers: data acquisition-augmentation (pre-processing) layer and decision layer. A case study was performed to validate the functionality of the platform.**

Keywords–Context Awareness; Middleware; Mobile Applications.

I. INTRODUCTION

Most today smart devices have built-in sensors that measure motion, orientation, and various environmental conditions. These sensors are capable of providing raw data with high precision and accuracy. Every smartphone nowadays is equipped with a set of small sensors. These sensors can be hardware-based embedded in the smartphone (e.g., acceleration sensor) or software-based that derives their data from one or more hardware-based sensors (e.g., linear acceleration sensor). A third category of sensors could be introduced, which is the logical sensors (e.g., calendar events).

Providing mobile applications with high level sensor information can enhance the efficiency of these applications toward power saving and user experience. Examples are many. For instance when the user is traveling, application that needs synchronization with cloud services (data upload/download through mobile network), can postpone these jobs until the user is at home or at work in order to save battery power even if the mobile network provides a high bandwidth data connection over LTE (Long Term Evolution) for instance. Because once the battery of the mobile is drained, recharging the phone is difficult while traveling. Another example is making the phone silent when the user is sleeping, which enhances the user experience. A power friendly operating system process scheduler can swap processes from memory to persistent storage while taking into consideration the user state. A reminder application can notify the user events not only based on time and dates but also based on his location and according to what he/she is doing. For instance, an alarm can be set based on date time and user states; one can choose ringing when sleeping only, awake only, or both. Traditionally, applications have input from the user, persistent storage, and recently from the network via Remote Procedure Calls (RPC) for message passing. But offering meaningful user context information originally gathered from different physical and virtual sensors, provides a new input source that, if standardized, will provide a new brand of applications. Furthermore, the history of these high-level user context information can be used as a user signature to authenticate the user to his/her device.

In this work, we present the architecture of a middleware platform that provides an open standard interface offering high-level information to the application layer. This platform maintains user context information in a finite state machine through state engines. State engines, that represent user states, can be added and removed any time, and hence the openness of the platform originates. The platform uses layered approach and is composed of two relocatable layers: data acquisition-augmentation (pre-processing) layer and decision layer. We also present a case study that decides about the current user location, its purpose is to test the validity of the middleware by mapping the different components of the application to the different SCAMMP modules.

The rest of this document is organized as follows: Section II presents the architecture of the middleware platform. Section III describes the different blocks of the decision layer. Section IV explains the modules of the Data Acquisition-augmentation Layer. Section VI explores a set of related work and compares them with our work. Section V presents a case study as an evaluation of the functionalities of the different SCAMMP components. Finally, Section VII concludes the paper and presents the future work.

II. ARCHITECTURE

Fig. 1 depicts the general architecture of the middleware platform. The application layer represents any mobile ap-

 ISBN: 978-1-61208-353-7

plication that might embed context-aware information. The Decision Layer located at the top of the middleware platform (SCAMMP) provides the application layer with an Application Programming Interface (API) to access high-level context-aware information. This information is stored in a finite state machine and represents in a real time fashion the different user states. Hence, any application can easily integrate user context information. At the bottom is the Acquisition Layer, its main role is to represent the data captured from different sensors in a unified XML format.

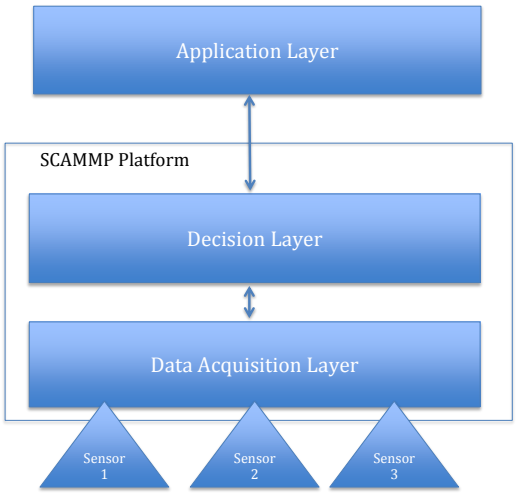


Figure 1. SCAMMP Architecture.

Each layer provides services to the layer above through a well-defined set of commands (protocol). At each layer, modules can be added and removed dynamically, hence guaranteeing the openness and the scalability of the platform. The communication between the different layers is accomplished as follows: The lower layer sends notification (using push mechanism) to the upper layer each time a sensor notification is received signaling the availability of valid data. If the upper layer is interested, it issues a request asking for the new update using a predefined protocol. Moreover, at any time the upper layer can issue a command to the lower layer asking for data updates. The separation of the system into different layers offers a great flexibility for layer hosting and hence the relocatabilty of the platform is originating.

III. DECISION LAYER

The main task of the decision layer is to maintain a finite state machine that reflects the different user states in a soft real time. Fig. 2 depicts the different components that build up this layer. All components should reside in the same address space and hence, intra-communication can be done through shared variables.

The API component represents the interface to the application layer and the controller component provides an interface to the lower layer, namely the data acquisition layer. All remaining components have no direct access outside this layer. This layer can be hosted on the mobile device or on the cloud; it can be relocated depending on the available bandwidth. The history of the user states can also be uploaded on the cloud.

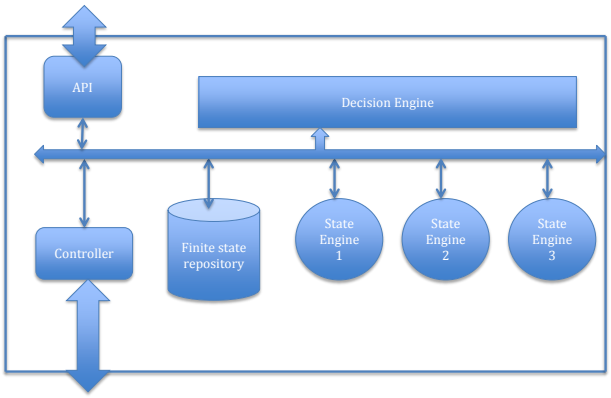


Figure 2. The Architecture of the Decision Layer.

A. API

The API component is the only interface provided to the application layer to communicate with SCAMMP. A set of pre-defined commands (protocol) are used as a communication mean between the application layer and the SCAMMP. In fact, the current and the history of the different user states will be made available to the application layer. For instance, currently the user is at home (state) and is sleeping (attribute). The API component has access to the state repository where the current and the history of the different user states can be found. The communication protocol can be simple, such as the HTTP, where any application at the application layer can send a request and receive response. There are two types of requests that can be sent to the API component, one that requests a list of the available user states and their attributes (names and descriptions) and one that requests the current or the history of the different user states for a specific period of time.

B. Decision engine

The decision engine is responsible for updating the different user states (along with the corresponding attributes). The kernel of the decision engine is based on a mathematical model that decides about the current user state. The decision is based on input from the different state engines. Each time a state engine makes an update, the decision engine is informed in order to recheck the user state and eventually update it. The outcome of this engine will be made available to the application layer through the API component.

C. Finite state repository

The finite state repository is a storage system. It contains three types of data, two of them are available to the application layer through the API component and the third one is only for internal use.

The first data is an XML entry list that contains an entry of every registered state engine. The second data is an XML entry list that stores the current and the history of the different user states. The third data is also an XML entry list that contains the output of every state engine. It is only for internal use and is available for the decision engine. Because the history of the user state could be huge after a while, it can be archived and

 ISBN: 978-1-61208-353-7

uploaded to the cloud, while still be available to the application layer.

D. State engine

Every user state can have many attributes, for instance, home is a user state and sleeping is its attribute. Every state engine is attached to one or more sensor agents of the lower layer. The finite state engines take input from the sensor agents through the controller, and produce output in the finite state repository. The data produced by the finite state engine are only for internal use, namely the decision engine use them as input. The main task of the state engine is to calculate the certainty of a certain user state and it is left to the decision layer to decide which is the current state of the user.

Each time a sensor agent sends a notification to announce the availability of new sensor data, the corresponding finite state engine will be informed via the controller. It is up to the state engine to decide whether to request or not the data. A new finite state engine should be introduced to the controller in order to be registered. During the registration of a new state engine, a list of the corresponding sensor engines of the lower layer needs to be specified.

E. Controller

The controller's main role is to maintain the communication with the lower layer and hence isolating the decision layer form the data acquisition layer. It receives notifications form the sensor agents of the lower layer and forwards them to the corresponding state engines. It sends requests to the lower layer on behalf of the different finite state engines and forwards the response to the corresponding finite state engines. The second role is to maintain a list of the active finite state engines. Each time a new state engine is introduced, it has to be registered at the controller.

IV. Data Acquisition-augmentation Layer

The main role of the Data Acquisition-augmentation layer (see Figure 3) is to provide a unified format of the data collected from different sensors.

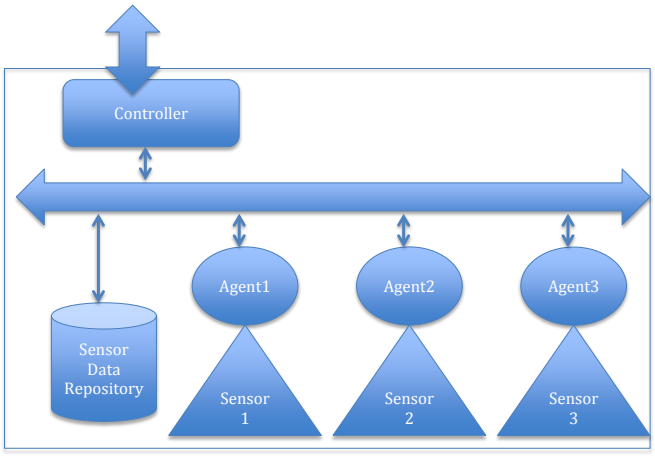


Figure 3. Architecture of the Acquisition layer.

Every sensor whether physical or virtual will be attached to a single dedicated agent. Once a sensor has produced a new data, the corresponding engine will decide according to a certain threshold whether to forward a notification to the upper layer or not; if yes, the agent will collect the data and store it in the data repository in a unified XML schema (see Figure 4).

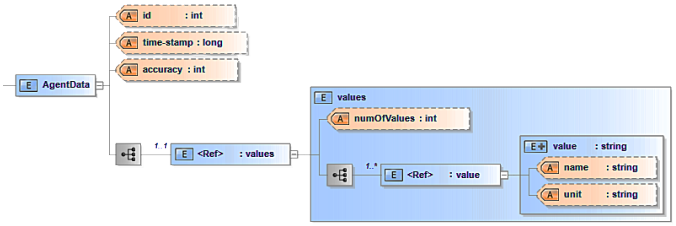


Figure 4. Unified Agent Data XML schema.

These data will be made available to the upper layer through the controller.

A. Sensors

Most smartphones nowadays are equipped with many small sensors. As previously mentioned, some of these sensors are hardware-based and some are software-based. Hardware-based sensors are physical components built into handsets or tablet devices. They derive their data by directly measuring specific environmental properties, such as acceleration, geomagnetic field strength, or angular change. Software-based sensors are not physical devices, although they mimic hardware-based sensors. Software-based sensors derive their data from one or more of the hardware-based sensors and are sometimes called virtual sensors or synthetic sensors. The linear acceleration sensor and the gravity sensor are examples of software-based sensors. Another category of sensors is logical sensors, such as the tweets and the calendar events.

B. Agents

Most popular mobile operating system (OS) provides sensor framework as an API. For instance, Android-powered mobile devices provide raw sensor data by using the Android sensor framework. The sensor framework is part of the android hardware package and includes many classes and interfaces (SensorManager, Sensor, SensorEvent, SensorEventListener, etc.). The agents encapsulate the OS framework to provide a homogeneous sensor data representation. A unified XML schema is used by all agents to represent the captured sensor data. The main role of the agent is to convert the raw sensor data into a unified XML format (see Figure 4). Examples are illustrated for the three categories of sensors (physical, virtual and logical) in Figs. 5 (for the accelerometer sensor), 6 (for the Battery sensor), and 7 (for the calendar sensor). This is done using a threshold that is based on the difference of two consecutive carried data. For every sensor (whether physical, virtual, or logical) there will be a dedicated agent. The internal implementation of the agent is sensor-dependent.

 ISBN: 978-1-61208-353-7

```
<AgentData id="1"
           time-stamp="10000"
           accuracy="1">
    <values numOfValues="3">
        <value name="x"
               unit="m/s^2">9.45</value>
        <value name="y"
               unit="m/s^2">1.34</value>
        <value name="z"
               unit="m/s^2">2.7</value>
    </values>
</AgentData>
```

Figure 5. Agent captured data of the Accelerometer sensor.

```
<AgentData id="2"
           time-stamp="10990"
           accuracy="1">

    <values numOfValues="2">
        <value name="Level"
               unit="unitless">88</value>
        <value name="Status"
               unit="unitless">discharging</value>
    </values>

</AgentData>
```

Figure 6. Agent captured data of the Battery sensor.

```
<AgentData id="3"
           time-stamp="00"
           accuracy="1">
    <values numOfValues="3">
        <value name="eventName"
               unit="unitless">Meet Manager</value>
        <value name="from"
               unit="hh:mm">11:30</value>
        <value name="to"
               unit="hh:mm">12:10</value>
        <value name="date"
               unit="dd/mm/yyyy">10/12/2014</value>
        <value name="reminderTime"
               unit="minutes">60</value>
    </values>
</AgentData>
```

Figure 7. Agent captured data of the Calendar sensor.

Each time a new agent is created, it has to be registered in the controller repository using a unified XML format (see Figure 8). Fig. 9 depicts the content of an agent's registration file containing three different agents.

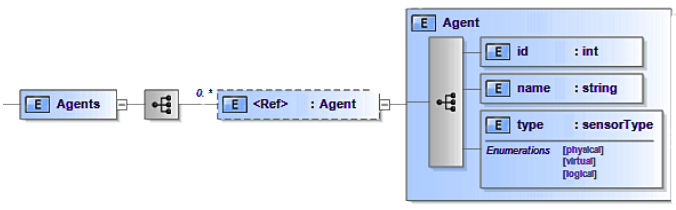


Figure 8. Agent directory XML schema.

```
<Agents>
    <Agent>
        <id>1</id>
        <name>Accelerometer</name>
        <type>physical</type>
    </Agent>
    <Agent>
        <id>2</id>
        <name>Camera</name>
        <type>logical</type>
    </Agent>
    <Agent>
        <id>3</id>
        <name>Calender</name>
        <type>virtual</type>
    </Agent>
</Agents>
```

Figure 9. Agents registration file.

C. Controller

The controller is the interface of the data acquisition-augmentation layer to the decision layer. The controller roles are to:

- Maintain a repository that has an entry for every registered agent using the unified XML format (see Figure 8)
- Receive notifications from the agents and issue simple commands to the agent, such as switch the sensor data acquisitions on and off.
- Forward notifications to the upper layer once received from the agents.
- Have read access to the repository that holds the data stored using a unified format (see Figure 4) collected from the different agents.

V. Case Study

The main goal of SCAMMP is to provide an open standard middleware that offers user context-aware information through an API to the application layer. Hence, any application that wishes to integrate context-aware information can use this API . In order to evaluate SCAMMP, we consider one case study that decides about the current user's location. By design, SCAMMP is intended to host decision logic . Thus, we decided to integrate user-location logic in order to evaluate SCAMMP by mapping the different components of the application in the different SCAMMP modules. It is important to mention that this mapping is done statically in order to evaluate the functionality of SCAMMP; The intra- and inter- communication of the different SCAMMP entities is simulated (done manually since the system is not fully implemented yet).

A. Data Acquisition-augmentation layer mapping

We define three agents required to determine the user's location. These agents embody the following sensors: Location

 ISBN: 978-1-61208-353-7

(hardware/software sensor), network connection (software sensor) and calendar (logical sensor) sensors. They are considered as input for the "Location State Engine". These agents convert the raw data generated by the embodied sensors into a unified XML data as per the schema defined in Fig. 4.

Location Agent: Both Android and IOS operating systems offer a location framework that determines the location of the device. It is a software-based sensor that uses the GPS, cell tower information, and connected Wifi network to detect the user's location. It returns the location using the attributes: longitude, latitude, altitude, accuracy in meters, and time. Fig. 10 is a sample data produced by the Location Agent.

```
<?xml version="1.0" encoding="utf-8"?>
<AgentData id="4"
           time-stamp="1000"
           accuracy="20">
    <values numOfValues="3">
        <value name="longitude" unit="degree">35.528873</value>
        <value name="latitude"  unit="degree">33.8662331</value>
        <value name="altitude"  unit="meter">0</value>
    </values>
</AgentData>
```

Figure 10. Location Agent Data.

Network Connection Agent: This agent determines the type of network the user is currently connected to (e.g., WiFi and Mobile Data network). This information can be obtained from the "Connectivity Manager" in the mobile operating system. The agent's role is to send a notification whenever the user switches from one type of connection to another. The returned data includes: connection type, connection SSID for Wifi networks, and the set of cell towers the device is connected to. Fig. 11 represents real sample data for a Wifi network connection with SSID 'Alfa'.

```
<?xml version="1.0" encoding="utf-8"?>
<AgentData id="5"
           time-stamp="1000"
           accuracy="100">
    <values numOfValues="2">
        <value name="type" unit="unit-less">Wifi</value>
        <value name="ssid" unit="int">Alfa</value>
    </values>
</AgentData>
```

Figure 11. Network Connection Agent Data.

Calendar Agent: This is a logical agent that can be either local or hosted on the cloud. The information collected from the embodied sensor can be used (at the decision layer) to raise the certainty of the location obtained from other agents. Fig. 12 is a sample real data presenting a calendar event named 'Meet Manager'.

B. Decision layer mapping

The agents presented above are attached to a single engine called "Location State Engine" at the decision layer. This engine's kernel can determine user's location (Home, Work, or elsewhere) using Relational Markov Model [7], and K-Nearest Neighbors (KNN) Algorithm [8].

```
<?xml version="1.0" encoding="utf-8"?>
<AgentData id="3"
           time-stamp="93262"
           accuracy="100">
    <values numOfValues="5">
        <value unit="unit-less"  name="eventName">Meet Manager</value>
        <value unit="hh:mm"      name="from">11:30</value>
        <value unit="hh:mm"      name="to">12:10</value>
        <value unit="dd/mm/yyyy" name="date">10/12/2014</value>
        <value unit="minutes"    name="reminderTime">60</value>
    </values>
</AgentData>
```

Figure 12. Calendar Agent Data.

C. Simulation

The engine will remain for a period of time collecting the locations that the user frequently visits (learning phase). It associates for each location the active network connection of the device and the time of identifying this location. After collecting data, the location history is analyzed using a simple heuristic to determine the user's home and work locations. This heuristic will work only for users with fixed work location, since it is most likely that a user is at home at night time, and at work in weekdays in the middle hours of the day. To overcome this drawback many works have been done to obtain personal significant places from raw location data using Relational Markov Model, and K-Nearest Neighbors (KNN) Algorithm. We choose to use KNN as a classification method, for this sake we recorded the locations of a mobile holder for 3 days in a frequency of one hour. Each location is classified by one of the labels {0,1,2} corresponding to {Home, Work, Elsewhere} respectively. Table I presents a sample of the collected data.

TABLE I. TRAINING DATA SAMPLE

Date	Time	Longitude	Latitude	Location
4-9-2014	10:06:45	35.5263591	33.8657569	1
4-9-2014	11:05:37	35.5181525	33.8368607	2
4-9-2014	12:45:07	35.5638238	33.8653605	1
4-9-2014	14:45:40	35.5291208	33.8660798	1
4-9-2014	15:03:26	35.528873	33.8662331	1
4-9-2014	16:19:06	35.5146795	33.8498035	0
4-9-2014	17:21:33	35.514577	33.849815	0

During the training phase the "Location State Engine" uses only the Location Agent, it transforms the training data (longitude, latitude) into a matrix. This matrix is the input of the processing phase where the KNN classifier is obtained. As a simulation, we used Matlab [9] to create the classifier. The produced classifier predicted incorrectly 3% of the training data for k=3. The output of the state engine is an XML file dedicated for internal use (see Figure 13), it is used by the decision engine to aggregate outputs from different state engines and provide the final users state through the API.

After setting the classifier, and in order to save battery power, the "Location State Engine" can use the active network connection to decide the user's location without using the Location Agent. Since the device usually connects, the accuracy of the decision is raised by the calendar events. So if an event points that the user created a shopping checklist or have a meeting at a specific hour, the engine could confirm his location in the shop or at work according to the time.

 ISBN: 978-1-61208-353-7

```
<StateEngineData id="1"
                 time-stamp="293722"
                 accuracy="80">
    <values numOfValues="1">
        <value name="Location">Home</value>
    </values>
</StateEngineData>
```

Figure 13. State Engine output sample.

VI. Related Work

The multiplicity and diversity of sensors embedded within mobile devices makes context aware applications difficult to develop, so middleware solutions were proposed to provide an abstraction layer between the operating system and applications. In this section, we review some of proposed middleware solutions of the context-aware systems.

Baldauf et al.[1] in their survey over context aware systems, concluded that there is a common layered conceptual framework for most systems: Starting from the low level (Sensors Layer) passing through the Raw Data Retrieval Layer and Preprocessing Layer that raise the level of abstraction of context data. Then the Storage and Management Layer that provides an interface for the Application Layer to obtain what is needed from the collected data. Although most systems have a common architecture, but they differ in the kind of target applications they serve. Some of the architectures gather information for general context-aware applications, such as smart homes, intelligent vehicles, and context aware hospitals. Other systems, including SCAMMP, are specialized for mobile devices.

Henricksen et al. [2] proposed the **PACE** middleware that supports heterogeneity, mobility, traceability and control, and deployment and configuration of new components, which are some of the requirements for context-aware middleware. On the other hand, it doesn't achieve the scalability requirement. This middleware is developed for context-aware systems in general rather than mobile devices. The middleware is divided into 3 layers: *Context Repositories Layer*, *Decision Support Tools Layer*, and *Application Components Layer*. Dey et al. [3] presents a framework that supports the rapid development of general context-aware applications. Using the *Context Widget*, *Interpreters*, *Aggregators*, and *Services*, it separates the context acquisition from the use of context in the application. The **Context Toolkit** was implemented to instantiate the framework, but it is limited in scalability and ease of deployment and configuration. Another generic context-aware framework is **CMF** [4]. It is a scalable context-aware framework that enables processing and exchange of heterogeneous context information. The *Context Source* uses reasoning techniques to integrate data collected from different sensors, and offers them to the *Context Provider*. The framework takes benefits from user profiles stored in the *User Management* component. The sentient object model is proposed by the **CORTEX** [5] project for the development of context-aware applications in mobile ad hoc environments. A sentient object is a mobile intelligent software component that senses the surrounding environment via sensors and other sentient objects. It consists of three parts: *Sensory Capture*, *Context Hierarchy*, and *Inference Engine*. The model was improved more in [6] by adding the reflection capability and the Service Discovery component. It was also tested by building an intelligent vehicle application.

All the referenced approaches are similar to **SCAMMP** in the way they collect data from different sensors, raise its abstraction level, and offer context information for applications. But SCAMMP is a standard, relocatable, and scalable middleware for applications targeting handheld devices. Using the layered approach allows any of the layers to be hosted on cloud. In addition, the collected data can be stored on a remote server to overcome the storage limitation. The scalability of the middleware is obtained by using a unified XML schema to register additional state engines and physical, virtual, or logical sensors. The provided API feeds applications with high level context information in a finite state model that represents the user's state (Home, Work, Traveling, etc.).

VII. Conclusion and Future Work

The main goal of SCAMMP is to provide an open and standard middleware, that offers user context-aware information to the application layer through a well-defined API, which can be accessible by any future application wishing to integrate context-aware information. SCAMMP is open in a way that new user state engines can be added dynamically. It is relocatable allowing, for instance, the decision layer to be hosted on the cloud. It uses standard protocol for inter-layer communication and URI for name spacing. These different aspects (standard, open, and relocatble) distinguish SCAMMP from the currently proposed middlewares. Our current and future works are the detailed design and implementation of SCAMMP, and its evaluation through different case studies.

References

[1] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems," International Journal of Ad Hoc and Ubiquitous Computing, vol. 2, no. 4, 2007, pp. 263–277.

[2] K. Henricksen, J. Indulska, T. McFadden, and S. Balasubramaniam, "Middleware for distributed context-aware systems," in On the Move to Meaningful Internet Systems 2005: CoopIS, DOA, and ODBASE. Springer, 2005, pp. 846–863.

[3] A. K. Dey, G. D. Abowd, and D. Salber, "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications," Human-computer interaction, vol. 16, no. 2, 2001, pp. 97–166.

[4] H. Van Kranenburg, M. Bargh, S. Iacob, and A. Peddemors, "A context management framework for supporting context-aware distributed applications," Communications Magazine, IEEE, vol. 44, no. 8, 2006, pp. 67–74.

[5] G. Biegel and V. Cahill, "A framework for developing mobile, context-aware applications," in Pervasive Computing and Communications, 2004. PerCom 2004. Proceedings of the Second IEEE Annual Conference on. IEEE, 2004, pp. 361–365.

[6] C.-F. Sørensen et al., "A context-aware middleware for applications in mobile ad hoc environments," in Proceedings of the 2nd workshop on Middleware for pervasive and ad-hoc computing. ACM, 2004, pp. 107–110.

[7] C. Zhou, N. Bhatnagar, S. Shekhar, and L. Terveen, "Mining personally important places from gps tracks," in Data Engineering Workshop, 2007 IEEE 23rd International Conference on. IEEE, 2007, pp. 517–526.

[8] L. Liao, D. J. Patterson, D. Fox, and H. Kautz, "Building personal maps from gps data," Annals of the New York Academy of Sciences, vol. 1093, no. 1, 2006, pp. 249–265.

[9] Matlab the language of technical computing. http://www.mathworks.com/products/matlab/index.html. (2014 (accessed April 10, 2014))

 ISBN: 978-1-61208-353-7

A Bayesian Tree Learning Method for Low-Power Context-Aware System in Smartphone

Kyon-Mo Yang, Sung-Bae Cho
Dept. of Computer Science
Yonsei University
Seoul, Korea
kmyang@sclab.yonsei.ac.kr, sbcho@yonsei.ac.kr

***Abstract*—Context-aware services using smartphone have been proliferated for ubiquitous computing. However, the capacity of smartphone battery is extremely limited so that the services cannot be effectively used. In this paper, we propose a low-power context-aware system using tree-structured Bayesian network. Bayesian network, one of the probabilistic models, is known to handle the uncertainty flexibly. A well-known problem of the probabilistic model, however, is high time complexity, which leads to significant consumption. To reduce the time complexity, we propose a tree-structure learning method. The key idea lies in how to consider the relation of each node. For the reason, we conduct the spanning tree based on the mutual information among nodes. The data for experiment were collected from Android phone for two weeks. The amount of the collected data is 7,464. The accuracy of proposed method achieves 94.13%. The energy consumption is measured using the power tutor application.**

Keywords-Low-power consumption, context-awareness, tree-structure Bayesian network, Structure learning.

I. INTRODUCTION

Recent proliferation of the smartphones leads to developing a large variety of applications and investigating on the use of various sensors through context-awareness. Previous researchers focus on the accuracy of context-awareness using all possible sensors [1]. The battery capacity of smartphone is well behind the development of service application. In a typical case, the user has to carry an extra battery or charge it frequently. There is the critical issue of how to reduce the battery consumption for the context-awareness in smartphone.

In this paper, we propose a low-power context-aware system using tree-structure Bayesian network. Bayesian network is one of the powerful probabilistic methods for context-awareness [2]. It can infer context in uncertain situation or with the incomplete data. However, the probability model generally has high time complexity, because the model has to calculate the probability of each node every time. It causes the significant consumption for context-awareness in smartphone. We propose a tree-structure learning method to reduce the time complexity.

We compare the accuracy using different structure learning methods and evaluate the time complexity of the proposed method. In addition, we verify the low-consumption feature of the proposed method in a real smartphone environment.

The paper is organized as follows. Section 2 presents the related works for context-awareness, battery problem, and Bayesian networks. Section 3 describes in detail the proposed low-power context-aware system. Finally, section 4 reports the experiments conducted to compare the power consumption of the proposed method and the monolithic BN.

II. RELATED WORKS

A. Context-aware service in smartphone

Context-aware services aim to provide the convenient services to users who are in the contexts recognized. Context is all the information related to the interactions between user and applications [3]. Interactions are becoming important as research in pervasive computing progresses. Context-aware services in smartphone recognize the situation and provide services. The applications are developed using various sensors.

Lee and Cho proposed a method using the KeyGraph and Bayesian network to infer mobile life log [4]. Ravi, et al. proposed battery management service [5]. The service recommends the battery charging time depending on current state. Phithakkitnukoon and Dantu proposed a three-step approach in designing the model based on the embedded sensor data for controlling alert mode [6]. Lester, et al. presented the approach to building a system that exhibits these properties and provided evidence based on data for 8 different activities; Sitting, standing, walking, walking up stairs, walking down stairs, riding elevator down, riding elevator up, and brushing teeth [7]. Santos, et al. described the architecture, operation and potential applications of a prototype system developed within the User-Programmable Context-aware Services (UPCASE) project [8]. A lot of applications have applied to the context-awareness. However, these previous works lack of dealing with the power consumption of smartphone.

B. Bettery lifetime problem

Fig. 1 shows the key features of mobile devices. According to the TIME Mobility Poll, conducted in June and July 2012, the main issue for mobile users is battery life. 62 percent of American mobile users wish for improvements in that area. Device size does not seem to be a problem for most

 ISBN: 978-1-61208-353-7

of them, as only 5 percent of American users want a smaller device. Because the battery life is short, the user has to carry an extra battery or charge it frequently. There is the critical issue of how to reduce power consumption for context-awareness.

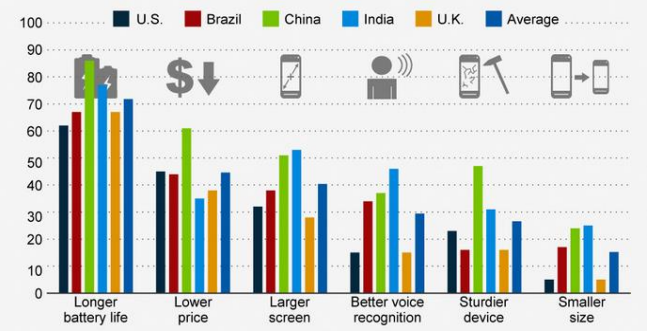


Figure 1. Key features of mobile devices by TIME Mobility Poll.

Many researchers have proposed low-power application using context-awareness for solving the problems. Seo, et al. proposed a context-aware configuration manager for smartphones [9]. The system changes the configuration of a smartphone according to the user-defined policy rules. Bareth and Kupper proposed a hierarchical positioning algorithm [10]. The algorithm dynamically deactivates different positioning technologies and only activates the positioning method with the least energy consumption. Miraoui, et al. proposed limited resources-aware service [11]. The system changes the services considering the current resources of mobile phone. These researchers lack of reducing the power consumption of the context-aware module. In this paper, the learning method maintains the advantage of probabilistic model and reduces the time complexity for the power consumption.

C. Bayesian networks

Bayesian network is devised as a powerful technique for handling the uncertainty. Bayesian network has a structure of a directed acyclic graph which represents the link relations of the node, and has conditional probability tables (CPT). Assume that a node is independent of each other.

$$\begin{aligned} P(U) &= P(A_1, A_2, \ldots, A_n) \\ &= P(A_1)P(A_2 \mid A_1)\ldots P(A_n \mid A_1, A_2, \ldots, A_{n-1}) \\ &= \prod_{i=1}^{n} P(A_i \mid pa(A_i)). \end{aligned} \qquad (1)$$

The conditional probability distribution of variable A can be represented as $P(A \mid pa(A))$, where $pa(A)$ denotes the set of parent variables of variable A, where U is a set of node, and the joint probability distribution is computed by the chain rules as (1).

There are two approaches to identify the structure and parameter of Bayesian network. The first approach is to learn model from the data on problem domains. The structure learning is useful if we have a lack of understanding about the system. The method requires the sufficient amount of data, but it is not easy to obtain reliable data in many real-world problems. The second one can be construct it based on the domain knowledge. The experts identify the structure and set of parameters according to their knowledge, if we do not have enough data in the domain.

The time complexity of the Bayesian network is calculated using the LS algorithm as follows [12]. Here, n represents the number of nodes, k represents the maximum number of parents for each node, r denotes the number of values for each node, and w represents the maximum number of clique.

$$CMPX = O(k^3 n^k + wn^2 + (wr^w + r^w)n) \qquad (2)$$

D. Structure design of Bayesian network

TABLE I. RELATED WORKS FOR REDUCING INFERENCE TIME

Authors	Necessity of knowledge	Extra procedure	Description
Pearl [13]	X	O	Using only relevant CPT about current inference
Heckerman and Breese [14]	O	X	Removing the uncertain interaction between causes and effects
Zhang and Poole [15]	X	X	Removing weak dependencies before inference
Kjaerulff [16]	X	O	Removing weak dependencies before inference
Koller and Pfeffer [17]	O	X	Applying object concept to BN
Oude, et al. [18]	O	X	Hierarchical modular approach designed for multiple agents

Bayesian network is a robust tool for practical problems which involve high level of uncertainty. However, utilizing it in the large-scale domains is difficult because considerable effort is put on designing and maintaining the network. Besides, it is unable to entirely apply on ubiquitous devices since lots of computation power and resources are required in the inference process. For these reasons, there have been many studies on reducing the time complexity. Table I shows two types of the related works for reducing inference time. First, the necessity of knowledge implies that the network is not designed automatically. Second, when the system infers, it needs extra procedure for modifying the network structure.

The Noisy-OR model was proposed by Pearl [13]. The model can compute the distributions required for the CPT from a set of distributions, elicited from the expert, and the magnitude which grows linearly with the number of parents. Heckerman and Breese proposed an extended version of the method called Noisy-MAX Gate [14]. This method showed a collection of conditional independence assertions and functional relationships and removed the representation of

 ISBN: 978-1-61208-353-7

the uncertain interactions between cause and effect. Zhang et al. removed weak dependencies before inference [15]. The method evaluated the relation of each node with query node and modified the structure of network through removing the nodes and edges. Kjaerulff presented a method for reducing the computational complexity through removal of weak dependences [16]. Koller and Pfeffer proposed a method where Bayesian network is applied to an object concept called OOBN [17]. This method used a Bayesian network fragment to describe the probabilistic relations between the attributes of an object. Oude divided BN model into several smaller multi-level modules and inferred each module sequentially from the low level to the high level [18]. Its composition is similar to MBN, but it restricts the networks in hierarchical structure.

The previous works in the table have problems. If the design method needs domain knowledge, it requires the time to analyze the domain. If the system needs extra procedure, it consumes extra battery. To solve these problems we propose a tree structure learning method.

III. LOW-POWER CONTEXT-AWARE SYSTEM

In this paper, we propose the low-power context-aware system. The proposed method considers for the power consumption of inference module.

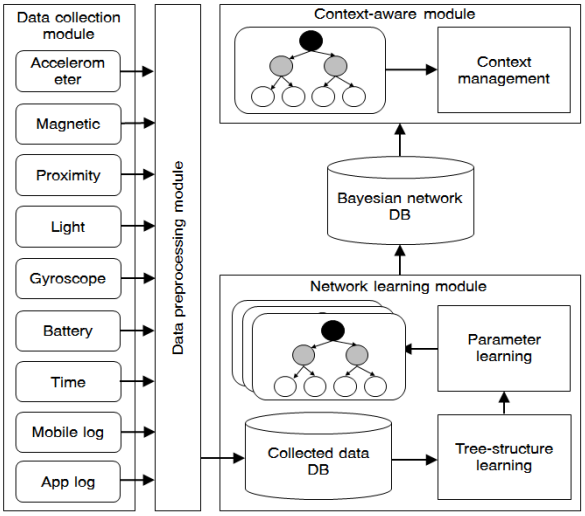


Figure 2. System architecture

Fig. 2 illustrates the system architecture. The system consists of four modules: Sensor collection, data preprocessing, network design, and context-awareness. In this system, we do not use the sensors that require high power consumption. The sensor collection module obtains the continuous sensor data in smartphone. The data are sent to the data preprocessing module that discretizes them using decision tree. The network learning module trains the structure and parameter of the BN. The context-awareness module infers user situation using the tree-structure Bayesian network. If the result of inference is higher than the threshold, it is the current situation.

A. *Sensor data preprocessing*

The proposed system focuses on how to reduce the power consumption for context-awareness using sensor information. Therefore, the other source that generates the energy consumption such as memory, synchronization, and so on, is not considered in the system. Abdesslem, et al. measured the energy consumption of different sensors as shown in Table II [19]. Each sensor runs continuously on a Nokia N95 8GB smartphone until the battery was depleted. In this research, the power consumption of GPS is 623mW, and it is 6 times more power consumption than the accelerometer sensor.

TABLE II. POWER CONSUMPTION IN SENSORS

Sensors	Battery life (hrs)	Average power consumption (mW)
Camera	3.5	1258
IEEE 802.11	6.7	661
GPS	7.6	623
Microphone	13.6	329
Bluetooth	21.6	211
Accelerometer	45.9	96

We select the sensors which can use during half-day because of the battery life time.

X_axis	Y_axis	Z_axis	Orientatic	Pitch	Roll	Magnetic1	Magnetic2	Magnetic3	Proximity	light	Gyro1	Gyro2	Gyro3
-0.22984	8.389283	8.283937	78.09244	-57.7119	-6.56882	-37.38	-18.48	-28.5	8	4435	0.060781	-0.67562	-0.12126
2.27928	7.967903	5.477933	82.18597	-49.7656	-3.5679	-41.7	-13.5	-24.12	8	7461	0.376598	3.172834	0.534202
0.191536	5.48751	5.822699	63.09831	-43.964	-8.35287	-31.2	-0.36	-42.12	8	3754	0.460898	2.666427	-0.57879
-3.00712	5.315128	11.21444	71.23709	-18.1507	-19.5634	-17.34	1.44	-44.88	8	4353	-2.03082	-1.85245	-1.62215
1.436521	3.533842	5.554548	55.31022	-30.242	-6.72346	-37.56	-2.7	-35.58	8	10003	-0.91935	0.572381	0.939511
-0.16281	-3.75411	9.844957	63.49014	8.344509	3.910431	-41.76	30.24	-25.8	8	11076	1.686599	0.640798	0.586125
1.647211	7.31668	-1.78129	18.79346	-2.84995	3.336133	-30.6	31.68	-32.76	0	121	-2.83105	-2.25104	1.138653
-3.5913	2.911349	-2.62405	250.4009	-31.2422	-54.0916	27.66	25.62	7.74	0	304	-2.16613	1.440115	1.304503
-7.67102	10.92714	-1.94409	244.0992	-52.0354	-36.8056	35.4	1.56	24.48	0	147	1.385442	-4.84844	1.579697
-2.78685	13.81933	-6.6463	284.0452	-74.0556	-9.47077	41.22	-16.5	-7.56	0	8	1.054048	-1.51098	1.498758
-5.77481	9.433155	-3.47638	272.1701	-68.8965	-21.1027	37.8	-22.08	8.22	0	80	-3.43551	-2.70033	-1.18752
-8.19775	4.520253	3.926491	267.1496	-77.119	-12.7805	40.44	-14.64	-11.64	0	4	1.169502	5.180137	-0.91996
-13.3692	9.356541	0.938527	291.223	-56.512	-26.1563	37.08	-15.06	10.68	0	31	-2.49813	-6.2409	-0.98471
-7.76679	8.465898	-2.16436	248.3102	-66.6607	-23.3265	42.36	-4.86	-7.8	0	2	1.469436	1.490206	-0.40897
-14.1545	13.54161	-3.07416	258.3507	-45.9125	-44.0226	36.9	-6.42	14.58	0	12	-1.87749	-3.80203	-1.88299
-7.87214	5.181053	-1.10133	284.9868	-57.2688	-23.4543	42.6	4.5	-5.04	0	3	-2.37657	-1.95752	2.866791
-6.42604	11.98059	-1.87705	256.9636	-38.9338	-49.3785	37.38	-1.02	18.36	0	4	-0.99632	0.311541	-0.74037

Figure 3. An example of the sensor data collected

Fig. 3 shows an example of the sensor data collected. If these continuous data were used as the states of the input values, the time complexity of network would be very high. It affects to increase the size of CPT because it is related to the number of the states of values [2]. In other word, if the continuous values map to the states of the value, the number of state is almost infinite. For this reason, the states of sensor data are preprocessed.

TABLE III. DEFINITION OF INPUT AND OUTPUT

Type	Sensors	Values
Input	Sensor: Accelerometer: X_axis	{Low, Middle, High}
	Sensor: Accelerometer: Y_axis	{Low, Middle, High}
	Sensor: Accelerometer: Z_axis	{Low, Middle, High}
	Sensor: Accelerometer: Orientation	{Low, Middle, High}
	Sensor: Accelerometer: Pitch	{Low, Middle, High}
	Sensor: Accelerometer: Roll	{Low, Middle, High}
	Sensor: Magnetic:1	{Low, Middle, High}
	Sensor: Magnetic:2	{Low, Middle, High}

 ISBN: 978-1-61208-353-7

	Sensor: Magnetic:3	{Low, Middle, High}
	Sensor: Proximity	{Low, Middle, High}
	Sensor: Light	{Low, Middle, High}
	Sensor: Gyro:1	{Low, Middle, High}
	Sensor: Gyro:2	{Low, Middle, High}
	Sensor: Gyro:3	{Low, Middle, High}
Output	User: Situation	{Sleeping, Exercising, Moving street, Having meal, Shopping, Studying, Viewing}

There are two discretization techniques. First, the range of input values can be divided into a predefined number of intervals of equal width. Second, it can be divided using statistical methods. A decision tree is one of the powerful and popular tools for making rules. All the continuous input such as accelerometer, gyroscope, and so on make the rules with a range of the division using the decision tree, because the input data do not need to change into the semantic data. It just needs to divide three ranges: Low, middle, and high.

The General Social Survey (GSS) collected the data on social trends in order to monitor changes in the living conditions [20]. The survey defined the category of situation. The output of network is defined with referring to the survey. Table III represents the input values and output values.

B. Tree-structure learning

The purpose of tree-structure learning is to reduce the time complexity by considering the relation of each node. The proposed method does not need the extra computational time for modifying the structure. In addition, the method does not need expert's knowledge because of learning from the collected data. Fig. 4 shows the flowchart of the proposed method. The learning method consists of six steps.

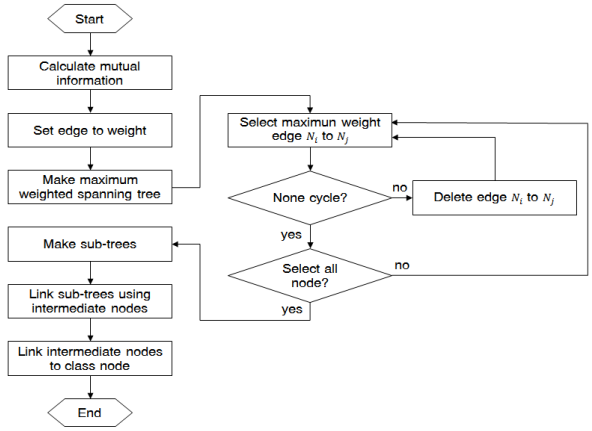


Figure 4. Flowchart of tree-structure learning method

First, we calculate mutual information with the relation of class. The mutual information is to calculate the relation between the attribute *X* and attribute *Y* from data as (3).

$$I_{ij}(N_i;N_j/C) = \sum_{n_i \in N_i, n_j \in N_j, c \in C} P(n_i,n_j,c)\log\frac{P(n_i,n_j/c)}{p(n_i/c)\,p(n_j/c)} \quad (3)$$

The attributes N_i and N_j are used for input nodes in the network. The class C represents the output node in the network.

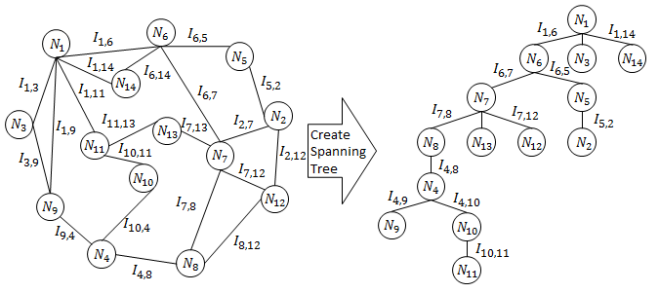


Figure 5. Creation result of maximum spanning tree

Next, maximum spanning tree is created using the calculated mutual information. The mutual information sets the weight of spanning tree. The tree randomly selects one node. Then, it selects another node that has maximum weight from the node. If the link does not create cycle, the node is selected. This procedure is repeated until all nodes are selected. Fig. 5 shows this step. In this figure, node 1 is root node. That is selected randomly. Nodes 6, 3, and 14 have the higher degree of association to the node 1 than other nodes. Nodes 7 and 5 have the highest degree of association to node 6.

Input:
Node *N*, Mutual information I_{ij}

Output:
Maximum spanning tree

Setting the link weight to I_{ij} ;

Select node N_i ;

While (Not all nodes *N* are selected)
Link N_i to N_j , I_{ij} has the maximum weight;
If (Cycle is created by the link)
Delete link N_i to N_j ;
Return to the first step;

Figure 6. Creation algorithm of maximum spanning tree

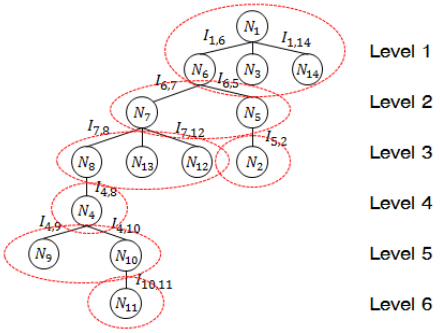


Figure 7. Grouping the maximum spanning tree

Fig. 6 shows the creation algorithm of maximum spanning tree. The network can maintain the relation of each node though this algorithm. Third, the method constructs a sub-tree through grouping some nodes by considering the relation of each node. Fig. 7 shows how to make sub-trees when the group level is 1. The group level means how many depths can be grouped.

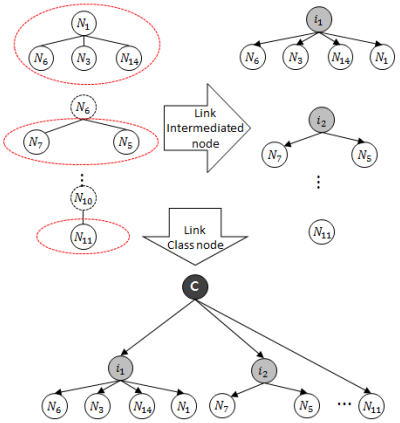


Figure 8. Creating tree-structured network

The group links to an intermediated node, which is used for considering relation of the grouped nodes. Finally, the class node links all intermediate nodes. Fig. 8 shows the final step. The parameters of the network are trained using Maximum Likelihood Estimation (MLE) [2].

IV. EXPERIMENTAL RESULTS

A. *Experimental setting*

The data were collected from three graduate students for two weeks. We used the Samsung Galaxy S3. Android phone collects sensor data twice per a second, and the amount of the collected data is 7,464.

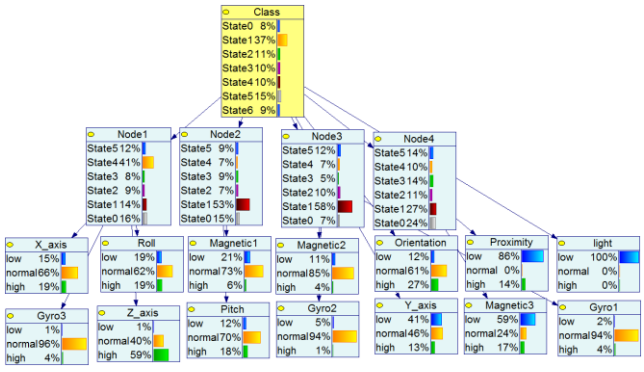


Figure 9. Learned tree-structured network

We collected on seven situations: Sleeping, exercising, moving street, having meal, shopping, studying, and viewing. When they collected the data, the smartphone was put into their pocket. The students selected the situation and conducted it. We learned the network using seven sensor data: Accelerometer, magnetic, gyroscope, light, orientation, pitch, and roll. Fig. 9 shows the tree-structured network using the proposed structure learning method. The network consists of fourteen input nodes, four intermediate nodes, and one output node. The monolithic BN is also trained using EM algorithm [2].

B. *Time complexity*

This experiment verifies that the proposed method has lower time complexity than other methods. We calculate it using LS algorithm. We compare monolithic BN (BN), Tree-agumented BN (TAN) and the proposed method. It is assumed that the number of clique w equals numbers of parents k . The maximum number of states is 7. The maximum number of parent node of monolithic BN is 7 while the maximum number of parent node of the proposed BN is 1. The number of state of the intermediate module changes 2 to 15. Fig. 10 shows the time complexity of each method.

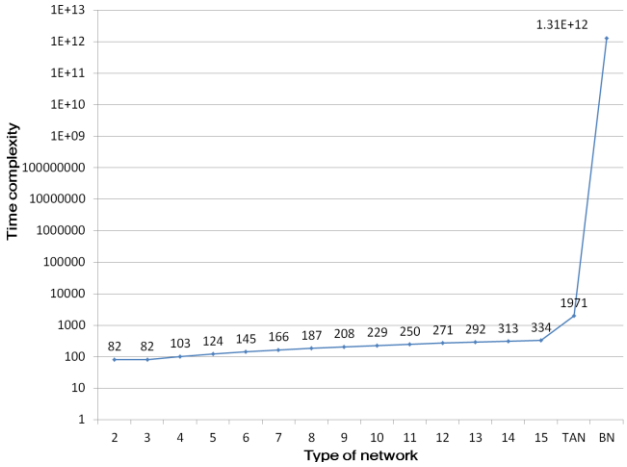
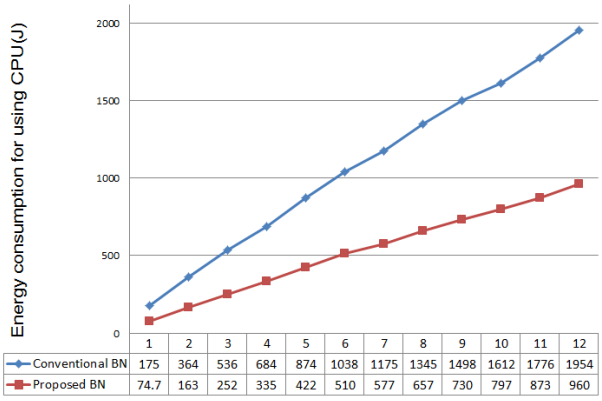


Figure 10. Time complexity usng EM algorithm



	1	2	3	4	5	6	7	8	9	10	11	12
Conventional BN	175	364	536	684	874	1038	1175	1345	1498	1612	1776	1954
Proposed BN	74.7	163	252	335	422	510	577	657	730	797	873	960

Figure 11. Comparison of energy consumption

The experimental result shows that the time complexity is slightly increased according to the number of states in the proposed network. However, the time complexity of the monolithic BN is dramatically increased in comparison with the proposed network. To verify the relation between

 ISBN: 978-1-61208-353-7

reducing the time complexity and reducing the energy consumption, we measured the energy consumption using power tutor application [21]. The application infers 100 times per a second. Fig. 11 confirms the difference of the time consumption of the monolithic BN (BN) and the proposed BN (PBN). Although the monolithic BN consumes the 1,954*J* for an hour, the proposed BN consumes the 960*J* for an hour.

C. *Comparison of accuracy*

We conduct 10-fold cross validation to calculate the accuracy of each network as shown in Fig. 12.

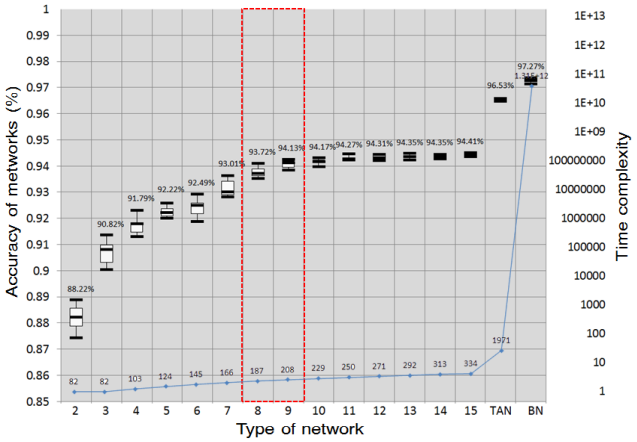


Figure 12. Accuracy of the networks

As a result, although the proposed method has slightly lower accuracy than monolithic BN, there is the relatively small difference of the time complexity. If the system selects eight or nine as the number of states, the accuracy of the network is 93.72%, although the system can have better battery life than monolithic BN.

V. CONCLUDING REMARKS

In this paper, we have proposed tree-structure learning method for a low-power context-awareness. The method does not require the extra computational time and the domain knowledge considering the relation of each node. The system is aware of seven situations. To verify the efficiency of the proposed system, we compare the accuracy of the proposed method against the monolithic Bayesian network and calculate the time complexity. In addition, we confirm the power consumption using power tutor application and verify that the system has lower consumption than the monolithic BN. We will improve the method through modular approach by considering the relation of nodes. The system will be applied to various context-aware service applications.

ACKNOWLEDGMENT

This work was supported by Samsung Electronics, Inc.

REFERENCES

[1] G. Chen, D. Kotz, "A survey of context-aware mobile computing research," Technical Report, Dept. of Computer Science, Dartmouth College, vol. 1, no. 2, pp. 1–16, 2000.

[2] F. V. Jensen, Bayesian Networks and Decision Graphs," Springer, 2007.

[3] A. K. Dey, "Understanding and using context," Personal and Ubiquitous Computing, vol. 5, no. 1, pp. 4–7, 2001.

[4] Y. S. Lee and S.-B. Cho, "Extracting meaningful contexts from mobile life log," Intelligent Data Engineering and Automated Learning, vol. 4881, pp. 750–759, 2007.

[5] N. Ravi, J. Scott, L. Han, and L. Iftode, "Context-aware battery management for mobile phones," Pervasive Computing and Communications, pp. 224–233, March, 2008.

[6] S. Phithakkitnukoon and R. Dantu, "Context-aware alert mode for a mobile phone," IEEE Pervasive Computing and Communications, vol. 6, no. 3, pp. 1–23, 2010.

[7] J. Lester, C. Tanzeem, and G. Borriello, "A practical approach to recognizing physical activities," Pervasive Computing, vol. 3968, pp. 1–16, 2006.

[8] A. C. Santos, G. M. P. Cardos, D. R. Ferreira, P. C. Diniz, and P. Chainho, "Providing use context for mobile and social networking applications," Pervasive and Mobile Computing, vol. 6, no, 3, pp. 324–341, 2010.

[9] S.-S. Seo, A. Kwon, J. M. Kang, J. Strassner, and J. W. Hong, "PYP: Design and implementation of a context-aware configuration manager for smartphones," Proc. of Int. Workshop on Smart Mobile Applications, pp.12–15, June, 2011.

[10] U. Bareth and A. Kupper, "Energy-efficient position tracking in proactive location-based services for smartphone environments," Computer Software and Applications, pp. 516–521, July, 2011.

[11] M. Miraoui, C. Tadj, J. Fattahi and C. B. Amar, "Dynamic context-aware and limited resources-aware service adaptation for pervasive computing," Software Engineering, vol. 2011, pp. 1-11, July, 2011.

[12] V. K. Namasivayam and V. L. Prasanna, "Scalable parallel implementation of exact inference in Bayesian networks," Conf. Parallel and Distributed Systems, vol. 1, pp. 8–16, July, 2006.

[13] J. Pearl, Bayesian Networks, University of California, Los Angeles, CA 90095, 2011.

[14] D. Heckerman and J. S. Breese, "Causal independence for probability assessment and inference using Bayesian networks," IEEE Trans. on Systems, Man and Cybernetics, vol. 26, no. 6, pp. 826–831, 1996.

[15] N. L. Zhang and D. Poole, "A simple approach to Bayesian network computations," Proc. of Conf. on Artificial Intelligence, pp. 171-178, May, 1994.

[16] U. Kjaerulff, "Reduction of computational complexity in Bayesian networks through removal of weak depedencies," Proc. of Conf. on Uncertainty in Artificial Intelligence, pp. 374-382, 1994.

[17] P. Weber, and L. Jouffe, "Complex system reliability modelling with dynamic object oriented Bayesian networks (DOOBN)," Reliability Engineering and System Safety, vol. 91, no. 2, pp. 149–162, 2006.

[18] P, De Oude, G. Pavlin, and T. Hood, "A modular approach to adaptive Bayesian information fusion," Information Fusion, pp. 1–8, 2007.

[19] F. B. Abdesslem, A. Phillips, and T. Henderson, "Less is more: Energy-efficient mobile sensing with senseless," Proc. of Workshop on Networking, Systems, and Applications for Mobile Handhelds, pp. 61–62, 2009.

[20] "General Social Survey," http://www23.statcan.gc.ca/, 2010.

[21] "A power monitor for android-based mobile platforms," Available from: http://powertutor.or

 ISBN: 978-1-61208-353-7

Immersive Virtual Environment and Artificial Intelligence: A proposal of Context Aware Virtual Environment

Fabrício Herpich*, Gleizer Bierhalz Voss†, Felipe Becker Nunes†, Rafaela Ribeiro Jardim* and Roseclea Duarte Medina*

* PPGI - Computer Science Post Graduate Program
Federal University of Santa Maria (UFSM)
Networks Group and Applied Computing (GRECA)
Santa Maria, RS, Brazil

† PPGIE - Computer Education Post Graduate Program
Federal University of Rio Grande do Sul (UFRGS)
Porto Alegre, RS, Brazil

(fabricio.herpich, gleizer.voss, nunesfb, rafa.rjardim, roseclea.medina)@gmail.com

***Abstract*—Virtual worlds has emerged as an environment of great interaction and immersion, where students have at their disposal different types of tools which are necessary for carrying out its activities. The objective of this paper is to present a proposal for developing an immersive environment for teaching Computer Networks that fits to the context and cognitive profile of the student. For this, the OpenSim [1] virtual world and the environment Moodle [2] are connected by the technology Sloodle [3]. In the immersive environment, agents with rules of Artificial Intelligence (AI), which offer support to learners according to these cognitive characteristics and their level of expertise.**

Keywords–Virtual World; Artificial Intelligence; Cognitive Profile; Expertise; Computer Networks.

I. INTRODUCTION

With the increasing use of Information and Communication Technologies (ICT) in the educational scenario, several needs were emerging and making necessary a reflection on the new paradigms of computing in education. In this context, many studies have been conducted in immersive virtual environments to be able to provide the student interaction with learning objects and the ability to be immersed in the environment [4][5][6].

Through the virtual worlds, for example, Second Life [7], OpenSimulator [1] and OpenWonderland [8] it is possible to create immersive virtual environments. However, the creation of immersive environments focused on education requires many factors to be considered, such as educational objectives and teaching strategies based on well-defined learning theories, friendly design and objects that are able to encourage the interaction and collaboration among users.

This paper aims to present an immersive virtual environment in development, which not only includes features of an educational environment, but too characteristics of context awareness, because it provides personalized assistance to students according to their cognitive profile and their level of expertise. For that, Intelligent Pedagogical Agents (IPA), has been implemented through rules of AI, since, according to Soliman and Guetl [9], in an Immersive Virtual Learning environments, it is expected that the learner will have great flexibility, faced with numerous learning opportunities and therefore it requires intelligent support and guidance.

Immersive experiences tend to further the engagement between students with the objectives established in the environment; thus, it is also possible to claim that the IPA could contribute significantly in helping the students, because according to Soliman and Guetl [9], IPA can act as a teacher, learning facilitator, or even a student peer in collaborative settings. The IPA will guide the learner in the virtual environment, explain topics, ask questions, give feedback, help the learner collaborate with others, provide personalized learning support, and act upon the learner in different times and in virtual places.

This paper is organized as it follows: theoretical references are presented in Section II, which exposes some concepts about immersive virtual environments and artificial intelligence; in Section III, a metodology is presented. Section IV presents the proposal and related work. Conclusion and future work are discussed in Section V.

II. THEORETICAL FOUNDATION

This Section aims to identify the main concepts related to the use of AI in Immersive Environments for supporting the processes of teaching and learning.

A. Virtual Worlds

Also known as immersive virtual environments or metaverse, virtual worlds are tools that simulate the real world environment in three dimension (3D), providing the user with a controlled environment experience with many possibilities. According to Bainbridge [10], they are defined as persistent online computer-generated environments where people can interact in a comparable way to the real world, either for work or for leisure.

Virtual worlds allow performing many activities, including training and tasks of educational character. According to Valente et al. [11], 3D virtual worlds enable the inclusion and practice of activities for experiential learning, simulation, modeling of complex scenarios, among others, with opportunities for collaboration and co-creation that cannot easily

 ISBN: 978-1-61208-353-7

be experienced on other platforms. Medina [12] reinforces this by stating that the learning gained through the personal experiences of the participants and their interactions with other participants, becomes more productive, dynamic and consolidated.

We can quote as examples of metaverse Second Life, OpenSim and OpenWonderland. This project will work with OpenSim because it is open source, it has extensive documentation, and also, according to Voss et al. [13], it allows the creation of the virtual world, in which all desired objects are placed, such as the creation of classroom, chairs, interactive scripts, among others. Also, this tool is used by the research group in which the authors of this paper belong to.

B. Artificial Intelligence

Artificial Intelligence is an area of computing that for years has been devoted to propose methods, techniques and tools that may be able to represent human knowledge in artificial systems. Besides, according to Liu et al. [14], artificial intelligence is the science of research, design and application of intelligent machines or intelligent system to simulate intelligent human capabilities and extension of human intelligence.

Filho [15] suggested that these methods and techniques should allow the computer to simulate the behavior aspects of intelligence, such as playing chess, proving logical theorems, understanding specific parts of a natural language, for example, Portuguese, among others.

However, other authors such as Liu et al. [14], Pollock [16], and Singh and Gupta [17], state that the AI should also be able to be aware of and demonstrate cognitive skills (problem solving, reasoning, and be autonomous to the point of being self-taught) and not just replicate knowledge.

C. Intelligent Agents and IPAs

Intelligent agents in the educational context are widely used as tools to support students with the goal of supporting the student interaction with the environment they are situated, providing personalized learning. In this context, Tyugu [18] understands that intelligent agents are software components that possess some features of intelligent behavior that makes them special: proactiveness, understanding of an agent communication language (ACL), reactivity (ability to make some decisions and to act).

Some authors such as Guetl and Soliman [9] and Garrido et al. [19] define intelligent agents used in education as IPAs, which, according to Guetl and Soliman [9], IPAs combine different characteristics including artificial intelligence capabilities to enrich the learning environment. Already Garrido et al. [19], states that they are software agents, which have educational purposes. They are able to communicate, cooperate, discuss, and guide other students or agents.

Moreover, Soliman and Guetl [20], elect five Pedagogical Agents Functional Requirements, which are: Learner interface requirements, Autonomy, Cognitive abilities, Agent Social Abilities, Environment and Context Awareness. For the development of this research, were approached three of these concepts: Autonomy, Cognitive Abilities, Environment and Context Awareness. The autonomy of the IPA is critical because that is how you will give origin to the processes of learning in virtual worlds, through interactions, explanations of classes, 3D objects or scenes. Another interesting point related to this research is the question of Context Awareness, because, according to Soliman and Guetl [20], inside a virtual world, this is related to the ability of the agent being able to discovering, constructing or suggesting learning resources, scenarios or scenes that are suitable to learner abilities and goals.

D. Context Aware

Context aware computing is characterized by performing the collection of various information involving the user, i.e., computational context of the user, physical and time. Thus, information is collected about the environment, in which its location and computational device used. According to Dey [21], context is any information that can be used to characterize the situation of entities that are considered relevant to the interaction between an user and an application.

Systems that use context information to provide personalized services to users, such as the adaptation of content and tools according to the user preference, may be considered a sensitive environment to the context. For Baldauf et al. [22], these are able to adapt their operations to the current context without explicitly requiring the user intervention, thus seeking to maximize their usability and effectiveness, taking into account the environmental context. Possible applications are the tour guides, restaurants, smart homes, among others.

According to Knappmeyer et al. [23], context area can be considered as an interdisciplinary field of research involving artificial intelligence, mobility, human-machine interaction, among others, in which, many researches have been conducted to overcome existing challenges.

In this aspect, information about the context of the cognitive profile of the user and their level of expertise will be used, which will be incorporated into the API from the implementation of rules on Artificial Intelligence, as described in Section IV.

III. METHODOLOGY

The development of this study arose from the need of a tool that can contribute and assist students in their learning process in the discipline of Computer Networks. This research proposes a different approach to the theory-practice relationship in the discipline of Computer Networks approach through an immersive virtual environment. For this purpose, a set of steps that the environment should suit were developed, as it follows:

The first phase was characterized by a survey of the theoretical reference about the topic, where also the technologies that would be used in the development of this work were defined, as shown in Figure 1.

The technologies discussed were WampServer [25], Moodle [2], OpenSimulator [1], Sloodle [3] and Firestorm [26]:

- WampServer was selected because it is free, hosts the necessary applications for the operation of technologies and includes three elements: MySQL, PHP and Apache. It creates a local server that will host the MySQL database application of OpenSim and the learning environment Moodle.
- Virtual Learning Environment (VLE) Moodle was selected because it is open source and widely used

 ISBN: 978-1-61208-353-7

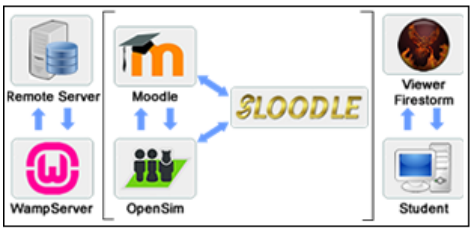


Figure 1. Technologies used. [24]

by the research institution of the authors, as well as OpenSimulator. Moodle materials and activities related to the discipline of Computer Networks were available and they were displayed to the students through Sloodle technology.

- Sloodle technology was chosen because it is open source and its performs integration between OpenSim and Moodle, thus enabling the display and interaction of materials available on the VLE for users of OpenSim. As for the Firestorm viewer, which is used in the projects of the research group of the authors, it has the function of making the connection to the virtual world, plus allowing importation of several objects, as already discussed by Nunes et al. [27], which highlighted the importance in choosing a viewer.

The second phase addressed the implementation of immersive virtual environment in OpenSim. In addition, a course was created in Moodle, inserting contents about Computer Networks.

As the research is still in progress, other phases will be discussed: lesson plan, learning objects in 3D, instructional design and theories of learning in immersive virtual environments, and finally, evaluating the environment with undergraduate students.

IV. PROPOSAL

This study presents a work in progress, which involves the development of an immersive virtual environment for teaching Computer Networks. In this environment, it will be used context information of students for the adaptation of materials, tools and activities to their cognitive profile. This same context information will also be used to define the level of expertise of the student. Such context information is very important for the process of teaching and learning of students, because through the adaptations of contents to the students and the support offered by the intelligent agent, it is possible to offer an appropriate, personalized and objective support to the learner.

In addition to the cognitive profile information and level of expertise, this environment also aims to gather information about the progress of the process of student learning through Sloodle Tracker, where it is possible to monitor the progress of students within the immersive environment. With this tool, it is possible to obtain location information of the student within the environment and monitor their activities. In this sense, it is intended to make the intelligent agent follow the student activities and offer him support in cases of difficulty.

Thus, the environment will advance to collect information about the cognitive profile and level of expertise (Figure 2). Later, it will collect information about the student learning process. From this information and with the formulated context, it will be possible to implement via AI, rules so that the IPA can offer adequate support to the needs of the student, thus providing a personalized education.

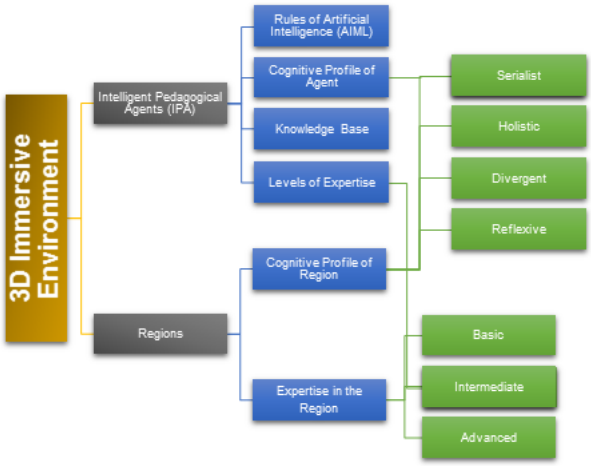


Figure 2. Structure of the proposal.

To this end, in addition to the IPA, the immersive environment as a whole offers five regions, which are: Serialist, Holistic, Reflective, Divergent and Computer Networks. From this perspective, when the student accesses the immersive virtual environment, he will be directed to the area of Computer Networks (Figure 3), in which the person has to answer a questionnaire, so one can set his cognitive profile. Then, the student will be teleported to the region that fits its cognitive profile; but, it is important to note that the student will not be prevented from viewing the other regions, providing freedom of choice to the user.

Figure 3. Immersive virtual environment.

Furthermore, immersive environment will also use information from the user's context to make adjustments to the immersive environment and IPA characteristics. This is essential to the process of personalized teaching and learning. Also, it is essential to identify the cognitive profile of the student and his level of expertise. In addition,in order to provide an adequate level of education based on the experience of the student, it is also possible to focus the teaching on the student's preferences and/or needs.

 ISBN: 978-1-61208-353-7

V. Conclusion and Future Work

With the integration of new educational technologies in the teaching context, new paradigms of teaching and learning that are transforming traditional education scenario emerged. Thus, the creation of new methods and instructional strategies that address the extent of teaching beyond the classroom environment and the use of technology in education become necessary.

Given this context, this paper has proposed a work in progress toward the teaching of Computer Networks. To provide resources and establish an immersive virtual environment to the students, integration of Moodle with OpenSim has been accomplished by Sloodle tool. These immersive environments allow educators to create new teaching alternatives, through simulations of equipment and performing experiments.

In this research, it was observed as a general result, the existing potential in immersive virtual environments focused on education. This case could corelate theoretical concepts with practice, not only because of the immersive environment, but also because of the personalized support offered by the IPA, the content presented in Computer Networks, and the tools available to the student.

Moreover, the fact that the immersive environment uses context information of the user to perform the adaptation of their characteristics is fundamental to the process of personalized teaching and learning, as well as to identify the cognitive profile of the student. It is also capable to identify their level of expertise and thus, in addition to it, provide an adequate education for the level of the student experience. It is also possible to do a more focused teaching based on their preferences and/or needs.

According to what has been presented, this study proposed to cover as future work the implementation of other features and functionality in the immersive virtual environment.

It will be validated with the use of the students that are taking Computer Networks at the undergraduate level. A comparison will be performed based on the knowledge level of the students before and after the use of the environment. This way we can show and prove the efficiency of the use of virtual worlds for education. The control group for this comparison will be the students who did not use the virtual lab. There is also the intention to tailor the user interface environment using instructional design concepts supported by theories of learning and their use in virtual worlds.

References

[1] "Open Simulator," 2014, URL: http://www.opensimulator.org [accessed: May/2014].

[2] "Moodle," 2014, URL: http://www.moodle.org.br [accessed: April/2014].

[3] "Sloodle," 2014, URL: http://www.sloodle.org [accessed: May/2014].

[4] M. Callaghan, K. McCusker, J. L. Losada, J. Harkin, and S. Wilson, "Integrating virtual worlds & virtual learning environments for online education," 2009 International IEEE Consumer Electronics Society's Games Innovations Conference, Aug. 2009, pp. 54–63.

[5] F. M. Schaf, S. Paladini, and C. E. Pereira, "3D AutoSysLab Prototype A Social , Immersive and Mixed Reality Approach for Collaborative Learning Environments," iJEP, vol. 2, no. 2, 2012, pp. 15–22.

[6] F. B. Nunes, S. Stieler, G. B. Voss, and R. D. Medina, "Virtual Worlds and Education: A Case of Study in the Teaching of Computer Networks Using the Sloodle," 2013 XV Symposium on Virtual and Augmented Reality, May 2013, pp. 248–251.

[7] "Second Life," 2014, URL: http://secondlife.com/ [accessed: July/2014].

[8] "Open Wonderland," 2014, URL: http://openwonderland.org/ [accessed: July/2014].

[9] M. Soliman and C. Guetl, "Intelligent pedagogical agents in immersive virtual learning environments: A review," MIPRO, 2010 Proceedings of the 33rd, 2010, pp. 827–832.

[10] W. S. Bainbridge, "Online Worlds: Convergence of the Real and the Virtual," Human-Computer Interaction Series. Springer-Verlag London Limited, 2010, p. 318.

[11] C. Valente and J. Mattar, "Second Life e web 2.0 na educao: o potencial revolucionrio das novas tecnologias," Novatec, 2007.

[12] R. D. Medina, "ASTERIX: Aprendizagem significativa e tecnologias aplicadas no ensino de redes de computadores: integrando e explorando possibilidades," Tese de Doutorado. Universidade Federal do Rio Grande do Sul, Porto Alegre, 2004, p. 174.

[13] G. B. Voss, F. B. Nunes, A. R. Muhlbeier, and R. D. Medina, "Context-Aware Virtual Laboratory for Teaching Computer Networks: A Proposal in the 3D OpenSim Environment," 2013 XV Symposium on Virtual and Augmented Reality, May 2013, pp. 252–255.

[14] Q. Liu, L. Diao, and G. Tu, "The Application of Artificial Intelligence in Mobile Learning," 2010 International Conference on System Science, Engineering Design and Manufacturing Informatization, Nov. 2010, pp. 80–83.

[15] C. F. Filho, "Histria da Computao. O caminho do pensamento e da tecnologia," Porto Alegre: EDIPUCRS, 2007, p. 205.

[16] J. L. Pollock, "How to Build a Person: A Prolegomenon," Cambridge: MIT Press, 1989.

[17] V. K. Singh and A. K. Gupta, "From artificial to collective intelligence: Perspectives and implications," 2009 5th International Symposium on Applied Computational Intelligence and Informatics, May 2009, pp. 545–550.

[18] E. Tyugu, "Artificial Intelligence in Cyber Defense," 3rd International Conference on Cyber Conflict, vol. 3, 2011, pp. 1–11.

[19] P. Garrido, F. J. Martinez, C. Guetl, and I. Plaza, "Enchancing Intelligent Pedagogical Agents in Virtual Worlds," in Proceedings of the 18th International Conference on Computers in Education, 2010, p. 8.

[20] M. Soliman and C. Guetl, "Evaluation of intelligent agent frameworks for human learning," 14th International Conference on Interactive Collaborative Learning (ICL2011) 11th International Conference Virtual University (vu'11), 2011, pp. 191–194.

[21] A. K. Dey, "Understanding and Using Context," Personal and Ubiquitous Computing, vol. 5, no. 1, 2001, pp. 4–7.

[22] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems," Int. J. Ad Hoc Ubiquitous Comput., vol. 2, 2007, pp. 263–277.

[23] M. Knappmeyer, S. L. Kiani, E. S. Reetz, N. Baker, and R. Tonjes, "Survey of Context Provisioning Middleware," IEEE Communications Surveys and Tutorials, vol. 15, no. 3, 2013, pp. 1492–1519.

[24] F. Herpich, R. Ribeiro, F. B. Nunes, G. B. Voss, L. M. Fontoura, and R. D. Medina, "Virtual Lab: An Immersive Tool to Assist in the Teaching of Software Engineering," XVI Symposium on Virtual and Augmented Reality (SVR), 2014, pp. 118–126.

[25] "Wamp Server," 2014, URL: http://www.wampserver.com/en [accessed: April/2014].

[26] "Firestorm," 2014, URL: http://www.firestormviewer.org [accessed: May/2014].

[27] F. B. Nunes, G. B. Voss, F. Herpich, A. R. Muhlbeier, C. C. Possobom, and R. D. Medina, "Viewers para ambiente virtuais imersivos: uma análise comparativa teórico-prática," RENOTE - Novas Tecnologias na Educação, vol. 11, no. n. 1, July, 2013, pp. 1–10.

 ISBN: 978-1-61208-353-7

Recommendation System for Assisting the Management of Information Technology

Taciano Balardin de Oliveira
Lutheran University of Brazil
Cachoeira do Sul, Brasil
Email: taciano@ulbra.edu.br

Felipe Becker Nunes
Computer Education Post Graduate Program
Porto Alegre, Brasil
Email: nunesfb@gmail.com

Gleizer Bierhalz Voss
Computer Education Post Graduate Program
Porto Alegre, Brasil
Email: gleizer.voss@ufrgs.br

Roseclea Duarte Medina
Federal University of Santa Maria
Santa Maria, Brasil
Email: roseclea.medina@gmail.com

Jose Valdeni de Lima
Computer Education Post Graduate Program
Porto Alegre, Brasil
Email: valdeni@inf.ufrgs.br

***Abstract*—Management of the problems occurred in environments that make use of the Information Technology (IT), together with the need for the quick response from the support teams area, a challenge for today. With this, organizations require systems to manage these incidents, as a Service Desk to centralize these records. The objective of this work is to integrate a system of Mobile Service Desk to a recommendation system that stores past interactions and automatically suggests as a possible solution for new similar incidents in the managed environment. As a contribution of this work, an algorithm was compared for similarity analysis and it has been integrated to the tool that showed the best results.**

Keywords–Mobile Service Desk; Recommendation System; Similarity Analysis Algorithms

I. INTRODUCTION

The increasing dependence of organizations on the use of Information Technology (IT) is making the management of IT services within these environments an increasingly important activity [1]. In case of any problem in these managed local places (e.g., computer, printer, software, networks, or any device that causes abnormal functioning of IT services), the expectation is that the user has a quick response of the team support, so the damage can be minimized [2].

The concept of service to users of IT service, which was originally named Help Desk [3], was created; this way, the problems could be centralized and subsequently solved by technicians responsible for these tasks. However, today, this area has absorbed other services and proceeded to call up of Service Desk, in the case of an extended version of the Help Desk and offering a greater amount of services [4].

A challenge that reaches those responsible technicians for these management environments is that, in many organizations, there is a high turnover of human resources. In 2010, according to the Research Institute Gartner [5], the turnover of IT personnel around the world was only 3%. In 2011, it jumped to 5%. Thus, the departure of an employee is a loss of human capital and generates a replacement cost (i.e., recruitment, selection, hiring and training) which can be high. In addition, there is the difficulty of transferring knowledge and experience among employees of this area [6].

A possible alternative to the problem of turnover is the integration of a Service Desk tool to a recommendation system, where the technical solutions applied in previous situations are retained in the database system, and when a new calling with similar features occurs, they are presented as a possible solution to the incident.

This work is part of a project that aims to design and implement a Service Desk tool mobile, called Mobile Service Desk, which has features of context awareness (e.g., geographical position, expertise and time), and it equipes this tool with a recommendation system. Thus, the intellectual capital generated through the services performed by the support staff which is retained in the system and it is recommended as a possible solution for new similar callings. This paper outlines the design and validation of the recommendation system integrated into the tool.

The paper is structured as follows: Section 2 presents the related work in the field of Service Desk. Section 3 explains the concepts of Service Desk, recommendation systems and pre-processing text. Section 4 discusses the proposal for the relationship between calls, which serves as a basis for recommendation system. Section 5 presents the methodology employed in the development of this work. Section 6 treats the implementation of the recommendation system integrated into the Service Desk. Section 7 presents the results obtained with the work. Finally, Section 8 contains final considerations of this paper and future work.

II. BIBLIOGRAPHICAL REVISION

This section is a literature review of the terms related to this research and also serves as a basis for the development of this work, such as Service Desk systems, pre-processing techniques and algorithms analysis of similarity, that provides support to the system that recommends similar calls.

A. Service Desk Systems

With the increasing business demand and globalization, more and more organizations need to ensure the quality of services performed to obtain better chances on the market. Thus, the goal of a Service Desk is to provide IT users with a Single Point of Contact (SPOC), vital to the realization of effective communication between users and teams that manage IT in an organization [7].

 ISBN: 978-1-61208-353-7

Its primary mission is to restore the normal operation of services of the users the fastest as possible, minimizing the business impact caused by IT failures [8]. In addition to it, the customer service keeps users informed about progress in solving incidents, changes and related events [4].

Operation of a Service Desk system occurs through the opening of calls or tickets. From this point on, whenever there is an open call, it is managed to be serviced. Moreover, these systems can also be based on some practical methodology for maintaining IT services, for example, the Information Technology Infrastructure Library (ITIL) [9].

B. Recommender Systems

A recommendation system combines computational techniques to select custom items based on users' interests and as the context in which they live. According to Adomavicius and Tuzhilin [10], this issue has become an important area of research from the early work on collaborative filtering emerged in the 90s. According to the authors, the interest in this area remains high, because it has a large number of research problems and also for the abundance of applications that help users deal with information overload and provide recommendations, customized content and services to them.

Based on how the recommendations are made, these systems are generally classified as follows: (i) content-based, which seeks to recommend items similar to those that have been an interest to the user, (ii) collaborative filtering, which operates identifying users with similar preferences to present and recommend items that were of interest of that similar user, and (iii) hybrid-approach, resulting from the combination of collaborative and content-based methods [11].

The method based on content (i) calculates the utility of an item s for user c, based on the utility of "similar" items to the same user c. The calculation of similarity between items is performed through the use of a set of attributes that characterize each item [10]. To enable this type of recommendation it is necessary to find associations between these items [12].

The goal of collaborative filtering (ii) is to recommend new items or predict the utility of an item for a given user, based on the data from the similar users to it. Thus, the user will receive recommendations for items that people with similar preferences to it, preferred in the past. This method is divided into two categories: the first is called memory-based, and the second is called model-based. The calculation of the value of an item *s* in relation to a user *c* is made from the utility of the same item for other users *c*, similar to the user *c* [10].

The hybrid method (iii) is defined as a method that combines both strategies based on content recommendation, as for the collaboration-based strategies [11]. The advantage of an approach that unifies the others is to significantly increase the chances of getting correct answers on their recommendations and to eliminate the limitation of both approaches.

C. Pre-Processing Text

Text mining techniques, which can also be found in the literature as text data mining or knowledge discovery of textual data bases, in general, refer to the process of discovering knowledge in unstructured text documents. This technique can be seen as an extension of data mining or knowledge discovery in a structured databases [13].

In a Service Desk system, text mining techniques can be applied in the description of open calls by the user, in order to perform a pre-processing of texts to later analyze the similarity between them and identify similar cases of past interactions stored in database. This can be accomplished by using some text mining techniques, for example, by removing stopwords, and stemming algorithms for the similarity analysis.

A set of strings that compose a document consists of a few words (tokens) that have no semantic value, being useful only for the text that can be understood in general. In a system of data mining, such as words that are considered stopwords and belong to a stoplist. With a well organized stoplist is possible to eliminate up to 50% of the total words in a text [14].

An example can be seen through the phrase "I have problem in my printer."; by applying the technique of removing stopwords results in "Printer problem" resulting in around 66% of the words that compose the phrase being removed.

Stemming aims to reduce each word until its radical is obtained by removing suffixes that indicate variation in form of the word as plural, verb tense, adverbs, gender and accentuation. According to Krovetz [15], the use of stemming improved in 35% the recovery of information in some datasets.

Radicalization is a process that involves different algorithms according to the language in question, as there are differences in how words are formed, so that the application of a specific technique can produce mixed results according to the language of the texts [16].

According to Viera and Virgil [16], the approach that is best known for the Portuguese language is that of Orengo and Huyck [17]. There is also the algorithm of Porter [18], for the Portuguese, following the same rules developed by the same author for the English language.

1) Orengo Algorithm: The algorithm Orengo and Huyck is developed specifically for the Portuguese [17]. This algorithm consists of a series of eight steps, performed in accordance with a predefined order by the algorithm, such as the longest suffix that must be removed first. This algorithm was developed based on the most common suffixes found in Portuguese [19].

The Orengo algorithm presents eight steps that are as follows: (i) reduction in the plural, that removes the end *s* indicative of plural words that do not constitute exceptions to the rule, making modifications as necessary; (ii) reduction of the female, that removes the final *a* of female words based on the most common suffixes; (iii) reduction adverbial, that removes the final *minded* of words that do not constitute exception; (iv) reduction in grade, removes most common indicators of augmentative and diminutive; (v) nominal reduction, which are removed 61 suffixes for nouns and adjectives; (vi) verbal reduction, which reduces the number of verbal forms to their radicals; (vii) removing vowels, where it is removed the vowels a, e, o, of the words which were not addressed by the previous two steps; (viii) removal of accents, which are removed the diacritical signs of the words.

2) Porter Algorithm: Porter's algorithm, originally proposed for English, is based on the idea that the suffixes in

 ISBN: 978-1-61208-353-7

English are mainly composed of a combination of smaller and simpler suffixes [18]. The algorithm consists of a series of five steps, in which some rules are applied to remove suffixes in each step. If a suffix combines with the word, the suffix is removed if the rules that were defined for that step are applicable [19].

In consonance with Viera and Virgil (2007), Porter's algorithm adapted to the Portuguese language is also based on rules. Five steps are performed by him: removal of suffixes; removal of verb suffixes, if that first step is not carried out there isn't any changes; removing the suffix (i) if it is preceded by (c); removal of residual suffixes *os, a, i, o, á, í, ó*; removal of suffixes (and), (is) and treatment of cedilla; after all, the nasalized vowels should be return to its original shape [16].

D. Similarity Algorithms between Strings

In literature, there are several techniques for calculating the similarity between strings, such as the inverted index model, Levenshtein Distance algorithm, natural language processing, algorithms of Boyer-Moore, Karp-Rabin, Jaro-Winkler, among other techniques that can be seen in [13][14][17][20]. In the next sections, some of these algorithms are presented.

III. RELATIONS BETWEEN TICKETS: PROPOSAL

One of the features of Service Desk proposed in this paper is that it presents solutions of past problems for the new incidents that the support will have to answer, in order to find a way to solve this new incident by applying a method already used in another similar call.

Thus, the technical support team can have at your fingertips a possible solution to the problems that need to answer. For this to be done, when a user opens a call in Service Desk system, the system should automatically relate it with other previous records in the system that have already been solved.

In this work, five stages that compose the pre-processing text are applied, as proposed by Avila [21], which serve as an initial step for further analysis of similarity between strings.

The first is the removal of invalid characters, such as quotation marks, brackets, parenthesis, among others, that need to be removed. After this, the replacement of accented character is performed, which is substituted by the respective non-accented character. The third step is the exclusion of repeated words, to avoid unnecessary comparison of a duplicate word several times. Also, a lowercase is applied, to prevent words with the same meaning that start with uppercase characters, will be differentiated from a similar word starting with lowercase characters.

The fourth step is the removal of stopwords, to prevent words like articles, adverbs, pronouns, prepositions, among others that have no semantic value, being only useful for text that can be understood, in general; Finally, stemming algorithm from Orengo and Huyck [17] is applied; this was chosen because it has been developed specially for the Portuguese language, to reduce morphological variants of the words, as singular forms, plural, verb conjugations, for its root or radical, by removing suffixes and prefixes.

After the execution of the steps for pre-processing in the opening text of the call, there will remain a set of strings that will be analyzed with previous case in order to determine the similarity between calls. For this, the *similar_text()* algorithm was used; the justification of using this algorithm is detailed in Section VI, which presents the comparative analysis of the algorithms for similarity between strings.

With the determination of similarity between calls, the system recommends possible solutions to a new incident. Thus, at the time the technician will accomplish its service, the system must have the knowledge of solved cases where the opening text of the call has similarity to the current incident. This paper proposes a method for content-based recommendation system, where the contents and characteristics of the calls are analyzed in order to determine what level of similarity they have and then recommend cases already solved to a similar that has not been answered.

Figure 1 shows in flow chart form the operation of the recommendation system of solutions proposed in this work.

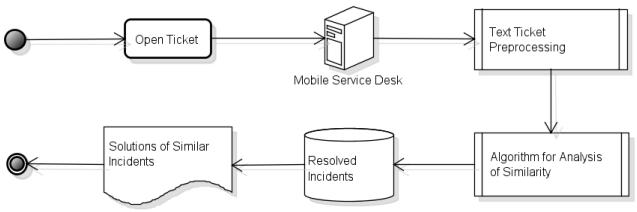


Figure 1. Operation of Recommendation Solution System

Then, the techniques of preprocessing text, as addressed in Section 4, are applied. These techniques analyze the similarity of the resulting text with cases already solved in order to obtain a possible solution to the incident and submit it to the technician at the time that it is to serve you.

Figure 2 shows the operation flow of the proposed system. From the opening of a call by the user, the system automatically links this with previous calls; with it, it is possible to suggest the support team possible solutions related to the call opened. To connect the calls, the use of text preprocessing algorithms and similarity analysis have been proposed.

User opens a call in the systemafter and the system searches if there are similar calls. If there is any information relating to that call, it will be displayed by the system. Otherwise, it transforms this new information in context to the system and writes the solution of the problem to be used in the future.

IV. METHODOLOGY

The recommendation concerns the use of mobile devices and systems research in order to propose a system for Service Desk were stimulated by the observation of the behavior of the User Service Center (USC) of a federal university. During the observation period (i.e., between October and December 2012), an informal interview each month, with some technical supervisors and the USC was performed.

Through interviews, some questions were raised, such as: (i) operation of the current system of USC as the opening of a call; (ii) the existence of some kind of prioritization and classification of calls; (iii) the existence of a division of the technicians as to their knowledge; (iv) how is the distribution

 ISBN: 978-1-61208-353-7

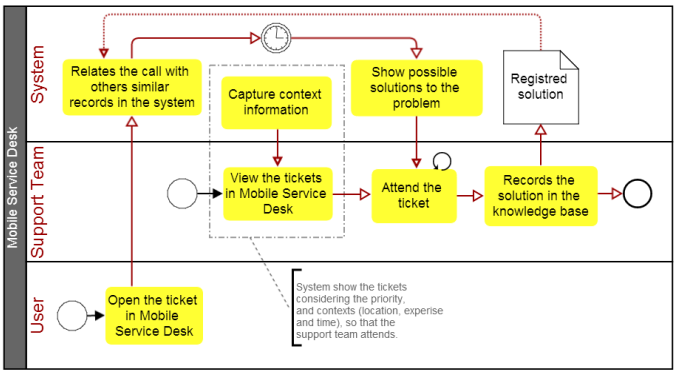


Figure 2. Operation of Mobile Service Desk

of the distance calls that the technicians need go to solve; (v) whether the solutions of the incidents are stored; (vi) number of technicians who work.

Besides these questions, along with USC, a dataset containing real calls of the system was fetched, in order to use them to compare the algorithms for similarity analysis. The dateset is useful to evaluate the time it takes to process a comparison of similarity with other records and evaluate the quality of the results obtained by each algorithm. There were more than thirty-five thousand records of calls imported into the system, which comprise the calls opened during the period between March 13, 2009 and November 20, 2012.

Four algorithms that calculate the similarity were compared: (i) Jaro-Winkler Algorithm [22], (ii) Levenshtein Distance Algorithm [23], (iii) String Similarity Algorithm, which is a class used to calculate the similarity between two text strings, created from the "diff" algorithm, that compares the difference between files under GNU (GNU Operating System) [24], and (iv) The *similar_text()* function, that is native from PHP for calculating the similarity [25].

For system modeling the Unified Modeling Language (UML) was used, which allows to represent application objects through a standardized graphical notation and diagrams. For creating UML diagrams, the software Astah Community [26] version 6.6 was used, which is a free UML modeling tool.

The database was modeled with the support of DBDesigner software, in version 4, which is a free program that integrates creation and graphical modeling of data [27]. The database used for data persistence was MySQL [28], and the tool used to access and maintain the database was HeidiSQL [29], which is also free.

The module recommendation of the Mobile Service Desk solution has been implemented in NetBeans, a free software, which is an Integrated Development Environment (IDE), and can be used with different programming languages, such as Java, PHP, HTML, JavaScript, C/C++, etc. [30].

To validate the proposal, a free programming language that has features such as ease of handling strings and Web programming was sought. Beyond the initial requirements, the PHP Hypertext Preprocessor (PHP) was also chosen for having the following characteristics: (i) server-side language, which performs multi-functions, (ii) compatible with MySQL, (iii) cross-platform, (iv) functions available for use, and (v) documentation.

To work on many mobile devices, the system has a unified user interface, i.e., an unique implementation for all mobile devices and operating systems. Thus, to provide this unified interface, the jQuery Mobile framework was used, which it is based on HyperText Markup Language 5 (HTML 5), and has libraries as the jQuery and jQuery UI, and has as characteristic to be optimized for touch interactions [31].

Finally, the Mobile Service Desk was installed on a server with the operating system Ubuntu, version 12:04, Long Term Support (LTS), using 32-bit architecture. In this, MySQL 5.5 database and Apache 2.2 web server, with support for PHP 5.3 programming language, were installed and configured.

V. Mobile Service Desk Integrated at the Recommendation System

Faced with the high use of mobile devices, such as smartphones and tablet, it was decided to study ways to use these technologies to apply in IT management. The idea of using the location information of the technician is to speed up the service of the tickets which is already expressed in the work of Lobo [2], who also considered that not only the location was important, but also the definition of a mechanism that utilizes the experience and practice of technicians to decrease the occurrence of a second service.

In this work, these practices were maintained and improved in order to consider not only the latitude and longitude of the technical but also their altitude at the time of setting a priority of service. Since the cases where the building is composed of several floors may occur. Thus, the question of the altitude is an essential factor to set the distance from one point to another.

Furthermore, other problems found in these IT environments, such as the issue of staff turnover, led to the study of the techniques for storing the solutions and later on the treatment of these data, so for those new team members can get suggestions for possible solutions to similar problems that have already been resolved. In order to have suggestions of similar solutions could be aggregated to the tool, the study of techniques in the area of recommender systems was necessary.

The Mobile Service desk is a system that has specific functional features, such as run on a variety of mobile devices, treat context sensitivity, and keep a history of the solutions of the problems so that later this solution may be suggested to some other similar ticket through of a system of recommendation.

To perform the service of tickets, from the moment when a “Support” level user accesses the system, the open tickets are listed. At the moment in which informs that will start attending to the incident, the system will list the solutions of similar cases to the current problem, so that a solution can be reused. Even at the time the technician will attend some ticket, theses cases are related to the current treatment are presented. For this, the relationship of tickets was implanted, as addressed in Sections V-A and V-B.

The implementation of the relationship of tickets, to be further suggested as a possible solution of a problem to be treated, is divided into two stages: (i) pre-processing text; (ii) similarity analysis.

 ISBN: 978-1-61208-353-7

A. Pre-Processing Text

In this step, to each opening text of ticket, the following actions are applied: (i) removal of invalid characters, (ii) application of lowercase, (iii) removing stop words, (iv) exclusion of repeated words, and (v) application of stemming algorithm.

In the second line of Figure 3, the variable call "string" receives off a POST method, the value of the text to be processed; in "limpaTexto()" function the text is passed to lower case and accents and other punctuation characters, dashes, quotation marks, brackets, among others are removed. The fourth line is responsible for mounting an array with all the words that compose the text.

```
1  <?
2      $string = $_POST['descr'];
3      $string = limpaTexto($string);
4      $stpw = explode(' ',$string);
5      $retorno = removeStopwords($stpw);
6      $retorno = array_unique($retorno);
7      arquivoStemming($retorno);
8      $temp = exec('python stemmer\stm.py');
9      $preprocessado = arquivoStemming('','ler');
10 ?>
```

Figure 3. Algorithm Code of Pre-Processing Text

The “removeStopwords()” function in line 5 is responsible for the removal of the stopwords of the text, in turn, the “array_unique()” function has the task of removing the terms duplicated. In the line 7, the “arquivoStemming()” function generates the text file with the key words of the text, so that in line 8 to run the Python script, which applies the technique of stemming in Portuguese on the file previously generated. This, the result at 9, is stored in the "preprocessado" (pre-processed in English) variable.

Other information to be highlighted about the pre-processing text step is regarding to the stemming; Ptstemmer script [32] was used because it possesses both the algorithms presented in [17] and [18], implemented for the Portuguese language. In this work, we opted to use the Orengo and Huyck [17] algorithm, because its rules were specifically created for the Portuguese language; in turn, the Porter algorithm is an adaptation of another language to Portuguese. In case the opening of ticket is performed in another different language than the Portuguese, a specific stemming technique for the language should be applied to the system.

B. Similarity Analysis

To verify the similarity between the text of the ticket to be attended to and the texts stored in the system relating to, the comparing strings and returns the percentage of similarity between them was applied. This algorithm was chosen due to its superior performance in relation to the other algorithms tested. The complete comparison between the similarity algorithms is presented in the Section VI.

VI. RESULTS

In this section, the validation of the suggestion system integrated at the Mobile Service Desk with the test plan and results obtained through the evaluation of the analysis of similarity algorithms is described.

The tests of the similarity analysis algorithms were performed using real data captured from the database of User Service Center (USC) containing the tickets opened between January and November 2012, in a total of 7033 tickets. To test the running time of the algorithms as well as the quality of the results were opened several tickets in the system; however, some of them will not be shown in this paper; the others can be seen in Table I.

TABLE I. TEXTS TO TEST THE SIMILARITY ANALYSIS ALGORITHMS

Nº	Texto do Chamado
1	Meu computador está extremamente lento. Desconfio que seja algum vírus
2	Ocorre erro de spooler na impressão!!!

After pre-processing of texts opening tickets was obtained the results presented in Table II. This resulting text was used to compare the similarity between them and the USC tickets imported to the system.

TABLE II. RESULT OF PREPROCESSING TEXT (IN PORTUGUESE)

Nº	Preprocessed Text
1	comput extrem lent desconfi viru
2	ocorr err spool impressa

The results referring to the five cases with highest similarity obtained when running the test related to the first open ticket in the system "Meu computador está extremamente lento. Desconfio que seja algum vírus" ("My computer is extremely slow. There must be a virus" in English) are shown in Tables III, IV, V and VI.

In these tables, the first column shows the opening text for the ticket similar to how it was in the database of the USC (i.e., without spelling corrections or abbreviations), the second column is the percentage of similarity between these records and the case that has been verified.

Analyzing the data from the first test of similarity algorithms, it is possible to conclude that the five records with greater similarity captured by Jaro-Winkler algorithm are somewhat related to the problem of "computador lento" ("slow computer" in English). In the Levenshtein distance algorithm, two of the results are unrelated to the problem, since one it is just a slowness in browser and another is without network access. In the String Similarity, the result "Computador e retro-projetor desconfigurado" ("Deconfigured Computer and multimedia projector" in English) is unrelated to the problem. Finally, in *similar_text()* the top five results are related to the problem.

This way, it is possible to conclude that, in the first test, the Levenshtein distance and String Similarity algorithms do not return a good result, by seeing how these similar cases that were unrelated to the problem.

The second test was carried out by way of the so-called "Ocorre erro de spooler na impressão!!!", ("There are error of spooler at the printing!!!" in English); after the execution of the similarity algorithms, the results achieved are shown in Tables VII, VIII, IX and X.

Analyzing the data from the second test of the similarity algorithms, it is possible to conclude that the Jaro-Winkler algorithm does not obtain good results, given that the first

 ISBN: 978-1-61208-353-7

TABLE III. TEST OF TICKET 1 FOR JARO-WINKLER ALGORITHM

Text of Ticket	Preprocessed Text	Similarity
Computador extremamente lento e, eventualmente travando.	comput extrem lent event trav	89,75 %
Computador necessita ser formatado, pois está com excesso de vírus.	comput necessit format excess viru	88,83 %
Computador não liga normalmente, infectado por virus.	comput lig norm infect viru	88,28 %
Computador extremamente lento e trancando.	comput extrem lent tranc	87,94 %
Veio um técnico mas parece que ficou pior.	vei tecn fic pi	
Computador lento e configurar a impressora	comput lent configur impress	87,25 %

TABLE IV. TEST OF TICKET 1 FOR LEVENSHTEIN DISTANCE ALGORITHM

Text of Ticket	Preprocessed Text	Similarity
O computador está muito lento, e alguns arquivos da área de trabalho sumiram. Desconfiamos que esteja com vírus.	comput lent arqu are trabalh sum desconfi estej viru	79,17 %
Computador extremamente lento e trancando.	comput extrem lent tranc	78,41 %
Veio um técnico mas parece que ficou pior.	vei tecn fic pi	
Computador extremamente lento para entrar na internet. As vezes tem que ser reiniciado e mesmo assim não consegue entrar na internet. Demora para abrir os e-mails e não consegue enviar a resposta do e-mail.	comput extrem lent entr internet vez reinici assim conse dem abr email envi respost email	76,71 %
Computador não está entrando na internet e nem no SIE. Está sem rede. Após configurar impressora em rede.	comput entr internet sie red configur impress	73,65 %
Computador extremamente lento e, eventualmente travando.	comput extrem lent event trav	71,25 %

TABLE V. TEST OF TICKET 1 FOR STRING SIMILARITY ALGORITHM

Text of Ticket	Preprocessed Text	Similarity
Computador extremamente lento e, eventualmente travando.	comput extrem lent event trav	77,96 %
Computador lento e escaner.	comput lent escan	72,34 %
Computador extremamente lento e trancando.	comput extrem lent tranc	71,42 %
Veio um técnico mas parece que ficou pior.	vei tecn fic pi	
Computador e retro-projetor desconfigurado	comput retroproje desconfigur	71,18 %
Computadores com sistema muito lento	comput sistem lent	66,66 %

TABLE VI. TEST OF TICKET 1 FOR *SIMILAR_TEXT()* ALGORITHM

Text of Ticket	Preprocessed Text	Similarity
Computador extremamente lento e, eventualmente travando.	comput extrem lent event trav	74,19 %
Computador extremamente lento e trancando.	comput extrem lent tranc	71,23 %
Veio um técnico mas parece que ficou pior.	vei tecn fic pi	
computador lento, possível vírus	comput lent possi viru	68 %
Computador muito lento...deve ter virus....	comput lentodev viru	67,92 %

TABLE VII. TEST OF TICKET 2 FOR JARO-WINKLER ALGORITHM

Text of Ticket	Preprocessed Text	Similarity
Ocorreu um problema na galeria da página do centro de Educação, solicito pessoal especializado em Páginas.	ocorr problem gal pag centr educaca solicit especi pag	75,45 %
OBR. Aguardo.	obr aguard	
Esta ocorrendo enceramento dos programas, tanto do oficce quanto de páginas de internet. Em algumas vezes aparece a mensagem de memória insuficiente ou que ocorreu falha no sistema e em outras vezes os programas são simplesmente finalizados.	ocorr encer programas tant oficc pag internet vez aparec mens memor insufici ocorr falh sistem simples final	73,85 %
Não ocorre a inicialização. A fonte está funcionando normalmente.	ocorr inicializaca font funcion norm	73,15 %
Está ocorrendo um erro na inicialização do computador. Já ocorreu a visita do técnico do CPD o qual informou que está ocorrendo um erro em virtude da própria atualização do computador e que provavelmente terá que ser formatado.	ocorr err inicializaca comput ocorr visit tecn cpd inform virtud atualizaca prova format	72,81 %
Não consigo imprimir, erro na impressão.	consig imprim err impressa	72,65 %

four records more similar that he found have no relation with the problem. Levenshtein distance, String Similarity and *similar_text()* algorithms, all the results has to do with printing problems, moreover bring a result related to the spooler problem.

To measure the execution time of each algorithm, the timestamp server was detected at the start of processing and, at the end, this value it was subtracted from the current

TABLE VIII. TEST OF TICKET 2 FOR LEVENSHTEIN DISTANCE ALGORITHM

Text of Ticket	Preprocessed Text	Similarity
Impressora configurada e não responde às solicitações de impressão.	impress configur respond solicitaco impressa	75,47 %
Computador formatado ontem precisa ser colocado em rede. Inatalar impressora.	comput format ont precis coloc red inatal impress	74,06 %
Reinstalar e configurar na rede serpro uma impressora matricial Epson fx 2180 .Obs. para impressão de relatórios contínuos.	reinstal configur red serpr impress matric epson fx ob impressa relato continu	73,33 %
Instalar serviço de spooler, impressoras desativadas	instal serv spool impress desativ	72 %
Peço para retirar de uma sala e instalar em outra um computador e colocar em rede o mesmo com a impressora.	pec retir sal instal comput coloc red impress	70,99 %

TABLE IX. TEST OF TICKET 2 FOR ALGORITHM STRING SIMILARITY

Text of Ticket	Preprocessed Text	Similarity
Instalar serviço de spooler, impressoras desativadas	instal serv spool impress desativ	57,14 %
Não consigo imprimir, erro na impressão.	consig imprim err impressa	57,14 %
Solucionar problema com impressora	soluc problem impress	54,90 %
Solucionar problema com impressora	soluc problem impress	54,90 %
Demora p/ impressão.	dem p impressa	54,54 %

TABLE X. TEST OF TICKET 2 FOR ALGORITHM *SIMILAR_TEXT()*

Text of Ticket	Preprocessed Text	Similarity
Não consigo imprimir, erro na impressão.	consig imprim err impressa	62,74 %
Instalar serviço de spooler, impressoras desativadas	instal serv spool impress desativ	62,06 %
porblemas de impressão.	porblemas impressa	60,46 %
Não dá a ordem para impressão	ord impressa	59,45 %
Reinstalar impressora..	reinstal impress	58,53 %

timestamp. This way, the runtime in microseconds that was converted to the right measure in seconds was obtained and presented in the form of average between all the tests.

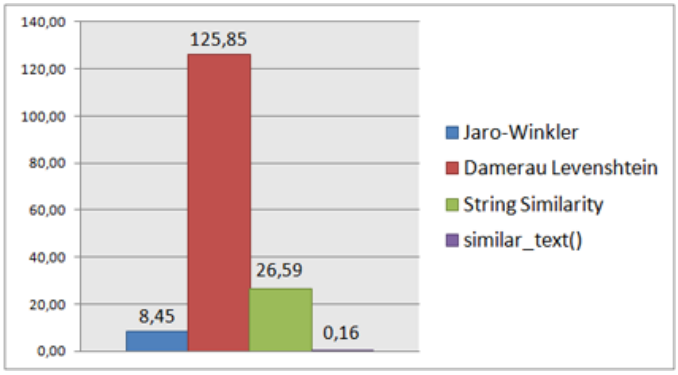


Figure 4. Performance Analysis of Algorithms Similarity (in seconds)

Comparing each of the opening texts of tickets with each record in the database system, the similarity analysis algorithms had very different performances. The lowest average time to perform all the calculations, as the graph shown in Figure 4, it was obtained by the similar_text() function algorithm with an average of 0.16 seconds. The Levenshtein Distance algorithm shown the worst performance, taking an average of 125.85 seconds to accomplish the task. In turn, Jaro-Winkler (8.45 seconds) and String Similarity (26.59 seconds) algorithms had intermediate performance.

VII. Conclusion and Future Work

The function of a Service Desk is to perform the process of incident management, dealing with all incidents of an organization linked to the IT area, such as software and hardware failures, communication networks or any device that causes the abnormal functioning of the IT services. The main objective of incident management is to restore the operation of the service as quickly as possible.

In addition, the Service Desk provides for IT users an IT SPOC, vital for an effective communication between users and teams that manage IT in an organization [7]. This helps managers, since it is not necessary to visualize numerous tools to access information concerning incidents of IT environment.

The results obtained by this research indicate that the similarity analysis algorithms and pre-processing text can be part of an integrated Service Desk to recommendation system solutions which, as seen in the Section VI, in the majority of cases return results referring to the problem in question. With these similar problems, it is possible that the solution adopted in the previous one should be reused by a technician in the new problem that needs to be solved; this way, by means of indications of the possible solutions, the system of recommendation may speed up to solve the incident.

Thus, the main contributions of this study are: (i) modeling of a recommendation system aggregate to the Service Desk tool; (ii) validation of the proposed solution by testing using several similarity analysis algorithms and using the algorithm with the best performance in the recommendation system.

In the light of the results, it can be considered that the use of recommendation for possible solutions might help the

 ISBN: 978-1-61208-353-7

technicians of the teams of assistance, especially for those that joined the team recently. It is further considered that these improvements also help to reduce the determining and resolution cost of incidents, which for Song [33] represent more than half of the operating IT costs.

REFERENCES

[1] I. L. Magalhaes and W. B. Pinheiro, IT Service Management in Practice: An approach based on ITIL. São Paulo: Novatec, 2007.

[2] J. Lobo, "Expertise location and context influencing the management of it," Master's thesis, Universidade Federal de Santa Maria (UFSM), 2011.

[3] G. Cavalari and H. Costa, "Modeling and development of a help desk system for the municipality of lavras," RESI - Revista Eletrônica de Sistemas de Informação, vol. 4, no. 2, 2005, pp. 1–18.

[4] M. Jantti and J. Kalliokoski, "Identifying knowledge management challenges in a service desk: A case study," in Proceedings of the 2010 Second International Conference on Information, Process, and Knowledge Management, ser. EKNOW '10. Washington, DC, USA: IEEE Computer Society, pp. 100–105.

[5] GARTNER, "Cio alert: U.s. it staff turnover trends and analyses," [Online]. Available from: http://envisat.esa.int/, retrieved: May, 2014.

[6] D. Wang, T. Li, S. Zhu, and Y. Gong, "ihelp: An intelligent online helpdesk system," Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, vol. 41, no. 1, Feb. 2011, pp. 173 –182.

[7] OCG, Office of Government Commerce, ITIL Service Transition. UK: The Stationary Office, 2007.

[8] R. Cohen, Help Desk and Service Desk Management. NOVATEC, 2011.

[9] ITIL, "Information technology infrastructure library," [Online]. Available from: http://www.itil.org/, retrieved: Jun., 2014.

[10] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," IEEE Trans. on Knowl. and Data Eng., vol. 17, no. 6, Jun 2005, pp. 734–749.

[11] M. Balabanović and Y. Shoham, "Fab: content-based, collaborative recommendation," Commun. ACM, vol. 40, no. 3, Mar 1997, pp. 66–72.

[12] E. B. Reategui and S. C. Cazella, "Recommendation systems," XXV Congresso da Sociedade Brasileira da Computação - V ENIA, vol. 17, 2005, pp. 306–348.

[13] C. Aranha and E. Passos, "Technology of text mining," RESI - Revista Eletrônica de Sistemas de Informação, 2006, pp. p.v. 2, p. 2.

[14] J. R. Junior, "Developing a methodology," Master's thesis, PUC-Rio, 2007.

[15] R. Krovetz, "Viewing morphology as an inference process," in Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, ser. SIGIR '93. New York, NY, USA: ACM, 1993, pp. 191–202.

[16] A. F. G. Viera and J. Virgil, "A review of algorithms radicalization in portuguese," [Online]. Available from: http://InformationR.net/ir/12-3/paper315.html, retrieved: Jun., 2014.

[17] V. Orengo and C. Huyck, "A stemming algorithm for the portuguese language," in String Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings. Eighth International Symposium on, Nov 2001, pp. 186 – 193.

[18] M. F. Porter, Readings in information retrieval. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, ch. "An algorithm for suffix stripping", pp. 313–316.

[19] M. V. B. Soares, R. Prati, and C. Monard, "Wci 02 improvements on the porter's stemming algorithm for portuguese," Latin America Transactions, IEEE (Revista IEEE America Latina), vol. 7, no. 4, Aug 2009, pp. 472 –477.

[20] M. Bendersky and B. Croft, "Finding text reuse on the web," in Proceedings of the Second ACM International Conference on Web Search and Data Mining, ser. WSDM '09. New York, NY, USA: ACM, 2009, pp. 262–271.

[21] R. de Avila and J. Soares, "The design of the correction of essay questions based on the adaptation of algorithms comparison and text search techniques combined with pre-word processing tool," RENOTE - Revista Novas Tecnologias, vol. 10, no. 3, Dez 2012.

[22] W. E. Winkler, "String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage," in Proceedings of the Section on Survey Research, 1990, pp. 354–359.

[23] V. I. Levenshtein, "Binary coors capable or 'correcting deletions, insertions, and reversals," in Soviet Physics-Doklady, vol. 10, no. 8, 1966.

[24] GNU, "Binary files and forcing text comparisons," [Online]. Available from: http://www.gnu.org/software/diffutils/manual/html_node/Binary.html, retrieved: Jun., 2014.

[25] PHP, "Php similar_text manual," [Online]. Available from: http://php.net/manual/pt_BR/function.similar-text.php, retrieved: Jul., 2014.

[26] Astah, "Astah community," [Online]. Available from: http://www.astah.net/, retrieved: Jul., 2014.

[27] DbDesigner, "Dbdesigner," [Online]. Available from: http://fabforce.net/dbdesigner4/, retrieved: Jul., 2014.

[28] MySQL, "Mysql," [Online]. Available from: http://www.mysql.com/, retrieved: Jun., 2014.

[29] HeidiSQL, "Heidisql," [Online]. Available from: http://www.heidisql.com/, retrieved: Jul., 2014.

[30] NetBeans, "Netbeans," [Online]. Available from: http://netbeans.org/, retrieved: Jul., 2014.

[31] M. S. Silva, jQuery Mobile - Develop web applications for mobile devices with HTML5, CSS3, AJAX, jQuery and jQuery UI. Novatec, 2011.

[32] PTStemmer, "Ptstemmer - a stemming toolkit for the portuguese language," [Online]. Available from: http://code.google.com/p/ptstemmer/, retrieved: May, 2014.

[33] Y. Song, A. Sailer, and H. Shaikh, "Problem classification method to enhance the itil incident and problem," in Integrated Network Management, 2009. IM '09. IFIP/IEEE International Symposium on, Jun 2009, pp. 295 –298.

 ISBN: 978-1-61208-353-7

Monitoring, Modeling and Visualization System of Traffic Air Pollution – A Case Study for the City of Skopje

N. Koteli[1], K.Mitreski[1]. D. Davcev[1]
[1]Faculty of Computer Science and Engineering
University "Sts. Cyril and Methodius",
Skopje, R. Macedonia,
nikola.koteli@finki.ukim.mk,
kosta.mitreski@finki.ukim.mk,
danco.davcev@finki.ukim.mk,

M.Ginovska[2]
[2]Faculty of Electrical Engineering and Information Technologies
University "Sts. Cyril and Methodius",
Skopje, R. Macedonia
gmarga@feit.ukim.edu.mk

Abstract - Air quality dispersion models can be used to provide information about the impact of individual emission sources or source categories on the air quality and to predict air quality as a result of changes in emissions, such as increase of traffic, emission control measures, etc. Dispersion models can be used to complement the data gained by monitoring as the spatial coverage of air quality information provided by monitoring is often limited. They are also an important tool for supporting air quality improvement plans and programmes. In this paper, based on combination of few existing air pollution models, we are presenting, as a main contribution of this paper, the first this-kind of study for the city of Skopje. This study uses measurements for emissions of many physical and chemical parameters from traffic sources in order to produce the general picture of the pollution on annual level. Our system based on real time air pollution visualization is easy extendible to national and trans-boundary levels, that is one of the most important EU recommendations. At the same time, this is the first step in building the real time decision (not only prediction) support system.

Keywords-dispersion; modelling;traffic;monitoring;system; visualization;air;pollution

I. INRODUCTION

Air pollution is a global threat to human health which is growing daily. It can be described as the pollution of the atmosphere with gases, or dust of solid materials, particulate matter as well as other substances that can endanger human health, animal life and plant life, reducing visibility and a number of other consequences.

Problems such as global warming, acid rain and ozone destruction are well known, although it may seem distant from our everyday life in urban environments. These are global issues affecting the entire world community. In addition to these global problems, these recent decades, an important and worrying issue for the experts for healthy environment, as well as for all residents in urban areas is the air pollution in the most populated urban places. The relationship of air pollution and health status of the people is top priority issue. It is estimated that worldwide, 2 million people and more than half of them are in developing countries, die every year from air pollution. In many cities worldwide there are health risks from exposure to particulate matter (PM) and Ozone (O3) [19]. The majority, 51% of the European population lives in these urban areas, and their daily activities and economic activities are concentrated in, or around that area [16]. Transport by motor vehicles across the road infrastructure close to the residential buildings is the main and immediate source of air pollution in this areas. The lack of knowledge of the health impacts from pollution, is a big obstacle in defining the actions and mobilizing local, and international resources. [19]

Although the network of monitoring stations is very important in such an urban environment because it provides information on actual concentrations of certain parameters of air, it cannot cover every point of interest. Consequently, only the most important points could be monitored.

Air dispersion modelling could be used to estimate and predict the concentration of the pollutants in air using mainly emission and meteorological data. Air dispersion models include mathematical algorithms based on combination of physical and chemical parameters so that they can simulate the spread of pollutants in the air as well as the complex processes of air pollution creation. Dispersion modelling of air will allow the implementation of effective control of pollution as well as the development of strategy to reduce emissions of harmful substances that pollute the air. In this way, it will be possible to develop a plan to reduce the environmental pollution and satisfy the EU environmental air quality standards [20].

In addition, we propose a particular system for urban environment air quality monitoring, modelling and visualization (applied for our city's environment because each environment has its own space, meteorological parameters and configurations) that:

1. Extends monitored air pollution data (that is the number of points representing monitoring stations in real-time [18]) to each particular point above the rooftops of the city and visualizes on map in continuous color coded layer, taking the limit values for color coding according the limits defined by EU environmental air quality standards per

 ISBN: 978-1-61208-353-7

parameter.

2. Predicts the air pollution from street network traffic, on street canyons level, in the most traffic jammed parts of the city's street sections applying the Operational Street Pollution Model (OSPM) [17] to our environment configuration.

3. Visualizes the air pollution as the network of sections, by developing effective and usable visualization tool as a part of our system.

The idea with this kind of system is to raise the public awareness and to help the city planners and regulatory institutions to get real-time feedback in one smart city concept.

In Section II of the paper, we will present the state of the art in Monitoring, Modeling and visualization of traffic air pollution. The case study will be described in Section III while the paper will be concluded in Section IV.

II. MONITORING, MODELING AND VISUALIZATION OF TRAFFIC AIR POLLUTION-STATE OF THE ART

Climate change of local, regional and global scale has an outstanding need for a systems that monitors, models and visualizes the distribution of air pollution emissions with high spatial resolution. There is a lot of interdisciplinary research in this area. In [11] and [13], the monitoring module of the system is based on network of sensor devices that among other parameters monitor air pollution and traffic data. Modelling module is always the required complement of a distributed data collection module for the prediction of the air pollution state. Air pollution monitoring data are correlated to health problems in [10]. In addition, in [12], an application that helps local government developing more accurate prevention and health care plans has been developed. Systems that use visualization of data and cloud solutions connected with air pollution data are described in [7], [8] and [21]. In [7], a geographical approach for air pollution map generation using parallel processing and cloud computing system is presented. In this way, the information is available anytime, anywhere. In [8] a system for processing large amounts of data is proposed. It uses GIS and air pollution visualization by introducing customized cloud computing technology with major goal on reducing the processing time of the visualization. In [21] BigSmog system using cloud computing framework and big spatio-temporal data for big smog analysis conducts parallel correlation analysis of the factors and scalable training of artificial neural networks for spatio-temporal approximation of the concentration of PM2.5. Global warming as air pollution related problem, especially for emission of greenhouse gases from traffic, is analyzed in [9]. A tool for emission estimation has been developed. Sensor-based Emissions Monitoring System is described in [14], while more general cloud computing for Internet of Things and sensing-based applications is presented in [15].

In our study, we combined and adopted many of these technics and technologies to develop a robust, long-term system for traffic air pollution monitoring, taking into account all the related parameters and providing in this way a case study as per EU recommendations with extensive number of experimental data. Our study is different from others because it was realized according to the EU Air quality directives as well as WHO (World Health Organization) recommendations, especially contributing in the process of real time air pollution visualization. In this way, it is easy extendible to national and trans-boundary levels, that is one of the most important EU recommendations. At the same time, this is the first step in building the real time decision (not only prediction) support system.

III. OUR CASE STUDY

A. *Real-time monitoring data in city of Skopje*

The real time monitoring data network of stations on the area of Skopje is composed of eight air quality measurement stations. These stations are:

- two *urban traffic stations*
  - "Centar"
  - "Rektorat"
- two *urban background stations*
  - "Karpos"
  - "Finki"
- one *suburban background station*
  - "Gazi Baba"
- one *urban industrial station*
  - "Lisice"
- two *rural industrial stations*
  - "Mrsevci"
  - "Miladinovci"

Seven of them are controlled and maintained by the "Ministry of environment and physical planning in Republic of Macedonia" (MOEPP) and one is controlled and maintained by our "Laboratory for Eco informatics at Faculty of computer science and engineering, Sts. Cyril and Methodius' University in Skopje" (Ecolab FINKI). The position and configuration of the network of the air quality measurement stations is as displayed on map in Figure 1.

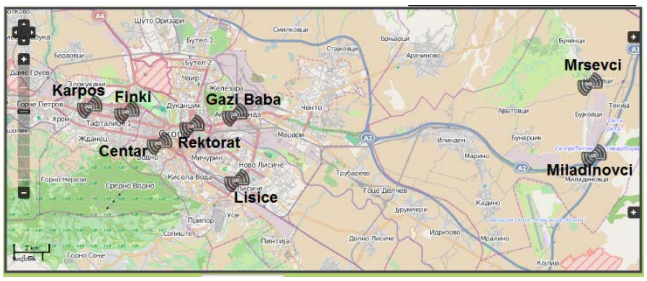


Figure 1. Map of air quality measuring stations

The measured air quality parameters are as follows:

- Particulate matter with diameter per particle less than 10 micrometers (PM10)
- Particulate matter with diameter per particle less than 2.5 micrometers (PM2.5)
- Ozone (O3)
- Carbon monoxide (CO)
- Nitrogen dioxide (NO2)
- Sulfur dioxide (SO2)

B. *Real-time visualization of the monitored data*

Workflow diagram of the visualization module of our system is presented in Figure 2.

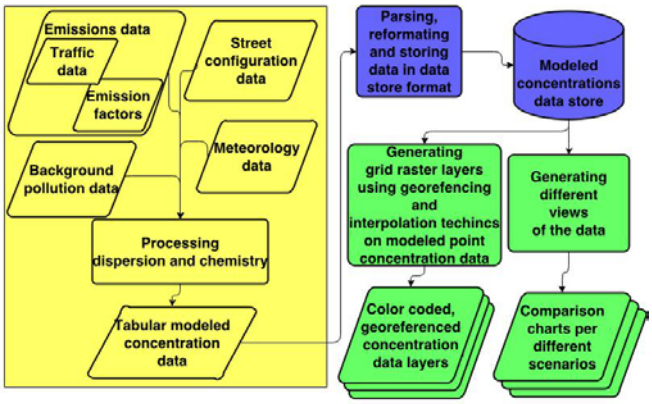


Figure 2. Schema of monitoring and real-time visualization of monitored data module of our system

The process of visualization is based on the newest 'up to time' data as the average of the air parameter concentration per hour for each parameter. The data are provided by the network described in section A and stored in the server database. They are used for generation of the grid raster layers by interpolation technique. The layers are color-coded and scaled from green to red depending of the concentration level. Then the geo- referenced interpolated concentration layers are automatically published as World Mapping Service (WMS) and publicly available at [18].

The client side that displays the web content layers uses the java script library that makes connection to the publicly available WMS server, catches the latest raster layer data and displays it. The view can be changed to view each air quality parameter layer separately or to view combination of two or more parameter layers together. An example of the real-time visualization on our web is shown in Fig 3.

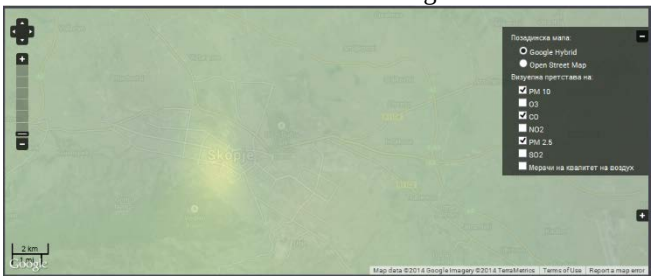


Figure 3. Web real-time concentration visualization

C. *Modelling part - Application of our collected data in the model*

The model used in our work was OSPM (Operational street pollution model) [17]. In Figure 4 an extended visual based modelling module in our system is presented.

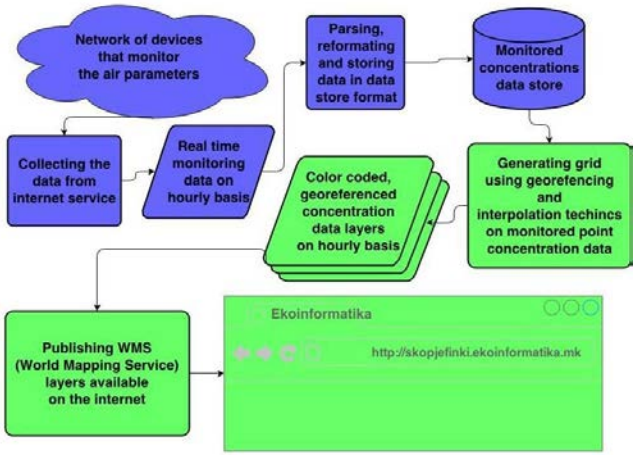


Figure 4. Extended visual based modelling module in our system

On the basis of the traffic data from one of the main streets in the city of Skopje, based on our extended visual based modelling module as described in Figure 4, we generated our modelled data charts in Part D of this Section.

The needed and collected information and data are on meteorological conditions and concentrations of air parameters in the area which is modelled in the duration of an entire year. This type of data is actually measured data or data previously obtained with a mean value over the years that actually reflect the meteorological situation at the level of a year. This data can also be data obtained with the prediction for next year which follows the use of modelling in order to obtain predictive model of pollution. The density of this type of data is on hourly level interval for one year timespan. These data contain data for temperature, speed and direction of wind, global solar radiation, and relative humidity, and also background concentrations of PM10, O3, CO, NO2, and PM2.5.

Emission information and data of the type of fuel used in transport and their composition in terms of substances important for pollution is taken (provided) from the largest supplier of fuel in the Republic of Macedonia, MAKPETROL.

Traffic information and data on the number of vehicles moving on the street are classified by the type of vehicle: bus, car, van, truck. The real distribution of flow per class is for 24 hours timespan. Also the distribution by volume and type/technology of motor vehicles and fuel has been determined. The classes of 'type/technology of motor vehicles and fuel' in our case are also provided. Data for the distribution of vehicles in terms of volume and type of engine and motor fuel are provided from city corresponding institutions.

 ISBN: 978-1-61208-353-7

Information and data of the average speed of vehicles that travel along the street. This parameter is taken to be 50 km / h (actual speed limit for the road).

Street configuration information and data about the height and placement of buildings along the 'Partizanski Odredi' street canyon (under investigation in this study) are presented in Figure 5.

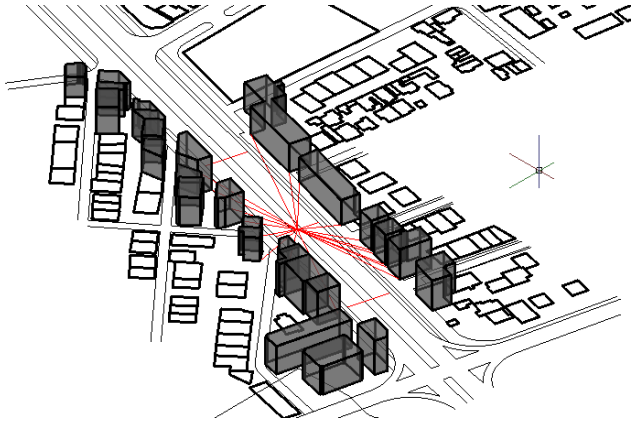


Figure 5. Part of the street 'Partizanski Odredi' together with the configuration of the buildings

D. Experimental work with model and results (charts with different views and grid raster layers)

In the model, we define three modelling scenarios about the flow of the vehicles. Each of these three scenarios has two sub scenarios (about the mean speed of the vehicles), resulting in total of 6 as follows.

Scenario 1 is a scenario that is based on the real measured data of flow of vehicles. In this scenario the vehicles per 24 hour period is 32000.

Scenario 2 is a scenario that is based on the double of (2x) real measured data of flow of vehicles. In this scenario the vehicles per 24 hour period is 64000.

Scenario 3 is a scenario that is based on the triple of (3x) real measured data of flow of vehicles. In this scenario the vehicles per 24 hour period is 96000.

The sub scenarios per every scenario refer to average speed of vehicles that travel along the street (50 km/h and 80 km/h).

In this way, we define the following six scenarios:

Scenario 1.1 (32000 vehicles/24h, 50 km/h)
Scenario 1.2 (32000 vehicles/24h, 80 km/h)

Scenario 2.1 (64000 vehicles/24h, 50 km/h)
Scenario 2.2 (64000 vehicles/24h, 80 km/h)

Scenario 3.1 (96000 vehicles/24h, 50 km/h)
Scenario 2.2 (96000 vehicles/24h, 80 km/h)

Results from the applied modeling for the above six scenarios are presented as charts on Figure 6-9.

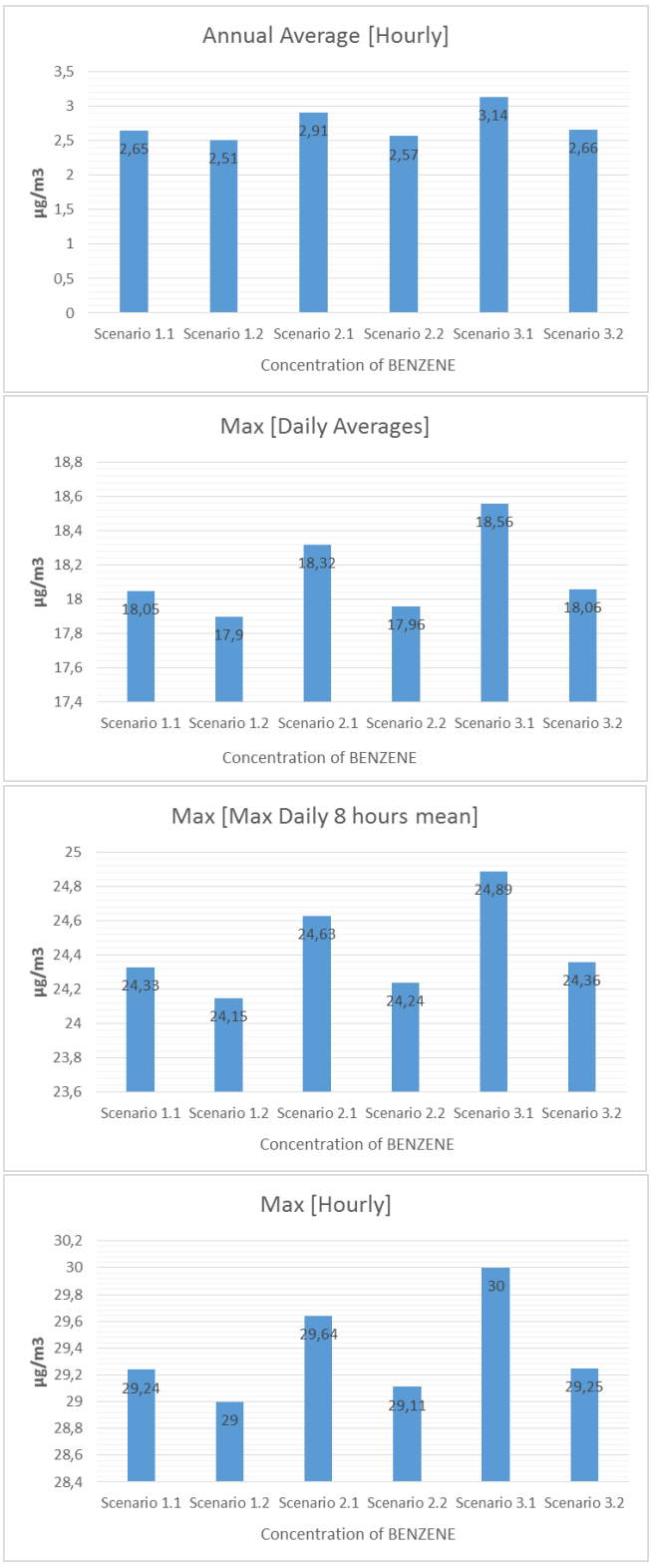
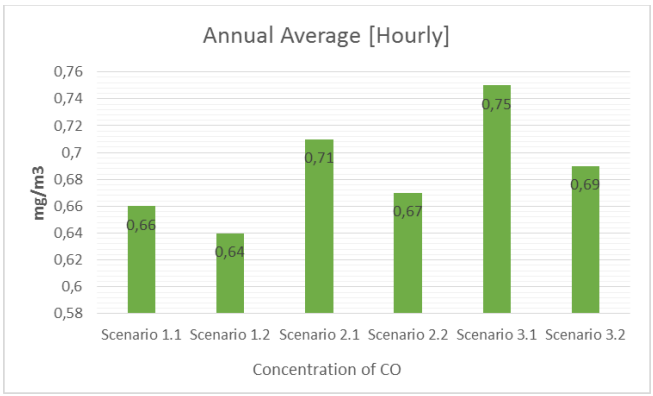
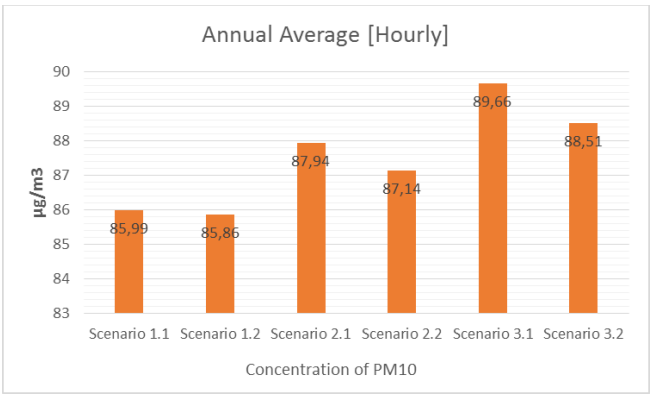
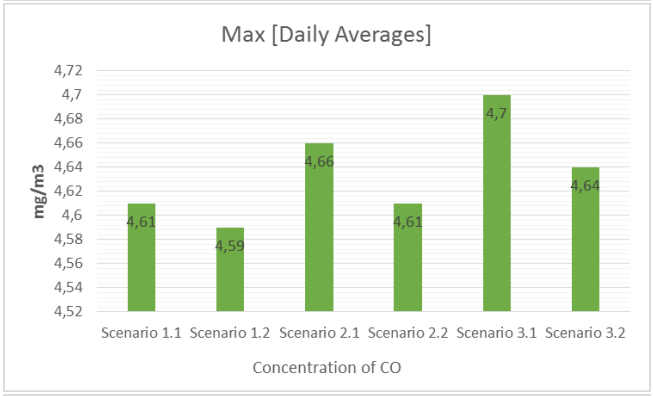


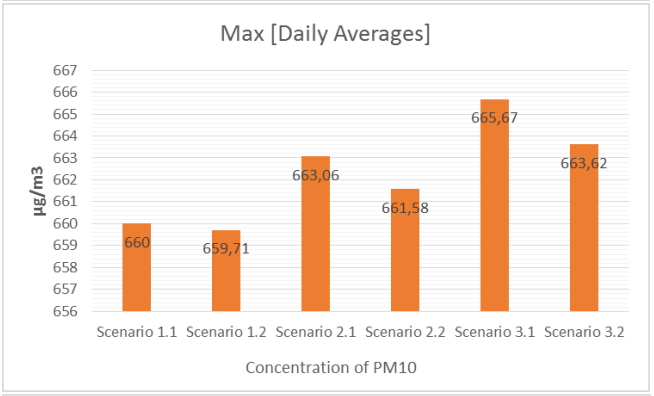
Figure 6. BENZENE Concentration Charts

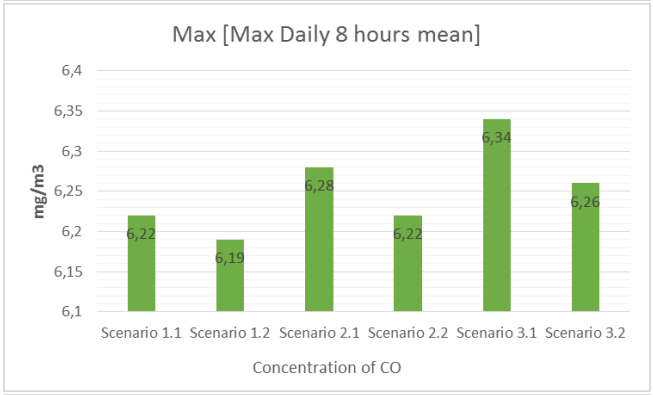
 ISBN: 978-1-61208-353-7

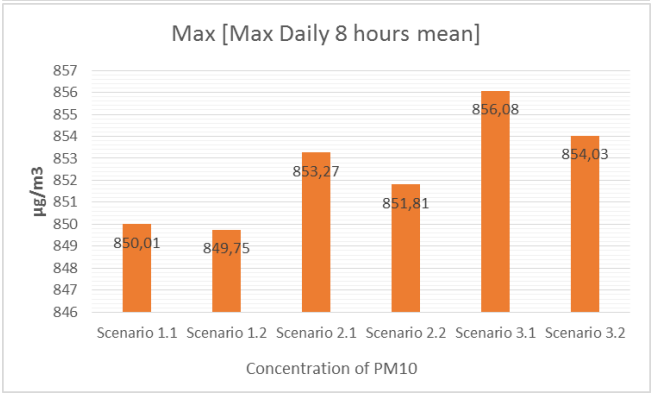


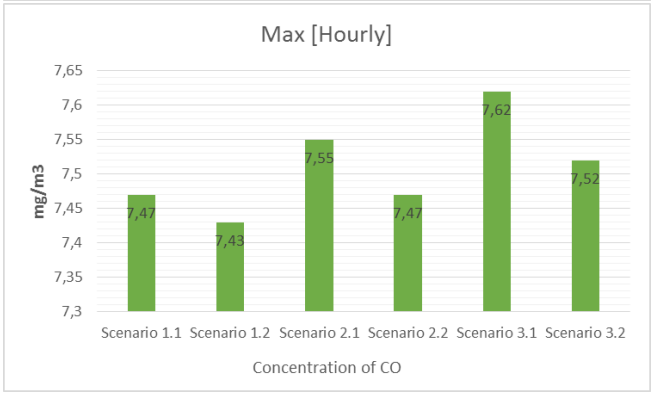












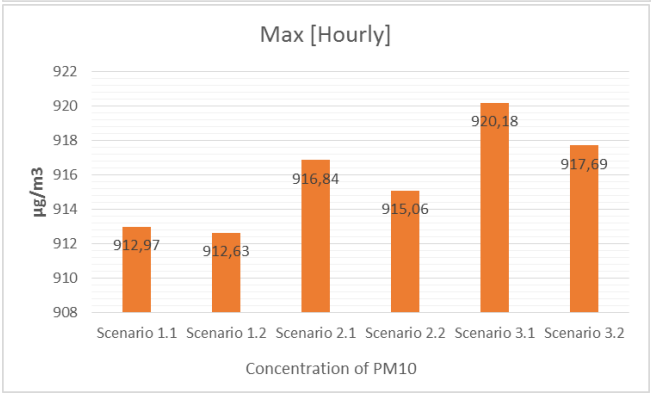


Figure 7. CO Concentration Charts

Figure 8. PM10 Concentration Charts

 ISBN: 978-1-61208-353-7

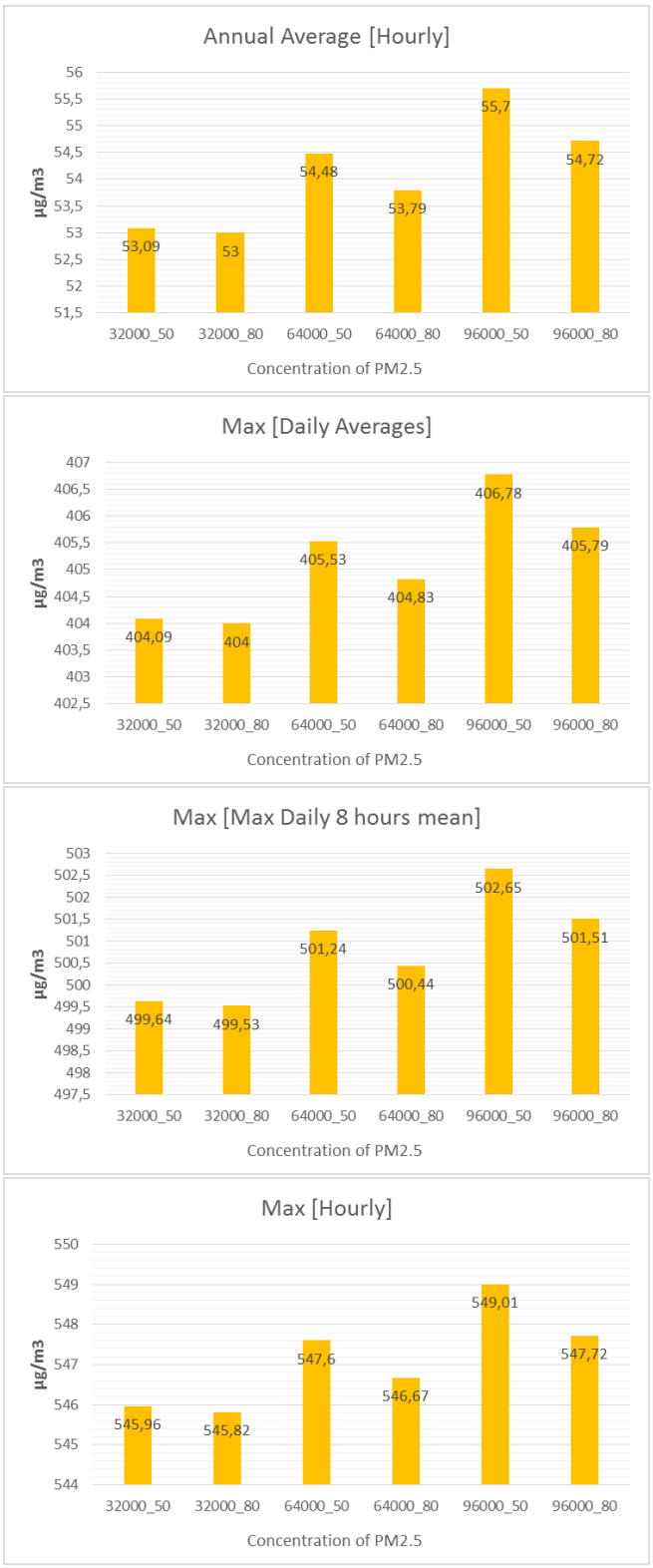


Figure 9. PM2.5 Concentration Charts

Visualization of the color coded concentration data layers on street network are given in Figure 10.

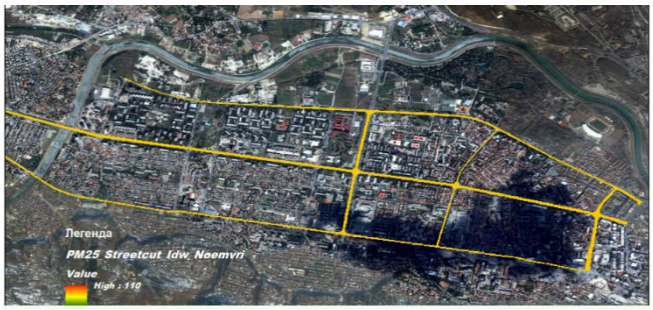


Figure 10. Color coded concentration data layers of street network

The suggested view is generated by interpolating measured data (as the average value for each parameter) and the layer is clearly represented by colors. In this way, the real situation of the street pollution in the last hour up to the present moment is presented. The data is collected and presented in real-time.

The aim for this kind of system is to be able to give not only automated real-time visualization of measured data, but also automated near-future modelled air pollution visualization, and make all of that available on the web like an internet service.

IV. DISCUSSION AND CONCLUSIONS

In this paper, we presented our system for monitoring modelling visualization of the traffic air pollution of the city of Skopje.

The benefit of this kind of system is the real-time aspect of the system with the combination of public availability of the generated color coded layers [18]. Another part that this system is incorporating inside is the prediction of the pollution by change of the traffic parameters using OSPM. This kind of visualization will be further improved in our future work.

From the experimental results we can conclude that the recommended limit values for health protection according to the latest EU Air quality standards (50 µg/m3 for pm10 daily average, 25 µg/m3 for pm2.5 yearly average) are significantly exceeded for PM10 and PM2.5 parameters concentration as it is visible from the Figures 8-9. The CO parameter values are in the normal range below the limit values of 30 mg/m3 according to WHO (World Health Organization). The Benzene is also in normal range below the 5 µg/m3 limit value as recommended by EU Air quality standards.

In the future, we plan to extend our study for other streets in the city of Skopje, as well as for other cities in our country and the Balkan region. We also plan to use the cloud computing for the process of modeling, visualization and other processing of these very large (big) air pollution data in real time.

The real time modelling and visualization of the air pollution is the first step that can lead to real time decision (not only prediction) support system. It will also allow better

 ISBN: 978-1-61208-353-7

planning and abating decisions from the authorities on the local, regional or trans-boundary global level.

V. References

[1] In Chae J., Guohua L., Sang B. L., Sensor-Based Emissions Monitoring System, IEEE lib. – pp. 336-339 (2013)

[2] Lee K., Murray D., Goodfield D., Anda M., Experiences and Issues for Environmental Engineering Sensor Network Deployments – Digital Ecosystems Technologies (DEST), 6th IEEE International Conference, E-ISBN: 978-1-4673-1701-6, pp 1-6 (2012)

[3] Pummakarnchana 0., Tripathi N., Dutta J., Air pollution monitoring and gis modeling: a new use of nanotechnology based solid state gas sensors - Science and Technology of Advanced Materials, pp 251–255 (2005)

[4] Pengfei You, Yuxing Peng, Hang Gao, "Providing Information Services for Wireless Sensor Networks through Cloud Computing," Services Computing Conference (APSCC), 2012 IEEE Asia-Pacific , vol., no., pp.362,364, 6-8 Dec. 2012

[5] Bae, W.D., Alkobaisi, S., Narayanappa, S., Liu, C.C., "A Mobile Data Analysis Framework for Environmental Health Decision Support," Information Technology: New Generations (ITNG), 2012 Ninth International Conference on , vol., no., pp.155,161, 16-18 April 2012

[6] Mell P., Grance T., The NIST definition of Cloud Computing, technical report, National Institute of Standards and Technology (2011)

[7] Jong Won Park, Chang Ho Yun, Shin-gyu Kim, Yeom, H.Y., Yong-Woo Lee, "Cloud computing platform for GIS image processing in U-city," Advanced Communication Technology (ICACT), 2011 13th International Conference on , vol., no., pp.1151,1155, 13-16 Feb. 2011

[8] Jong Won Park, Chang Ho Yun, Hae-Sun Jung, Yong-Woo Lee, "Visualization of Urban Air Pollution with Cloud Computing," Services (SERVICES), 2011 IEEE World Congress on , vol., no., pp.578,583, 4-9 July 2011

[9] Junyan Zhao, Junkui Zhang, Siqi Jia, Qi Li, Yue Zhu, "A MapReduce framework for on-road mobile fossil fuel combustion CO2 emission estimation," Geoinformatics, 2011 19th International Conference on , vol., no., pp.1,4, 24-26 June 2011

[10] Oliver Ling, Hoon Leh, Ting, Kien Hwa, Shaharuddin, Ahmad, Kadaruddin, Aiyub, Yaakob, Mohd Jani, "Air quality and human health in urban settlement: Case study of Kuala Lumpur city," Science and Social Research (CSSR), 2010 International Conference on , vol., no., pp.510,515, 5-7 Dec. 2010

[11] Suakanto, S., Supangkat, S.H., Suhardi, Saragih, R., "Smart city dashboard for integrating various data of sensor networks," ICT for Smart Society (ICISS), 2013 International Conference on , vol., no., pp.1,5, 13-14 June 2013

[12] Bae, W.D., Alkobaisi, S., Narayanappa, S., Liu, C.C., "A Mobile Data Analysis Framework for Environmental Health Decision Support," Information Technology: New Generations (ITNG), 2012 Ninth International Conference on , vol., no., pp.155,161, 16-18 April 2012

[13] Lee, K., Murray, D., Goodfield, D., Anda, M., "Experiences and issues for environmental engineering sensor network deployments," Digital Ecosystems Technologies (DEST), 2012 6th IEEE International Conference on , vol., no., pp.1,6, 18-20 June 2012

[14] In Chae Jeong, Guohua Li, Sang Boem Lim, "Sensor-based Emissions Monitoring System," Information Science and Service Science and Data Mining (ISSDM), 2012 6th International Conference on New Trends in , vol., no., pp.336,339, 23-25 Oct. 2012

[15] Rao, B.B.P., Saluia, P., Sharma, N., Mittal, A., Sharma, S.V., "Cloud computing for Internet of Things and sensing-based applications," Sensing Technology (ICST), 2012 Sixth International Conference on , vol., no., pp.374,380, 18-21 Dec. 2012

[16] World Health Organisation http://www.who.int/gho/urban_health/situation_trends/urban_population_growth_text/en/, last accessed 05.03.2014

[17] Operational Street Pollution Model, (www.au.dk/ospm), last accessed 05.03.2014

[18] http://www.skopjefinki.ekoinformatika.mk. last accessed 05.03.2014

[19] World Health Organisation
http://www.who.int/phe/air_quality_q&a.pdf?ua=1,
last acessed 06.07.2014

[20] Eu Air Quality Standards
http://ec.europa.eu/environment/air/quality/standards.htm,
last accessed 06.07.2014

[21] Jiaoyan Chen, Huajun Chen, Jeff Z. Pan, Ming Wu, Ningyu Zhang, and Guozhou Zheng. 2013. When big data meets big smog: a big spatio-temporal data framework for China severe smog analysis. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data* (BigSpatial '13). ACM, New York, NY, USA, 13-22. DOI=10.1145/2534921.2534924 http://doi.acm.org/10.1145/2534921.2534924

 ISBN: 978-1-61208-353-7

A Peer-to-Peer Model for Virtualization and Knowledge Sharing in Smart Spaces

Dmitry G. Korzun
Department of Computer Science
Petrozavodsk State University (PetrSU)
Petrozavodsk, Russia
e-mail: dkorzun@cs.karelia.ru
Helsinki Institute for Information Technology (HIIT)
and Department of Computer Science and Engineering (CSE)
Aalto University
Helsinki, Finland

Sergey I. Balandin
ITMO National Research University
St. Petersburg, Russia
FRUCT Oy
Helsinki, Finland
e-mail: sergey.balandin@fruct.org

***Abstract*—This paper is targeted to initiate discussion on how to describe and make formal definition of smart spaces in Internet of Things (IoT) environment by utilizing well-known models for Peer-to-Peer (P2P) networks. Indeed, when one starts studying smart spaces solutions and applying them in IoT environments then traditional models of P2P interaction come in mind as the first association. Every device, sensor, or network process could be represented by the corresponding agent (or peer). Services are emerged as a result of cooperative work of multiple agents in the smart space. Each agent contributes to the service by sharing its portion of knowledge in the smart space. We propose initial ideas on how a P2P model can virtualize physical objects and service construction processes by representing them as a network of interacting information objects in the smart space. Although interaction between objects is not physically direct communication, the model logically organizes direct interaction of objects as peers. This approach aims at higher interoperability in knowledge sharing and at an effective abstract level for service design.**

Keywords–Smart spaces; Peer-to-Peer; P2P; Multi-agent; Internet of Things; IoT; Services.

I. INTRODUCTION

We are about to see how hundreds of billions of interconnected devices envisioned by the Internet of Things (IoT) will finally become part of our daily life. Service networks build on IoT technology is become a reality of today and a strong call for making proper models and analysis of such networks. It is important to keep in mind that in the new generation of service networks most of communications will be handled between machines without direct contribution to some particular user service. Already, we see that Machine-to-Machine (M2M) communications are gaining momentum and soon will be mature enough to take significant share of real-world applications in various fields of our life.

Generally, IoT environments are becoming large, highly dynamic, hyperconnected, and functionally distributed, e.g., see [1]–[5] and references therein. Typically, an IoT environment consists of multiple heterogeneous networks with a large number of networked elements and users' devices. Further evolution of the IoT concept envisions increasing of the number of connections by yet another order of magnitude from currently connected approximately 10 billion "things". This will result in unprecedented challenges in network scalability, resource efficiency, privacy considerations, and overall management of this multitude of "things". The traditional models of networks organization would have serious problems to deal with it, so more and more often some alternative ways to network virtualization are considered [5].

Another key trend that we witness nowadays is a demand for making services be proactive and smarter to increase efficiency of IoT environment use and free more time for the user. Along this trend, over the past few years, we have seen many predictions and comments on importance and future perspectives of the smart spaces paradigm [4][6]–[9]. Despite of its elegance and clear advantages, we must admit that the paradigm still has very limited practical use. One of the problems is that its model of virtualization and knowledge sharing is still not so clear for service developers. On the other hand, we can see that these models are very close and even similar to what has been applied for many years in the Peer-to-Peer (P2P) systems area.

This study elaborates on how to apply models for virtualization and knowledge sharing in smart spaces deployed in IoT environment. We focused on the traditional approach to modeling P2P networks [10]. Our intention is to see how such models can be adopted for the problems of knowledge virtualization and sharing. As we know it is not a trivial task to make a useful model for the considered problem. In this paper, we are not constructing a finalized ready model that answers most of the questions, rather we are sharing results of study and analysis on how to adopt well-known P2P models for the emerging application area of smart spaces.

The rest of the paper is organized as follows. Section II states the problem of knowledge representation for smart spaces. Section III presents related work and enabler approaches to modeling for knowledge representation in the smart spaces area. Section IV describes our initial P2P model for knowledge virtualization and sharing. Section V discusses the use of P2P model for service construction and delivery. Section VI summarizes the paper.

II. SMART SPACES

Let us study specific features of smart spaces deployed in localized IoT environment. Such an environment consists of surrounding IoT devices (embedded in the physical environment or appeared as mobile entities), communication network that connecting them, plus it has access to the global Internet

 ISBN: 978-1-61208-353-7

with its diversity of services. Later we focus only on smart spaces that belong to this category.

In general, the smart spaces paradigm aims at development of ubiquitous computing environment, where participating entities acquire and apply knowledge to adapt services to the inhabitants in order to enhance user experience, quality and reliability of the provided information [6]–[8]. A primary operational element is a smart object—an autonomous information processing unit.

The term "smart" means [11] that the object is (i) active, (ii) digital, and (iii) networked. Any smart object (iv) operates to some extent autonomously, (v) is re-configurable, and (vi) has local control of resources it needs to utilize (e.g., energy and data storage). The IoT concept supports this vision on smart objects [2]. The most common view of IoT refers to the connection of physical objects, while the core of technology is in information interconnection and convergence. Operation of IoT solutions is based on continued processing of huge number of data flows, originated from various sources and consumed by multiple applications.

In contrast to this basic IoT vision, a smart object in the smart space is not necessarily attached to a fixed device, as any available device can host the object. This kind of virtualization provides a powerful abstraction for creating complex systems. For instance, the M3 concept for smart spaces employs the term "knowledge processor" (KP) to emphasize the processing responsibility of each object [4][12]. Services are constructed as interactions of smart objects in this shared space. The deployment flexibility is very high. For example, the smart space can be deployed using a cloud or on user's devices that interact with each other and use pertinent services regardless of the physical location.

The M3 concept further evolves this IoT-based fusion of physical and information worlds [12]. M3 stands for Multi-device, Multidomain, and Multivendor. An M3 smart space makes it possible to mash-up and integrate information between a wide spectrum of applications and domains spanning from embedded domains to the Web. Information from physical world (objects and devices in the physical environment) becomes easily available for participants in the shared smart space. The latter also is a hub linking the information to other services and solutions in the Internet. Therefore, smart spaces open embedded data kept in many surrounding devices to use by applications for creating local services in various physical places [8].

The multitude of participants (humans, machines, processes) obviously leads to the interoperability problem. The M3 concept provides the following conceptual solution. Ovaska *et al.* [7] defined a smart space a digital entity where the relevant real-world information (i.e., information about the physical environment, the objects therein located, and the recent situation) is stored in an interoperable, machine understandable format, kept up to date and made available to unanticipated and authorized situation dependent applications. Resource Description Framework (RDF) format from the Semantic Web provides a proper representation model to store the shared information [13][14]. SPARQL (Simple Protocol and RDF Query Language) is a query language to effectively retrieve and manipulate the information in the RDF format.

This definition supports three interoperability levels [8].

1) At the bottom, the communication level provides techniques for transmitting data between devices. It enables the device and network world to exchange bits.
2) At the middle, the service level provides technologies for devices to share services in the smart space. It enables the service world to use the services across device boundaries.
3) At the top, the information level allows the information to be understood similarly in all the smart objects. It equips the information world with the interoperability means to make the same meaning of information for different participants.

The notion of semantics is subject to various definitions, e.g., see Aiello *et al.* [15]. Since a smart space aims at encompassing (directly or indirectly) all information pieces the application system needs for service operation, we can characterize semantics as follows. Semantics is a relationship or mapping established between such information pieces. This definition also covers the case when relations are established implicitly, due to relating elements of the information structure. For instance, in ontology terms, such implicit relations appear between concepts (classes).

III. Related Work and Enabler Approaches

Let us discuss existing research on approaches to modeling for virtualization and knowledge sharing applicable in the case of smart spaces. The considered approaches will be adopted in the proposed P2P model later in Sections IV and V.

Halevy and Madhavan [16] introduced the corpus-based representation principle for large collections of knowledge fragments. Unlike a traditional knowledge base with careful ontological design, a knowledge corpus consists of independent uncoordinated contributions. This idea suits well to smart spaces where many autonomic participants share information and apply the collaboratively collected knowledge.

Bertossi and Bravo [17] considered virtual integration of many different data sources. A mediator (software system) offers a common interface to a set of autonomous, independent and possibly heterogeneous data sources. The same approach is applicable for organization of smart spaces content. The primary data are kept in their sources. The smart space acts as an informational hub to relate all the data and to provide to participants a single access point.

Patouni *et al.* [5] summarized recent virtualization trends for IoT environments development. The dynamics of such a hyperconnected and full of data telecommunications environment need moving the functionality to the network edges. For this purpose logical network services are distinguished from physical resources. Furthermore, Software Defined Networks (SDN) propose decoupling of the network control and data planes, moving the control of the network behavior to third party software. The idea is similar to solutions applied in P2P based large-scale network infrastructures [10]. Compared to our case, physical entities and resources are virtually represented in smart spaces, and the appropriate smart space supports making control decisions.

Aiello *et al.* [15] discussed the notion of emergent semantics. Local semantics from information agents are consolidated into a global, population-wide semantics. Knowledge representation structures emerge from continuous interaction of the agents. This incremental, bottom-up, semi-automatic

 ISBN: 978-1-61208-353-7

construction follows a P2P style, without relying on pre-existing, global semantic models. Emergent semantics supports virtual data integration in smart spaces: data of already existing sources may be updated, added, or deleted; new sources and services may appear and disappear dynamically.

Gorodetsky [9] studied smart space generic architecture composed of many agents interacting as peers in a P2P system. The agents are mediators for data integration in the smart space as well as they take care about construction of smart services and their delivery to users. The P2P approach is used for structuring agent interaction, i.e., establishing relations between agents for direct communication.

Pellegrino *et al.* [18] proposed a P2P-based infrastructure for distributed RDF storage and a publish/subscribe layer for storing and disseminating RDF events. The P2P approach allows constructing a large-scale distributed system for knowledge sharing based on existing Semantic Web technologies.

Matuszewski and Balandin [19] presented a P2P model and system architecture for knowledge sharing in mobile environments. Humans are treated as peers. Their collective knowledge is arranged into a distributed hierarchical structure based on user-defined relations between objects and references to the data sources of other peers.

IV. Content Representation Model

A characteristic property of any smart space is information sharing with knowledge self-generation from the collected content [4][6][9]. Ideally, all data a service needs should be accessed via its smart space: either the data are directly stored in the smart space or they are accessed indirectly by a kept reference. The property leads to many concurrent and low coordinated contributions, and we can consider information content of a smart space a large dynamic collection I of disparate knowledge fragments.

No careful design of a single comprehensive ontology or a database schema in advance is possible to represent finely tuned structure of the content. The corpus-based representation principle is used instead [16]. Smart space content I is structured dynamically, in ad-hoc manner. For its participants, the smart space provides search query interfaces to reason knowledge over I and its instant structure.

Based on the ontological modeling approach, we can consider I consisting of information objects and semantic relations among them [4][7][14]. Its basic structure is defined by problem domain and activity ontologies (classes, relations, restrictions), e.g., using the Web Ontology Language (OWL) from the Semantic Web. Factual objects in I are represented as instances (OWL individuals) of ontology classes and their object properties represent semantic relations between objects.

The well-known P2P approach [10] can be applied for modeling the virtualization of objects in the smart space and the derived knowledge representation. Any object $i \in I$ is treated a peer. Each i keeps some data (values of data properties) and has links to some other objects j (object properties). Therefore, a P2P network G_I is formed on top of I. Contributions from smart space participants (insert, update, delete) change the network of objects, similarly as it happens in P2P due to peers churn and neighbors selection. We shall also use the terms a node and a link when referring an element in G_I and its relation.

This P2P model extends the notion of ontology graph (interrelated classes and instances of them) to a dynamic self-organized system. The following model properties clarify this extension and show the role of enabler approaches from Section III.

Virtualization: Objects in G_I are self-contained pieces of information. It can be effectively described using OWL in terms of individuals and classes. Each object provides a digital representation of a real thing (sensor, phone, person, etc.) or of an artificial entity (event, service, process, etc.). This property suits well the IoT concept as well as its evolution to Internet of Everything [5]. Participants (agents) and information objects become equal nodes. From the point of view of applications, all essential system components become observed on "one stage" (with all semantic relations) and manipulated by changing their information representation (digital).

Hierarchy: The decomposition principle from ontological modeling allows defining semantic hierarchies of concepts, e.g., hierarchy of classes of an ontology. Objects in G_I becomes connected with hierarchical semantic links, as it happens in hierarchical P2P systems. In particular, this idea was applied by Matuszewski and Balandin [19] for P2P-like structuring personal information about a person and groups of persons.

Emergent semantics: There can be non-hierarchical semantic relations in G_I. They reflect the recent state of the dynamic system. For instance, relation "friendship" connects two persons or relation "is reading" appears between a person and a book. Object originals are autonomic and they constantly evolve. The representation of relations between them is also subject to frequent changes. Even global information is highly evolutionary: changes on the object's origin side (not in the representation in I) influence the semantics. That is, if an object corresponds to a database then updating its content can change the object's relations to others. This type of dynamic semantics consolidation from the local semantics held by participating objects follows the emergent semantics approach for knowledge management [15]. The property corresponds to the P2P network topology maintenance problem.

Composition: The granularity level of objects provides an additional degree of freedom. One can consider a group of objects in I as a node in G_I a self-contained element with own semantic relations. For instance, a group of persons forms a team or a service is constructed as a chain of simpler services. From the P2P point of view, the composition property is similar to peer clustering and aggregation, including superpeer-based P2P systems.

Data integration: A smart space can be considered a virtual data integration system [17] for multiple sources. Some objects in I represent external data sources (e.g., databases) and the means to access data (or even reason knowledge over these data) from the sources. This property is conceptually close to hybrid P2P architectures and P2P-based search problem, including semantic-aware P2P systems.

Based on this P2P model we can translate some well-known P2P problems for use in smart spaces.

1. Nodes heterogeneity. Objects in I are of different concepts (even incomparable) of the application problem domain. It provides basic restrictions on node linkage in G_I. For instance, some nodes cannot be connected with a direct link or cannot be clustered together, similarly as it happens in structured P2P networks. The same restrictions exist in practical deployments of P2P systems due to the Internet Protocol (IP) level reachability factors (e.g., a NAT prevents

 ISBN: 978-1-61208-353-7

establishing a direct IP connection between two P2P nodes).

2. Neighbor selection. Every knowledge fragment should serve the system goals. It means that any object of I relates some other objects to form local semantics (over the data attributes the object has). When an object has many relations the knowledge becomes less concretized, thus, similarly to the P2P case, a node in G_I preferably keeps a moderate number of direct links. In P2P networks, a node has short-range and long-range links: the former is for nearby nodes, the latter allows jumping to distant area of the network. A short-range link in G_I describes a kind of persistent or system-wide knowledge. Similarly, emergent semantics provide long-range links for G_I, representing less stable knowledge relations.

3. Network topology maintenance. Objects can apply certain system-level rules when selecting neighbors, as it happens in structured P2P networks. The aim is at maintaining knowledge representation that allows efficient knowledge reasoning over I based on existing technologies of Semantic Web (e.g., SPARQL). To some extent, the maintenance can also preserve the consistency of collected knowledge.

4. Routing. Knowledge reasoning over I is reduced to traversal in G_I, when semantic relations between objects allows interpreting and then forming derivate knowledge. Knowledge can be defined as a connected subgraph in G_I. In particular, such a subgraph consists of a node and some paths starting from this node. Routing algorithms provide a way to construct such graphs. An object (node in G_I) acts as a client when it needs knowledge, a server when it completes knowledge reasoning, and a router when it forwards the construction to subsequent objects for additional knowledge.

In summary, the model allows considering content I as interacting objects, which are active entities (make actions) on one hand and are subject to information changes (actions consequence) on the other hand. Result of interaction is derived knowledge in a graph-based form. This fact allows us to describe formally the conceptual processes of service construction and delivery.

V. Service Construction and Delivery

From the information-centric point of view, we can consider a service as knowledge reasoning over the content I and delivering the result to the users. Conceptual steps of the service construction are algorithmically formalized in Figures 1 and 2 (adapted from [20]). Let o be a particular ontology used by the service. Write $[q(o) \to I]$ to denote the action of content retrieval. The result is either existential (yes/no) or constructive (found piece of information). Write $I + y$ and $I - y$ to denote the insertion and removal of information piece y, respectively.

The algorithm in Figure 1 embodies actions an information service. Step 1 detects when the service is needed based on the current situation in the smart space. Step 2 makes selection of knowledge x to deliver to the user. Step 3 decides which UI elements are target devices. Step 4 updates recent visualization v_d to include x on device d.

Algorithm: Information Service

Require: Ontology o to access information content I of the smart space. The set U of available UI devices.

1: Await $[q_{\text{act}}(o) \to I] = \text{true}$ {event-based activation}
2: Query $x := [q_{\text{info}}(o) \to I]$ {information selection}
3: Select $d \in U$ {target UI devices}
4: Visualization $v_d := v_d + x$ {service delivery}

Figure 1. Actions in information service delivery.

Algorithm: Control Service

Require: Ontology o to access smart space information content I. The set U of available UI devices.

1: Await $[q_{\text{act}}(o) \to I] = \text{true}$ {event-based activation}
2: Query $x := [q_{\text{info}}(o) \to I]$ {information selection}
3: Decide $y := f(x, o)$ {formulation of control action}
4: Update $I := I + y$ {service delivery}

Figure 2. Actions in control service delivery.

The algorithm in Figure 2 embodies actions of a control service. Step 1 analyzes the space content to detect that a control action is needed. Steps 2 and 3 are reasoning in context of the current situation, and the service decides what updates (possibly without human intervention) are needed in the recent system state. The updates become available to the participants (original in the physical and information worlds).

From the architectural point of view, a service is made by interaction of software agents, when each agent makes its contribution by changing objects in I. Moreover, an agent can be represented as an object $i \in I$ itself, similarly as in [9].

Now let us formally define a smart space service as a step-wise branching process of changing objects in I. Information content of any object involved into this process can be used as service outcome to deliver to users. The definition captures the following properties, which we illustrate below using SmartRoom system [20][21] in the examples.

1. Information service. The simplest case of a service is reading the information content of a given object representation in I. The only step of the process is that someone has published or updated the object.

Example: SmartRoom keeps (as objects in I) all human participants (person profiles) and their presentations. The latter has links to the files with slides, e.g., in PDF (Portable Document Format). Information on participants and presentations is accessed via the corresponding objects and then visualized on appropriate user interface (e.g., SmartRoom client that a participant runs on her/his mobile computer).

2. Control service. An informational service can be extended with control functions due to the virtualization property. If a controlled entity is represented as $i \in I$ then changing i leads to appropriate actions at the i's original.

Example: Presentation-service of SmartRoom follows up changes in the slide number of currently shown presentation. Whenever the number is updated the new slide is displayed on the SmartRoom wall screen (media projector).

3. Step-wise process. Smart spaces are event-based: a change of $i \in I$ forms an event observed by other participants. When $i_1 \in I$ is changed it can course creating or updating $i_2 \in I$, and so on. The process can be branched, i.e., one change affects many objects.

Example: When a new participant joins SmartRoom then several objects appear in I: person profile, presentation, time restrictions, etc. In turn, the activity agenda is updated (a speaker is added), adapting to the current situation.

 ISBN: 978-1-61208-353-7

Let us consider how the P2P model supports the structural description of virtualization and knowledge sharing in smart spaces. As a result, service construction can be formulated in terms of flows of information changes, which is convenient for the use in service design.

Given a starting object $s \in I$ and its initial change. Let $D(s)$ be a graph routable from s in G_I. Construction of a service corresponds to a routing path $s \rightarrow^* d$, as schematically depicted in Figure 3 using thick arrows. Injection of the change starts the service, analogous to a P2P node starting a lookup query. The sequence of changes flows in G_I. Note that parallel paths are possible. Any point when an agent reads an object can be considered a final step of the service construction since the agent consumes an outcome.

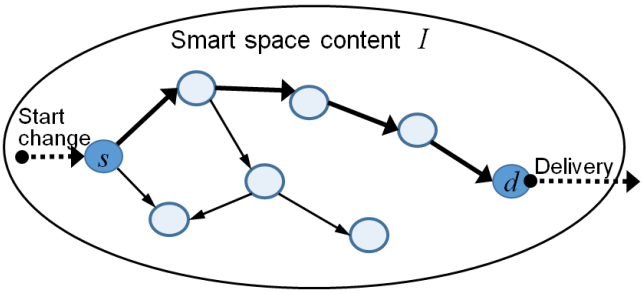


Figure 3. Service construction as P2P route $s \rightarrow^* d$.

This formalization is very flexible for various service constructions. There can be a large number of services due to freedom in selection of starting and final points.

Consider a path $s \rightarrow^* d$. There can be two types of links: ontological and mediatorial. An ontological link represents an object property from the ontology. Such a link is kept directly in I and is used in search queries of knowledge reasoning. A mediatorial link $i \rightarrow j$ is a result of actions of a participant (software agent): it analyzes object i and, as a consequence, changes another object j.

Figure 4 illustrates an example service "show me a slide". Thick arrows visualize the service construction process. Let user B browse available presentations kept in SmartRoom.

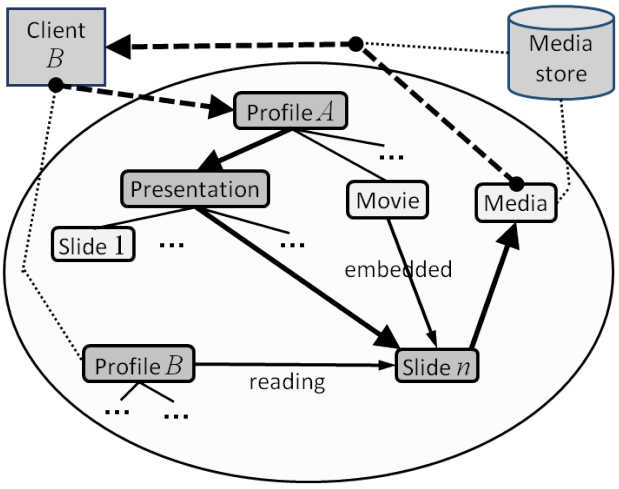


Figure 4. Example service construction for SmartRoom.

The B's client can find the A's presentation starting a path from Profile A and then running over hierarchical links till a given slide. The real slide (one-page PDF file) is physically located in SmartRoom media store (e.g., implemented as a web server). That is, Client B has to resolve the link from the smart space to the real content in the media store. Furthermore, if the recent slide embeds a movie, the latter is available for displaying by Client B from either the media store or the movie is located somewhere in public Internet and referenced by a global URL (Uniform Resource Locator, in particular use for web addressing).

As an additional effect of this service construction, more semantic relations can emerge in the smart space. In the example from Figure 4, the relation "reading" establishes the emergent semantics between B and the slide she or he is now analyzing.

VI. Conclusion

This paper addressed the problem of virtualization and knowledge sharing in smart spaces deployed in localized IoT environments. By this publication we want to initiate broad discussion on how to utilize well-known models for P2P networks for describing and making formal definition of virtualization and knowledge sharing in smart spaces. The paper summarized previous research on applicability of P2P methods for smart spaces. We proposed ideas on P2P modeling for virtualization and knowledge sharing in smart spaces. The discussion aims at the use of P2P models for service construction and delivery, following the M3 concept of smart spaces. We provided examples on how some well-known P2P problems, including P2P nodes heterogeneity, neighbor selection, network topology maintenance, and routing, are translated to smart spaces problems. Our P2P model allows structural description of virtualization and knowledge sharing, resulting in the definition of a smart space service in terms of information change flows. This description simplifies the problem of smart spaces design by providing enhanced abstractions for service construction and delivery.

Acknowledgment

This work is financially supported by Government of Russian Federation, Grant 074-U01. The research is a part of project 14.574.21.0060 (RFMEFI57414X0060) of Federal Target Program "Research and development on priority directions of scientific-technological complex of Russia for 2014–2020". The authors are grateful for DIGILE IoT SHOK program that provided required support of this research. The research of D. Korzun was funded by project # 1481 (basic part of state research assignment # 2014/154) of the Ministry of Education and Science of the Russian Federation.

References

[1] D. Miorandi, S. Sicari, F. De Pellegrini, and I. Chlamtac, "Survey internet of things: Vision, applications and research challenges," Ad Hoc Netw., vol. 10, no. 7, Sep. 2012, pp. 1497–1516.

[2] G. Kortuem, F. Kawsar, V. Sundramoorthy, and D. Fitton, "Smart objects as building blocks for the internet of things," IEEE Internet Computing, vol. 14, no. 1, Jan. 2010, pp. 44–51.

[3] D. J. Cook and S. K. Das, "Pervasive computing at scale: Transforming the state of the art," Pervasive Mob. Comput., vol. 8, no. 1, Feb. 2012, pp. 22–35.

 ISBN: 978-1-61208-353-7

[4] D. Korzun, S. Balandin, and A. Gurtov, "Deployment of Smart Spaces in Internet of Things: Overview of the design challenges," in Proc. 13th Int'l Conf. Next Generation Wired/Wireless Networking and 6th Conf. on Internet of Things and Smart Spaces (NEW2AN/ruSMART 2013), LNCS 8121, S. Balandin, S. Andreev, and Y. Koucheryavy, Eds. Springer-Verlag, Aug. 2013, pp. 48–59.

[5] E. Patouni, A. Merentitis, P. Panagiotopoulos, A. Glentis, and N. Alonistioti, "Network virtualisation trends: Virtually anything is possible by connecting the unconnected," in Proc. 2013 IEEE Software Defined Networks for Future Networks and Services (SDN4NFS). IEEE, Nov. 2013, pp. 1–7.

[6] S. Balandin and H. Waris, "Key properties in the development of smart spaces," in Proc. 5th Int'l Conf. Universal Access in Human-Computer Interaction (UAHCI '09). Part II: Intelligent and Ubiquitous Interaction Environments, LNCS 5615, C. Stephanidis, Ed. Springer-Verlag, 2009, pp. 3–12.

[7] E. Ovaska, T. S. Cinotti, and A. Toninelli, "The design principles and practices of interoperable smart spaces," in Advanced Design Approaches to Emerging Software Systems: Principles, Methodology and Tools, X. Liu and Y. Li, Eds. IGI Global, 2012, pp. 18–47.

[8] J. Kiljander, A. Ylisaukko-oja, J. Takalo-Mattila, M. Eteläperä, and J.-P. Soininen, "Enabling semantic technology empowered smart spaces," Journal of Computer Networks and Communications, vol. 2012, 2012, pp. 1–14.

[9] V. Gorodetsky, "Agents and distributed data mining in smart space: Challenges and perspectives," in Agents and Data Mining Interaction (ADMI 2012), LNAI 7607, L. Cao, Y. Zeng, A. Symeonidis, V. Gorodetsky, P. Yu, and M. Singh, Eds. Springer-Verlag, 2013, pp. 153–165.

[10] D. Korzun and A. Gurtov, Structured Peer-to-Peer Systems: Fundamentals of Hierarchical Organization, Routing, Scaling, and Security. Springer, 2013.

[11] S. Poslad, Ubiquitous Computing: Smart Devices, Environments and Interactions. John Wiley & Sons, Ltd, 2009.

[12] J. Honkola, H. Laine, R. Brown, and O. Tyrkkö, "Smart-M3 information sharing platform," in Proc. IEEE Symp. Computers and Communications (ISCC'10). IEEE Computer Society, Jun. 2010, pp. 1041–1046.

[13] D. Khushraj, O. Lassila, and T. W. Finin, "sTuples: Semantic tuple spaces," in Proc. 1st Annual Int'l Conf. Mobile and Ubiquitous Systems (MobiQuitous 2004). IEEE Computer Society, 2004, pp. 268–277.

[14] M. Palviainen and A. Katasonov, "Model and ontology-based development of smart space applications," Pervasive Computing and Communications Design and Deployment: Technologies, Trends, and Applications, May 2011, pp. 126–148.

[15] C. Aiello, T. Catarci, P. Ceravolo, E. Damiani, M. Scannapieco, and M. Viviani, "Emergent semantics in distributed knowledge management," in Evolution of the Web in Artificial Intelligence Environments, SCI 130, R. Nayak, N. Ichalkaranje, and L. Jain, Eds. Springer-Verlag, 2008, pp. 201–220.

[16] A. Y. Halevy and J. Madhavan, "Corpus-based knowledge representation," in Proc. 18th Int'l Joint Conf. on Artificial Intelligence (IJCAI'03). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 1567–1572.

[17] L. Bertossi and L. Bravo, "Consistent query answers in virtual data integration systems," in Inconsistency Tolerance, LNCS 3300, L. Bertossi, A. Hunter, and T. Schaub, Eds. Springer-Verlag, 2005, pp. 42–83.

[18] L. Pellegrino, F. Huet, F. Baude, and A. Alshabani, "A distributed publish/subscribe system for RDF data," in Data Management in Cloud, Grid and P2P Systems, LNCS 8059, A. Hameurlain, W. Rahayu, and D. Taniar, Eds. Springer-Verlag, 2013, pp. 39–50.

[19] M. Matuszewski and S. Balandin, "Peer-to-peer knowledge sharing in the mobile environment," in Proc. 5th Int'l Conf. on Creating, Connecting and Collaborating Through Computing (C5 '07). IEEE Computer Society, 2007, pp. 76–83.

[20] D. Korzun, I. Galov, A. Kashevnik, and S. Balandin, "Virtual shared workspace for smart spaces and M3-based case study," in Proc. 15th Conf. of Open Innovations Association FRUCT, S. Balandin and U. Trifonova, Eds. ITMO Univeristy, Apr. 2014, pp. 60–68.

[21] D. Korzun, I. Galov, and S. Balandin, "Smart room services on top of M3 spaces," in Proc. 14th Conf. of Open Innovations Association FRUCT, S. Balandin and U. Trifonova, Eds. SUAI, Nov. 2013, pp. 37–44.

 ISBN: 978-1-61208-353-7

Cost-Optimized Location and Service Management Scheme for Next-Generation Mobile Networks

Chulhee Cho
IT Strategy Department, IT Security Team
Seoul Guarantee Insurance Co. Ltd.
Seoul, Republic of Korea
E-mail: tgb017@nate.com

Jun-Dong Cho and Jongpil Jeong
Department of Human ICT Convergence
Sungkyunkwan University
Suwon, Republic of Korea
E-mail: {jdcho, jpjeong}@skku.edu

***Abstract*—We propose a cost-optimized location and service management scheme for next-generation mobile networks (NGWN), where a per-user service proxy is created in order to serve as a gateway between the mobile user and all client-server applications engaged by the mobile user. The service proxy is always co-located with the mobile user's location database during a location handoff, a service handoff also ensues to co-locate the service proxy with the location database. This allows the proxy to know the location of the mobile user at all times in order to reduce the network communication cost for service delivery. We analyze four integrated location and service management schemes. Our results indicate that the centralized scheme performs the best when the mobile user's service to mobility ratio (SMR) is low and ν (session to mobility ratio) is high, while the fully distributed scheme performs the best when both SMR and ν are high. Through analytical results, we demonstrate that different users with vastly different mobility and service patterns should adopt different integrated location and service management methods to optimize system performance.**

Keywords-Location Management; Service Management; LTE Networks; SMR.

I. INTRODUCTION

Location and service managements have often been separately addressed in literature [1]-[3]. For location management, the most popular scheme in Long-Term Evolution (LTE) networks is the MME-Cell scheme where each Mobile User (MU) has a Mobility Management Entity (MME). Whenever a MU enters a Cell, the system updates its MME location database so that when a call arrives, the MME location database knows exactly which Cell contains the MU. Variations to the basic MME-Cell scheme have been proposed in recent years to process location update and search operations more efficiently, e.g., Local Anchor (LA) [4], Forwarding and Resetting [5], Two-Level Pointer Forwarding [6], and Hybrid Replication with Forwarding [7], etc. These location management schemes are designed to handle location update and search operations without consideration to service management.

In this paper, we investigate the notion of integrated location and service management for minimizing network cost without making the assumption of fully replicated servers within cell in the LTE network. Instead, we target general personalized services in the LTE network including personal banking, stock market and location-dependent services for which the MU will communicate with a backend server.

Based on the concept of using a per-user service proxy as a gateway between the MU and all client-server applications engaged by the MU concurrently [8], the proxy keeps track of service context information such as the current state of the execution for maintaining service continuity. Similarly to Chen et al. [9], we always co-locate the MU's service proxy with the MU's location database, which stores the current location of the MU, so that the service proxy knows the current location of the MU at all times so as to eliminate the cost associated with tracking the user location on behalf of the server applications for data delivery. Whenever the MU moves across a registration area boundary, a location handoff occurs for the location management system to update the location database. If a location hand-off results in moving the MU's current location database to stay closer to the MU, then the associated service handoff will also move the service proxy to the same location [10].

In this paper, we investigate and analyze integrated location and service management schemes. These schemes derive from the basic MME-Cell and LA schemes for location management, and the personal service proxy scheme for service management in the LTE network. We are motivated to investigate and identify the best cost-optimized location and service management scheme that can be applied on an individual user basis to minimize the overall cost incurred to the LTE network per time unit for the servicing location and service operations of all users. The amount of cost saving is relative to the speed of the LTE network and is proportional to the number of users, so the benefit is especially pronounced for slow and congested networks with a large number of mobile users.

The rest of the paper is organized as follows. Section 2 provides a description of the related work. Section 3 describes in detail the four integrated schemes to be investigated and analyzed in the paper. Section 4 analyzes the cost incurred under each our schemes and presents analytical results with simulation validation. Finally, Section 5 summarizes the paper.

 ISBN: 978-1-61208-353-7

II. RELATED WORK

The seamless management of user mobility is an issue that involves every OSI [1]-[2] layer, from layers 1 and 2 (handover between cells), through layer 3 (routing updates in the network core), up to the application layer (persistence of transport connections and user state, delay-tolerant operation). The Internet Protocol suite did not originally include any support for end-point mobility. Over the years, a whole family of Mobile IP (MIP) procedures were introduced in an attempt to provide mobility support in a backward-compatible way. On the other hand, current cellular standards, such as LTE, have all been designed with mobility in mind and integrate the appropriate support in the core network. The cellular control plane includes elements that store and maintain the state of the terminal while its association to the network persists, and oversees the creation of appropriate bearers to seamlessly provide applications with the illusion of a constant connection between the mobile terminal and the network.

TABLE I. TERMINOLOGY.

Acronym	Meaning
LTE	Long Term Evolution 3/4G cellular network
MME	Mobility Management Entity
UE	User Equipment (cellular terminal)
eNB	Extended Node B (base station w/ controller)
SGW	Service Gateway (interface to IMS / phone system)
TA	Tracking Area (scope for initial UE paging attempt)
TAL	Tracking Area (TA) List
HSS	Home Subscriber Server
ECM	Evolved Packet System Connection Management
DHT	Distributed Hash Table

The MME for LTE network supports the most relevant control plane functions related with mobility: it authenticates the User Equipment (UE) as it accesses the system, it manages the UE state while the users are idle, supervises handovers between different base stations (extended Node B, eNB), establishes bearers as required for voice and Internet (packet data network, PDN) connectivity in a mobile context, generates billing information, implements so called lawful interception policies, and oversees a large number of features defined in its extensive 3GPP specifications. Table 1 summarizes the relevant acronyms that will be used throughout this paper.

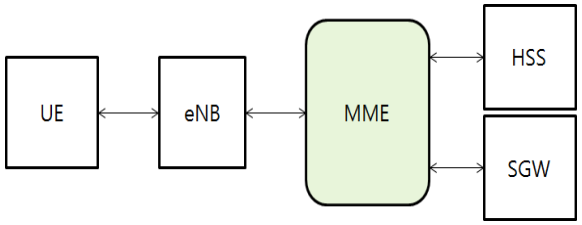


Figure 1. Schematic representation of the main logical MME interface.

All network events involve control plane messaging procedures require the interaction of one or more entities, besides the UE and the MME. The ones relevant to our interests are the eNB, which manages the air interface toward the UE, and the Service Gateway (SGW), a control plane element that acts as a global mobility anchor, managing the entire data plane within a large geographic region (usually spanning several TAs). The messaging sequences are codified by the 3GPP standards as logical interfaces, such as S-1 (eNB to SGW) and S-11 (SGW to MME). Fig. 1 schematically illustrates the interfaces supported by the MME, which are detailed in [11]-[12].

Mobile IP [13] allows a MU to maintain ongoing connections while roaming among IP subnets and requires the MU to inform its Home Agent (HA) of the new Foreign Agent (FA) address whenever it moves from one subnet to another. The function of a HA within a Mobile IP is similar to a Home Location Register (HLR) in Personal Communication System (PCS) networks for location management. Similar to the LA scheme in PCS networks, a variant of Mobile IP, called Mobile IP dynamic regional registration [14], has been proposed to group FAs into a gateway foreign agent (GFA) dynamically to minimize signaling costs in Mobile IP. These solutions, although elegant, solve only location management issues. For service management, a delivery protocol using a service proxy has been proposed to provide the reliable delivery of messages to MUs. However, the proxy used to forward messages to a MU must explicitly track the location of the MU, so extra communication costs are incurred to notify the proxy when the MU moves across a location registration area boundary.

III. COST-OPTIMIZED LOCATION AND SERVICE MANAGEMENT SCHEME

A. *Network Architecture*

We first describe a LTE system model for location management services. Then we describe an extended system model for integrated location and service management. We consider the LTE network architecture as shown in Fig. 2 where the LTE service areas are divided into Registration Areas (RAs).

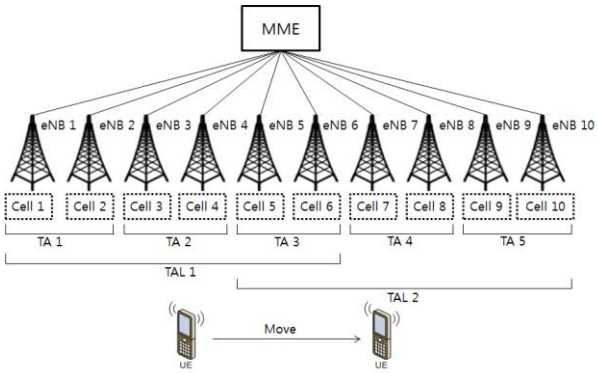


Figure 2. LTE Mobility Management Architecture.

We assume that a particular MU will remain in a Cell before moving to another. For simplicity, the residence time is assumed to be exponentially distributed with an average rate of σ. Such a parameter can be estimated using the approach described by Yang and Lin [15] on a per-user basis. We also assume that the inter arrival time between two

consecutive calls to a particular MU, regardless of the current location regarding the MU, is exponentially distributed with an average rate of λ.

TABLE II. PARAMETERS.

Parameter	Description
λ_s	The average rate at which the UE is being called.
λ_m	The average rate at which the UE moves across Cell boundaries.
γ	The average rate at which the UE requests services.
ν	call to mobility ratio, e.g., λ_s/λ_m
SMR	service request to mobility ratio, e.g, γ/λ_m
T	The average round trip communication cost between a Cell and the MME (or between a Cell and the server) per message.
τ_1	The average round trip communication cost between the anchor and a Cell in the anchor area per message.
τ_2	The average round trip communication cost between two neighboring anchor areas per message.
τ_3	The average round trip communication cost between two neighboring Cells per message.
M_{CS}	The number of packets required to transfer the service context.
N_S	The number of server applications concurrently engaged by the UE.
P_{InA}	The probability that a UE moves within the same anchor area when a Cell boundary crossing movement occurs.
P_{OutA}	The probability that a UE moves out of the current anchor area when a Cell boundary crossing movement occurs.

When applying the anchor scheme to the cost-optimized location and service management, the cost model must include not only location update/search costs, but also the communication cost between a UE and its servers. Also, to deliver responses from a server to a UE through the proxy, the proxy must know the UE's current location. It is desirable not to query the MME to obtain the location information because of the high communication cost. Thus, for an integrated local anchor scheme to serve both location and service handoffs, whenever the UE moves to a new anchor area, it may be desirable to also migrate the service proxy to the new anchor area to be "co-located" with the new anchor in an anchor area, so that the service proxy can query the anchor to know the current location of the UE without going to the MME. Consequently, both a location handoff and a service handoff would occur when the UE crosses an anchor boundary in the integrated scheme. A service handoff that migrates the service proxy involves two operations, namely, an address-change operation to inform all application servers of the location change, and a service context transfer. The cost of the address change operation per server is T. The service context transfer is unique for the service handoff operation, with the amount of context information being application dependent. The context transferred may include both static context information such as user profile and authentication data as well as dynamic context information such as files opened, objects updated, locks and time-stamps, etc. Let τ_2 be the average communication cost between two neighboring anchor areas (per packet), and M_{CS} be the number of packets required to transfer the service context. We list the system parameters considered in the paper in Table 2, including user parameters and application-specific parameters (such as M_{CS}). Their effects on the performance of Cost effective location and service management schemes are to be analyzed in the paper.

Note that for the case in which a UE concurrently interacts with multiple servers, there would still only be one per-user service proxy co-located with the location database under our proposed integrated schemes. In this case, the service rate parameter γ would reflect the aggregate rate at which the UE makes requests to these multiple services, while the context transfer cost parameter, M_{CS}, would reflect the aggregate context transfer cost for moving the service context information of multiple concurrent services from one location to another.

B. Operational Structures

In this section, we discuss four possible schemes, i.e., centralized, fully distributed, dynamic anchor, and static anchor for integrated location and service management.

We illustrate the centralized scheme in Fig. 3 (left). As the MU moves from Cell A, Cell B and subsequently to Cell C, the MME and the service proxy are updated to point to Cell B and then to Cell C sequentially.

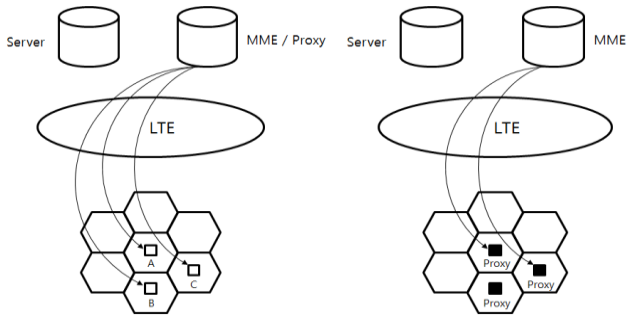


Figure 3. Centralized (left) and Full Distributed (right) Schema.

We illustrate the fully distributed scheme in Fig. 3 (right). When the MU moves from Cell A to Cell B, the service proxy migrates from Cell A to Cell B, and the MME and the server are updated to point to Cell B. The subsequent move to C behaves similarly. To service a location search request (not initiated from the current Cell), the MME database is accessed first to know the current Cell (A, B, or C) and then the MU is found within the current Cell. When the service proxy needs to forward replies to the MU, no additional searching cost is required to find the current Cell, since the service proxy is located in the current Cell.

Under the dynamic anchor scheme, a location anchor is used for location management such that the anchor changes whenever the MU crosses an anchor boundary. In addition, the anchor may also change its location within an anchor area when a call delivery operation is serviced. The service proxy dynamically moves with the anchor and is always collocated with the anchor. In Fig. 4, when a MU moves

 ISBN: 978-1-61208-353-7

within anchor area 1 from Cell A to Cell B, only the local anchor in Cell A is updated to point to the current location. Thus, the location update to the MME and application servers is avoided. Suppose that a call arrives after the MU moves into Cell C. The call will invoke a search operation in the MME database and a subsequent search operation in the anchor. Once the call is serviced, the MME database will be updated to point to Cell C; the anchor and the service context are moved from Cell A to Cell C; and the application servers are informed of the address change. Later, if the MU subsequently moves from Cell C to Cell D due to an inter-anchor movement, the MME database will be updated to point to Cell D, which will subsequently become the new anchor after the service context is transferred to it. Data delivery from the server will pass through the service proxy co-located with the anchor to reach the MU.

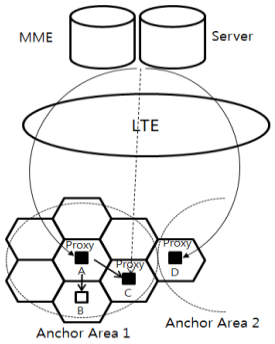


Figure 4. Dynamic Anchor Schema.

Under the static anchor scheme, the service proxy is again co-located with the anchor. However, the anchor will remain at a fixed location as long as the MU remains in the same anchor area. The only condition under which the anchor would move (along with the service context transferred) is when the MU moves across an anchor boundary. The procedures for processing the location update, call delivery, and service requests are the same as in the dynamic anchor scheme except that upon a successful call delivery, the anchor's location remains unchanged. Thus, there is no need to migrate the service proxy to the current serving Cell (if they are not the same) after serving a call delivery operation.

IV. PERFORMANCE ANALYSIS

In this section, we develop analytical models for evaluating and comparing various integrated schemes introduced in Section 3. We first define the communication cost analysis model for two states in the LTE system.

For analysis, the two-dimensional hexagonal random walk model [11]-[13] has been adopted. The LTE system can be assumed to be configured as a hexagonal network with a cell having radio coverage of an eNB. The UE moves from one cell to another, and its movement is modeled based on the two-dimensional hexagonal random walk model. In this model, a hexagonal cell structure is modeled and the cells are classified in a 6-layer cluster shown in Fig. 5. We assume that an UE resides in a cell unit for a specified time period and then moves to any of the neighboring cells with equal probability. Using this, a one-step transition matrix of this random walk can be derived by letting $P_{(x,y)(x',y')}$ be the one step probability from state (x,y) to (x',y'). Table 3 describes the system parameters for performance analysis.

TABLE III. SYSTEM PARAMETERS FOR PERFORMANCE ANALYSIS.

Parameter	Description
$C_{ServInM}$	The average cost of performing an intra-anchor location update operation when the UE changes its Cell within the same anchor area.
$C_{ServOutM}$	The average cost of performing an inter-anchor location update operation when the UE moves out of the current anchor area.
$C_{ServCvdC}$	The cost to handle a call delivery operation when the current Cell is the same as the anchor Cell.
$C_{ServNonCvdC}$	The cost for handling a call delivery operation when the current Cell is different from the anchor Cell.
$C_{ServCvdS}$	The cost to handle a service request when the anchor resides in the current serving Cell.
$C_{ServNonCvdS}$	The cost to handle a service request when the anchor is different from the current serving Cell.
$C_{servInM}$	The average cost of performing an intra-anchor location update operation when the UE changes its Cell within the same anchor area.
$C_{servOutM}$	The average cost of performing an inter-anchor location update operation when the UE moves out of the current anchor area.
C_{ServC}	The cost to handle a call delivery.
C_{ServS}	The cost to handle a service request.

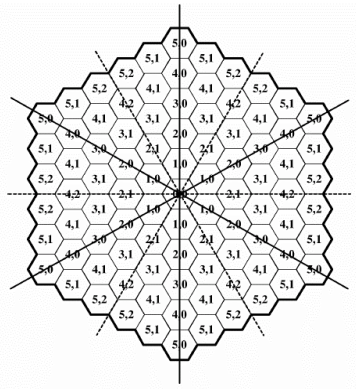


Figure 5. Hexagonal cell structure for performance analysis.

From [11]-[13], we can model the signaling cost of two mobility states in an LTE system: LTE_ACTIVE where the network directs UE to the serving cell and the UE is ready to perform Uplink/Downlink transport with very limited access delay, and LTE_IDLE where the UE in a low power consumption state, could be tracked in the Tracking Area and be able to travel to LTE_ACTIVE at approximately 100ms.

We first parameterize the performance models developed by means of a hexagonal network coverage model for describing a LTE network to evaluate the performance of the cost-optimized location and service management schemes proposed in order to identify conditions under which one scheme could perform the best when given a set of parameters characterizing a UE's mobility and service

 ISBN: 978-1-61208-353-7

behaviors [15]. We use a hexagonal network coverage model to describe a LTE network where cells are assumed to be hexagonally shaped, with each cell having six neighbors. At the lowest level of Fig. 5, an n-layer Cell covers $3n^2-3n+1$ cells where n is equal to either two or three. For a LTE system described by the hexagonal network coverage model as such, it can be shown that [11] with random movements, the probability that a UE moves within the same anchor area, that is, the probability of an intra-anchor movement, as the UE moves across a Cell boundary, is given by (1):

$$P_{InA} = \frac{3n^2-5n+2}{3n^2-3n+1} \quad (1)$$

Thus, the probability of an inter-anchor movement, when the UE moves across a Cell boundary, is given by (2):

$$P_{OutA} = 1-\frac{3n^2-5n+2}{3n^2-3n+1} = \frac{2n-1}{3n^2-3n+1} \quad (2)$$

Without loss of generality, consider $n=2$ for *n*-layer Cells, TAL, TA and MME composing the LTE. Then, the probability P_{InR} that a UE moves within the same TAL, that is, the probability of an intra-TAL movement, when the UE moves across a Cell boundary, is given by (3):

$$P_{InR} = \frac{21n^2-27n+10}{7\left(3n^2-3n+1\right)} \quad (3)$$

Let C_{ta} be the cost of searching UE in a Cell. Let C_{mme} be the cost of transmitting a message between TA/TAL and MME. The communication between MME and a Cell will traverse through the Cell-TA/TAL-MME path sequence. Let C_{lte} be the cost of transmitting a message between a Proxy and Application Server, or a MME and Application Server. So, we define that network cost T between a specific Cell and Application Server will be equal to C_{lte}. For the centralized scheme, there are no additional parameters to parameterize. For the fully distributed scheme, we need to parameterize τ_3 standing for the average communication cost between two neighboring Cells. With reference to the LTE network shown in Fig. 1, the communication cost between two Cells within the same TA/TAL (with probability P_{InA}) is $2C_{ta}$; the communication cost between two Cells out of the same TA/TAL but within the same MME (with probability $P_{InR}-P_{InA}$) is $2(C_{ta}+C_{mme})$; the communication cost between two Cells out of the same MME (with probability $1-P_{InR}$) is $2C_{ta}+2C_{mme}+C_{lte}$. Therefore, τ_3 can be parameterized as (4):

$$\tau_3 = 2C_{ta}\times P_{InA} + 2(C_{ta}+C_{mme})\times(P_{InR}-P_{InA}) + (2C_{ta}+2C_{mme}+C_{lte})\times(1-P_{InR}) \quad (4)$$

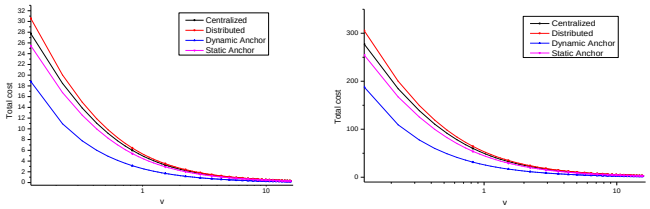
For the dynamic anchor scheme, we need to parameterize τ_1 for the average communication cost between the anchor Cell and another Cell (other than the anchor Cell itself) in an anchor area, as well as τ_2 for the average signaling communication cost between two neighboring TA/TAL areas. In (5), τ_1 is equal to the communication cost between two Cells within the same TA/TAL. To calculate τ_2, two scenarios are considered: the communication between two Cells within the same TA/TAL with cost $2(C_{ta}+C_{mme})$ and the communication between two Cells out of the same TA/TAL with cost $2C_{ta}+2C_{tal}+C_{lte}$.

$$\tau_1 = 2C_{ta}$$
$$\tau_2 = 2(C_{ta}+C_{mme})\times\frac{P_{InR}-P_{InA}}{1-P_{InA}} + (2C_{ta}+2C_{mme}+C_{lte})\times\frac{1-P_{InR}}{1-P_{InA}} \quad (5)$$

For the static anchor scheme, we need to parameterize τ_1 for the average communication cost between the anchor Cell and any Cell (including possibly the static anchor Cell itself) in an anchor area, as well as τ_2 for the average signaling communication cost between two neighboring TA/TAL areas, in as (6). Since the static anchor scheme does not track the location of the MU within an anchor area, the MU can reside in each Cell with equal probability. Thus, for a LTE network with $n=2$ where each TA/TAL has 7 Cells.

$$\tau_1 = 2C_{ta}\times\frac{6}{7}+0\times\frac{1}{7} = C_{ta}\times\frac{12}{7}$$
$$\tau_2 = 2(C_{ta}+C_{mme})\times\frac{P_{InR}-P_{InA}}{1-P_{InA}} + (2C_{ta}+2C_{mme}+C_{lte})\times\frac{1-P_{InR}}{1-P_{InA}} \quad (6)$$

We present numerical data obtained based on our analysis for a LTE network consisting of a 2-layer Cell, TA/TAL and MME as shown in Fig. 1 modeled by the hexagonal network coverage model. Performances of the centralized, fully distributed, dynamic anchor, and static anchor schemes in the LTE network in terms of the communication cost incurred to the network per time unit as a function of CMR and SMR under identical network signaling-cost conditions, whereby all costs are normalized with respect to the cost of transmitting a message between a Cell and its MME, i.e., $C_{ta}=0.5$, such that $C_{mme}=1$ and $C_{lte}=6$.



(a) $\lambda_s=0.1$ (b) $\lambda_s=1$

Figure 6. Total Cost Under Different SMR (Session) Values.

Fig. 6 shows the cost incurred to the LTE network per second as a function of the UE's v for cost effective schemes. The *X* coordinate represents the value in the range [0.1 16] with the call arrival rate λ_s fixed at 0.1, 1 while changing the mobility rate λ_m. When the v value is low, both the centralized and fully distributed schemes perform worse than the dynamic and static anchor schemes. This is attributed to the fact that the total cost rate is dominated by mobility-related cost factors at low v at which the mobility rate is much higher than the call arrival rate. Specifically, the centralized scheme performs badly in this condition because of the high cost for servicing location update operations as these operations need to access the MME in the centralized scheme. The fully distributed scheme does not performs well at low v because with a high mobility rate, the location update cost and the context

 ISBN: 978-1-61208-353-7

transfer cost are high in the fully distributed scheme. At very high ν, the centralized scheme performs the best followed by the dynamic anchor over fully distributed and in the last place the static anchor scheme. The dynamic anchor performs better than the static anchor in this extreme case because in the dynamic anchor scheme the anchor collocated with the service proxy is close to the UE. Thus, the cost for service requests and location updates due to movements within an anchor area is low. Another reason is that when a call arrives and the anchor Cell is not the current serving Cell, the dynamic anchor scheme will update the MME after the call is serviced and move the anchor to the current Cell. This keeps the MME database up-to-date and keeps the anchor close to the UE. As a result, it reduces the call delivery cost since the system is able to find the UE quickly on subsequent calls, the effect of which is especially pronounced when ν is high.

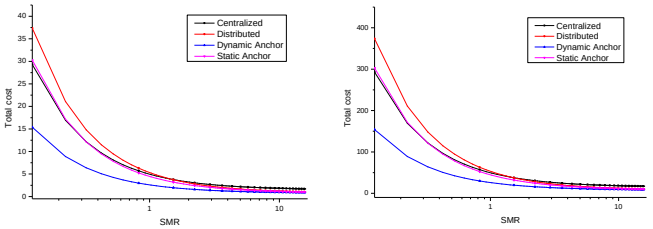


Figure 7. Total Cost Under Different SMR (Service) Values.

In Fig. 7, as the SMR increases, the cost rate under all four schemes increase because when the mobility rate γ is fixed. At very high SMR, however, the fully distributed scheme performs the best among all followed by the dynamic anchor over the static anchor and centralized because in the fully distributed scheme, the UE's service requests can be serviced quickly by the local service proxy located in the current Cell database.

V. CONCLUSION

In this paper, we investigated the concept of cost effective location and service management with the objective to reduce the overall communication cost for servicing mobility-related and service-related operations by the integrated LTE network environment. Our analysis result shows that the dynamic anchor scheme performs the best in most conditions except when the context transfer cost is high (when the server is heavy). The centralized scheme performs the best at low SMR and high ν. Also, the fully distributed scheme performs the best at high SMR and high ν. The static anchor scheme is a relatively stable scheme, performing reasonably well under a wide range of parameter values examined in the paper. These results mean that different users with vastly different mobility patterns should adopt different cost-optimized location and service management scheme for the better system performance.

ACKNOWLEDGEMENT

This research was supported by the Ministry of Trade, Industry and Energy (MOTIE), KOREA, through the Education Support program for Creative and Industrial Convergence (Grant Number N0000717). Also, this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2010-0024695).

REFERENCES

[1] M. H. Dunham and V. Kumar, “Impact of mobility on transaction management”, Proceedings of the International Workshop on Data Engineering for Wireless and Mobile Access, MobiDE '99, 1999, pp. 14-21.

[2] Y. Fan, “General modeling and performance analysis for location management in wireless mobile networks”, IEEE Trans. on Computers, vol. 51, no. 10, October 2002, pp. 1169-1181.

[3] I. Widjaja, P. Bosch, and H. La Roche, “Comparison of MME signaling loads for long-term-evolution architectures”, In VTC Fall, September 2009, pp. 1-5.

[4] J. S. Ho and I. F. Akyildiz, “Local anchor scheme for reducing signaling costs in personal communications networks”, IEEE/ACM Transactions on Networking, vol. 4, no. 5, October 1996, pp. 709-725.

[5] R. Jain, Y. B. Lin, C. Lo, and S. Mohan, “A forwarding strategy to reduce network impacts of PCS”, 14th Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE INFOCOM '95), April 1995, pp. 481-489.

[6] W. Ma and Y. Fang, “Two-level pointer forwarding strategy for location management in PCS networks”, IEEE Transactions on Mobile Computing, January 2002, pp. 32-45.

[7] I. R. Chen and B. Gu, “Quantitative analysis of a hybrid replication with forwarding strategy for efficient and uniform location management in mobile wireless networks”, IEEE Transactions on Mobile Computing, January 1998, pp. 3-15.

[8] K. Kitagawa, T. Komine, T. Yamamoto, and S. Konishi, “A Handover Optimization Algorithm with Mobility Robustness for LTE systems”, Personal Indoor and Mobile Radio Communications (PIMRC), 2011 IEEE 22nd International Symposium on, September 2011, pp. 1647-1651.

[9] I. Chen, B. Gu, and S. Cheng, “On Integrated Location and Service Management for Minimizing Network Cost in Personal Communication Systems”, IEEE Trans. on Mobile Computing, February 2006, pp. 172-192.

[10] M. Roussopoulos, “Personal-level routing in the mobile people architecture”, Proceedings of the USENIX Symposium on Internet Technologies and Systems, Boulder, CO, USA, October 1999, pp. 165-176.

[11] 3GPP standardization, “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) Overall description Stage 2”, TS 36.300 v9.2.0, http://www.3gpp.org/, [retrieved: January, 2010].

[12] 3GPP standardization, “General packet radio service enhancements for evolved universal terrestrial radio access network access”, TS 23.401, http://www.3gpp.org/, [retrieved: March, 2011].

[13] C. E. Perkins, “Mobile IP”, IEEE Comm. Magazine, May 1997, pp. 84-99.

[14] J. Xie and I. F. Akyildiz, “A novel distributed dynamic location management scheme for minimizing signaling costs in Mobile IP”, IEEE Trans. on Mobile Computing, vol. 1, no. 3, 2002, pp. 163-175.

[15] S. R. Yang and Y. B. Lin, “Performance Evaluation of Location Management in UMTS”, IEEE Transactions on Vehicular Technology, November 2003, pp. 1603-1615.

 ISBN: 978-1-61208-353-7

New First - Path Detector for LTE Positioning Reference Signals

Paweł Gadka
Gdańsk University of Technology
Faculty of Electronics, Telecommunications and Informatics
Department of Radio Communication Systems and Networks
Gdańsk, Poland
pawgadka@pg.gda.pl

***Abstract*—In today's world, where positioning applications reached a huge popularity and became virtually ubiquitous, there is a strong need for determining a device location as accurately as possible. A particularly important role in positioning play cellular networks, such as Long Term Evolution (LTE). In the LTE Observed Time Difference of Arrival (OTDOA) positioning method, precision of device location estimation depends on accuracy of the Positioning Reference Signal (PRS) first-path detection, what is particularly challenging in multipath environment. There are a few algorithms, available in the literature, that are dedicated to detect the first-path of PRS signals, often basing on estimation of the strongest path or which do not adapt to continuously changing environmental conditions. The paper presents a new algorithm, called First-Path Estimator (FPE), which detects the first path of received PRS signal in the LTE system. Simulations showed that proposed algorithm reduces Received Signal Time Difference (RSTD) measurement error compared to the well-known Maximum Likelihood Estimator (MLE) in urban scenario.**

Keywords-RSTD; LTE; OTDOA; PRS.

I. INTRODUCTION

Long Term Evolution (LTE) system specification driven by the 3rd Generation Partnership Project (3GPP) consortium [1] defines several methods for positioning, i.e., locating of mobile terminals in the network coverage area. These methods may be particularly useful in harsh environments such as densely populated urban areas or indoor locations, where Global Navigation Satellite Systems (GNSS) [2][3] are not working with accuracy demanded for some applications. One of the specified method is Observed Time Difference of Arrival (OTDoA) [4][5], for which special reference signals, called Positioning Reference Signals (PRS) [6], were defined. The OTDoA bases on measurements of the reception time differences between PRS signals arriving from multiple base stations to the User Equipment (UE). These measurements, along with knowledge of the geographical coordinates of the measured base stations and their relative timing, allow for estimation of the UE location in the network. Despite the frequency reuse factor in LTE equals one, transmissions of the PRS signals take place with frequency reuse factor of six in order to avoid near-far effect. In time domain, PRS signals are located on so-called positioning occasions, which are periodically repeating. Every positioning occasion consists of the set of consecutive PRS subframes. Furthermore, it is possible to mute PRS transmission in chosen positioning occasion, what, with scalable bandwidth of PRS signals, gives a sophisticated tool for flexible management of radio resources. The PRS signal main parameters are shown in Table I [7].

TABLE I. PRS SIGNAL PARAMETERS

PRS Bandwidth	1.4, 3, 5, 10, 15, 20 MHz
PRS Periodicity	160, 320, 640, 1280 ms
Consecutive PRS Subframes	1, 2, 4, 6

Accuracy of UE positioning in OTDOA depends mainly on PRS signals time of arrival estimation accuracy, therefore it is crucial to detects PRS reception time as precisely as possible. In this paper, new Time Delay Estimation (TDE) algorithm, called First-Path Estimator (FPE), detecting the first path of PRS signal, is proposed. The aim of its elaboration is to fulfill the requirements imposed by 3GPP on the accuracy of UE location estimation [8].

There are several algorithms available in the literature that detect arriving signal, e.g., Maximum Likelihood Estimator (MLE) or Fitz estimator [7]. The MLE algorithm detects the strongest path in the received signal, what may causes large errors in multipath environments, where in majority of cases the first arriving signal path is not the strongest one. The Fitz estimator is a low complex algorithm that moves signal reception time estimation into frequency domain. Its performance is close to the MLE in terms of Root Mean Square Error (RMSE) of time of signal arrival estimation. Another worth to mention algorithm is an estimator that detects the first arriving signal by the first occurrence of signal level above a -30 dB detection threshold relative to the strongest peak in the Power Delay Profile (PDP). If the dynamic range of the PDP is less than 30 dB, the first occurrence of signal level above the noise floor in PDP is used [9]. This algorithm however, does not adjust threshold to the noise floor, what may results in errors of first path detection in different power of noise relative to the power of arriving signal scenarios. FPE algorithm adjusts dynamically threshold accordingly to the ratio of signal power to noise power, what makes it more precise in the estimation of the first path time of arrival.

The rest of this paper is organized as follows. Section II describes MLE and proposed FPE estimator. Section III gives an overview on simulation model and section IV presents results of the simulations. At the end of the paper conclusions were drawn.

 ISBN: 978-1-61208-353-7

II. TIME DELAY ESTIMATION ALGORITHMS

Time delay estimation algorithms, with regard to positioning process in LTE, are in charge of reception time measurements of the PRS signals, performing in order to calculate the Received Signal Time Difference (RSTD) value. RSTD defined by 3GPP in [10] is a parameter specifying the relative timing difference between the neighbour cell *j* and the reference cell *i* and is defined as

$$RSTD = T_{SubframeRxj} - T_{SubframeRxi} \quad (1)$$

where $T_{SubframeRxj}$ is the time when the user equipment receives the start of one subframe containing PRS signal from cell *j* and $T_{SubframeRxi}$ is the time when the UE receives the corresponding start of one subframe (also containing PRS signal) from cell *i* that is the closest in time to subframe received from cell *j*.

Among other things, accuracy of the RSTD measurements depends on the size of search window, where the TDE algorithm searches for the desired signals. LTE system provides a tool for prior estimation of the expected RSTD and expected RSTD uncertainty values, both determining a search window for PRS signals. Properly evaluated search window is necessary for accurate estimation of the position in time of PRS signals receiving from neighbour base stations. Generally, the smaller the search window, the more accurate estimation could be performed. Referring to the Fig. 1, minimal and maximal values of RSTD can be evaluated as [11]:

$$RSTD_{\min} = \frac{\left| d_{nei1} - d_{ref1} \right|}{c} \quad (2)$$

$$RSTD_{\max} = \frac{\left| d_{nei2} - d_{ref2} \right|}{c} \quad (3)$$

where *c* is the radio waves propagation velocity. Values d_{ref1} and d_{ref2} could be estimated based on the cell size or Timing Advance (TA) measurements [10]. Then, it can be written:

$$RSTD_{\min} = \frac{\left| d_{ref-nei} - 2 \cdot d_{ref} \right|}{c} \quad (4)$$

$$RSTD_{\max} = \frac{\left| d_{ref-nei} \right|}{c} \quad (5)$$

where $d_{ref-nei}$ is the distance between $\text{eNodeB}_{\text{ref}}$ and $\text{eNodeB}_{\text{nei}}$ and $d_{ref} = d_{ref1} = d_{ref2}$. Expected RSTD value is

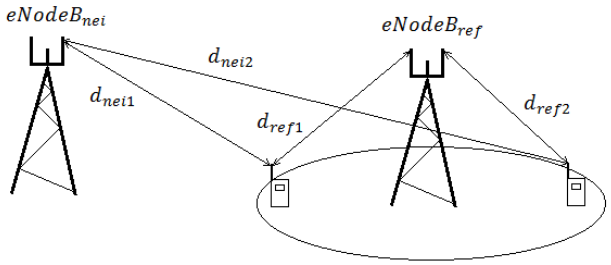


Figure 1. Illustration to expected RSTD and expected RSTD uncertainty values computation

computed as a mean value of (4) and (5):

$$RSTD_{\exp} = \frac{RSTD_{\min} + RSTD_{\max}}{2} \quad (6)$$

Finally, the search window, in which TDE algorithm searches for PRS signals arriving from a base station, is $\langle RSTD_{\min}, RSTD_{\max} \rangle$ centered at (6).

A. Maximum Likelihood Estimator

Maximum Likelihood Estimator finds the position of the strongest path in the received signal and can be defined as [7]:

$$T_{PRS} = \arg\max \left\{ \left| \sum_{i=0}^{P-1} r[i+m] \cdot s_{PRS}^{*}[i] \right|^2 \right\} \quad (7)$$

where T_{PRS} is a position in time of PRS signal, *i* is a time index, *m* refer to delays of correlation function, *P* is a PRS signal length in time domain, $r[i]$ is a received signal, $s_{PRS}[i]$ is a PRS signal and $(\cdot)^*$ is a complex conjugation operation.

B. Proposed First - Path Estimator

Proposed First-Path Estimator searches for the first path in the receiving signal through analyzing the correlation between receiving PRS signal and signal pattern stored in the memory. Algorithm, using knowledge about parameters of the highest correlation peak and all correlation peaks that lies before the highest one, significantly reduces search window and chooses from the new window the first peak, assuming that it correspond to the first path of the received signal.

The problem of detecting the first path in the correlation function is to set the threshold above which it may be assumed that detected peaks corresponds to the paths of the useful signal. Setting too low threshold THR_{low} (Fig. 2) induces detection of peak *P1*, which does not represent path of the useful signal. On the other hand, setting too high

 ISBN: 978-1-61208-353-7

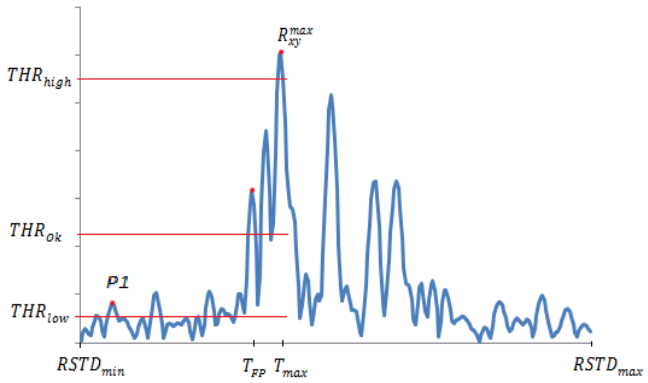


Figure 2. Correlation function of received PRS signal and PRS signal pattern stored in the receiver memory

threshold THR_{high} causes that peak detector misses the first path, in the worst case, becomes MLE estimator.
Therefore, it is necessary to find an appropriate value THR_{Ok}, which is done here through the estimation of the difference between maximum value of the correlation function and the mean value of all peaks that lies before this maximum value.

In order to find an estimate of the first path, algorithm firstly evaluate position $T_{\max}$ and value $R_{\max}$ of the highest peak in the correlation function $R_{xy}[m]$:

$$R_{xy}[m] = \sum_{i=0}^{P-1} r[i+m] \cdot s_{PRS}^{*}[i] \tag{8}$$

$$T_{\max} = \arg\max\left\{ \left| R_{xy}[m] \right|^2 \right\} \tag{9}$$

$$R_{\max} = \max\left\{ \left| R_{xy}[m] \right| \right\} \tag{10}$$

A position of the peak T_{FP} in function (8) that corresponds to the first path of received signal satisfies $T_{FP} \leq T_{\max}$, thus searching out the highest peak in the correlation function (8) allows to reduce the search window to $\langle RSTD_{\min}, T_{\max} \rangle$. Then, algorithm finds all the correlation peaks in a new interval, i.e., set R^p containing all pairs $\{m, R_{xy}[m]\}$ for which the conditions (11), (12) and (13) are satisfied.

$$R_{xy}[m-2] < R_{xy}[m-1] < R_{xy}[m] \tag{11}$$

$$R_{xy}[m] > R_{xy}[m+1] > R_{xy}[m+2] \tag{12}$$

$$m \leq T_{\max} \tag{13}$$

The mean value $\overline{R}_{xy}$ of all peaks within set R^p is given by

$$\overline{R}_{xy} = \frac{1}{\left| R^p \right|} \sum_{R_{xy}[m] \in R^p} R_{xy}[m] \tag{14}$$

where $\left| R^p \right|$ is a number of elements within set R^p.
The value of the threshold THR_{Ok} then is computed as

$$THR_{Ok} = \begin{cases} A \cdot R_{\max}, \ if \ 0.8 \leq R_{xy}^{con} \leq 1 \\ B \cdot R_{\max}, \ if \ 0.7 \leq R_{xy}^{con} < 0.8 \\ C \cdot R_{\max}, \ if \ 0.5 \leq R_{xy}^{con} < 0.7 \\ D \cdot R_{\max}, \ otherwise \end{cases} \tag{15}$$

where

$$R_{xy}^{con} = \left(R_{\max} - \overline{R}_{xy} \right) / R_{\max} \tag{16}$$

Position of the first peak in the correlation function (8) above THR_{Ok} is assumed as the first path position in the received signal. Algorithm changes THR_{Ok} threshold accordingly to the difference between the highest and the mean value of the remains peak components in the correlation function from $\langle RSTD_{\min}, T_{\max} \rangle$ adjusting threshold properly to the environmental conditions. When the difference $R_{\max} - \overline{R}_{xy}$ in (16) decreases, THR_{Ok} value has to be increased in order to avoid detection of the peak component that do not relate to the useful signal path. Values of the weights A = 0.33, B = 0.5, C = 0.9 and D = 1 in (15) were chosen empirically to minimize RSTD measurement error.

III. SIMULATION MODEL

In the implemented simulation model, UE performs RSTD measurements between subframes containing PRS signals transmitted from two base stations. In order to receive reference signals, UE firstly synchronizes with the reference cell using Primary Synchronization Signals (PSS) and then is informed about expected RSTD and expected RSTD uncertainty values, which determine position and size of the search window for PRS signals. In the prior RSTD value estimation process, the distance between $eNodeB_{ref}$ and UE is estimated through TA measurements. Due to an inaccuracy of timing advance measurements, measured distance $\underline{d_{TA}}$ is a random variable described by:

$$\underline{d_{TA}} = d + \underline{d_{TA}^{'}} \tag{17}$$

where d is the real distance between base station and UE and $\underline{d_{TA}^{'}}$ is a random variable generated from the normal distribution with mean value of 0m and standard deviation

 ISBN: 978-1-61208-353-7

48.83m corresponding to UE Rx-Tx time difference [10] measurements accuracy requirements given by 3GPP [5] and equals to $\pm 10 \cdot T_s$, where $T_s = 1/(15000 \cdot 2048)$ is the LTE basic timing unit [6]. Therefore, it can be written:

$$d_{ref} = d + \underline{d_{TA}^{'}} \tag{18}$$

Distance $d_{ref-nei}$ between $\text{eNodeB}_{\text{ref}}$ and $\text{eNodeB}_{\text{nei}}$ is given in Table III, among other simulation parameters.

A. Base station model

For the purposes of the present simulations, physical layer of the implemented base station model consists of blocks creating PRS and PSS signals. Extract of transmitted signal resource grid is presented on Fig. 3. In the first subframe, PSS signal is transmitted in order obtain coarse synchronization between transmitter and receiver. In the second one PRS signals are placed.

B. Channel model

Channel model was realized as a Finite Impulse Response (FIR) filter, which scheme is shown on Fig. 4. Signal $s(t)$ from the base station antenna passes through the set of delay units and multipliers. Individual taps $\{\tau_i, a_i, b_i(t)\}$ represent consecutive paths of receiving signal, where τ_i is delay of *i-th* path, a_i is attenuation of *i-th* path and $b_i(t)$ is Rayleigh coefficient modeling motion of the receiver relative to the transmitter for *i-th* path. To the sum of signals from all taps complex Additive White Gaussian Noise (AWGN) is added forming signal $r(t)$ as an input of the receiver.

Values of the delays $\{\tau_o, \tau_1, \ldots, \tau_{L-1}\}$ are generated through random process defined as [12]:

$$\tau_i = -\gamma \cdot \log(1 - \underline{P_\tau}) \tag{19}$$

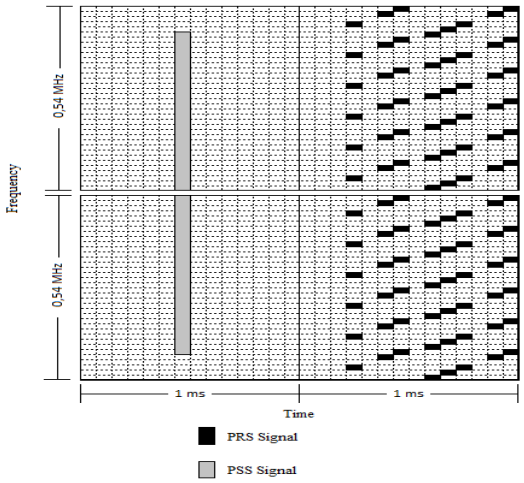


Figure 3. Extract of the resource grid of transmitted signal

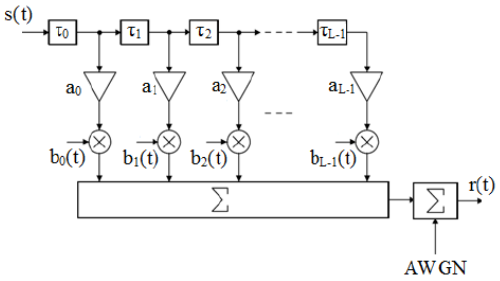


Figure 4. Channel model as a FIR filter

where $\underline{P_\tau}$ is a random variable uniformly distributed within $\langle 0,1 \rangle$ and $\gamma = 0.83$. Process (19) is normalized to the value of rms delay spread τ_{rms}, which limits values of random delays. The value of rms delay spread τ_{rms} is computed in accordance with [13][14]:

$$\tau_{rms} = T_1 d^{\varepsilon} y \tag{20}$$

where T_1 is a median value of τ_{rms} at $d = 1km$, ε is an exponent depended on environment, $\underline{Y} = 10\log(y)$ is a Gaussian random variable having zero mean and standard deviation σ_y. Values of parameters T_1, σ_y and ε can be found in Table II.

Overall attenuation of the channel g is a sum of attenuations of individual paths and is defined as [13]:

$$g = \frac{G_1}{d^{\alpha}} x = a_0 + a_1 + \ldots + a_{L-1} \tag{21}$$

where d is the distance between receiving and transmitting antenna, G_1 is the median value of g at distance $d = 1km$ determined by Hata model [15], α is an exponent which lies between 3 - 4 dB, $\underline{X} = 10\log(x)$ is a Gaussian random variable having zero mean and standard deviation σ_x of value 6 - 12 dB. Basing on (21) and function:

$$f(\tau_i) = 8.5 \cdot 10^{-1} + 1.3 \cdot 10^{-3} \tau_i - 2.1 \cdot 10^{-6} \tau_i^2 + 10^{-9} \tau_i^3 - 2.1 \cdot 10^{-13} \tau_i^4 + 1.5 \cdot 10^{-17} \tau_i^5 \tag{22}$$

which is polynomial interpolation of 3GPP Extended Typical Urban (ETU) delay profile model [16] for LTE system, values of individual paths attenuations $\{a_1, a_2, \ldots, a_{L-1}\}$ could be computed.

TABLE II. PARAMETERS VALUES OF FUNCTION (20)

T_1	0,4μs (urban microcells) 0,4–1.0μs (urban macrocells) 0,3μs (suburban areas) 0,1μs (rural areas)
σ_y	2 – 6 dB
ε	0,5 (urban, suburban, rural areas) 1,0 (mountainous areas)

 ISBN: 978-1-61208-353-7

Relationship (22) defines relative powers of individual, accordingly delayed paths of received signal. Plot of the function (22) is presented on Fig. 4. The dots on Fig. 4 are the interpolation nodes.

Rayleigh coefficients $\{b_0(t), b_1(t), \ldots, b_{L-1}(t)\}$ are generated with the use of sum-of-sinusoids method [17]:

$$b_i(t) = b_r(t) + jb_u(t) \tag{23}$$

$$b_r(t) = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} \cos[\omega_d t \cos(\alpha_n) + \varphi_n] \tag{24}$$

$$b_u(t) = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} \sin[\omega_d t \cos(\alpha_n) + \varphi_n] \tag{25}$$

$$\alpha_n = \frac{2\pi n + \Theta_n}{N} \tag{26}$$

where Θ_n and φ_n are independent random variables uniformly distributed within $<-\pi, \pi)$, N is a number of summations. For $N \geq 8$ model is highly convergent with desired channel characteristics [17].

C. User equipment model

Determining the position of PRS signals in time is preceded by cell synchronization process. It is assumed that UE and base stations are fully synchronized in frequency domain. The first action undertaken by the UE is to search for PSS signal in samples of the received signal. Finding of PSS signal allows to determine the position of time-frequency signal structure and ensure coarse synchronization. Cell synchronization is performed by maximum likelihood estimator which finds PSS position m_0 in time [18]:

$$m_0 = \arg\max \left\{ \left| \sum_{i=0}^{M-1} r[i+m] \cdot s_{PSS}^{*}[i] \right|^2 \right\} \tag{27}$$

where i is a time index, m refers to delays of correlation function, M is a PSS signal length in time domain, $r[i]$ is received signal, $s_{PSS}[i]$ is a PSS signal.

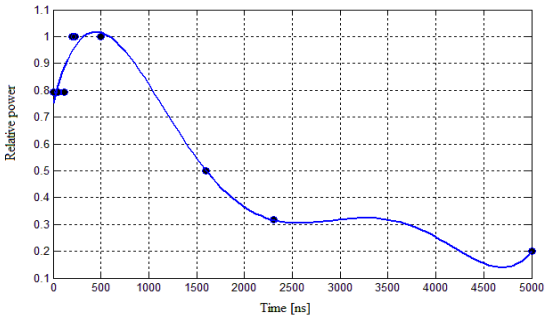


Figure 4. Plot of the function (22)

After successful cell synchronization process, receiver removes Cyclic Prefix (CP) from the received symbols minimizing Inter-Symbol Interferences (ISI) impact. Through consecutive Discrete Fourier Transform (DFT) and Inverse Discrete Fourier Transform (IDFT) operations and appropriate signal processing, all unnecessary Orthogonal Frequency Division Multiplexing (OFDM) symbols are removed, keeping only PRS symbols. Received PRS symbols are correlated with the known pattern in the receiver and time position of PRS signal is determined on the basis of a given algorithm.

IV. Simulation Results

Simulations were carried out in compliance with the conditions specified by 3GPP for RSTD measurements accuracy requirements [8]. The main parameters of the simulation is presented in Table III, where $(\text{PRS Es/Iot})_{\text{ref}}$ and $(\text{PRS Es/Iot})_{\text{nei}}$ are ratios of received energy per Resource Element (RE) during the useful part of the symbol to received power spectral density of the total noise and interference for a certain RE, respectively for PRS signals transmitted from reference and neighbour base station. During the simulations, mean RSTD measurement error and number of trials that satisfy requirements given by 3GPP was evaluated. Mean RSTD measurement error is given by

$$\overline{RSTD}^{err} = \frac{1}{N_{err}} \sum_{i} \left| RSTD_i^{err} \right| \tag{28}$$

where $RSTD_i^{err}$ is an RSTD measurement error on *i-th* trial and $N_{err} = 5000$ is the number of trials. Comparison of the results obtained with MLE and FPE algorithms is shown in Table IV. $\text{LOC}_{\%}$ is a percentage of trials number which falls into the requirements of RSTD measurements accuracy imposed by 3GPP [8]. For 5 and 10 MHz bandwidth and 1 ms duration of PRS signal the required accuracy is $\pm 5 \cdot T_s$ and $\pm 6 \cdot T_s$ respectively. The use of the FPE estimator significantly reduces mean RSTD error compared to the MLE estimator.

TABLE III. Simulation Parameters

PRS bandwidth	5 MHz, 10 MHz
PRS duration	1 ms
$(\text{PRS Es/Iot})_{\text{ref}}$	≥ -6 dB
$(\text{PRS Es/Iot})_{\text{nei}}$	≥ -13 dB
$d_{ref\text{-}nei}$	3 km
Expected RSTD uncertainty	< 5 µs
Max. Doppler frequency shift	50 Hz
Receiver sampling frequency	50 MHz
Environment	Urban macrocells
α	3 dB
σ_x	8 dB
T_1	1.0µs
σ_y	4 dB
ε	0.5
N_{err}	5000

TABLE IV. SIMULATION RESULTS

	$\overline{RSTD}^{err}$ [T_s]		LOC$_{\%}$ [%]	
	5 MHz	10 MHz	5 MHz	10 MHz
MLE	9.01	8.85	42	44.5
FPE	5.2	3.41	76.4	80.5

This improvement may be even higher at longer distances between base stations and mobile terminal due to larger rms time delay spread of receiving signals. It should be noticed that gain from using FPE estimator is obtained only when it is possible to separate more than one path of the received signal in the correlation function of received signal and pattern stored in the receiver. This could be done only if bandwidth of PRS signal is sufficiently large relating to the delays of consecutive signal paths arriving to the receiver. The larger the PRS signal bandwidth is, the narrower autocorrelation function of PRS signal becomes and the greater possibilities in extracting distinct received signal paths from correlation function. Further assessments showed that for 1.4 MHz bandwidth of PRS signals, it is not possible to distinguish any two paths of receiving PRS signals in the correlation function due to small mutual delay between consecutive signal paths. It means that for 1.4 MHz PRS bandwidth, use of FPE algorithm does not improve accuracy of RSTD measurements. Therefore, for such small bandwidth it is justified to use MLE algorithm due to its smaller processing consumption.

V. CONCLUSIONS

A new First-Path Estimator FPE was proposed for detecting the first path in receiving PRS signal. Simulations have shown that it significantly reduces RSTD measurement error in urban environment, comparing to the well-known Maximum Likelihood Estimator. Moreover, it ensures that about 80% of RSTD measurement falls into accuracy range defined by 3GPP for 5 and 10 MHz PRS bandwidth, becoming a reliable tool for positioning purposes.

REFERENCES

[1] 3GPP official website, Available from: http://www.3gpp.org/about-3gpp, retrieved: June 2014.

[2] R. Katulski, J. Magiera, J. Stefański, and A. Studańska, „Research study on reception of GNSS Signals in presence of intentional intereference", 2011 34th Internetional Conference on Telecommunications and Signal Processing (TSP), Aug. 2011, pp. 452-456, doi:10.1109/TSP.2011.6043691.

[3] A. Angrisano, S. Gaglione, and C. Gioia, „RAIM algorithms for aided GNSS in urban scenario", Ubiquitous Positioning, Indoor Navigation, and Location Based Service (UPINLBS), Oct. 2012, pp. 1-9, doi:10.1109/UPINLBS.2012.6409786.

[4] 3GPP TS 36.355, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); LTE Positioning Protocol (LPP), ver. 9.2.1, 2010.

[5] 3GPP TS 36.305, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Stage 2 functional specification of User Equipment (UE) positioning in E-UTRAN, ver. 11.2.0, 2013.

[6] 3GPP TS 36.211, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation, ver. 11.0.0, 2012.

[7] J. del Peral-Rosado, J. López-Salcedo, G. Seco-Granados, F. Zanier, and M. Crisci, „Achievable localization accuracy of the positioning reference signal of 3GPP LTE", 2012 International Conference on Localization and GNSS 2, June 2012, pp. 1-6, doi:10.1109/ICL-GNSS.2012.6253127.

[8] 3GPP TS 36.133, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Requirements for support of radio resource mamgement, ver. 11.6.0, 2013.

[9] J. Medbo, I. Siomina, A. Kangas, and J. Furuskog, „Propagation channel impact on LTE positioning accuracy: A study based on real measurements of observed time difference of arrival", 2009 IEEE 20th Internetional Symposium on Personal, Indoor and Mobile Radio Communications, Sept. 2009, pp. 2213-2217, doi:10.1109/PIRMC.2009.5450144.

[10] 3GPP TS 36.214, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements, ver. 11.1.0, 2013.

[11] I. Siomina and Y. Zhang, Method and arragement of determining timing uncertainty, Patent Application Publication, PCT/CN2010/000217, 2011.

[12] H. Asplund, A. Glazunov, and J.-E. Berg, „An investigation of measured and simulated wideband channels with applications to 1.25 MHz and 5 MHz CDMA systems", IEEE Vehicular Technology Conference, vol. 1, May 1998, pp. 562-566, doi:10.1109/VETEC.1998.686637.

[13] L. Greenstein, V. Erceg, Y. S. Yeh, and M. Clark, „A new path-gain/delay-spread propagation model for digital cellular channels", IEEE Trans. on Vehicular Technology, vol. 46, May 1997, pp. 477-485, doi:10.1109/25.580786.

[14] J. Stefanski, „Method of location of a mobile station in the WCDMA system without knowledge of relative time differences", IEEE 65th Vehicular Technology Conference, April 2007, pp. 674-678, doi:10.1109/VETECS.2007.149.

[15] M. Hata, „Empirical formula for propagation loss in land mobile radio services", IEEE Trans. on Vehicular Technology, vol. 29, Aug. 1980, pp. 317-325, doi:10.1109/T-VT.1980.23859.

[16] 3GPP TS 36.521-1, LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) conformance specification; Radio transmission and reception; Part 1: Conformance testing, ver. 8.2.1, 2009.

[17] C. Patel, G. Stüber, and T. Pratt, „Comparative analysis of statistical models for the simulation of Rayleigh faded cellular channels", IEEE Trans. On Communications, vol. 53, June 2005, pp. 1017-1026, doi:10.1109/TCOMM.2005.849735.

[18] F. Shuh, Synchronization and cell search, LTE Seminar: Future of mobile terminals, 2010, Available from: http://www.lmk.lnt.de/fileadmin/Lehre/Seminar09/Vortraege/Vortrag_Schuh.pdf, retrieved: June 2014.

 ISBN: 978-1-61208-353-7

Wearable Sensor System Prototype for SIDS Prevention

Gustavo López
Research Center on Information and Communication Technologies
University of Costa Rica
San José, Costa Rica
gustavo.lopez_h@ucr.ac.cr

Mariana López, Luis A. Guerrero
School of Computer Science and Informatics
University of Costa Rica
San José, Costa Rica
{mariana.lopez, luis.guerrero}@ecci.ucr.ac.cr

***Abstract*— Sudden Infant Death Syndrome (SIDS) causes unexpected death of infants; a variety of risk factors for SIDS have been detected through the years. A significant number of deaths occur when the children are being cared by non parental caregivers. We found that, using wearable systems some of those risk factors can be constantly monitored and the gathered information can be sent to the parents through a mobile application. In this paper, we present and evaluate a prototype of an augmented object and a mobile application that could help the prevention of SIDS. A wizard of Oz validation helped us determine the feasibility of developing and implementing the prototype. Testing results of the prototype showed positive parent reception to wide range monitoring and acceptable performance of the application.**

Keywords- Wearable system; health monitoring; HCI; augmented objects

I. INTRODUCTION

Sudden Infant Death Syndrome (SIDS) is an unknown phenomenon that causes the death of infant from birth up to the first year of age [1]. The American Academy of Pediatrics issued in 1992 a series of recommendations to prevent SIDS. Although the mortality rate has decreased, there are still SIDS cases reported worldwide [2][3][4].

According to the Institute for Clinical Systems Improvement, stomach or side sleeping are major risks for SIDS [5]. Therefore, they advise that children should sleep on their backs. Side sleeping was an alternative position, however this position is no longer recommended.

Besides SIDS, there have been reports that show a rise in the number of children deaths by accidentally suffocation or strangulation in bed [6].

Tintinalli et al. [7] stated that SIDS main risks can be categorized into two types of factors extrinsic and intrinsic. On one hand extrinsic factors include: prone or side sleeping, bedclothes overhead, sleeping on sofa or soft furniture, high ambient temperature, soft bedding, bed sharing, postnatal smoke exposure and prenatal smoke, alcohol or drug exposure. On the other hand, intrinsic factors include, but are not limited to: prematurity, family history of SIDS, gender, race and poverty.

In the United States, approximately 20% of SIDS deaths occur while the infants are in the care of a non-parental caregiver [1]. This may be due to a change on the sleeping position.

Nannies or other caregivers could place the baby in the stomach sleeping position and the babies are not accustomed to being placed in that position [7]. To address this issue, we develop an application that allows wide range remote monitoring using Wi-Fi. Common approaches of baby monitoring mainly use short range communication.

The main recommendations of the American Academy of Pediatrics to prevent SIDS are [1][3][4]:

- Supine sleeping position
- Firm sleeping surface
- No loose objects on the crib
- Sleep close but separate of the baby
- Avoid overheating
- Do not use home monitors as a strategy to reduce the risk of SIDS.

Even though authors explicitly say that parents should not focus on a single risk factor [1][3], technology nowadays allows us to verify the position of the baby and the environment's temperature very easily using wearable sensors.

Some wearable systems are based on augmented objects. Augmented Objects are basically everyday objects provided with additional characteristics through hardware and software in order to allow users to interact with computation systems through those objects [8].

Our contribution in this paper is to provide a description of the state of the art in wearable systems for child security on SIDS. We studied academic and industry efforts to create devices to help prevent SIDS. Also, we took into account the reported recommendations of experts in medicine and pediatrics.

We present the description and evaluation of an augmented object that is able to extract the baby´s information. The information is send to a computational device that analyzes it and sends it to the parent's phone through a mobile application. We used Wi-Fi as our wireless technology to enable constant wide range monitoring.

Also, we provide some data on the time frames required to provide assured information to parents when monitoring their babies. Performance testing was conducted to assert the possibility of managing multiple monitoring and notification devices concurrently. We present the benefits of product validation techniques used in the development process of ubiquitous computing applications.

 ISBN: 978-1-61208-353-7

In the following, Section II of this paper shows related work to this research, some different approaches in SIDS prevention. Then, Section III describes the design and implementation of the augmented object, the mobile application and all the configurations required. Section IV presents the results of the evaluation carried with the object. Finally, Section V presents conclusions and further work.

II. RELATED WORK

Over the past decades there has been an extensive research on bio-monitoring techniques. In 2003, Budinger [9] stated that the first success on remote monitoring systems for babies was the use of wireless breathing monitor using radio signal. Author also presented additional research however results produced up to 50% false positive signals.

In 2006, Linti, Horter, Österreicher and Planck [10] developed a sensory baby vest used to continuously monitor respiration, temperature and humidity. Authors used a personal computer to run the monitoring software which also alerts when parameters exceed the threshold established. The main drawback of this research was the lack of remote communication in order to allow monitoring of gathered data.

In 2007, Cao and Hsu [11][12] presented a non-invasive and remote infant monitoring system using CO2 sensors that uses exhaled air from an infant to reduce the potential risks for SIDS. The proposed system uses sensors placed on the crib´s edge. The main advantage of Cao and Hsu proposal is that it incorporates wireless communication capabilities. Other large advantage is that the system is non invasive. However, several sensors are required to approach the problem and even though authors mention the possibility of risky sleeping positions they do not addresses the recommendations given by the American Academy of Pediatrics (AAP).

Rimet et al. [13] presented a surveillance system for infants using a specially designed infant shoe to carry on a pulse oximetry in order to determine the CO2 saturation. The designed shoe also has an integrated 3-axes accelerometer.

Other examples of systems used to monitor children are Sensory Baby Vest, baby suit, baby glove and the already mentioned BBA bootee all these systems are wearable. There have been of course, efforts in the development of non-wearable systems such as augmented cribs, instrumented toys and others. However, toys or other loose objects in the crib go against the recommendations to help prevent SIDS [14].

Another interesting work was presented by Ziganshin, Numerov and Vygolov [15]. The authors present an ultra-wide band baby monitoring system that unlike the common sound or video baby monitor, constantly monitors the babies. This system is a peer to peer system with a sensing unit and a parent unit.

Most of the mentioned prototypes were developed on academic environments and laboratories and never reached the industry or were massively sold.

Besides academic prototypes, there are some instruments offered in the industry, such as Snuza Breathing Monitor[16], which is a small device that can be clipped to a baby´s diaper and vibrates or sounds when it does not detect breathing. It also has visual indicators.

Babysense V Hisense is a monitor that detects the baby´s movement using pads that are placed under the mattress. It incorporates communication to mobile phones however the information that it sends is a breathing graphic [17].

TiltMon Baby Sleeping Posture Monitor [18] is another peer to peer sensing monitor that alerts parents when the baby´s tilt is dangerous. WeMo [19] by Belkin is a monitor that helps to analize sleeping and crying patterns and notifies parents through a mobile application; however, it notifies only through audio as a common monitoring system.

Also, some non-technological efforts have been made including: Baby Sleeps Safe that is a two piece sleep system that replaces loose bedding and prevents the baby from turning around [20].

The last product that we would like to address is Sensible Baby. This product was first globally seen at the International Consumer Electronics and Consumer Technology tradeshow, CES 2014. Sensible Baby monitors position, temperature and movement or breathing sending alerts when a risk is detected [21].

Some issues with the actual approaches in industry are:

- They use loose objects to monitor the baby´s status.
- Communication is peer to peer: one monitoring device and one notification device.
- Communications are short range: communication protocol normally allows only a few meters between the monitor and the notification devices.

Our prototype addresses all the presented issues. By being an augmented object rather than a loose object it addresses the recommendations of the American Academy of Pediatrics. The used wireless technology (Wi-Fi) and architecture allow multiple notification devices to monitor one sensing device. Moreover, Wi-Fi allows large range communication and monitoring as well as short range. This feature provides more functionality to our proposal and addresses the fact that 20% of SIDS deaths occur while the infants are in the care of a non-parental caregiver.

III. PROTOTYPE DEVELOPMENT

This section presents the design and development of an augmented object that senses motion, position and temperature to prevent SIDS.

The augmented object presented is complemented with a mobile-Web application that displays the information gathered by the augmented object.

The augmented object prototype was developed following an Augmented Object Development Process (AODeP) [22][23]. This method ensures that all the stages in the development are focused on the problem to solve and from an engineering perspective.

AODeP proposes six main stages: (1) problem definition; (2) context of use definition; (3) requirement definition; (4) selection of the object; (5) development; and (6) testing with users. We will explain the development through these six steps.

Through studying the reality that parents live and the commercial systems we observed that, the problem we must

 ISBN: 978-1-61208-353-7

address is the continuous monitoring of babies, providing nonstop feedback to the parents when possible or at least locally in order to prevent some of the risk factors of SIDS.

The system usage context actually depends on the user. We addressed the context of a particular baby and at least one person monitoring the baby´s condition.

The requirements of the system -defined by the authors by studying the users- are the following:

- The system must be able to detect the baby´s position and send the gathered data to the mobile application.
- As it is been confirmed by medical pediatric institutions the back sleeping position is the best one to prevent SIDS. So, the system must be able to detect any other position and trigger an alarm on all monitoring artifacts.
- The system must be able to detect ambient temperature and trigger alarms if thresholds are violated.
- The systems must be configurable.

The sensing device (augmented object) needs to be selected. Following the proposed methodology presented by Guerrero, Ochoa and Horta [23] a set of candidate objects were identified and we selected a clothing patch. Table 1 shows all the possible candidates for monitoring and notification. The notification message would be through a mobile application and the local notifications will be addressed through future analysis.

TABLE I. CANDITATE OBJECTS TO BE AUGMENTED

Objects		
Monitoring Object		***Notification Object***
Blanket	Tank top	Baby monitor
Crib	Clothes patch	Phone Application
Pijama	Belt	Alarm clock

A. System´s Architecture

Architectures have been presented by several authors through the years to approach wearable health monitoring systems [24]. Normally raw data is not processed locally at the smart device but at the "cloud" [24]. We decided to follow that approach.

The system is composed by: (1) the sensing device that would be attached to the baby´s clothes, (2) the notification system that displays the main information gathered on the parent's phone and (3) a central Information Management Unit that processes and manages the information and displays the information in a Web site. Figure 1 shows the main modules of the system.

The communication technology between the sensing patch and the information management unit is Wi-Fi using a local area network. We assume that the parent's phone has access to Internet, either using mobile network or wireless network.

The main application flow consists of the monitoring objects sampling the data gathered by the sensors, sending it to the information management unit that process and stores the information gathered, and publishes it on a Web site.

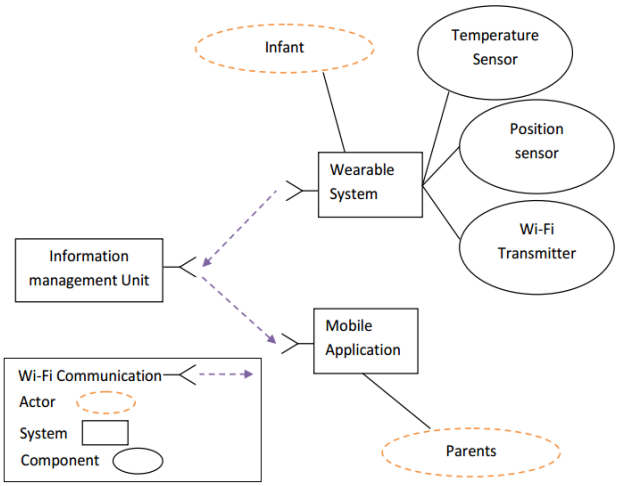


Figure 1. System´s Architecture.

The Web site uses Google's Open Id as the way to login. This information can be accessed through any web browser in case you do not have your phone but the main flow will use a phone so that if an alert is triggered the phone capabilities are used to alert the parent.

B. System´s Configuration

The application configuration is quite simple. It requires the user to configure the web site URL, when the information management unit is configured. This process is carried using a wizard approach on the server side. This wizard guides the process to set up the monitoring device and provides the URL that must be set on the parents or caregivers phone.

When the information management unit is configured an emergency phone number is required. This number is accessible through a panic button on the application interface. Usually this number would be the one of the caregiver or house where the child is located.

The next configuration is temperature. This is required in order to allow the correct functioning of the application in different latitudes and altitudes. The parent is required to set the "normal" range of temperature and the threshold acceptable depending on the variations.

Finally, the tilt acceptable for the application is configurable because some doctors indicate that the crib mattress should be tilted to about a 30° to 45° angle. If there is no tilting this parameter is not required but it is available to counteract the surface tilting.

C. Monitoring Device Design

We built the prototype of the sensing object using Phidgets sensors [25]. The sensors used in the system are the temperature sensor and a gyroscope with accelerometer.

We set all the parts inside a cloth patch (shaped as a bear) adhered to the child's clothes and that can be removed when needed. However, in order to remove it caregiver interaction is required.

 ISBN: 978-1-61208-353-7

Figure 2. Doll with attached sensors used for lab testing

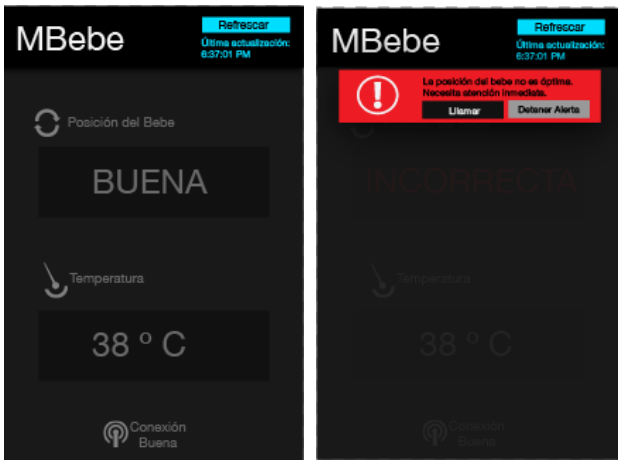


Figure 3. Mobile application screenshot

The sensors used were a temperature sensor and a 3-axis accelerometer. Figure 2 shows both sensors on a doll to demonstrate size. Using Phidgets to create the prototype did not allow us to construct a smaller sensing device but did allow us to test the product. In order to incorporate in the industrial setting more specific and smaller sensors must be used.

The incorporation of complex processing algorithms in wearable systems is very limited because of autonomy and capability of the components. Therefore we decided to use a centralized approach. The components only read and send the data and the Information Management Unit process it.

The Information Management Unit is a computer that process the information gathered by the sensors and publishes it on a web page that can be accessed through the internet.

Temperature sensor included in the prototype allows measurement rated at -30°C to +80°C with and error threshold of ± 2°C with a current consumption of 1 miliAmperes. The error threshold is taken in count when monitoring since ± 2°C are subtracted and added to the reported data and compared with the configured parameters.

The 3-axis accelerometer allows a measurement of acceleration up to ± 8 g. The gyroscope information allows a speed of ± 2000°/s (degrees per second). 0.07°/s is the minimum required speed that can be detected. The compass allows a 5.5 Gauss as the maximum magnetic field measured.

The maximum speed of sampling is 4 s/s (samples per second). However, to avoid server saturation we use one sample per second on our prototype. Wireless communication depends on the routers and server speed.

D. Mobile Web Application Design

The mobile Web application was built using a WebView to embed the web page displaying the information into the phone, instead of using mobile web browser because browsers has already known vulnerabilities that attackers are targeting [24].

Other important aspects that lead us to develop the system using a mobile-web application was the access that these applications get to the main components of the phone as the Ringtone Manager, Internet access, vibration, and the ability to bring to foreground the application when the alarm is triggered.

The mobile-web approach provides the parents with the possibility of accessing the system via Internet browser; of course, this approach limits the capabilities of the system.

Figure 3 shows screenshots of the application. The application was developed in Spanish. However, it allows multiple languages.

The left part of Figure 3 is the application on a normal state, which provides the position and temperature information and the connection status. It also allows a manual refresh of the information for a confidence boost, however the data is updated automatically every few seconds, this parameter can be configured.

When the alarm is triggered the image on the right is displayed. The alarm states: “The position of the baby is not optimal. Attention needed immediately”. This second screen has two buttons the one on the left is used to call the emergency number previously configured, the concept behind this emergency call number is that it will contact the person in charge of the baby at the moment; the one on the right “Detener Alerta” stops the alarm.

We considered using a Trusted Execution Environment for this application to increase the security of the application and the information.

The authentication on the application (actually the Web application) is done by Oauth2with Google credentials.

IV. PROTOTYPE EVALUATION

This section presents the evaluations performed to out prototype, starting with a concept validation through a Wizard of Oz approach [26]. Afterwards, we performed functional testing and performance testing. We measure the response time of the application on a load demand setting and on stressed settings with up to a 1000 connections.

 ISBN: 978-1-61208-353-7

A. Concept Validation

The first part of the evaluation process was a concept validation. We interviewed 5 users in their actual context who were related to care giving for children. During these interviews, some interrogations emerged: What is the safest position for a baby? Which exact tilt degree is dangerous? Is the phone app important for the notification or are notifications only necessary near the baby's location? Are users going to be looking at the monitoring app frequently enough? What is the correct size for the object given that it will be adhered to a baby? Where is the correct place to put the patch in the body? What would happen if the baby moves a lot?

To answer these questions we decided to conduct a Wizard of Oz validation. Wizard of Oz is a technique for validating new ideas and evaluating prototypes. It is a simulation in which participants are given the impression that they are interacting with an actual system. However, participants are actually interacting with humans, which pretend to be the system [26].

The Wizard of Oz validation consisted of 5 different people that were given the task of monitoring a "baby" through a mobile application. We had people in charge of trigger the alarm on the phones to simulate a threshold violation. From this validation we decided to reduce information on the application interface to improve readability and make it easier to use.

This was done to verify the ability to detect movements of the baby, to validate the connection with the notification system. Our Wizards were given the task to tilt the dolls position at random times and our participants were tasked with keeping their babies alive, calling to check if they noticed that something was wrong in the babies' environment.

Each session consisted of a timeframe where the caregiver was away from the baby and had to monitor the environment remotely with the mobile device, which as we previously stated is used as an interface for the information the augmented object is sending. We would validate after each session that the prototype for sensing was working correctly, and the response from the user was appropriate.

In the post-session interviews, we verified if the information appeared to be timely and enough, and validated the information displayed times by performance testing. To allow quick response the application was designed to update the data periodically although the page only requires 1.9 KB of data per update.

These validations helped us to determine the acceptance of wide range monitoring from the parents, and that they are able to react to an alert from our prototype.

B. Functional Testing

In parallel to our study about our user's confidence and information needs, we also had to test some physical characteristics about our prototype. We developed a clothpatch. We tested the prototype on a real baby, who was monitored at all times. The purpose of this experiment was to determinate the best position for the augmented object and to evaluate if it was comfortable for the baby, despite is prototype limitations. We conclude that the baby is not able to take the object off or turn it off, and that they are able to sleep comfortably without the object interfering.

We tested the sensing device on few people in order to make sure that the position was comfortable when sleeping in the correct position and that the babies could not take off the device. However, at our university and country there is a real challenge getting official permission to run clinical tests. Therefore, we conducted the tests with people related to the project and volunteers.

C. Performance Testing

We conducted a series of test using both online and offline tools to check our application's performance. The first test was carried using a website performance monitoring tool: CA APM Cloud Monitor. Our information management unit was a Quad-Core AMD Opteron™ Processor 2382 at 2.59GHz with 1.00 GB of installed RAM and Windows Server 2008 R2 Enterprise 64-bit.

We designed the mobile-web application so that it would perform as fast as possible. Table 2 and Figure 4 show the results of an evaluation using CA APM Cloud Monitor [27]. The main results show that the average result is 180ms on the download time. An interesting assessment that can be viewed in Figure 4 is that all the countries in Europe have a difference of 36 ms and Asia has the biggest download time. These results were expected due to the communication latency.

All this data shows a normal behavior considering the geolocation. It is obvious that, if the system was to be launched at an industry level many servers would be placed all over the world in order to provide reliability and stability therefore reducing the times to probably a similar rate of the one showed by U.S.A of 92 ms.

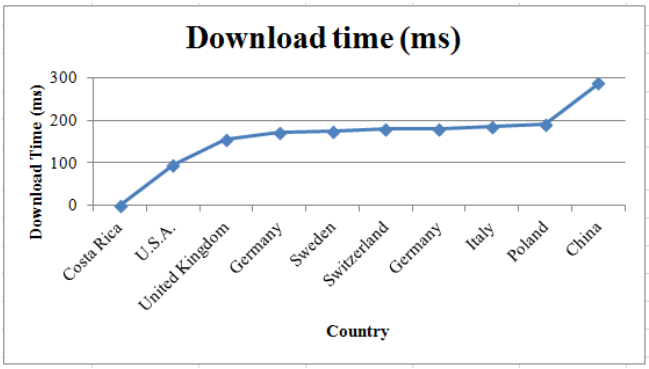


Figure 4. Graphical represention of data gathered with CA APM Cloud Monitor

The main defects found, that if fixed would help speed the process for download is:

- Some issues with JavaScript could be fixed by eliminating render-blocking and CSS.
- It could be also helpful to compact the JavaScript code.
- Sizing need to be improved depending on the displaying device

 ISBN: 978-1-61208-353-7

TABLE II. CA APM CLOUD MONITOR RESULTS

Test Perform from	Measured Parameters		
	Connect Time (ms)	*Download time (ms)*	*Download Size (Bytes)*
Costa Rica	0.0032	0.0036	2059
U.S.A.	92	96	2065
United Kingdom	152	155	2066
Germany	169	172	2068
Sweden	170	173	2065
Switzerland	176	180	2069
Germany	181	180	2068
Italy	183	185	2071
Poland	188	191	2076
China	283	287	2075

Figure 4 shows that the farther the parent is from the baby the higher the time to get feedback, however this time could be reduced by improving communications depending on the location of usage.

We propose a content delivery network approach to achieve high performance on our system, although this could increase the costs of operation. However, this approach could not be tested.

We test the systems performance through stress testing using two different configurations. The first configuration was: Users Count (40-600), user increase rate (10), network configurations (LAN and 3G). Table 3 shows the results of this configuration.

TABLE III. RESULTS OF PERFORMANCE TEST USERS 40-600

Counter	*Min*	*Max*	*Average*
User Load	40	600	320
Pages/Sec	14	291	153
Response time	0.019	0.30	0.048

The second configuration was: users count (1-1000), user increase rate (100), network configurations (LAN and 3G). Table 4 shows the results of this configuration.

TABLE IV. RESULTS OF PERFORMANCE TEST USERS 40-600

Counter	*Min*	*Max*	*Average*
User Load	1	1000	634
Pages/Sec	1	424	271
Response time	0.0046	1.1	0.36

The main results of the performance evaluation are that average response time is lower than 0.36 seconds either with 1 or 1000 users accessing the server. The inputs ranged from one to 357 sensing devices.

V. CONCLUSION AND FUTURE WORK

SIDS is a cause for concern amongst parents of infants, however there are risk indicators that given the proper technology could notify parents in time that their baby's environment is not optimal.

In this paper, we have presented the development of an augmented object prototype using a combination of the AODeP and HCI methodologies that can successfully notify caregivers of changes in the baby's environment.

The prototype here presented focuses mainly on sensor-based interfaces due to the requirements of the case study and is successful in monitoring the environment. It is important to note that the development of this augmented object required sensitivity due to the fact that this object will reside in the baby's clothes and that the study conducted to test the viability of this object being present when a baby was sleeping was extremely important.

Future work for this prototype could include incorporating the feedback received during the Wizard of Oz study to improve the accuracy and confidence level of the device in the eyes of the baby's primary caregivers.

We also need to perform more tests in order to evaluate reliability of the data.

One of the main improvements identified would be to select another object to augment in the local environment to address short range communications. This object would be used to notify the care-giver in the house, having a redundant system of notifications with hopes that either the parents outside of the house or the care-giver in the house will take immediate action upon the alert.

ACKNOWLEDGMENT

This work was supported by CITIC (*Centro de Investigaciones en Tecnologías de la Información y Comunicación*) at Universidad de Costa Rica, grand No. 834-B2-228 and 834-B4-159 and by the School of Computer Science and Informatics at Universidad de Costa Rica.

REFERENCES

[1] American academy of pediatrics, Task Force on Sudden Infant Death Syndrome, "The Changing Concept of Sudden Infant Death Syndrome: Diagnostic Coding Shifts, Controversies Regarding the Sleeping Environment, and New Variables to Consider in Reducing Risk", Pediatrics, Nov. 2005, pp. 1245-1255, doi: 10.1542/peds.2005-1499.

[2] American Academy of Pediatrics, "SIDS and Other SleepRelated Infant Deaths: Expansion of Recommendations for a Safe Infant Sleeping Environment", Pediatrics, Nov. 2011, pp. 1030-1039, doi: 10.1542/peds.2011-2284.

[3] American SIDS Institute, http://www.sids.org, [retrieved: June, 2014].

[4] National SIDS/Infant Death Resource Center, http://www.sidscenter.org, [retrieved: June, 2014].

[5] Institute for Clinical Systems Improvement, https://www.icsi.org/, [retrieved: June, 2014].

[6] J. Wilkinson, C. Bass, S. Diem, A. Gravley, L. Harvey, R. Hayes, K. Johnson, M. Maciosek, K. McKeon, L. Milteer, J. Morgan, P. Rothe, L. Snellman, L. Solberg, C. Storlie, and P. Vincent, "Preventive Services for Children and Adolescents", Ninteenth Edition, Institute for Clinical Systems Improvement, 2013.

 ISBN: 978-1-61208-353-7

[7] J. Tintinalli, J. Stapczynski, O. John Ma, D. Cline, R.. Cydulka, and G. Meckler, "Tintinalli's Emergency Medicine: A Comprehensive Study Guide", Seventh Edition, McGrawHill Medical, 2011.

[8] H. Ishii and B. Ullmer, "Tangible bits: towards seamless interfaces between people, bits and atoms", ACM Conference on Human factors in computing systems, April. 1997, pp. 234-241, doi: 10.1145/258549.258715.

[9] T. F. Budinger, "Biomonitoring with wireless Communications", Lawrence Berkeley National Laboratory, 2003.

[10] C. Linti, H. Horter, P. Österreicher, and H. Planck, "Sensory baby vest for the monitoring of infants", International Workshop on Wearable and Implantable Body Sensor Networks, Jan. 2007, pp. 135-137, doi: 10.1109/BSN.2006.49

[11] H. Cao, L. C. Hsu, T. Ativanichayaphong, J. Sin, and J. C.Chiao, "A non-invasive and remote infant monitoring system using CO2 sensors", IEEE Sensors, Oct. 2007, pp. 989-992, doi: 10.1109/ICSENS.2007.4388570.

[12] H. Cao, L. C. Hsu, T. Ativanichayaphong, J. Sin, H. E. Stephanou, and J. C. Chiao, "An Infant Monitoring System Using CO2Sensors", IEEE International Conference on RFID, March. 2007, pp. 134 – 140, doi: 10.1109/RFID.2007.346161.

[13] Y. Rimet, Y. Brusquet, D. Ronayette, C. Dageville, M. Lubrano, E. Mallet, C. Rambaud, C. Terlaud, J. Silve, O. Lerda, L.I. Netchiporouk, and J. Weber, "Surveillance of infants at risk of apparent life threatening events (ALTE) with the BBA bootee: a wearable multiparameter monitor", IEEE International Conference on Engineering in Medicine & Biology Society, 2007, pp. 4997-5000, doi: 10.1109/IEMBS.2007.4353462.

[14] L. Zhang, L. Lao, K. Wu, Q. Liu, and X. Wu, Research in Development on Wireless Health Care of Infants, AsianPacific Conference on Medical and Biological, Springer, April. 2008, pp. 580-583, doi: 10.1007/978-3-540-79039-6_146.

[15] E. G. Ziganshin, M. A. Numerov, and S. A. Vygolov, "UWB Baby Monitor", IEEE Ultrawideband and Ultrashort Impulse Signals, Sept. 2010, pp. 159-161, doi: 10.1109/UWBUSIS.2010.5609156

[16] Snuza, mobile baby monitors, http://www.snuza.com/, [retrieved: June, 2014].

[17] Hisense, Health Monitoring Techniques, http://www.hisense.co.il, [retrieved: June, 2014].

[18] TiltMon Baby Sleeping Posture Monitor, http://www.digio2.com/, [retrieved: June, 2014].

[19] WeMo baby monitor, http://www.belkin.com/us/, [retrieved: April, 2014].

[20] Baby Sleeps Safe, Infant Sleep Safety System, [retrieved: June, 2014].

[21] Sensible Baby, http://mysensiblebaby.com, [retrieved: June, 2014].

[22] G. López, M. López, and L. A. Guerrero, "Improving the Process for Developing Augmented Objects: An HCI Perspective", Ubiquitous Computing and Ambient Intelligence. ContextAwareness and Context-Driven Interaction, Springer International Publishing, Dec. 2013, pp. 111-118, doi: 10.1007/978-3-319-03176-7_15.

[23] L. A. Guerrero, S. Ochoa, and H. Horta, "Developing Augmented Objects: A Process Perspective". Journal of Universal Computer Science, Jun. 2010, pp. 1612 – 1632, doi: 10.3217/jucs-016-12-1612.

[24] V. Custodio, F. J. Herrera, G. López, and J. I. Moreno, "A Review on Architectures and Communications Technologies for Wearable Health-Monitoring Systems", Sensors, Oct. 2012, pp. 13907 – 13946, doi:10.3390/s121013907.

[25] Phidgets, http://www.phidgets.com, [retrieved: June, 2014].

[26] J. F. Kelley, "An iterative design methodology for user-friendly natural language office information applications", ACM Transactions on Office Information Systems, Jan. 1984, pp. 26-41, doi: 10.1145/357417.357420.

[27] APM Cloud Monitoring, Website Performance Monitoring, http://cloudmonitor.ca.com/en/, [retrieved: June, 2014].

 ISBN: 978-1-61208-353-7

On Secure-Smart Mobility Scheme in Proxy Mobile IPv6 Networks

Jae-Young Choi
College of Information and Communication Engineering
Sungkyunkwan University
Suwon, Republic of Korea
E-mail: jaeychoi@skku.edu

Jun-Dong Cho and Jongpil Jeong
Department of Human ICT Convergence
Sungkyunkwan University
Suwon, Republic of Korea
E-mail: {jdcho, jpjeong}@skku.edu

***Abstract*—IPv6-based mobility management techniques for Proxy Mobile IPv6 (PMIPv6) system are proposed to improve the performance of a variety of Fast Handover of Proxy MIPv6 (F-PMIPv6). However, F-PMIPv6 cannot be better than PMIPv6 in all scenarios. Therefore, selecting a proper mobility management scheme between PMIPv6 and F-PMIPv6 is an interesting issue, for its potential in enhancing the capacity and scalability of a system. We developed an analytical model to analyze the applicability of PMIPv6 and F-PMIPv6. Based on this model, we designed a Secure Smart Mobility (SSM) scheme that selects the better alternative between PMIPv6 and F-PMIPv6 for user according to changing mobility and service characteristics. When F-PMIPv6 is adopted, SSM chooses the best mobility anchor point and regional size to optimize the system performance. Numerical results illustrate the impact of key parameters on the applicability of PMIPv6 and F-PMIPv6. SSM has been proven to show better results than both PMIPv6 and F-PMIPv6.**

Keywords-PMIPv6; F-PMIPv6; Secure Smart Mobility.

I. INTRODUCTION

Proxy Mobile IPv6 (PMIPv6) [1] lets Mobile Nodes (MNs) connect to the PMIPv6 domain with various interfaces at the same time, and supports inter-equipment handover. Even though PMIPv6 reduced the handover delay time compare to Mobile IPv6 (MIPv6) [2] and its extensions, it is inferior in regard to applications with requirements for real-time communications such as Voice of IP (VoIP). Moreover, the handover interrupt time of the vertical handover process is longer than that of a horizontal handover process, because a Duplicate Address Detection (DAD) process has to occur the new interface of the MN receives the packet. For these reasons, Fast Handover for PMIPv6 (F-PMIPv6) [3] and Fast Handover for Hierarchical MIPv6 (FH-PMIPv6) [4] is proposed to improve the handover performance of PMIPv6.

PMIPv6 allows for maintaining the existing connection even if MIPv6 is not applied. It is a network-based mobility management technique to manage node mobility. Also, PMIPv6 supports multiple interfaces. Handover needs to be considered when Mobile Access Gateways (MAGs) have different interfaces. PMIPv6 must also go through a Local Mobility Anchor (LMA). If a Corresponding Node (CN) and MN are in the same area, packets inefficiently have to go through the LMA. Due to the network-based nature of PMIPv6, interfaces of the terminals can be known. Thus, additional signaling processes and MAGs have to be proceeded. If an MN connects through a new interface, the PMIPv6 domain does not have the information about the intensions of the MN regarding whether its connection is for undergoing handovers or multi-roaming. In the PMIPv6 domain, packets converge to LMA and cause a bottleneck state. Additional LMAs with an inter-LMA redirection function can help with load balancing and network stabilizing. Therefore, the establishment of a protocol for inter-LMA redirection is needed.

Based on [5]-[9], our SSM scheme is proposed to resolve two of the existing problems. It chooses the better alternative between PMIPv6 and F-PMIPv6 according to its mobility transitions and service conditions. When F-PMIPv6 is chosen, SSM selects the optimized mobility anchor point and its regional size, as well as the better protocol by analyzing the applications of PMIPv6 and F-PMIPv6. This paper proposes a reference analysis model based on two functions. First, Internet architectures are modeled using an MIP network. The MIP network is based on the cellular architecture used in Personal Communication System (PCS). PCS is region-oriented, while the Internet is space-oriented. In a region-oriented network, the distance between two terminals is measured by their physical spaces. Therefore, the Internet architecture is suitable for an abstract MIP network. The proposed analysis model considers both registration and packet transmission capabilities. Previous research preferred statistics about delayed handovers according to a registration record of the mobility management. However, packet transmission capability is also an important statistic in delay-sensitive services like mobile networks. Thus, considering both registration and packet transmission capability is important in analysis research.

The rest of the paper is organized as follows. Section 2 of this paper covers the handover techniques of PMIPv6 and F-PMIPv6. Section 3 proposes SSM and explains the security certification processes. Section 4 presents the numerical values and results, and Section 5 summarizes and concludes the study.

 ISBN: 978-1-61208-353-7

II. RELATED WORK

A. Handover of Network-based Mobility Protocols

To resolve the problems of the existing host-based mobility protocols, PMIPv6 [1] is proposed. When mobile terminal equipment tries to perform a handover, PMIPv6 deals with the situation on the network, without concern for any of the IPv6 mobility protocol-related signals. In the current state of MIPv6, when mobile terminal equipment tries to do a handover, it has to register its location. But, in the case of PMIPv6, internet application services are available only with the IPv6 stacks.

F-PMIPv6 [3] was proposed to reduce the packet losses from the MAG-LMA handover delay. Like the existing FMIPv6, F-PMIPv6 sets tunnels between the Previous MAG (PMAG) and destination NMAG to reduce the packet loss before the mobile terminal equipment moves to a new network.

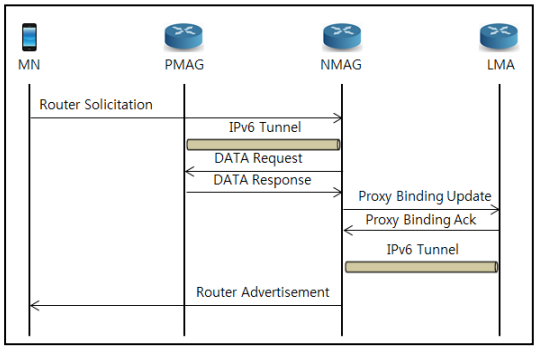


Figure 1. Handover of F-PMIPv6.

Fig. 1 shows the handover process of F-PMIPv6. Through tunneling, packets can be transmitted between PMAG and NMAG, even if the link is separated. However, this simultaneously overloads the network since the tunnel exchanges only the tunnel-related signals are exchanged between the MAGs, with no transmission of BU to LMA. Also, the transmission of BU from the MAG to the LMA to achieve Routing Optimization (RO) can cause order disturbance of packets.

B. Security Certification Process

Major operations of the security certification process are the initial registration process and the certification process. To join a regional mobile domain, an MN has to register at the AAA server and carry out a certification process to connect to the internet through the MAG [10].

III. SECURE-SMART MOBILITY (SSM) SCHEME

In this section, the Secure-Smart Mobility (SSM) method is proposed. SSM resolves the two following issues. It selects the better alternative between PMIPv6 and F-PMIPv6 for the users. And, when F-PMIPv6 is selected, SSM optimizes the LMA and system performance by choosing a proper regional size. Fig. 2 shows the structure of SSM and the operation of SSM. SSM is made up of four parts: LMA, MAG, MN, and a certifier AAA server. LMA and MAG are connected through a bi-directional tunnel using Proxy CoA. In SSM, a MAG does not have to be in the control of one LMA.

SSM is a protocol that supports mobility in a limited domain without the additional functional modification of MNs. The MN in a relevant domain can be distinguished from its MN-ID. If MN operates in the domain, connection certification is performed, and when it is completed, the MN receives a Home Network Prefix (HNP). The network supports mobility by maintaining the HNP of the MN statically, so that the MN can operate as if it is in the same Layer 3 wherever it is. LMA acts as a kind of home domain of the MN in the domain. It is usually located in the gateway location in the domain, and assigns HNP and sends it to the MN. LMA ensures the connection by maintaining the location and address information of every terminal of its range of management. Every packet sent from outside of the domain to inside is designed to be received by the LMA, and the packets are sent to the MN through the tunnel with the MAG. In contrast, packets sent from inside to outside of the domain are tunneled from MAG to LMA, and LMA sends those to the outside.

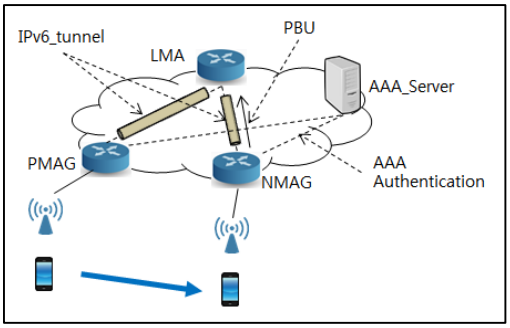


Figure 2. Operation of SSM.

MAG is the first hop that is directly connected to the MN, and instead of the MN, it undergoes mobility support signaling. Also, MAG performs the network connecting and routing functions of the MN. If the MN connects to the MAG, the MAG sets for the connection with LMA using the information of MN, and receives packets from the LMA for MN. Policy profiles include the address and its setup method of the LMA in charge, HNP information of the MN, the service policy, etc. MAG and LMA can be informed about the HMP information of the MN, and complete MN certification through these profiles.

SSM concerns every factor related to the network and MN. The most important factor so far is reflected on LMA, which affects the values of A, B, and δ. The discovery of a mobile LMA needs an LMA option of the router advertisement propagated into the MN from the LMA by a specific router interface. The LMA option includes a *preferences* field which is utilized in reflecting loads in LMA, and it ranges from 0 to 15. Another important network factor is the average distance between an LMA and its reachable MAG($l_{LMA \rightarrow MAG}$). $l_{LMA \rightarrow MAG}$ is made manually

 ISBN: 978-1-61208-353-7

in the LMA, and can be included in the extended LMA option offered to the MN. To obtain the number of hops to deliver the packets or signals, an MN or LMA can use the TTL field in the IP header. Then, the average number can be used when calculating C_T. The most important MN-related factors are the average time (T) that an MN stays in the MAG, and the average ratio of packet arrival (α). These parameters can be collected regularly by each MN using statistical analysis.

Suppose that the regional size of the MN's relocation is K. When K has an increasing value, F-PMIPv6 can obtain more average registration while the average packet transmission expense is increasing. However, K cannot be increased unlimitedly because of the bottleneck phenomenon of LMA. The overall average packet processing delay occurs due to $\alpha \cdot (A\omega K + BlgK + \delta)$, and can be differentiated by its load [9]. Thus, the optimized K which minimizes the overall cost C_T can be represented as K_{opt}, and it optimizes the overall performance of F-PMIPv6 compared to PMIPv6.

$$min\ C_T(K)$$
$$s.t.\ \alpha \cdot (A\omega K + BlgK + \delta) < \varphi \qquad K \in Z^+ \quad (1)$$

where ψ is a constant which limits the overall packet process delay of LMA.

$C_{FPIMPv6}$, the cost function of F-PMIPv6, represents the absolute performance of F-PMIPv6 in view of the average registration and packet transmission delay.

$$C_{FPMIPv6} = n_1 + \frac{(m-1)D_{intra}+D_{inter}}{mT} + n_2 \cdot T_{PH} \quad (2)$$

Due to the restriction of $\alpha \cdot (A\omega K + BlgK + \delta) < \psi$, C_T and $C_{FPIMPv6}$ are minimized which lets K_{opt} represent the absolute performance and the relative performance.

$$C_{FPMIPv6} = C_T + n1 \cdot D + n2 \cdot T_{PM} \quad (3)$$

Since it is independent of K, $C_{FPMIPv6}$ moves through the Y axis by C_T in the scale of $C_T + n_1 \cdot D + n_2 \cdot T_{PM}$. Consequently, the K value that minimizes C_T and $C_{FPMIPv6}$ is from the restriction $\alpha \cdot (A\omega K + BlgK + \delta) < \psi$. Therefore F-PMIPv6 can achieve both its absolute performance and its relative performance. Since (1) is too complex, it needs to be simplified. Since K is not limitless, suppose the maximum value is N, the same as the number of MAGs that MN can relocate. As K increases, the following will occur:

Case 1: If the differential function of C_T, C_T', shows the trend of $C_T'(N) > 0$ on the site of its first origination from C_T, then $C_T(K)$ increases instead of decreasing. In this case, $C_T(1)$ is the minimum.

Case 2: In the case of $C_T'(K)$, the increasing and decreasing domain of $C_T(K)$ is first altered from over zero to below zero. $C_T'(K)$ is the minimum when $K' = K{:}min(C_T(K=1), C_T(K=Kmax)$. This analysis simplifies the solution of K_{opt}, as shown below.

Clearly, K_{opt} can optimize the performance of F-PMIPv6. Yet, $C_T(K_{opt}) > 0$ indicates that the optimized performance of F-PMIPv6 is still poorer than that of PMIPv6. Therefore, PMIPv6 is the most adequate alternative if $C_T(K_{opt}) > 0$. However, F-PMIPv6 is more adequate when there are many MAGs in LMA. Since $C_{FPMIPv6}$ represents the absolute performance of F-PMIPv6 and K_{opt} minimizes both C_T and $C_{FPMIPv6}$, LMA should be designated as an optimized regional mobility management entity with the minimization of $C_T(K_{opt})$, and K_{opt} should be the optimized regional size.

IV. PERFORMANCE EVALUTION

A. Cost analysis of PMIPv6 and F-PMIPv6

This section compares the register performance and defines the average registration profit D_R as the average registration time using F-PMIPv6 instead of PMIPv6. If D_R has a positive value, the average registration delay of PMIPv6 is shorter than that of F-PMIPv6. Independent from the handover delay time, the MN does not consider the regular binding update, which refreshes and delivers its binding record when analyzing CN or LMA. The major symbols of the subdivided section are displayed in Table 1.

TABLE I. PARAMETERS FOR REGISTRATION ANALYSIS

Parameter	Description
D_{RM}	The Average Registration Delay of PMIPv6
D_{LMA1}	The Average Delay of Registration Signaling through LMA Before Handover
D_{LMA2}	The Average Delay of Registration Signaling through LMA After Handover
D_{MAG}	The Average Delay of Registration Signaling through MAG
D_{intra}	The Average Delay of Registration Process during intra-LMA Handover
D_{inter}	The Average Delay of Registration Process during inter-LMA Handover
$D_{LMA1\rightarrow MAG}$	The Average Delay of Registration Signaling from LMA to MAG
$D_{LMA2\rightarrow LMA1}$	The Average Delay of Registration Signaling from the new LMA to the previous LMA After Handover
$D_{MAG\rightarrow MN}$	The Average Delay of Registration Signaling from MAG to MN
$l_{LMA\rightarrow MAG}$	The Average Distance between LMA and MAG
$l_{MAG\rightarrow MN}$	The Average Distance between MN and MAG
μ	The Signaling Cost per Unit distance of Wired Link
$\frac{MinInt + MaxInt}{2}$	The Average Delay of RA (Router Advertisement) Transmission

While calculating D_R, it is hypothesized that the signal transmission delay of the uplink and downlink is the same for simplicity. The registration of PMIPv6 includes home registration. However, in F-PMIPv6, when the MN tries roaming to another region, the process includes regional registration as well as home registration.

As such, D_{RM}, D_{intra}, D_{inter} can be calculated with (4)-(6).

$$D_{RM} = \frac{MinInt+MaxInt}{2} + 2D_{MAG} + 2D_{LMA1\rightarrow MAG} \quad (4)$$

$$D_{intra} = \frac{MinInt+MaxInt}{2} + 2D_{MAG} + 2D_{LMA1} + 2D_{LMA1\rightarrow MAG} \quad (5)$$

$$D_{inter} = D_{intra} + D_{LMA1} + D_{LMA2} + 2D_{LMA2\rightarrow LMA1} \quad (6)$$

The number of handovers required in moving an MN, *m(m≥1)* means that the MN relocates to a new area on the m^{th} handover trial. Therefore, the overall average delays D_{FPT} and D_{PT} from the m^{th} handover of MN in F-PMIPv6 and PMIPv6 are given in (7)-(8).

$$D_{FPT} = (m-1)D_{intra} + D_{inter} \quad (7)$$

$$D_{PT} = mD_{RM} \quad (8)$$

D_R can be calculated as follows (9).

$$D_R = \frac{(D_{FPT} - D_{PT})}{mT} = \frac{((m-1)D_{intra} + D_{inter} - mD_{RM})}{mT} \quad (9)$$

Suppose that the average signal transmission delay of a wired link is proportional to the measured distance between the moving numbers of a hop. Let the cost of the unit distance signal transmission be μ, which includes the unit distance propagation delay and cuing delay of each hop. Since the wireless bandwidth is usually narrow, suppose the average signal transmission delay of a wireless link is $\theta \cdot \mu$ when $a > 1$. The average signal transmission delay is different between the core network and the access network. To simplify the analysis, μ reflects the average level of a signal in the core network and the access network. Therefore, D_R can be altered to (10).

$$D_R = \frac{\mu(2\theta + 2ml_{LMA\to MAG}) - 2\left(\left(\frac{MinIn+MaxInt}{2}\right)(m-1) + mD_M\right)}{mT} \quad (10)$$

Using (10), high registration profit from the distance between the MN and MAG and the distance between the LMA and MAG can be obtained from F-PMIPv6. When $D_R < 0$, higher average registration profit can be obtained from F-PMIPv6. According to (9)-(10), m has to satisfy the following (11) to achieve $D_R < 0$.

$$m > \frac{D_{inter} - D_{intra}}{D_{RM} - D_{intra}} = \frac{2\mu\left(\theta + \frac{MinInt+MaxInt}{2}\right) + D_L}{2\mu\left(l_{LMA\to MAG} - \frac{MinInt+MaxInt}{2}\right) + D_L - D_M} \quad (11)$$

The value of m has a close relation with the size of the region. Suppose each MN transfers to a random LMA, and the regional size is K.

When MN always roams in the assigned area, F-PMIPv6 has better performance than PMIPv6 in registration, and the average registration profit can be calculated as $|2\mu \cdot (l_{LMA\to MAG} - (MinInt + MaxInt)/2|$. $K \geqslant N$ when MN roams in its region. If F-PMIPv6 is selected, the number of handovers of intra-LMA and inter-LMA are m-1 and 0, respectively. However, if PMIPv6 is selected, the number of handovers would be m-1. Therefore, D_R can be calculated as below.

$$D_R = \frac{(m-1)D_{intra} - (m-1)D_{RM}}{m-1} = 2\mu\left(l_{LMA\to MAG} - \frac{MinInt+MaxInt}{2}\right) + D_M \quad (12)$$

Generally, $l_{LMA\to MAG} - (MinInt + MaxInt)/2 < 0$, and there are no big differences between MAG and MN in the average registration signal process delay time. In this case, D_R can be simplified as $2\mu \cdot (l_{LMA\to MAG} - (MinInt + MaxInt)/2$. Since $2\mu \cdot (l_{LMA\to MAG} - (MinInt + MaxInt)/2 < 0$, F-PMIPv6 has better performance than PMIPv6 in registration. Also, the average registration profit is related to $|2\mu \cdot (l_{LMA\to MAG} - (MinInt + MaxInt)|$.

When MN roams between indifferent areas, D_R is dependent on the regional area K, and their relations are arranged in (13). Also, F-PMIPv6 can achieve average registration profit only when (14) is satisfied.

$$D_R = \frac{(2\mu \cdot \theta \cdot D_H) \cdot (2N - 2K - 1) + 2\mu \cdot l_{LMA\to MAG} \cdot (1 - 2K)}{(2N-2) \cdot T} + \frac{4\mu \cdot (N-1) \cdot \frac{MinInt+MaxInt}{2} + 2D_M(N-1)(D_M - D_L)}{(2N-2)T} \quad (13)$$

$$\frac{2N-2}{2N-2K-1} > \frac{2\mu(\theta + l_{LMA\to MAG}) + D_L}{2\mu(l_{LMA\to MAG} - \frac{MinInt+MaxInt}{2}) + D_L - D_M} \quad (14)$$

When MN roams between indifferent areas, $K < N$. In this case, when condition I indicates an MN which enters LMA*(i = 1, 2, …, N)*, the movement of a roaming MN through different MAGs can be modeled by a Markov chain. As in Fig. 3, it is predicted that the MN can move in each direction (except for the boundary MAGs) with a probability of 1/2.

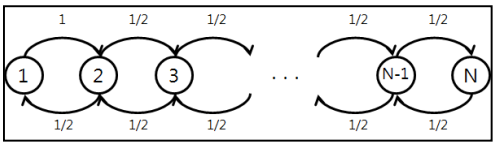


Figure 3. State transmission diagram.

The normal condition probability of condition I is defined as *π(I = 1, 2, …, N)*. By Fig. 4, the balance equation of the Markov chain is given as follows.

$$\begin{cases} \pi_1 = \pi_2 \times \frac{1}{2} \\ \pi_2 = \pi_1 + \pi_3 \times \frac{1}{2} \\ \pi_i = (\pi_{i-1} + \pi_{i+1} \times \frac{1}{2} & i = 3,4,\dots,N-2 \\ \pi_{N-1} = \pi_{N-2} \times \frac{1}{2} + \pi_N \end{cases} \quad (15)$$

(16) can be recreated repeatedly as follows.

$$\begin{cases} \pi_1 = 0.5 \times \pi_2 \\ \pi_2 = \pi_{i+1} & i = 3,4,\dots,N-2 \\ \pi_N = 0.5 \times \pi_{N-1} \end{cases} \quad (16)$$

By $\sum_{i=1}^{\infty} \pi = 1$, the normal condition probability is calculated as follows.

$$\begin{cases} \pi_1 = \pi_N = 1/(2 \times (N-1)) \\ \pi_i = \frac{1}{N-1} \end{cases} \quad (17)$$

This gives the probability of the MN's regional roaming.

$$P_{intra} = \sum_{i=1}^{K} \pi_i = \frac{2K-1}{2N-2} \qquad K < N \quad (18)$$

Therefore, the regional roaming probability is $P_{inter} = 1 - P_{intra}$. The regional relocation probability after the m^{th} handover (P_{out}^{in}) is expected as below.

$$P_{out}^{m} = {P_{intra}}^{m-1} \times P_{inter} = (\frac{2K-1}{2N-2})^{m-1} \times (1 - \frac{2K-1}{2N-2}) \quad (19)$$

$$E(m) = \sum_{m=1}^{\infty} m\, P_{out}^{m} = \frac{2N-2}{2N-2K-1} \quad (20)$$

With (20), (10) and (11) can be converted to (13) and (14), respectively. Clearly, D_R is dependent on the regional size by (13). A high K value indicates a high average registration profit. Also, the average registration profit of F-PMIPv6 can be obtained only in case if (14) is satisfied.

The packet transmission performance is compared and defined. The average packet transmission cost T_P is defined as the average consumed time to deliver packets from CN to MN through F-PMIPv6 instead of PMIPv6.

TABLE II. PARAMETERS FOR PACKET DELIVERY ANALYSIS

Parameter	Description
T_{PM}	The Average Packet Delivery Latency of MIPv6
α	The Average Packet Arrival Rate
$T_{CN\rightarrow LMA}$	The Average Latency for Packet Delivery from CN to LMA
$T_{LMA\rightarrow MAG}$	The Average Latency for Packet Delivery from LMA to MAG
T_{PF}	The Average Packet Delivery Latency of PMIPv6
T_L	The Average Packet Processing Delay of LMA
T_M	The Average Packet Processing Delay of MAG
$l_{LMA\rightarrow MAG}$	The Average Distance Between LMA and MAG

The major symbols used in analyzing T_P are shown in Table 2. In PMIPv6 and F-PMIPv6, packets can be delivered in two modes. One is transmitting packets through the MAG. In this mode, the MAG receives every packet instead of the MN and delivers it to the MN. In the other mode, packets are directly delivered to the MN. In the following analysis, the average packet transmission cost is modeled by the former mode, but its implied method is the latter. The average delay of packet transmission from CN to MN through PMIPv6 and F-PMIPv6 can be depicted as follows.

$$T_{PM} = \alpha \cdot (T_L + T_{CN\rightarrow LMA} + T_{LMA\rightarrow MAG} + T_{MAG\rightarrow MN} \quad (21)$$

$$T_{PF} = \alpha \cdot (T_L + T_M + T_{CN\rightarrow LMA} + T_{LMA\rightarrow MAG} + T_{MAG\rightarrow MN} \quad (22)$$

The average packet transmission cost is as below.

$$T_P = T_{PF} - T_{PM}$$
$$= \alpha \cdot (T_M + T_{CN\rightarrow LMA} + T_{MAG\rightarrow MN} - T_{LMA\rightarrow MAG}) \quad (23)$$

The average processing delay of LMA(T_M) is established in a similar way to that in the previous study. Since the average number of regional MN is ωK, suppose that the MAG can provide an average ω of MN. Therefore, the complexity of finding the binding cash in LMA is proportional to ωK. Also, since an inquiry of an IP routing table is generally based on the corresponding longest prefix, it is realized using a Patricia Trie [9]. The complexity of an IP routing table inquiry is proportional to the length of the routing table log. The average delay of packet encapsulation in LMA is δ. So, when A and B are clearly defined coefficients, T_M can be calculated by (24).

$$T_M = A\omega K + BlgK + \delta \quad (24)$$

Suppose that the average packet transmission delay of a wired link is proportional to the number of relocated hops with coefficient η. Then, (23) can be converted to (25).

$$T_P = \alpha \cdot (A\omega K + BlgK + \delta$$
$$+\eta(l_{CN\rightarrow LMA} + l_{LMA\rightarrow MAG} - \frac{MinInt+MaxInt}{2}) \quad (25)$$

From $l_{CN\rightarrow LMA} + l_{LMA\rightarrow MAG} \geq \frac{MinInt+MaxInt}{2}$, it is clear that (25) represents the average packet transmission cost T_P. $T_P < 0$ indicates that the average packet transmission delay of F-PMIPv6 is higher than that of PMIPv6. This is based on the fact that the regional propagation of LMA is the result of a triangle routing problem. The route of packet transmission is converted from PMIPv6 to an outer network to the outer network of an MN F-PMIPv6 and then the LMA, and finally, the MN.

The overall cost function expressed as C_T describes the general performance of F-PMIPv6 compared to PMIPv6 in every point of view, including registration and packet transmission. n_1 and n_2 are coefficients defined in (26).

$$C_T = n_1 \cdot D_R + n_2 \cdot T_P \quad (26)$$

It reflects the application of F-PMIPv6 and PMIPv6. When $C_T < 0$, F-PMIPv6 is applied rather than PMIPv6 or F-PMIPv6 is more adequate.

B. Numerical Results

The numerical analysis of several major parameters of F-PMIPv6 and PMIPv6 is shown. SSM, PMIPv6, and F-PMIPv6 are compared. The values of parameters used here are shown in Table 3.

TABLE III. PARAMETER VALUES USED IN PERFORMANCE ANALYSIS

Parameter	Value	Parameter	Value
μ	0.008	n_1	10
η	0.008	n_2	1
θ	2	D_{LMA1}	0.008
ω	15	D_{MAG}	0.008
N	30	A	0.00003
D_{LMA2}	0.005	$MinInt$	1
B	0.00007	$MaxInt$	5
δ	0.00005	T_L	0.008
$l_{CN\rightarrow LMA}$	18	φ	0.015

The registration delay in PMIPv6 or F-PMIPv6 affects the handover delay directly. This is an important statistic to evaluate the quality of service in the network. $n_1 > n_2$ due to the importance of handover delay.

Suppose that the MN does not relocate the access point more frequently than once every second. So, $T \geqslant 1$. The TTL field of the IP header is generally 32 or 64. That is, the

limit of the number of hops through which packets can be transmitted is 32 or 64. $l_{CN \to LMA} = 25$ and $l_{LMA \to MAG} = 10$.

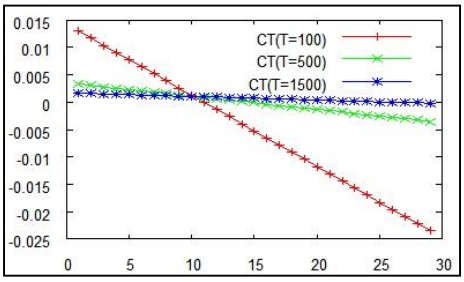


Figure 4. Impact of T on C_T.

Fig. 4 explains the relations between C_T and T. In this scenario, $\alpha = 0.05$ and $l_{LMA \to MAG} = 6$. C_T is less than zero when $K > 9$, and it decreases with T since T reflects the velocity of MN. In a small region ($K \leqslant 9$), the MN moves fast, and the ratio of MNs that show fast relocation is high. Due to the double registration in F-PMIPv6 would cause a long registration delay. In this case, F-PMIPv6 cannot deliver the average registration profit. As the MN movement becomes faster, the F-PMIPv6 registration performance is more degenerated. On the other hand, when K is high enough ($K \geqslant 9$), the probability of an N that moves outside of the region is low, even if the MN is moving fast. In other words, most of the mobility is micro mobility. In this case, F-PMIPv6 can yield the average registration profit.

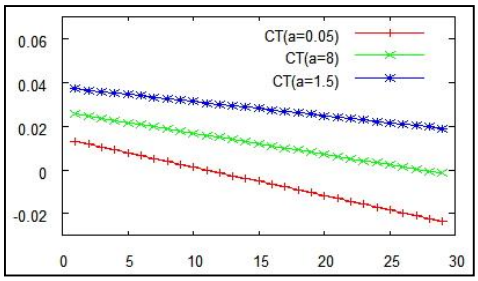


Figure 5. Impact of α on C_T.

Fig. 5 depicts the effect of α on C_T. In this scenario, T = 100 and $l_{LMA \to MAG} = 6$. It shows the proportional trend of C_T and α. The fact that the average packet transmission cost increases as α increases leads to an increasing value of C_T.

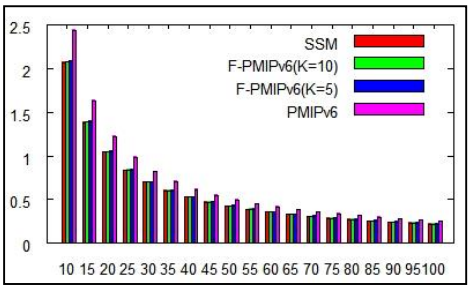


Figure 6. Cost *vs.* *T*.

In Fig. 6, the cost is a combination of the registration delay and packet transmission delay. The Cost of PMIPv6 can be calculated as $n_1 \cdot D_{RM} + n_2 \cdot T_{PM}$.

V. CONCLUSION

Both PMIPv6 and F-PMIPv6 are the mobility management solutions for IPv6 networks. However, F-PMIPv6 is an extension of PMIPv6, and surpasses PMIPv6 in some aspects, not every aspect. This study proposed an analytical model for an improved protocol, F-PMIPv6, and compared it to PMIPv6. Based on this analytical model, the SSM method selects the most adequate protocol and MAG. The mathematical results explain the effects of several key parameters based on the application ranges of PMIPv6 and F-PMIPv6. With the SSM method, an adequate protocol between PMIPv6 and F-PMIPv6 is chosen. Also, SSM showed better performance compared to PMIPv6 or F-PMIPv6.

ACKNOWLEDGEMENT

This research was supported by the Ministry of Trade, Industry and Energy (MOTIE), KOREA, through the Education Support program for Creative and Industrial Convergence (Grant Number N0000717) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2010-0024695). Also, this research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No.2010-0020737).

Corresponding author: Jongpil Jeong.

REFERENCES

[1] S. Gundavelli, K. Leung, V. Devarapalli, K. Chowdhury, and Patil, B., "Proxy Mobile IPv6", IETF RFC 213, August 2008.

[2] D. Johnson, C. Perkins, J. Arkko, "Mobility Support in IPv6", IETF RFC 3775, June 2004.

[3] H. Yokota and R. Koodli, "Fast Handovers for Proxy Mobile IPv6", IETF RFC 5949, September 2010.

[4] M. C. Chuang and J. F. Lee, "FH-PMIPv6: A fast handoff scheme in Proxy Mobile IPv6 networks", IEEE Consumer Electronics, May 2011, pp. 1297-1300.

[5] Guan, J. Zhou, H. Yan, Z. Qin, Y. and Zhang, H., "Implementation and analysis of proxy MIPv6", Wireless Communications and Mobile Computing, vol.11, April 2011, pp. 477-490.

[6] N. Neumann, J. Lei, X. Fu, G. Zhang, "I-PMIPv6 : An Inter-Domain Mobility Extension for Proxy-Mobile IP", IWCMC' 09, June 2009.

[7] G. Kim, "Low latency cross layer handover scheme in proxy mobile IPv6 domain", NEW2AN 2008, pp. 110-121, September 2008.

[8] S. Yoo and J. Jeong, "Analytical Approach of Fast Inter-Domain Handover Scheme in Proxy Mobile IPv6 Networks with Multicasting Support", Journal of Korea Information Processing Society, vol. 19-C, no. 2, April 2012, pp. 153-166.

[9] S. Han and J. Jeong, "Intelligent Hierarchical Mobility Support Scheme in F-PMIPv6 Networks", Jornal of the Korean Institute of Communications and Information Sciences", vol. 38A, no. 04, April 2013, pp. 337-349.

[10] I. Im, Y. Cho, J. Choi, and J. Jeong, "Security-Effective Fast Authentication Mechanism for Network Mobility in Proxy Mobile IPv6 Networks," ICCSA 2012, vol. 4, June 2012, pp. 543-559.

 ISBN: 978-1-61208-353-7

Mobile Transactions over NFC and GSM

Muhammad Qasim Saeed
and Colin Walter

Information Security Group (ISG)
Royal Holloway University of London
Egham, Surrey, UK
Email: muhammad.saeed.2010@live.rhul.ac.uk
colin.walter@rhul.ac.uk

Pardis Pourghomi
and Gheorghita Ghinea

School of Information Systems
Computing and Mathematics
Brunel University, Uxbridge, Middlesex, UK
Email: pardis.pourghomi@brunel.ac.uk
george.ghinea@brunel.ac.uk

***Abstract*—Dynamic relationships between Near Field Communication (NFC) ecosystem players in a monetary transaction make them partners in a way that they sometimes require to share access permission to applications that are running in the service environment. One of the technologies that can be used to ensure secure NFC transactions is cloud computing. This offers a wider range of advantages than the use of only a Secure Element (SE) in an NFC enabled mobile phone. In this paper, we propose a protocol for NFC mobile payments over NFC using Global System for Mobile Communications (GSM) authentication. In our protocol, the SE in the mobile device is used for customer authentication whereas the customer's banking credentials are stored in a cloud under the control of the Mobile Network Operator (MNO). The proposed protocol eliminates the requirement for a shared secret between the Point of Sale (PoS) and the MNO before execution of the protocol, a mandatory requirement in the earlier version of this protocol. This elimination makes the protocol more practicable and user friendly. A detailed analysis of the protocol discusses multiple attack scenarios.**

Keywords–Near Field Communication; Security; Mobile Transaction; Cloud.

I. INTRODUCTION

Agreed technical standards and fundamental interoperability are essential basics to achieve for industries working with NFC technology in order to establish positive cooperation in the service environment. Lack of interoperability in the complex application level has resulted in the slow adoption of NFC technology. Current service applications do not provide a unique solution for the ecosystem: many independent business players are currently making decisions based too closely on their own advantage over other players. Consequently, the service environment does not meet the optimal conditions for take-up. This has motivated us to extend current NFC ecosystem models to accelerate development. Our goal is to provide a concept for an NFC ecosystem that is technically feasible, accepted by all parties involved and provides an improved business case for each of the players. One of the main players in the NFC ecosystem is the Mobile Network Operator (MNO). The advantage an MNO has over other parties is that it owns a Secure Element (SE), the Subscriber Identity Module (SIM) card, that stores and protects the security parameters. Unlike other forms of SE, the SIM card can be easily managed by the MNO over-the-air. Thus, we foresee that the MNO will play a major role in future in the NFC ecosystem.

A. Our Contribution

Here, we extend the earlier proposed mobile transaction mechanism mentioned in [1]. The major contribution of our work is the elimination of the requirement for a shared secret between the Point of Sale (PoS) and the MNO, a prerequisite in the initially proposed protocol. This makes our work more flexible and it can even be used for monetary transfer between two individuals provided that the payer has registered a bank account with his MNO. We partition the SE into two sections: one stored in the SIM for authentication of a customer and the other stored in the cloud to hold cutomer account details. The authentication of the customer by the MNO is based on a GSM authenticating mechanism. The GSM standard, although not so secure, is still widely used for mobile communication, accounting for more than five billion subscriptions [2]. The idea is to reuse the existing cryptographic functionalities of the GSM standard thus reducing a need of additional cryptographic modules. Our protocol works on a similar pattern to that of 'PayPal': the MNO, acting in the same way as PayPal, registers multiple banking cards against a user for monetary transactions. The user then selects a single card at the time of the payment. But, unlike PayPal, our system uses the existing features of GSM standard for secure transactions. An overview of this model was proposed in [3].

This paper is structured as follows: Section II introduces the SE with a discussion of its management issues. We also highlight some advantages of having a cloud environment for mobile payment transactions. Section III describes related literature while Section IV recalls the essentials of GSM authentication. Then Section V introduces our proposed transaction protocol in detail followed by its analysis. Finally, Section VII places our solution in context, summarises how it operates, and draws some conclusions.

II. MANAGEMENT OF THE SE

The security of NFC is supposed to be provided by a component called the "security controller" that is in the form of an SE. The SE is an attack resistant microcontroller that can be found in a smart card [4]. It provides storage within the mobile phone and contains hardware, software, protocols and interfaces. It provides a secure area for the protection of payment assets (e.g., keys, payment application code, and payment data) and the execution of other applications. In addition, the SE

 ISBN: 978-1-61208-353-7

can be used to store other applications which require security mechanisms and it is involved in authentication processes. To be able to handle all these, the installed operating system has to have the capability of personalizing and managing multiple applications that are provided by multiple Service Providers, preferably over-the-air. Still, the ownership and control of the SE within the NFC ecosystem may result in a commercial and strategic advantage, as well as reluctance to participate from other providers. However, some solutions are already in place [4] and researchers are developing further models to overcome this problem.

A. Advantages of the Cloud-Based Approach

Our NFC cloud-based approach introduces a new method of storing, managing and accessing sensitive transaction data by storing data in the cloud rather than in the mobile phone. When a transaction is carried out, the required data is retrieved from a remote virtual SE which is stored within the cloud environment. The mobile phone SE provides temporary storage and authentication assets for the transaction to take place, and all communication between the cloud provider and the vendor terminal is established through the NFC phone.

An issue with SEs is that companies have to meet the requirements of organisations such as EMVco [5] to provide high level security to store personal data. This makes the SE expensive for companies. However, a cloud-based approach would transfer this cost. Then the SE in the NFC phone need only be responsible for user/device authentication and not for storing personal data. This improves the cost efficiency of the SE compared with the present, enabling many more secure applications to be supported because of the reduced space. Also, the NFC controller chips could be smaller and cheaper as they no longer have to support all previous functionality.

The NFC cloud-based approach makes business simpler for companies in terms of the integration of SE card provisioning. It would be much easier for businesses to implement NFC services without having to perform card provisioning for every single SE. The NFC phone user will be able to access many more applications as they are no longer stored in a physical SE. In terms of flexibility, all users would be able to access all their applications from all their devices (e.g., phones, tablets or laptops) since the applications are stored in a cloud environment that provides a single, shared, secure, storage space. Moreover, fraud detection would be instantaneous as the system runs only in a fully online mode.

III. Related Work

One of the major companies which operates the concept of a Mobile Wallet is Google, whose name for this service is "Google Wallet". The communication between the mobile phone and the PoS is carried out through NFC technology that transmits the payment details to the merchant's PoS. The Google Wallet is in the form of an Android application with a Secure Element (SE) on the customer's mobile phone and an SE in the cloud. The customer will have an account with Google Wallet which includes the relevant registered credit/debit cards whose details are stored in the online SE using secure servers. The transaction takes place in the form of a virtual prepaid credit card which is transfered to the merchant's PoS when the customer taps his phone on the PoS. For this, the Google Wallet app initializes the SE on the mobile device and a secure channel is established between the SE of the device and the SE in the cloud [6].

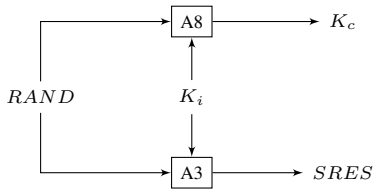


FIGURE 1. GENERATION OF K_c AND $SRES$ FROM $RAND$.

Our model differs from Google Wallet in the context of the location of the SE. We use the SIM as the SE in the mobile device whereas, the Google wallet requires an embedded SE. If a customer changes his handset, our approach still works as only the SIM needs to be replaced in the new device. This makes the approach more flexible.

Gerald Madlmayr and Josef Langer presented a purse-based micro-payment system [7]. They designed a pre-paid wallet where the money is stored in the Secure Element in the mobile device. The user can top-up his account Over-The-Air (OTA), anywhere and at any time.

Other solutions include "MasterPass" [8]. This service was developed by MasterCard as an extended version of the PayPass Wallet Services [9] that provides a digital wallet service for safe and easy online shopping.

IV. GSM Authentication

When a mobile device signs into a network, the MNO first authenticates the device (specifically the SIM). The authentication stage verifies the identity and validity of the SIM and ensures that the subscriber has authorized access to the network. The Authentication Centre (AuC) of the MNO is responsible for authenticating each SIM that attempts to connect to the GSM core network through a Mobile Switching Centre (MSC). The AuC stores two encryption algorithms, A3 and A8, as well as a list of all subscriber identities along with their corresponding secret keys K_i. The key K_i is also stored in the SIM. The AuC first generates a random number, denoted by $RAND$. This is used to generate two responses: a signed response $SRES$ and a key K_c as shown in Figure 1, where $SRES = E_{K_i}(RAND)$ uses the A3 encryption algorithm and $K_c = E_{K_i}(RAND)$ uses the A8 encryption algorithm [10].

$(RAND, SRES, K_c)$ is known as the *Authentication triplet* generated by the AuC. The AuC sends this triplet to the MSC. On receiving a triplet, the MSC forwards $RAND$ to the mobile device. The SIM computes the expected response $SRES$ from $RAND$, using A3 and the key K_i which is stored in the SIM. The mobile device transmits $SRES$ to the MSC. If this $SRES$ matches the $SRES$ in the triplet, then the mobile is authenticated. K_c is then used for communication encryption between the mobile device and the Base Station (BS).

 ISBN: 978-1-61208-353-7

TABLE I. ABBREVIATIONS

Acc_{ID}	Account ID of the customer
App_{ID}	Transaction approval message for customer account ID
AuC	Authentication Center (subsystem of MNO)
BS	Base Station
Cr_{req}	Credit Request Message
Cr_{app}	Credit Approved Message
IMSI	Internet Mobile Subscriber Identity
K_i	SIM specific key. Stored at a secure location in SIM and at AuC
K_c	$E_{K_i}(RAND)$ using A8 algorithm
K_1	Encryption key generated by the SIM
K_2	MAC key generated by the SIM
K_3	Encryption key generated by shop (the PoS)
K_4	MAC key generated by shop
K_{pub}	Public key of MTD
K_{pr}	Private key of MTD
K_{sign}	Signing key of MTD
K_{ver}	Verification key of MTD
LAI	Local Area Identifier
MD	Mobile Device
MNO	Mobile Network Operator
MSC	Mobile Switching Centre
MTD	MNO Transaction Department
PI	Payment Information
$RAND$	Random Number (128 bits) generated by MNO
R_s	Random number generated by SIM (128 bits)
SBAD	Shop Bank Account Detail
SE	Secure Element
SRES	Expected Response
TEM_u	Transaction Execution Message for user
TEM_s	Transaction Execution Message for shop
TMSI	Temporary Mobile Subscriber Identity
TP	Total Price
T_{SID}	Temporary Shop Identifier
TS_a	Approval Time Stamp
TS_s	Shop Time Stamp
TS_{tr}	Transaction Time Stamp
TSN	Transaction Serial Number

V. THE PROPOSED PROTOCOL

This section describes our proposed protocol for micro-payments based on NFC and cloud architecture. The assumptions are outlined as follows:

Our proposal is based on a cloud architecture where the cloud is being managed by the MNO. The cloud is used to store sensitive information about customers. A customer, who is a user of a cell phone, opens up a payment account with the respective MNO prior to use of the proposed payment feature. Each account is identified by a unique identity, the *Account ID* or Acc_{ID}. The account is either a *pre-paid* or a *pay-as-you-go* account. In the former type of account, the customer needs to top-up his account by either pre-paid vouchers or cash. This feature is more suitable for customers who do not have bank accounts. The amount a customer has in his pre-paid account is stored in the cloud against respective Acc_{ID}. In the *pay-as-you-go* type of account, the customer provides his banking credentials, like credit/debit card details to the MNO. The banking credentials are verified in the registration process by the MNO from respective bank and are stored against the respective Acc_{ID} in the cloud. A customer can have one or more accounts of either type and has the option to select one account while payment.

We suggest a dedicated department, the MNO Transaction Department (MTD), to manage the monetary transactions. A virtual secure tunnel is established between the mobile device and the MTD to ensure the security of the messages. The virtual tunnel is of special significance when the BS of some other network is used for the transaction as, in such scenario, the MNO responsible for monetary transaction does not want to reveal any sensitive information to the BS.

The mobile device communicates with the MNO over the standard GSM link. Communication over the GSM link between the mobile device and the BS is encrypted as specified in the GSM standard. Otherwise, communication between different entities of the GSM network is not considered to be secure and so encryption needs to be added where appropriate.

The MNO may be linked to the customer through its own BS or through a BS of some other network. Especially in the latter case, the proposed protocol should not disclose any sensitive information to the other network. The shop communicates with the MNO through the customer's mobile device using NFC and the GSM link. The shop PoS terminal does not require to be registered with the MNO. This makes the protocol more flexible and can also be used for monetary transactions between two individuals (the payer and the payee are analogous to the customer and the shop respectively).

The mobile device is connected to the shop terminal over an NFC link, but note that, although the NFC link is generally regarded as secure because of its short range of operation, yet it can be eavesdropped [11] or vulnerable to relay attacks [12]. A recent study suggested that any metallic object in the vicinity of an NFC link, even a shopping trolley, can act as 'rogue' antenna to eavesdrop the communication [13].

The shop does not use any direct link with the MTD for transactions. However, it needs to trust the MTD, i.e., a message digitally signed by it should be considered authentic and its contents trusted.

For simplicity, we refer to the mobile device and SIM as a single unit called the 'Mobile Device' (MD). K_{sign}, K_{ver} are the signing and verification keys respectively of the MTD, whereas K_{pr}, K_{pub} are the private decryption and public encryption keys respectively of the MTD.

Before any transaction can take place, the customer must have registered at least one account with his MTD and the merchant must have downloaded an (MNO-independent) application for performing our protocol. The merchant's application must be able to form several messages, such as PI (detailed below), in a format acceptable to the MTD of any MNO. In addition, the merchant needs to obtain and store trusted certificates for the keys of the MTDs he is willing to trust.

The protocol executes in three different phases: customer identification and credit check, customer authentication, and transaction execution. The steps of the protocol are illustrated in Figure 2, with numbering as in the text. Table I describes the abbreviations used in the proposed protocol.

A. Phase I: Customer Identification and Credit Check

This phase is initiated when the store owner sends the payment request to his NFC reader and the customer places his MD on the shop's NFC enabled point.

Step 1: The MD and shop terminal establish an NFC connection.

 ISBN: 978-1-61208-353-7

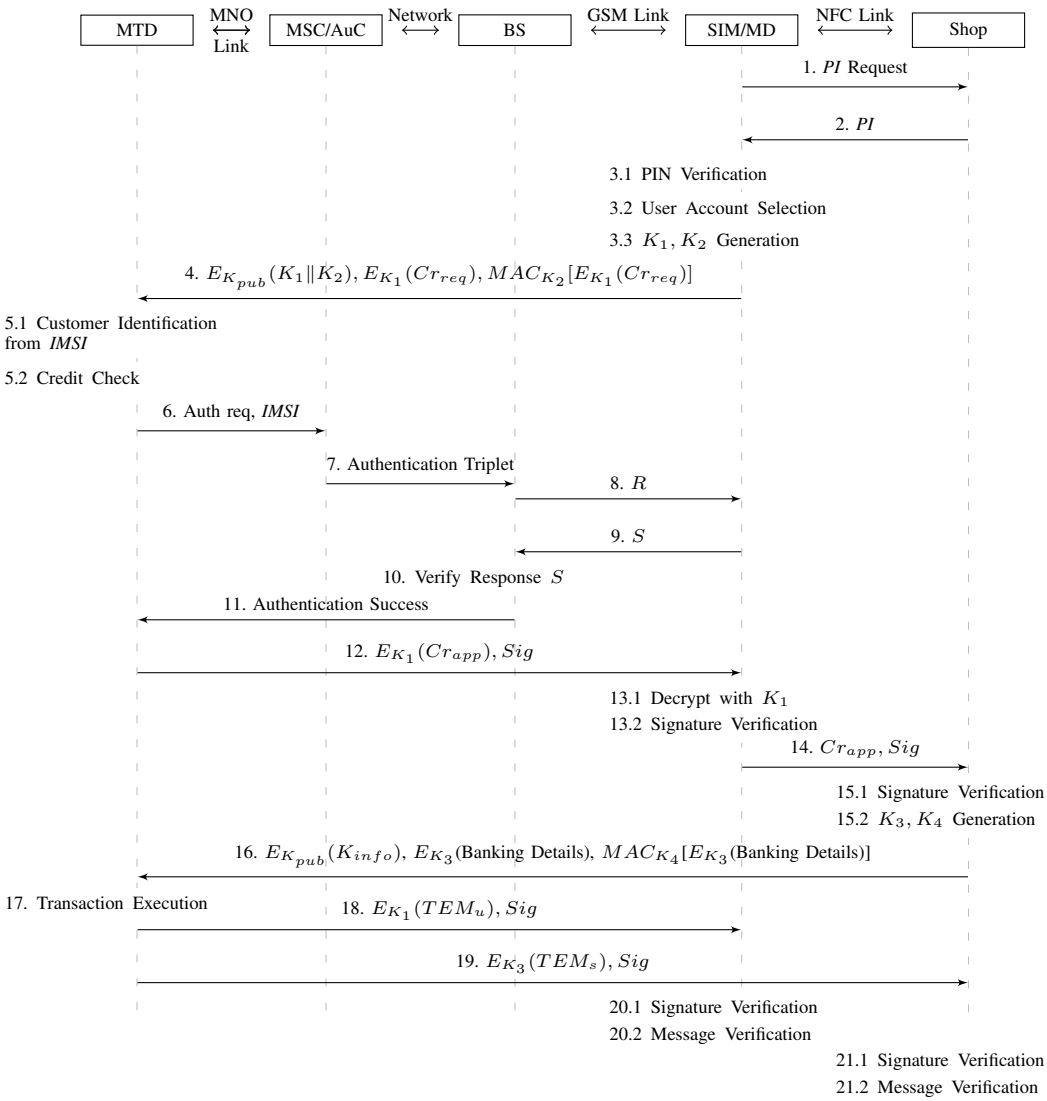


FIGURE 2. THE PROPOSED CUSTOMER AUTHENTICATION & PAYMENT PROTOCOL

Step 2: The shop terminal forms the Payment Information message *PI* containing at least the Total Price TP, a temporary shop identity T_{SID}, and the shop's Time Stamp TS_s, and sends it to the MD:

$$PI = TP \| T_{SID} \| TS_s \tag{1}$$

The T_{SID} acts as one time identifier used by the shop to identify the transaction. It is updated and fresh for each transaction. Optionally, PI may also contain a description of the shop and the goods which would appear on the customer's credit/debit card account statement.

Steps 3-4: Once the payment information is received from the shop, the application installed on the MD displays the transaction amount TP to the user and asks him to select a payment account and provide PIN authentication. This is for assurance that the customer is the legal owner of the mobile device, and therefore also the owner of the account which will be used for payment. It also provides confirmation that the amount and account details are accepted by the user.

After successful PIN verification, the mobile device needs to obtain a credit approval certificate for the shop from the respective MTD indicating that the customer has sufficient funds in his account and has agreed to pay the required amount. The information in this exchange should not be accessible to the BS or any other entity of GSM network other than the MTD. To provide a secure connection for this exchange between the MD and the MTD, the former generates two keys K_1, K_2 for symmetric encryption and MAC respectively. The actual encryption process used here is irrelevant, but will most likely be specified by the card provider and EMV requirements. It should not depend only on quantities known to either the BS or the MNO since only the MTD should be able to perform the decryption. The mobile device forms a credit request message Cr_{req} for credit approval from the MTD, namely,

 ISBN: 978-1-61208-353-7

$$Cr_{req} = PI \| IMSI \| Acc_{ID} \quad (2)$$

This is encrypted with K_1 and a MAC is computed on the ciphertext using K_2 to provide data integrity. Then the keys, K_1 and K_2, are encrypted with the MTD's public key K_{pub}. The entire message, consisting of the encrypted keys, the encrypted credit request and the MAC value, is sent to the MTD as in message 4 of Figure 2.

Step 5: Upon receipt of this message, the MTD starts by decrypting the first part of the message with its private key K_{pr} to extract the encryption and MAC keys, K_1 and K_2. It then verifies the MAC and in case of successful verification, it decrypts the second part of the message, containing Cr_{req}, and checks the freshness of the shop's time stamp in PI. The MTD identifies the customer from the IMSI in Cr_{req} and performs a credit check against the named account Acc_{ID}.

B. Phase II: Customer Authentication

Steps 6-11: Whether or not the credit and freshness checks are successful, the MTD sends an authentication request message to the MSC/AuC to authenticate the MD. The MD has already been identified by its IMSI. However, since the IMSI is not a secret, it may be used by a malicious party. To counter such threat, the MD needs to be authenticated under the IMSI claimed in Cr_{req} prior to any monetary transaction. With this IMSI, the MSC follows the usual procedure to authenticate an MD and it does not required further user interaction. So, in the case of successful authentication, the usual success message is sent from the BS to the MTD.

Step 12: If the credit check fails or the authentication success message is not received, the protocol is terminated with the sending of a fail message from the MTD to the MD. Termination does not occur before the authentication in order to hide the result of the credit check from an unauthenticated attacker. Otherwise, when both the credit check and the authentication are successful, a credit approval identifier App_{ID} is generated by the MTD. This acts as an index to a table in which the MTD stores information about the debit account, the amount to be transferred, the destination shop identity, a time stamp and the MD identity (IMSI). This identifier helps in resolving any disputes in the future but the details of the transaction are not contained therein.

The MTD now forms a new string Cr_{app} indicating credit approval for the Payment Information PI, namely,

$$Cr_{app} = PI \| TS_a \| App_{ID} \quad (3)$$

The MTD computes a signature with its signing key K_{sign} over the hashed plaintext and encrypts the string Cr_{app} with the key K_1. The encrypted Cr_{app} along with its signature is transmitted to the mobile device. The former cannot be decrypted in transit as the encryption key K_1 is unavailable, nor is Cr_{app} revealed by applying the verification key to the signature because of the hashing. Moreover, because of TS_s, TS_a or App_{ID}, the message differs each time even if the user buys the same goods on successive occasions.

Steps 13-16: The mobile device decrypts the message with the encryption key K_1 to obtain Cr_{app} and forwards it to the shop along with the corresponding signature. The shop verifies the signature using K_{ver} and compares the PI content in the Cr_{app} message to the one it initially sent in message 2. In the case of an invalid signature or a mis-match with PI, the shop discards the message, rejects the payment, and withholds the goods or services from the customer. A successful verification indicates that the customer is legitimate and that the MTD has obtained agreement from the customer to pay. This is like a three party contract where a middle party (the MTD), trusted by both other parties, provides assurance that the other party is willing to pay the specified price.

The shop now needs to send its banking details to the MTD to complete the transaction. The banking details may include the account name and number, the bank and branch codes, etc. This is sensitive information and should not be disclosed to any entity other than the MTD, not even to the MD. The shop therefore generates encryption and MAC keys, K_3 and K_4 to secure its banking details. It encrypts the banking details with the key K_3, and computes a MAC over the ciphertext with the key K_4. It also form a string, K_{info}, containing the information about the keys as follows:

$$K_{info} = K_3 \| K_4 \| App_{ID} \quad (4)$$

The role of the approval identifier App_{ID} in this step is to enable the MTD to connect the authentication phase to the transaction execution phase. The shop encrypts the string K_{info} with the public key K_{pub} of the MTD and sends it to the MTD via the MD. This forms a virtual tunnel between the shop and the MTD through the MD, as the latter cannot decrypt the message content. Note, however, that the shop needs to be certain it has K_{pub} correctly from the MTD, and not a key substituted by an attacker.

C. Phase III: Transaction Execution

Step 17: The MTD associates the App_{ID} received in the step 16 with the already stored App_{ID} (step 12). It decrypts the banking details of the shop with keys K_3, K_4 and transfers the approved amount, stored against corresponding App_{ID}, to the shop account. The MTD flags the App_{ID} indicating that the transaction has been executed to ensure that the same App_{ID} could not be used again.

Step 18-21 After a successful transaction, the MTD generates a Transaction Serial Number (TSN) and forms Transaction Execution Messages, TEM_u and TEM_s for the MD and the shop respectively.

$$\begin{aligned} TEM_u &= PI \| TSN \| TS_{tr} \| Acc_{ID} \\ TEM_s &= PI \| TSN \| TS_{tr} \| SBAD \end{aligned} \quad (5)$$

The MTD computes a signature on the hashed plaintext, encrypts TEM_u with the key K_1, and sends it to the MD. The MD decrypts the message and verifies the signature. An invalid signature indicates that the transaction confirmation has been accidentally or deliberately corrupted en route. In such a case, the MD enquires about the transaction from the MTD. If the transaction has already been executed, the MD asks for a fresh confirmation message. Otherwise, it is obvious that message 16 has not been delivered to the MTD. This may happen if a malicious party has blocked the message

 ISBN: 978-1-61208-353-7

from reaching the MTD and has instead transmitted a fake transaction confirmation message. Of course, such a fabricated message cannot go undetected as it is signed by the MTD. In such scenario, the MD asks the shop to resend message 16.

The MTD also forms TEM_s for the shop by appending the Shop's Banking Details as shown in Eq (5). The MTD computes a signature over the hashed plaintext and encrypts TEM_s with the key K_3. The MTD sends this encrypted message along with its signature to the customer MD which relays it to the shop. The customer's MD can neither decrypt this message as it does not possess K_3, nor alter any contents as they are protected by the signature. The shop decrypts the message, verify its contents and the signature, thereby confirming that his account (rather than an attacker's) has been credited correctly. The contents consist of important transaction information exchanged during the transaction. Hence, if the shop wants any subsequent clarification, it can approach the MNO quoting the TSN and the App_{ID} received in step 14. Finally, if the shop is satisfied, it produces a receipt together with the goods or services for the customer.

VI. Analysis

In this section, we analyze the protocol from multiple perspectives to ascertain the strength of our protocol. This analysis encompasses the authentication and security of the messages. We assume that the MNO is trustworthy, whereas the customer or the shop can be dishonest, and there may be an active attacker listening to any of the NFC or other messages.

A. Dishonest Customer

Scenario 1. A dishonest customer plans to buy some products, making the payment from someone else's account. The PIN requirement in step 3 should force the customer to use his own mobile device to enact the protocol. Indeed the protocol depends on the strength of this PIN, just as is the case with credit card withdrawals. However, rogue applications on the MD could have already sniffed the PIN.

Assume that the attacker uses his own mobile and knows the IMSI and account numbers ($IMSI'$, Acc'_{ID}) of the target victim. He must fabricate Eq (2) as:

$$Cr'_{req} = PI \| IMSI' \| Acc'_{ID} \tag{6}$$

As this message can be decrypted only by the MTD, the malicious contents remain undetected by all other entities. The MTD decrypts the message and identifies the customer from $IMSI'$. Assuming the protocol does not fail here because the target victim is not a legitimate customer or the account has insufficient funds, the MTD proceeds to the fresh authentication of $IMSI'$. So the MSC/AuC provides the authentication triplet in step 7 corresponding to $IMSI'$. However, the attacker cannot compute the valid response S' as his mobile device lacks the necessary key K'_i. So, the authentication check fails and the protocol terminates. Thus, an incorrect identity cannot be successfully used in the protocol.

Scenario 2. Suppose a dishonest customer plans to buy goods without payment. He could accomplish this by providing his own banking details, instead of the shop's, to the MTD for the payment recipient. He then blocks the legitimate message 16, and replaces it as follows. Using his own keys K'_3 and K'_4, he fabricates message 16 with own banking details and sends it to the MTD. The MTD performs the transaction against this information, deducting the amount from the customer's account but paying it back into the same or another account of the customer. (These may be distinct in an attempt to avoid detection). After executing the transaction, the MTD sends 'receipts' in messages 18 and 19. The MD must block message 19 as this message contains the substituted bank details which the shop checks. So the dishonest customer needs to replace the banking details in this message with the shop's banking details. He can decrypt message 19 as it is encrypted with his own malicious key K'_3. However, he must now change the banking details and encrypt them with the shop's key K_3. As he lacks this key, he cannot generate a valid ciphertext. Moreover, the original message is protected by the digital signature. If the customer were to make any alteration to the banking details, it would void the signature which the shop verifies next. In neither case is the shop able to verify the transaction and a failure message is reported to the shopkeeper. Hence, the dishonest customer is again unsuccessful.

There may be another approach to accomplish the above attack where the dishonest customer plans to buy some goods without payment. The dishonest customer does not communicate with the MTD since he could not succeed in the way described above; rather, he masquerades as the MTD to the shop. The target of the customer is to send fake but acceptable receipts to the shop at the end of the protocol by replaying old legitimate, messages or fabricating new messages. Since the customer is not communicating with the MTD, his account will not be debited. In the original protocol, the shop receives three messages from the MD: messages 1, 14 and 19. Message 1 originates from the MD, whereas messages 14 and 19 actually originate from the MTD but are relayed by the MD to the shop. The dishonest customer needs to construct or replay the latter two messages in such a way that they are acceptable to the shop. Both messages are digitally signed by the MTD. They contain the Shop Identitifier T_{SID} and Time Stamp TS_s. T_{SID} is a random value generated by the shop every time at the start of the protocol. This value not only serves as a shop identifier during the protocol, but it also adds freshness to the protocol messages. TS_s is updated too in every protocol round, but it may be predictable to some extent. A combination of these two values, along with the digital signatures of the MTD, does not allow either replay or alteration of the messages to succeed. Hence, the dishonest customer is again unsuccessful. Of course, as usual in PKI, the shop should check the digital certificates of the MTD keys to justify its trust in them.

Scenario 3. Assume now that the dishonest customer plans to pay less than the required amount but claim payment of the full amount. To accomplish this, the MD sends TP' in the Credit Request message Cr_{req} of step 4 to the MTD, where $TP' < TP$. The MD receives the Credit Approval message, Cr_{app}, in step 12 from the MTD confirming that the initially requested amount TP' has been approved for transaction. But the MD needs to confirm to the shop in step 14 that the original amount, TP, is approved for transaction. Since the approved price is digitally signed by the MTD, it cannot be amended by the MD. So the actual price that is approved by the MTD is transmitted to the shop. As the shop application checks the

 ISBN: 978-1-61208-353-7

approved amount against that requested, this attack also fails.

Scenario 4. Here, a dishonest customer wants to pay through a mobile device which he does not own. He might have stolen that device or found it as lost property. If the SIM is still valid and the credit/debit cards have not been cancelled, it can still be used for transactions. After the device receives the payment information *PI* from the shop in step 2, the application installed on the mobile device requires PIN verification from the customer. Since the customer does not own the mobile device, he should not have knowledge of the PIN. So the protocol does not proceed further. Additionally, the application can be designed to be blocked in the MD and by the MTD after a limited number of failed attempts at PIN verification. This provides an assurance to the customers that their lost mobile device could not be used for any monetary transactions even while the SIM remains active.

B. A Dishonest Shop

Scenario 5. The shop is dishonest and plans to draw more than the required amount without intimation to the customer. The information about the amount to be transferred is sent to the MTD by the MD in the Credit Request message, Cr_{req}, in step 4. A mobile device cannot send more than the price contained in PI and approved by the user in step 3 unless the device itself is compromised. Therefore, a shop cannot obtain more than the agreed amount if, as requested, the customer checks the amount before entering his PIN.

Scenario 6. The shop is dishonest and denies receipt of the transaction execution message in step 19. In this way, the shop decides not to deliver the goods or services despite receiving the required amount. However, the MD has the signed receipt from the MTD with the TSN from Eq (5). This is linked to the approval App_{ID} generated in step 12. As both are digitally signed by the MTD, the customer can approach the MTD regarding any dispute. With knowledge of the account credited during the transaction and the shop receipt from the customer, the MTD can take action to identify the criminal and refund the customer.

C. Message Security

Apart from the above-mentioned scenarios, we also analyzed our protocols from various other angles. The data over the GSM link (between the MD and the BS) is encrypted according to the GSM specification. The data sent over the NFC link in steps 1, 2 and 14 are sent in the clear. This data does not contain any particularly sensitive information except perhaps for the TP. However, the range within which this data can be captured is very limited, and it is occupied by the shop keeper and the customer, at least one of whom should notice unwelcome devices (such as other NFC capable mobile phones) in the vicinity. The read range of the price displayed on both the shop till and the user's MD is much more than the range of the NFC link. Therefore, we considered PI as not sufficiently sensitive to need protection over the NFC link. Nevertheless, we should consider this in a little more detail.

Other information that is sent in clear over the NFC link includes the App_{ID} in the Cr_{app} message. At this point the attacker can hi-jack the protocol by blocking the communication of message 16, replacing it with his own forged message which contains his own bank details. There is no relevant data which is not known to the attacker. This results in a successful transfer of funds to the criminal and also a successful acknowledgment in step 18 to the legitimate customer. However, the shop owner will either not receive the transfer message in step 20, or will receive one which fails his verification. Thus, although the shop keeper will not then release the goods, the attacker will have obtained the funds. The solution is to include a means for the MTD to verify that message 18 comes from the same source as message 2.

We therefore propose the inclusion of a Diffie-Hellman key agreement (DH) between the MD and Shop during messages 1 & 2 in situations where the NFC link may be compromised. Then step 18 can include a proof of origin. Step 1 would include the public parameters for DH, and the MD's exponentiated value, while step 2 would include the shop's response of the other DH exponentiated value. As message 16 contains a MAC of the other components of message 16 using the DH shared key, the MD can check the authenticity of message 16, ensuring that the protocol has not been hi-jacked. However, an attacker who can hi-jack the protocol at step 18 could equally easily hi-jack it at step 1. This requires blocking the legitimate message PI and replacing it as necessary with PI' so that the MD agrees a shared key with the attacker instead of the shop and, later, the forged message 16 is authenticated by the MD. Since TP is not known to the attacker until PI is transmitted, the attacker needs to collect the legitimate PI first in order to include TP in PI', this being necessary to obtain the customer's agreement over the price. However, for this to succeed, the attacker must prevent the correct PI from reaching the MD. Consequently, the success of Diffie-Hellman key exchange between shop and MD cannot be prevented unless the attacker can guess TP correctly or the customer fails to check the amount carefully. An attacker may use a hidden camera which can read the shop's till display, then his NFC hi-jack device can know TP in advance and so determine a value for PI which the customer will accept. He can therefore block the legitimate message 2 and replace it with his own. The threat from this is similar to that of a camera capturing PIN values. Payment Card Industry Security Standard Council (PCI SSC) prohibits use of cameras near a PIN entering device to avoid monitoring of displays, PIN pads, etc., [14]. Moreover, there are two methods to block RF communication on the NFC link and neither of the methods can easily be adopted in our scenario. The first is to cover the transmitter or receiver with some shielding material. The other method is to produce a high noise on the same operating frequency resulting in a significant decrease in the signal-to-noise ratio. For the former the attacker must shield the MD or the shop terminal. The latter requires noise generating hardware in close proximity to the MD and the shop sales terminal. Both approaches are visibly detectable. This means there is little scope for a successful attack when the MD also verifies the authenticity of message 16. It should therefore be an acceptably small risk.

App_{ID}, which is sent in the clear over the NFC link, is a random string generated by the credit approval authority. From an attacker's perspective, its only significance is its assurance that the customer had, at least before the transaction, the amount TP in his account. This assurance can also be achieved if a customer successfully pays for some goods. Therefore, App_{ID} is not sensitive information in this scenario.

 ISBN: 978-1-61208-353-7

The Role of the Approval Identifier in message 16. App_{ID} acts as a bridge between phase II and III. It also adds freshness to message 16, so it cannot be replayed in future. Any alteration in the K_{info} results in invalid keys and an invalid App_{ID}. Hence it is detectable.

Non-repudiation of Transaction Execution Messages. TEM_u and TEM_s are digitally signed by the MTD. In case of any dispute over payment, the MTD has to honour both messages. So, both the customer and the shop are completely assured of the transaction payment taking place.

Disclosure of Relevant Information. The Cr_{req} containing price information is not disclosed to the base station or any other GSM entity apart from the MTD. The SBAD is sensitive information. It is encrypted not only over the GSM links but also over the NFC link. It is transmitted through the mobile device to the MTD, yet the former cannot decrypt this information. The Acc_{ID} of the customer is not disclosed to the shop. The MNO does not need to know the shopping details of the customer. Therefore, only the total amount is communicated to the MNO for transaction.

New Keys for every Transaction. The encryption and MAC keys for the message Cr_{req}, namely K_1 and K_2, are freshly generated by the mobile device in each round. Similarly, the keys K_3 and K_4, generated by the shop, are fresh for each transaction. Of course, these should not be predictable, especially if previous such keys become known.

Encryption and MAC Keys. Separate keys are used for encryption and MAC calculation making the protocol more secure. *Encrypt-then-MAC* is an approach where the ciphertext is generated by encrypting the plaintext and then appending a MAC of the encrypted plaintext. This approach is cryptographically more secure than other approaches [15]. Apart from its cryptographic value, the MAC can be verified without performing decryption. So, if the MAC is invalid for a message, the message is discarded without decryption. This results in computational efficiency.

D. Monetary Transaction Between Two Individuals

The proposed protocol can be used for monetary transactions between two individuals. The payee acts as a shop PoS terminal, and uses his own mobile phone for this. The added advantage in our proposal is that the payee does not need to register himself with the payer's (= customer's) MNO. This eliminates dependency of both parties to be on the same mobile network for monetary transactions. The payee needs only to provide his banking details in step 16 of the protocol.

VII. CONCLUSION

In this paper, we have proposed a transaction protocol for providing a secure and trusted communication channel for payment of goods using mobile devices. The proposed protocol was based on the NFC Cloud Wallet model [16] and the NFC payment application [1] on secure cloud-based NFC transactions. We considered a cloud-based approach for managing sensitive data, ensuring the security of NFC transactions by means of a virtual SE within the cloud environment, as well as considering and simplifying the role of the physical SE within the NFC phone architecture. The operations performed by the vendor's reader, by an NFC enabled phone and by the cloud provider (in this paper the MNO) are detailed. Such operations are possible using current technology as most of the functionality is already implemented to support other mechanisms. We considered the detailed execution of the protocol and showed our protocol performs reliably and securely in a cloud-based NFC transaction architecture. The main advantage of this paper is to demonstrate another payment method for those who do not have bank accounts in the normal sense. This way of making payments eases the process of purchasing as participants only have to top up their MNO account without having to follow all the usual banking procedures.

REFERENCES

[1] P. Pourghomi, M. Saeed, and G. Ghinea, "A Proposed NFC Payment Application," in International Journal of Advanced Computer Science and Applications (IJACSA), vol. 4, no. 8. SAI, 2013, pp. 173–181.

[2] G. Cattaneo, G. Maio, P. Faruolo, and U. Petrillo, "A Review of Security Attacks on the GSM Standard," in Information and Communication Technology, ser. Lecture Notes in Computer Science, vol. 7804. Springer, 2013, pp. 507–512.

[3] P. Pourghomi, M. Saeed, and G. Ghinea, "Trusted Integration of Cloud-based NFC Transaction Players," in 9th International Conference on Information Assurance and Security. IEEE, 2013, pp. 6–12.

[4] P. Pourghoumi and G. Ghinea, "Challenges of Managing Secure Elements within the NFC Ecosystem," in 7th International Conference for Internet Technology and Secured Transactions. IEEE, 2012, pp. 720–725.

[5] J. Paillès, C. Gaber, V. Alimi, and M. Pasquet, "Payment and Privacy: A Key for the Development of NFC Mobile," in International Symposium on Collaborative Technologies and Systems. IEEE, 2010, pp. 378–385.

[6] M. Roland, J. Langer, and J. Scharinger, "Applying Relay Attacks to Google Wallet," in The 5th International Workshop on Near Field Communication, Zurich, Switzerland. IEEE, 2013, pp. 1–6.

[7] G. Madlmayr and J. Langer, "Near Field Communication Based Payment System," in Konferenz Mobile und Ubiquitäre Informationssysteme (MMS), Germany, ser. LNI, vol. 123. GI, 2008, pp. 81–93.

[8] "MasterPass," 2013, URL: https://masterpass.com/Wallet/Home [accessed: 2014-06-04].

[9] Sarah Clark, "MasterCard Enters the Mobile Wallet Market," in NFC World, 9 May 2012, URL: http://www.nfcworld.com/2012/05/09/315600/mastercard-enters-the-mobile-wallet-market/ [accessed: 2014-06-04].

[10] ETSI Specification of the Subscriber Identity Module - Mobile Equipment (SIM - ME) interface (GSM 11.11), European Telecommunications Standards Institute (ETSI) Std. Version 5.0.0, December 1995.

[11] G. P. Hancke, "Practical Eavesdropping and Skimming Attacks on High-Frequency RFID Tokens," in J. Comp. Sec., vol. 19, no. 2. IOS Press, 2011, pp. 259–288.

[12] L. Francis, G. Hancke, K. Mayes, and K. Markantonakis, "Practical Relay Attack on Contactless Transactions by Using NFC Mobile Phones," in IACR Cryptology ePrint Archive, 2011, pp. 618:1–618:16.

[13] T. W. C. Brown and T. Diakos, "On the Design of NFC Antennas for Contactless Payment Applications," in 5th European Conference on Antennas and Propagation (EUCAP). IEEE, April 2011, pp. 44–47.

[14] PIN Transaction Security (PTS) Point of Interaction (POI), Payment Card Industry (PCI) Std. Version 3, June 2013, URL: https://www.pcisecuritystandards.org [accessed: 2014-06-04].

[15] M. Bellare and C. Namprempre, "Authenticated encryption: Relations among notions and analysis of the generic composition paradigm," in Journal of Cryptology, vol. 21, no. 4. Springer, 2008, pp. 469–491.

[16] P. Pourghoumi and G. Ghinea, "Managing NFC Payments Applications through Cloud Computing," in 7th International Conference for Internet Technology and Secured Transactions (ICITST). IEEE, 2012, pp. 772–777.

 ISBN: 978-1-61208-353-7

Action Recogntion with Depth Maps Using HOG Descriptors of Multi-view Motion Appearance and History

DoHyung Kim, Woo-han Yun, Ho-Sub Yoon, Jaehong Kim
Intelligent Cognitive Technology Research Department
Electronics and Telecommunications Research Institute
Daejeon, Korea.
{dhkim008, yochin, yoonhs, jhkim504}@etri.re.kr

***Abstract*—The goal of this work is to recognize human actions only using depth maps without additional joints information. As a practical solution, we present a novel volumetric representation of global shape of depth motion, Depth Motion Appearance (DMA). The proposed framework also extracts dynamic information of the body movements called Depth Motion History (DMH), an extended version of motion history image. In the framework, a huge amount of data of an action video is summarized into concise action representation maps observed from multi-view. A histogram of oriented gradients then describes local appearances and shapes of the DMAs and DMHs, which results in more compact and discriminative action representation. The presented method has been compared with the state-of-the-art approaches on a public dataset. The experimental result demonstrates that our approach achieves a better and more stable performance with a relatively smaller feature maps and lower complexity.**

Keywords-Action recognition; Depth maps; Depth motion appearance; Depth motion history; Histogram of oriented gradients.

I. INTRODUCTION

Despite numerous research efforts and advances in the last decade, traditional human action recognition with the sequence of 2D color images is still a challenging problem. Human actions are in essential continuous evolution of dynamic motion of three-dimensional body parts and articulated joints. In addition, same action can be performed in various ways of body movements by each individual and two different actions having a similar trajectory of motion make it more difficult to distinguish correctly. So, the absence of depth information could lead to significant degradation of discriminating capability of an action recognizer and consequently limit its performance.

In recent years, the technology of action recognition has entered a new phase with the release of the low-cost depth cameras like Microsoft Kinect [1]. These depth cameras provide 3D depth data as well as color image sequences in real time, which makes it possible to explore the fundamental solution for traditional problems in human action classification. Recent studies taking advantage of 3D information have been showing advanced results compared to the traditional 2D video-based researches [2][3][5].

As it is well known, the human actions could be modeled by the motion of a set of three-dimensional articulated joints [4]. So, if we can obtain 3D positions of key joints in real time with reasonable accuracy, action recognition can be successfully accomplished. However, estimating 3D joint positions is still a challenging task. Although some consumer depth cameras provide body joints information, the estimated joint positions are coarse and sometimes have significant errors particularly when body parts are self-occluded like two hands crossing. Moreover, most depth sensors only provide a sequence of depth maps. For these practical reasons, the work presented here has focused on recognizing human actions only using depth maps without additional information of the joints of the skeleton.

The main contributions of this work include two aspects.

First, we propose the Depth Motion Appearance as a new way of describing the global 3D shape of a body movement. It is a 3D depth map which represents a region of forward depth motion stacked through all of the depth images of an action. Our method can be differentiated from the prior depth map-based studies. The work by Li et al. [5] only uses 2D projects of key poses instead of direct utilization of the 3D information, which could essentially lead to sub-optimal feature representations. While our method makes full use of 3D information of all depth maps in the sequence, which results in the improved discriminating power. Xiaodong et al. [6] generate a binary map of motion energy by computing and thresholding the difference between consecutive depth maps. But, their method crucially does not consider dynamic information of the body movements. On the contrary, our framework effectively combines the appearance feature with the temporal feature extracted by an extended framework of motion history image [7].

Second, the proposed approach yields the best accuracy when compared with many previous state-of-the-art action recognition methods based on 3D silhouettes or joints. Moreover, the result is achieved with relatively small feature sets. An entire sequence of depth maps can be encoded just to a 4096 dimensional HOG (Histogram of Oriented Gradients) descriptor [8]. This fact indicates that our action representation method is highly discriminative as well as computationally efficient.

This paper is organized as follows. In Section 2, the overview of the proposed framework is described. The detailed description of the proposed features is given in Section 3. An evaluation model and the experimental results

 ISBN: 978-1-61208-353-7

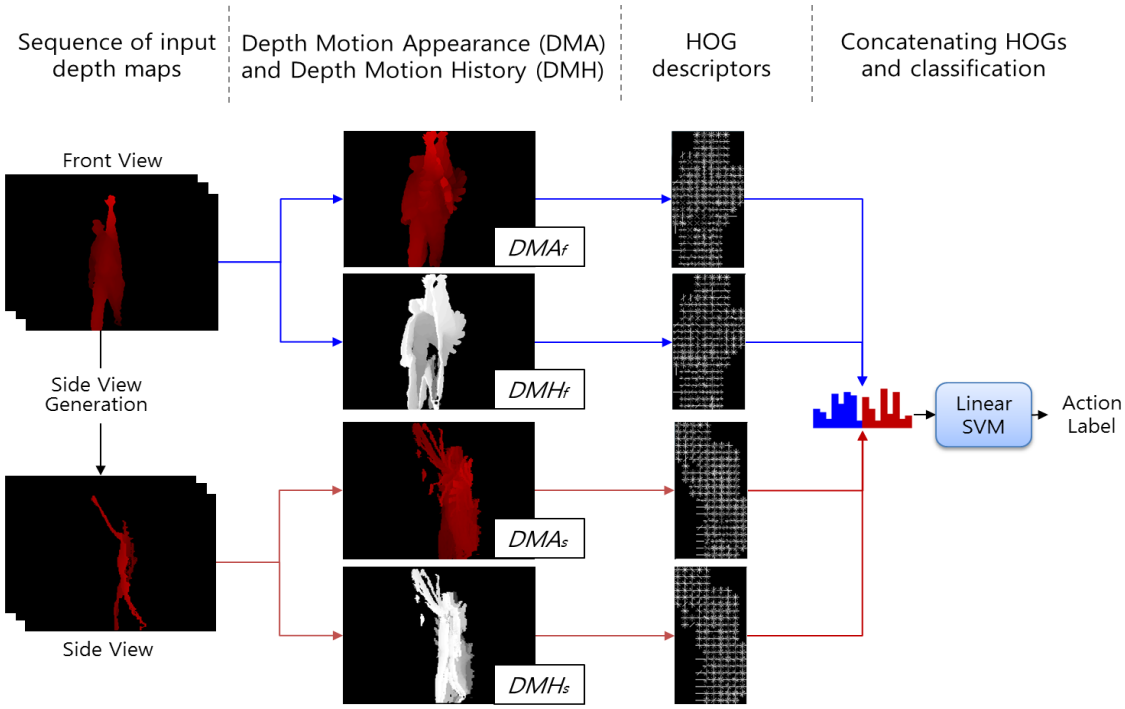


Figure 1. Overview of the feature extraction and action classification framework proposed in this paper.

are shown in Section 4. Finally, Section 5 concludes this paper.

II. Framework Overview

The proposed framework of the feature extraction and action classification is shown in figure 1. When the sequence of front-viewed depth maps is fed into the framework, it first generates a side-viewed depth map from the input depth map in order to acquire additional evidences. The framework then accumulates global activities through entire sequence of the depth images from each view and creates action representation maps called Depth Motion Appearance (DMA) and Depth Motion History (DMH). The DMA is an accumulated form of 3D depth information. It has no temporal information about the sequence of the motion, which can be complemented by the DMH that includes dynamic information of the entire motion region. In total four representation maps, two maps from each view, are generated for one action video. The system then calculates Histogram of Oriented Gradients (HOG) feature descriptors [7] for size-normalized DMAs and DMHs. The descriptors are concatenated into one single HOG descriptor which is fed into a linear Support Vector Machine (SVM) [9]. The linear SVM classifies the HOG descriptor and finally yields the action label of the query sequence.

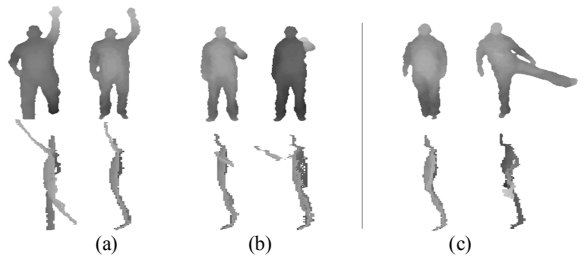


Figure 2. Original front-viewed depth maps (top row) and newly created side-viewed depth maps (bottom row): (a) and (b) depth maps with similar frontal shape but discriminable profile shape, (c) the opposite case of (a) and (b).

III. Action Representation

A. View generation

The side-viewed depth map provides an additional body shape and motion information different from that extracted from the frontal depth image. As shown in Figure 2, similar actions which are difficult to be distinguished from the front view might be easily discriminable in a lateral view and the opposite is true as well. Therefore, taking advantage of observations from various views can be an efficient and effective approach for 3D action classification. In order to capture full body actions, actors are commonly located at a long distance from depth sensors, which leads to a low depth resolution for the body region. So, interpolation methods are basically needed to estimate and produce new depth points when creating side-viewed depth images.

B. Depth Motion Appearance

The DMA is a volumetric representation of depth motion which describes the overall shape and appearance of a body movement forming an action. As for each view, we can

 ISBN: 978-1-61208-353-7

Figure 3. DMAs generated from different human actions: (a) sequences of input depth maps, (b) DMAs (side boxing, two hand wave, and tennis swing from top to bottom).

obtain the DMA by accumulating all depth maps of an action video from start to end.

$$DMA_v(i,j,t) = \begin{cases} D_v(i,j,t) & ,if\ DMA_v(i,j,t-1) = 0 \\ min(D_v(i,j,t), DMA_v(i,j,t-1)), & else. \end{cases} \quad s.t.\ D_v(i,j,t) > 0 \tag{1}$$

where v denotes the view, $D_v(i,j,t)$ is a depth value at a pixel position i, j of the t th input depth map under the view v, and $DMA_v(i,j,t)$ is a depth value at a pixel position i, j of the DMA_v generated from t input depth maps. The depth values of foreground region of an input depth map are only calculated for creating the DMA.

Figure 3 shows several action sequences and their DMAs respectively. For each action, the DMA represents its own distinctive appearance of body movement, which means it can be a strong feature for action classification. In addition, the DMA has an advantage in practical terms because it does not require any threshold values at all.

C. Depth Motion History

Although the DMA is a good method to represent appearance of a body movement, it does not include temporal information at all. Human actions are in essential continuous evolution of dynamic motion of body parts and articulated joints. Therefore, the absence of dynamic information on a sequence of movements can be a tremendous loss for an action recognizer. For extraction of temporal features, we present a method called the DMH, which is an extended form of Motion History Image (MHI) [8]. Traditional MHI can only cover the motion history occurred on the 2D image plane. With the depth information we can now encode the history of the motion along the depth changing directions.

$$DMH_v(i,j,t) = \begin{cases} \tau, & if\ |D_v(i,j,t) - D_v(i,j,t-1)| > \delta \\ max(DMH_v(i,j,t-1) - 1, 0), & else. \end{cases} \tag{2}$$

where $DMH_v(i,j,t)$ denotes a history value of depth motion at a pixel position i, j of the DMH_v created from t input depth map under the view v. τ is a time window for history and δ is a threshold value for depth difference between consecutive depth maps. The generated DMH is a two-dimensional image template where pixel intensity is a function of the recency of depth motion in a sequence.

D. Histogram of Oriented Gradients

The presented action representation method summarizes a great amount of depth data of the entire video into just four maps. We exploit the HOG method to describe local appearance and shape of the DMAs and DMHs. The HOG technique figures out the distribution of intensity gradients or edge directions in localized portions of an image [7]. Since the descriptor operates on localized cells, the method upholds invariance to geometric and photometric transformations.

For all the DMAs and DMHs, foreground regions are cropped and then normalized to a fixed size. Despite the same action, it can be variously performed by different actors. The size normalization can reduce intra-class variations including a human body type, a motion scale, and a distance between an actor and a sensor. We then achieve HOG descriptors by dividing each map into 8×16 non-overlapping cells and for each cell compiling a histogram of 8 gradient directions for the pixels within the cell. The local histograms are contrast-normalized using L2-norm measure. Each feature map is described as a HOG descriptor with the dimension of 8×16×8=1024 and we finally obtain a 4096 dimensional HOG descriptor from the entire action video. The HOG descriptor is fed into a multi-class linear SVM classifier that is implemented by using an open source library, LIBSVM [9].

IV. Experimental Results

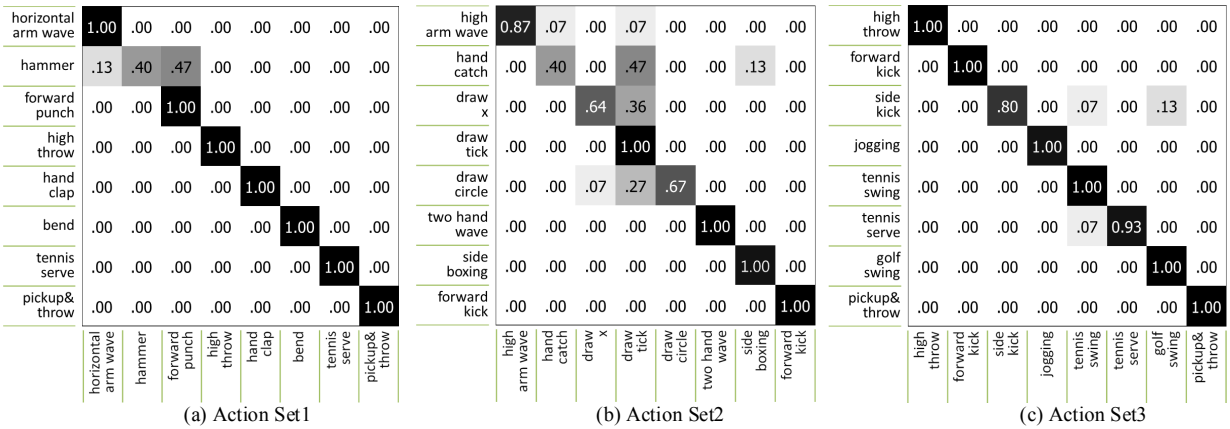
A. MSR Action3D dataset

The MSR Action3D dataset [5][10] is a public dataset on which a large number of methods have been experimented. The dataset provides sequences of depth maps captured by a depth sensor similar to the Kinect device. It contains 20 actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw.* The actions were chosen in the context of using the actions to interact with game consoles. They reasonably capture the various movements of arms, legs, torso and their combinations. In total, 567 depth map sequences are available. The resolution of the depth maps is 320×240. The dataset also provides the 3D joint positions extracted by the skeleton tracker [11]. Although the background of the dataset is clean, this dataset is still challenging due to the small inter-class variations among actions. Some actions of the dataset are shown in figure 3.

B. Evaluation of the proposed method

We evaluate our method with cross subject test setting [5][19], where the samples of the first five subjects are used

 ISBN: 978-1-61208-353-7



(a) Action Set1 (b) Action Set2 (c) Action Set3

Figure 4. Confusion matrices of the proposed method under the cross subject test setting on the MSR Action3D dataset

in training and the rest of the samples for testing. The cross subject test is more challenging and closer to the real world situation because the subjects used for training are different from those used for testing, which results in the considerable variations in the same action.

Table 1 shows the result of a comparative analysis of the proposed feature descriptors on each view and their combinations. Both the DMA and DMH show the competitive accuracy of 79.50% and 85.95%, respectively, just for the front view. It basically proves that our feature descriptors are appropriate to discriminate 3D human actions. We also achieved significant improvement on the recognition accuracy through combination of the observations from multiple views, 89.61% for the DMA and 87.64% for the DMH. This result means that reproducing new evidence from diverse views is an effective and practical approach to increase the discriminating power. We could finally obtain the outstanding recognition rate of 90.45% by combining the HOG descriptors of the multi-view DMA and DMH.

TABLE I. COMPARISON OF RECOGNITIONS RATES (%) FOR THE PROPOSED FEATURE DESCRIPTORS ON EACH VIEW AND THEIR COMBINATIONS ON THE MSR ACTION3D DATASET.

Feature Descriptors	Front view	Side view	Multi-view
DMA+HOG	79.50	69.66	89.61
DMH+HOG	85.95	70.78	87.64
DMA+DMH+HOG	85.95	71.07	**90.45**

The confusion matrices of the proposed method are illustrated in Figure 4. The recognition rates on Action Set1, Action Set2, and Action Set3 under the cross subject test setting were 92.37%, 82.35%, and 95.63%, respectively. The accuracy on Action Set2 containing many similar actions is relatively lower than those on the other two sets. The accuracies for hammer in Action Set1 and hand catch in Action Set2 are quite low compared to the other actions. This is because the way of performing these two actions varies depending on the subjects. In Action Set2, we observed that draw x, draw tick, and draw circle are mutually confused because they all have very similar trajectories of hand motion. For actions in Action Set3 in which body movements are quite different from one another, our method works very well.

C. *Comparison with the state-of-the-art methods*

We compared our approach with several previous methods. In terms of used primitives, previous 3D action recognition solutions could be categorized as 1) skeleton-based approaches that model the pose of the human body using motion of a set of 3D articulated joints [12][13][14], 2) depth map-based approaches that represent actions with volumetric and temporal features extracted from the entire depth maps in a sequence [6][15][16][17][18], and 3) hybrid solutions which combine information extracted from both the joints of the skeleton and the depth maps [19][20].

TABLE II. PERFORMANCE OF THE PROPOSED METHOD ON THE MSR ACTION3D DATASET COMPARED WITH THE PREVIOUS STATE-OF-THE-ART RESULTS.

Methods	Accuracy (%)
HOJ3D [12]	78.97
EigenJoint [13]	82.33
STOP [16]	78.20
DMM+HOG [6]	85.52
Random Occupancy Patterns [17]	86.50
Actionlet Ensemble [19]	88.20
HON4D+D_{disc} [18]	88.89
JAS+MaxMin+ HOG^2 [20]	94.84
DMA+DMH+HOG (*ours*)	**90.45**

As shown in Table 2, the proposed method clearly outperforms many well-known state-of-the-art approaches utilizing diverse primitives. It is also observed that the accuracy of our method is lower than that of one hybrid method [20] that exploits both joints and depth map information. Here, it is important to note that the goal of this work is to classify actions only using raw depth maps without additional joints information. Considering cost-effectiveness and extensibility, we believe our method has highly competitive performance.

 ISBN: 978-1-61208-353-7

V. Conclusion And Future Work

In this paper, we presented a practical and effective solution to three-dimensional human action recognition especially only using a sequence of depth maps. The method extracted a compact and discriminative HOG descriptor of the Depth Motion Appearances and Depth Motion Histories from multi-view. The experimental results on the public dataset showed that the proposed approach significantly outperformed the previous action classification methods.

As future work, we plan to investigate other descriptors based on both depth and skeleton information to manage the problem of human-object interaction and develop a dynamic classifier to reduce inter-class variations.

Acknowledgment

This work was supported by the IT R&D program. [10041610, The development of the recognition technology for user identity, behavior and location that has a performance approaching recognition rates of 99% on 30 people by using perception sensor network in the real environment]

References

[1] Z. Zhang, "Microsoft kinect sensor and its effect," Multimedia, IEEE, vol. 19, no. 2, 2012, pp. 4-10.

[2] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," Computer Vision and Image Understanding, vol. 115, no. 2, 2011, pp. 224-241.

[3] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," Pattern Recognition Letters, 2013, pp. 1995-2006.

[4] V. M. Zatsiorsky, Kinematics of Human Motion. Human Kinetics Publishers, 1997.

[5] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," Computer Vision and Pattern Recognition Workshops, 2010 IEEE Computer Society Conference on. IEEE, 2010, pp. 9-14.

[6] Y. Xiaodong, C. Zhang, and Y. Tian. "Recognizing actions using depth motion maps-based histograms of oriented gradients," Proceedings of the 20th ACM international conference on Multimedia. ACM, 2012, pp. 1057-1060.

[7] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 23, no. 3, 2001, pp. 257-267.

[8] N. Dalal, and B. Triggs, "Histograms of oriented gradients for human detection," Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, 2005, pp. 886-893.

[9] C. Chang and C. Lin, " LlBSVM: a library for support vector machines," ACM Tansactions on Intelligent Systems and Technology, vol. 2, no. 3:27, 2011.

[10] Microsoft Research. MSR Action Recognition Datasets, http://research.microsoft.com/en-us/um/people/zliu/ActionRecoRsrc/default.htm

[11] J. Shotton et al., "Real-time human pose recognition in parts from single depth images," Communications of the ACM, vol. 56, no. 1, 2013, pp. 116-124.

[12] L. Xia, C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," Computer Vision and Pattern Recognition Workshops, 2012 IEEE Computer Society Conference on. IEEE, 2012, pp. 20-27.

[13] Y. Xiaodong, and Y. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," Computer Vision and Pattern Recognition Workshops, 2012 IEEE Computer Society Conference on. IEEE, 2012, pp. 14-19.

[14] M. Raptis, D. Kirovski, and H. Hoppe, "Real-time classification of dance gestures from skeleton animation," In Proc. Of ACM SIGGRAPH/Eurographics Symp. on Computer Animation, 2011, pp. 147–156.

[15] E. Frigerio, M. Marcon, and S. Tubaro, "Improving action classification with volumetric data using 3D morphological operators," Acoustics, Speech and Signal Processing, 2013 IEEE International Conference on, 2013, pp. 1849-1853.

[16] A.W. Vieira, E.R. Nascimento, G.L. Oliveira, Z. Liu, and M.F Campos, "Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences," Progress in Pattern Recognition, Image Analysis, Computer Vision, and Application. 2012, pp. 252-259.

[17] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu, "Robust 3d action recognition with random occupancy patterns," Computer Vision–ECCV 2012, 2012, pp. 872-885.

[18] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," Computer Vision and Pattern Recognition, 2013, pp. 716-723.

[19] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," Computer Vision and Pattern Recognition, 2012, pp. 1290-1297.

[20] E. Ohn-Bar and M. M. Trivedi, "Joint Angles Similiarities and HOG2 for Action Recognition," Computer Vision and Pattern Recognition Workshops, 2013 IEEE Computer Society Conference on. IEEE, 2013, pp. 465-470.

 ISBN: 978-1-61208-353-7

Swoozy - An Innovative Design of a Distributed and Gesture-based Semantic Television System

Matthieu Deru and Simon Bergweiler

German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
Email: firstname.lastname@dfki.de

***Abstract*—In this article, we describe an innovative approach to an intelligent television system named *Swoozy* that enables viewers to discover extended information such as facts, images, shopping recommendations or video clips about the currently broadcasted TV program by using the power of technologies of the Semantic Web (Web 3.0). Via a gesture-based user interface viewers will get answers to questions they may ask themselves during a movie or TV report directly on their television. In most cases, these questions are related to the name and vita of the featured actor, the place where a scene was filmed, or purchasable books and items about the topic of the report the viewer is watching. Furthermore, a new interaction concept for TVs is proposed using semantic annotations called "Grabbables" that are displayed on top of the videos and that provide a semantic referencing between the videos' content and an ontological representation to access Semantic Web Services.**

Keywords–interactive television system; Semantic Web Technologies; Web 3.0; video annotation; gesture-based interaction.

I. INTRODUCTION

A study conducted by the German marketer for audiovisual media SevenOneMedia [1], reveals that in a viewer panel aged between 14 and 29, 45 % of them are surfing in parallel of watching television and that the main purpose of this browsing is to find out more information about the program, e.g., an actor's name or biography, a location or a depicted product. This search is likely done by either using a mobile or TV-app or by proactively typing in a keyword or complete phrase in a Web search engine.

The current development trend in interactive connected television systems is very app-oriented and forces users to install a lot of single apps, for example, one for searching videos another one for images. In cases when no suitable apps are found, users can interact with the TV's inbuilt Web browser to get additional information. Unfortunately the switch between several apps will oblige the user to leave his TV program and to interact several times with his remote controller before finally getting the information he was looking for.

The following approach discusses and shows a new way how viewers can interact with additional content while watching a TV program. They can search in parallel for information in the Web and easily browse through the found results without an interaction breach. In a first implementation, the developed prototype system relies on semantic annotations gained out of the analysis of a broadcasted video combined with gesture-based interactions that will enable users to directly start a search in the Web using Semantic Web technologies and get precise results in relation to the current scene like videos, text or news articles, pictures, and shopping recommendations.

Whereas systems like [2][3][4][5] are using the Semantic Web for detecting possible matches between the watched program and other Web-based contents and to only offer a personalized TV access, our approach uses semantics on several levels. The first level is the extraction of knowledge and concepts from an ordinary non pre-annotated Digital Video Broadcasting (DVB) signal, from a standard television provider. From this TV data stream, the required information is extracted and transferred by matching rules to annotations, which are necessary input to trigger a semantic search. Over an intuitive dedicated gesture-based graphical TV interface, presented in section IV, the viewer can then easily trigger a search using semantic queries. These queries are then finally processed by a specially designed and implemented engine called Joint Service Engine (JSE), which uses the Semantic Web, ontologies and semantic mappings to return context and domain sensitive results, as described in section V.

The prototype was implemented in form of setup box-based software solution to demonstrate the technically feasibility of a gesture based interactive television system combined with semantic processing, even if the current broadcasting infrastructures do not fully provide all annotations and information required for this task.

In section II, this paper gives an overview of existing and used Semantic Web technologies and shows how annotations and semantic information can be extracted after an audiovisual analysis of a TV signal. Section III presents in detail each implemented module, which is used during the extraction process. In section IV, the choices for the design of the user interface are motivated and the method how gesture interactions leads to a semantic search is presented. Section V, before the summary, will give an insight view on how the Semantic Web is used to query and deliver enriched multimedia results to the viewer.

II. RELATED WORK

A. Semantic Web technologies

The power of the Semantic Web (Web 3.0) [6] with its technologies resides in the fact that several information sources on the Web can be used in different combinations to establish new relations between conventional semantic representations of knowledge, such as ontologies, Resource Description Framework (RDF) triple stores [7], and common Web service interfaces in form of service mashups [8].

The World Wide Web Consortium (W3C) has declared ontologies as an open standard for describing information of

 ISBN: 978-1-61208-353-7

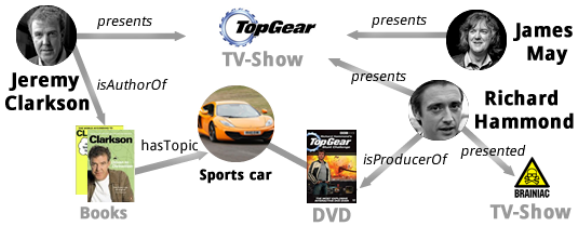


Figure 1. Discovering new semantic relations in a TV domain

an application domain and also defined appropriate ontological description languages such as RDF(S) [7][9] and OWL [10]. Ontologies, as specification languages have been specially developed for use in the Semantic Web and consists of concepts and relations. Relations organize concepts hierarchically and put them together in any relationship. These relations provide a quick access to important information in a given domain, like the biography of a presenter or speaker, interesting books or shopping items. Figure 1 shows an example of how those relations can be used to find out more information about the TV program TopGear. Starting from the TV show the three main characters, Jeremy Clarkson, Richard Hammond, and James May can be found, with further references to written books or produced DVDs. A further conclusion based on all of these relations leads to a science show named Brainiac that was also presented by Richard Hammond a few years ago.

But, in order to give viewers the access to these new relations and their contents, a relation between the video's content and its semantic representation must be established: the viewed video must be annotated or better said a mapping between what the viewer is currently seeing (e.g., *a person is speaking*) and the full scene description (e.g., *this person is a politician named Barack Obama, he is the President of the United States and is giving a speech*) along with semantic annotations must be achieved through semantic mapping. This mapping combines visual information from the current scene and ontological concepts like (person, fictional character, object, and monument). Through this assignment, extracted domain knowledge is classified [11]. This gain of knowledge out of a video can only be realized by video-based annotations: in our system we call these semantic terms.

Although several tools [12][13] and solutions exist for embedding metadata and annotations into a video - most of them are working with XML-based annotation formats like Broadcast Metadata Exchange Format (BMF) [14], Extensible Metadata Platform Format (XMP) [15], DCIM, or even MPEG-7 [16] - the core problem resides in the fact that all these metadata containing precious information are currently not transported as part of the DVB-stream, meaning that there is no possibility to reuse the semantic information of these metadata, mainly used during the production workflow. Television channels certainly could provide this semantic information over an additional interface (e.g., over a Web-based REST-API access), but unfortunately this is currently not the case.

B. Video annotation

Prior to any user interaction with the video stream, a processing mechanism is needed to be able to detect and analyze the actual video content. Here "analysis" describes the process of assigning a unique meaning to a video description and to be able to extract some key features such as who is presenting (name of show host, name of actor), the nature of the program (news, series, cartoon), the topic of the program ("Interview with", "News report", "Music Clip") and also objects or monuments along with their respective names and geo coordinates.

C. Video based analysis

The first straight forward solution is to use video and visual pattern recognition algorithms to do a pixel-based analysis of each video frame as described in [17][18][19] to get the intrinsic context [20][21] of the video (e.g., a plane is landing, or a person is speaking).

Although these approaches might be suitable, they will always need training sets [22] and computational time to consolidate the results by detecting and removing false positives and to, finally, get a fully semantically annotated video frame description [23][24][25]. The prototypical implementation uses the Open CV framework to realize the video based analysis. In order to refine the results, an additional source of information like a MPEG-2 stream is needed.

D. MPEG-2 stream-based analysis

Several types of possible additional sources of information that are embedded in the MPEG-2 stream [26][27][28][29] and used in broadcast systems like DVB were identified. As specified in [29][30], the MPEG-2 stream is delivered over DVB-T and contains several encoded tables and fields enabling contextual information the television receiver is able to decode:

- Electronic Programming Guide (EPG) information stored in the EIT table. Depending on the broadcaster, this information can be very detailed (full description of an episode including the actor's names) or very sparse: only the name of the program along with its schedule is transmitted.
- The channel's Hybrid Broadcast Broadband TV (HbbTV) endpoint URL. Usually a Web site or application URL that can be loaded and displayed by compatible television [31]. This information is extracted from the Application Information Table (AIT) [30].
- Content descriptors that are transmitted usually in form of nibbles which are 4-bit content descriptors that provide a classification of the broadcasted program type (movie, drama, news, sport).
- Teletext and closed captioning information in form of pixel tables (CLUTS) or textual information.

Depending on the country and the broadcaster's allocated bandwidth on a given frequency, the amount of content present in the aforementioned tables might vary, mostly due to the packet sizes in the transmission protocol: broadcasters will logically always privilege the image quality upon transmitting non-video related contents.

The Application Information Table (AIT) contains applications and related information that can be displayed on a

 ISBN: 978-1-61208-353-7

Figure 2. Generation process of a semantic term.

compatible receiver. Within its content descriptor loop, the AIT stores pointers to HbbTV specific information (in some cases also known as Red-Button Service). In most of the cases this pointer is an internet URL that refers to a TV-viewable Web page. By crawling this channel specific Web page additional context can be gained and extracted.

Beside the crawling and extraction of the MPEG-2 tables, another source for our semantic extraction engine is the analysis of Closed Captioning (CC) and subtitles. Subtitles and closed captions were initially introduced for the deaf community to assist them by giving a textual transcription of a scene in form of labels placed over the video. In cases like interviews or documentaries, the closed captioning is a 1:1 transcription of the narrator's spoken text.

All the textual information and extracted context information can be processed by a textual entailment [32] engine that will extract information and deliver semantic concepts and annotations.

E. Mapping of extracted information

Once extracted from the above mentioned streams, the system classifies the extracted terms into several concepts (Person, Object, Monument, etc.), organizes them ontologically (e.g., *[Person[Politician] name: Barack Obama] [isPresidentOf] [Country, name:United States of America]*) and displays them onto the user interface in form of semantic terms. Currently our system will use a classification with following categories: Person (Actor, Politician and Speaker), Object (Car, Building), Companies and fictional Characters. Figure 2 shows how extracted streams are used to generate a visual semantic term defined as *Grabbable*.

F. Audio-based analysis

While the video frame-based analysis is running, an analysis of the audio channel via speech-to-text engine can be used in order to get additional details about the content. The extracted text can then be saved or delivered as a transcript and reused for an information extraction engine. In the case the analysis of the original audio does not deliver enough information, the second possibility is to rely on the Audio Description (AD)

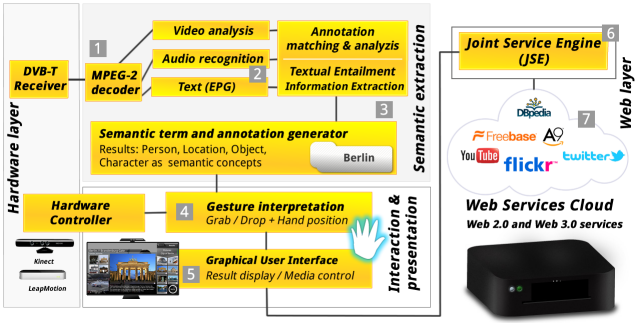


Figure 3. Architecture of the gesture-based semantic TV system.

channel. Along with the original sound of the program, an audio description provides similar to radio drama, a spoken scene description.

III. ARCHITECTURE

The implemented system prototype is based upon a setup-box plugged to a Digital Video Broadcasting Terrestrial (DVB-T) receiver, running a customized UI, and managing interaction hardware like a depth camera (Kinect), a gyration mouse or a finger tracking controller (LeapMotion Controller). The functionality of these components are represented in Figure 3.

The architecture of the prototype system is composed of several abstract processing steps. On the one hand there exists a user-hidden layer of signal analysis and evaluation, shown in the graphic as "Semantic extraction". This layer continuously performs an analysis of the DVB-T signal. As a result semantic terms are generated and can be used as input for a Semantic Web-based information search.

On the other hand, all user-visible processes are initiated by the user on the "Interaction and Presentation" level. This user-centered approach gives the viewer the possibility to access additional information in parallel to the TV program by interacting with the system via a non-disruptive gesture interaction. This gesture allows to trigger a search by simply grabbing a semantic term (e.g., an actor's name) in the system - these terms are called *Grabbables* onto a search field called *Dropzone*. This interaction can be achieved whenever the user wants to get additional information during a TV Program.

Furthermore, the "Web layer" handles the connections to Web-based content. Information from different knowledge domains can be addressed via this interface, as described in detail in Section V.

The following part lists every single processing step and task presented in Figure 3. The role of the complete solution is to:

- Display the DVB-T video signal and decode the information out of the MPEG-2/MPEG-TS stream (1).
- Analyze the MPEG-2 stream and extract information out of the tables to generate corresponding annotations for the broadcasted program (2).
- Create ontologically represented semantic terms and generate graphical equivalents in form of Grabbables (3).

 ISBN: 978-1-61208-353-7

- Interpret gesture interactions and translate them into fully formulated search queries (4).
- Use a graphical overlay principle, to enhance the user's graphical interface with additional Grabbables and multimedia annotated elements e.g., pictures, videos, or shopping items (4-5).
- Connect via Joint Service Engine to Web services, social services like Twitter, and Semantic Web Services such as Freebase and DBpedia (6-7).
- Display search results by using the interaction layer on the graphical user interface (5)

We have chosen this basis for our prototype as we are not restricted in the usage of certain APIs and have full control of both, the UI-side and the stream processing side contrary to closed proprietary solutions proposed by connected TV manufacturers.

IV. User Interface and Interaction

A. Motivation for user interface design

Although aggressively promoted by current TV manufacturers, the TV-app concept is not suitable for a quick search and browsing through the Web even less in the Semantic Web. Moreover if a Web search has to be realized directly from the television set, the painfully and frustrating typing or speaking of a keyword with a remote controller is hindering the interaction. And what happens, if the viewer does not know how to spell or pronounce the name of a building in an interesting reportage about a city? Or the viewer does not know the name of an actor, but can recall that he was starring in an American soap? Only a long search and several switches between TV-apps and the television program might help the curious and interested knowledge hungry viewer. In some cases, this problem can rapidly turn into a decisional problem, as each television broadcaster has its own app with own structures and corporate-designed interfaces leading the user to ask himself which app will be the most suitable for what he is looking for. The interaction problem is even higher when the user is zapping through several channels: must he also switch between different apps and retype his query string each time or change the context of the application manually? Unfortunately, this switching behavior brings a total interaction breach between watching the television program and getting information from the Web.

Starting from these observations, our approach tries to completely redefine the way viewers are interacting with the television by abandoning the current TV-app concept in favor of an intuitive user-centric graphical user interface.

B. User interface

The implemented graphical user interface of the created prototype system is purposely held very easy and follows all along its conception the "10 Feet Design paradigm" [33][34][35] by concentrating the efforts on having a positive trade-off between intuitive user experience, readability and easiness of interaction, so that non-computer specialists will also be able to use the system without having to cope with remote controllers and menus. Figure 4 depicts a screenshot of our

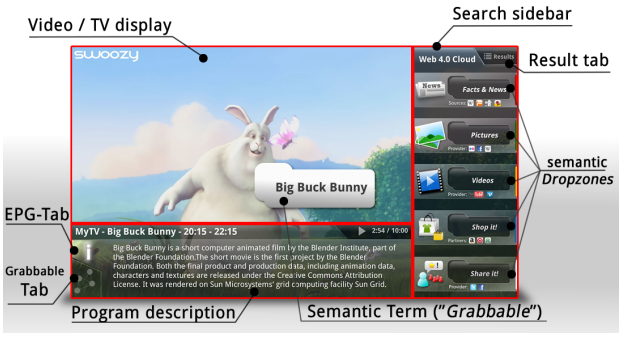


Figure 4. Screenshot of the User Interface.

current semantic television system graphical user interface. The interface consists of a graphical overlay that will be displayed over a video: in the middle of the interface, the regular television program (e.g., received over DVB) or video stream is played. On the right, the user will find a sidebar with five thematic slots (Facts & News, Pictures, Videos, Shop, Share) that internally corresponds to specific service queries. These slots are called *"Dropzones"* and they are able to receive the created semantic terms (*"Grabbables"*). Each displayed *Grabbable* can be grabbed and dropped by the user via gesture interaction. The metaphor of the *Dropzones* is an adaption of the *Spotlets* (graphical intelligent touchscreen-based search agents) mechanism - developed in a previous Web 3.0 based entertainment system [36][37][38][39].

The *Grabbable* dropped in one of the *Dropzones* is always annotated (Figure 5): this means a fact search about a person will have another internal meaning and output than an object search. When searching for facts about a person the search query is enriched by all extracted and represented information of the semantic description (first name, middle name, last name, gender, profession, etc.) which makes the search process of the connected Joint Service Engine (JSE) - described in V - more effective and precise by using better filter options. For example if the user is looking for detailed information about a building additional properties such as the location, its architecture or inauguration date can be returned as each result has a semantic visual representation. This approach follows the "no presentation without semantic representation" paradigm [40][41][42] in usage in numerous multimodal dialog systems [37]. At the bottom of the graphical user interface, the user can either choose one of the generated Grabbables (Figure 4) or switch to the traditional Electronic Program Guide (EPG) view.

This approach breaks with the philosophy of TV-Apps that every app needs its own services. In this implementation, the attached Joint Service Engine (JSE), is able to integrate different Web services, like Wikipedia, DBpedia, Freebase, Flickr or YouTube, simultaneously and it also delivers an orchestration of combined result structures. This means that the viewer will always get a unified result list, as depicted in Figure 6, where combined personal data, such as zodiac sign or portrait pictures of DBpedia and Flickr, is shown as part of the biography. In Figure 6, detailed facts about the famous football player "David Beckham" are displayed on the right side of the user's interface.

Figure 7 shows the results of a search for pictures that was

 ISBN: 978-1-61208-353-7

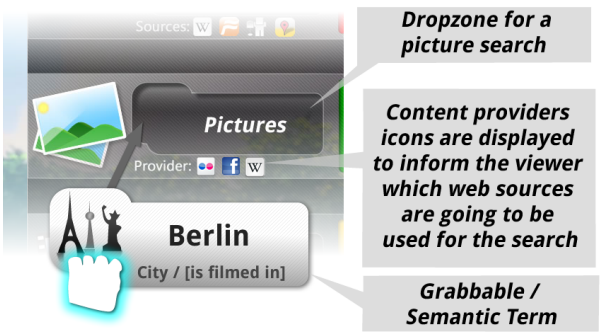


Figure 5. Close-up of a *Dropzone*.

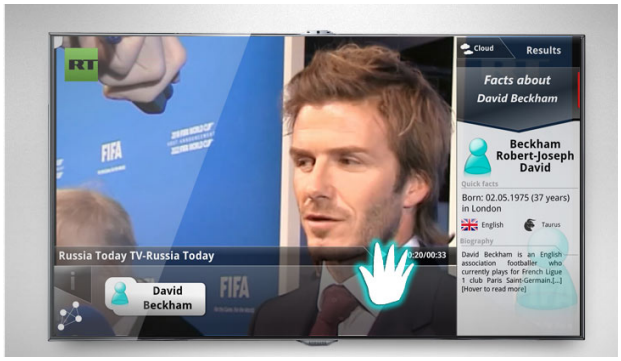


Figure 6. Display David Beckham's biography.

Figure 7. Picture request during a report about Dubai with results coming from different Web sources.

triggered by a location concept named "Dubai". The pictures are retrieved from different databases and extracted by a mashup of Web services (Flickr, Wikipedia and Freebase)

C. Interactions by gestures

Following the same principle of simplicity and easiness of use, we have inbuilt the possibility for the user to interact with the system over gestures: the user only needs to move his hand towards the television screen. At this precise moment, a virtual hand is displayed (Figure 8). The position of the hand can be either tracked over a depth camera like the Microsoft Kinect, or for smaller living rooms by using a finger tracking solution, like the LeapMotion controller device [43].

We have deliberately implemented only two gesture types: *Grab'n'drop* and the *Push*-gesture, as these interactions are

Figure 8. User gesture interaction: a virtual hand allows the user to grab out a semantic term from a video.

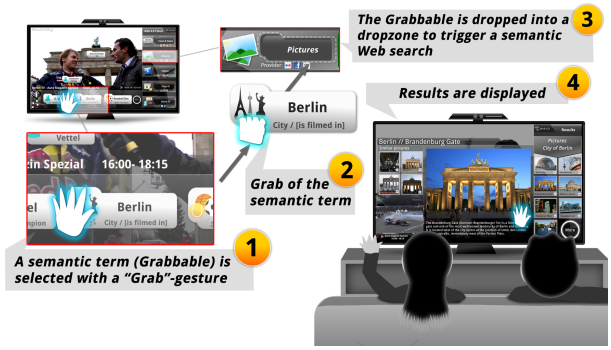


Figure 9. Grab'n'drop interaction steps to start a picture search for the city of Berlin.

simple to realize and do not need a specific user training and do not cause fatigue over time. The *Push* interaction is needed to make a selection and is a simplified metaphor of the traditional mouse click.

Figure 9 describes the interaction workflow. Step #1 shows how the user can grab a semantic term (*Grabbable*) from a sport report featuring Sebastian Vettel during a car show in front of the Brandenburg Gate in Berlin. In our case, the user has selected the term "Berlin" that is internally represented as a location with geo coordinates.

The user now would like to look for pictures of "Berlin". To achieve this, she will take the *Grabbable* (Step #2) and drop it into the Picture *Dropzone* (Step #3); within a few seconds first results coming from the Semantic Web are displayed in form of push-able elements in the right side bar (Step #4).

Beside the easiness of usage of such a system through gesture interaction, the main originality resides also in the fact that without having to type on a keyboard, or to start an additional app, any viewer will be able to rapidly get facts, video or even shopping recommendations during his favorite TV program.

D. Mobile client application

With the mobile application of our approach, depicted in Figure 10 - the mobile Swoozy App (for Android and iOS) -

 ISBN: 978-1-61208-353-7

Figure 10. Swoozy - mobile client application.

multiple users can simultaneously view the same TV program but interact with their own device in parallel. If viewers like to share interesting videos, pictures or facts with the other viewers, they can use the simple "sling-gesture" on their mobile device to transfer these interesting results to the TV with its large display, similarly to the 3D frisbee interaction approach presented by Becker et al. [38], where multimedia content is transferred from mobile devices to a kiosk system.

V. RETRIEVAL OF FACTUAL KNOWLEDGE BASED ON SEMANTIC TECHNOLOGIES

According to the system design, the viewer is supplied with new facts, pictures and videos while watching TV. Therefore, it is absolutely essential to access external sources and to quickly find information that exactly matches to the shown scenery. The presented approach uses a combination of techniques of the Semantic Web to create matching answers, whereas a composition of standard Web services and services of the Semantic Web is serving as knowledge source. However, the heterogeneous aspect of the services and their different Application Programming Interfaces (APIs) represents a challenge for building a correct query and retrieving matching contents. The latter must be adapted in an additional step, so that they can be correctly displayed onto the user's interface.

A. Motivation

As mentioned at the beginning, the video, audio and text analysis extracts knowledge concepts and adds them to predefined ontological structures which can define persons, fictional characters, objects or locations. By the procedure described in this approach and with these prepared input structures, the viewers are able to trigger queries to Web services or Semantic Web Services over simple gesture interaction without the need of special skills, such as programming Web services API or the need of specific database query languages or an RDF(S) query language like SPARQL [44][45]. For non-specialists it would be very hard to formulate such queries. These query languages are primarily used to access the full power of the Semantic Web by allowing a navigation through semantically annotated data sets and enabling the search for instances that corresponds to a given request.

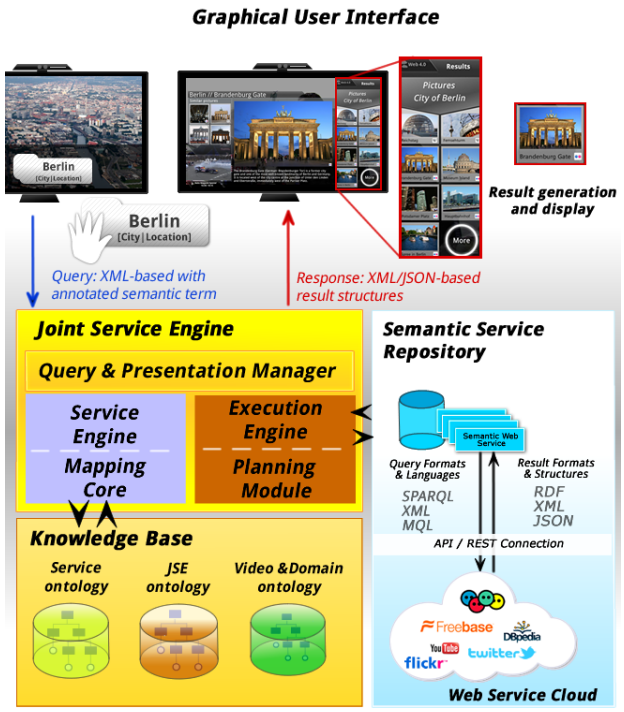


Figure 11. Architecture of the Joint Service Engine.

We assume that the typical viewer does not want to explicitly formulate his search in one of the above-mentioned query languages. That is why the search will be done in the background by using semantically annotated data sets that will be then mapped to the dropped *Grabbable*.

B. Retrieving semantic content

In order to start a search with a *Grabbable*, a dedicated engine was implemented to better solve the tasks of calling heterogeneous services and providing unified semantic results. This engine called *Joint Service Engine* (JSE) is involved in the retrieval of semantic content. The basic idea of the JSE is to use the joint potential of different services to focus information and knowledge. It provides and manages semantic descriptions of various pre-annotated information sources in a local *Semantic Service Repository* that opens up access to sources of different domains. This question answering component internally realizes a judicious orchestration and mashing up of Web 2.0 and Web 3.0 services and provides aggregated results coming from several sources - Web services - as a final result. This result is returned to the client and displayed on the respective user interfaces (television UI and second screen app).

One advantage of this component is that new sources can be added, removed or replaced without hard programmatic dependencies and without stringent dependencies on specific providers of information and their interfaces. Figure 11 shows a specific overview of the architecture design of the JSE.

C. Query processing

The "Query" module of the *Service Engine* [46] retrieves and decomposes the user's query. The produced query structures are formulated according to a terminology defined by domain

 ISBN: 978-1-61208-353-7

```
search topic:
generic concepts = {object (car, building),
                    person (actor, speaker, ..),
                    company,
                    location,
                    fictional character}
query-for:
  similar videos AND/OR pictures
  personal facts AND pictures AND/OR videos
  location AND/OR pictures AND/OR videos
  object facts AND pictures AND/OR videos
  shopping facts AND pictures AND videos
  sharing facts AND pictures AND videos

given properties:
  // depending on concept type
  {first name, middle name, last name, title},
  {gender, profession},
  {characterizing keywords},
  {geo-data (latitude, longitude)},
  {city-name, country-name}
  {building-name}
  {company-facts, company-name, keywords}
```

Figure 12. Query search topics and properties.

ontologies and expressed using a generic template-based query structure as shown in Figure 12. Each individual decomposed query part is mapped to a local meta-representation, the JSE ontology, modeled in OWL [10]. According to the user's query, basic ontological components like individuals are created based on the defined vocabulary of the JSE ontology and the planning module looks for adequate plans that fulfill all of the requested properties. The resolving internal query specifies the *input type* (object, person, fictional character, company, location) specified by *properties* (complete name, keywords, etc.) and *implicit relations* and *search topics* (similar pictures, shopping facts, etc.).

One crucial point in this scenario is the discovery and execution of services. This task is executed by an execution plan which describes the discovery process by specifying which type of services are needed, what kind of domain is addressed, in which order the services have to be executed and all the requirements needed for the matchmaking process occurring in the connected *Semantic Service Repository*. Results of the matchmaking process are ordered lists with adequately ranked information sources. The sequence of individual service calls that must be executed are listed in a scheduling table that needs to be processed by the *Planning Module* and the *Execution Engine*. The *Execution Engine* provides connectors and encapsulates the calls to the REST or API interfaces, by reformulating and using specific query formats like XML or languages, like SPARQL and Metaweb Query Language (MQL). Once all results of different called services are received by the *Execution Engine* an internal mapping process starts a review and reasoning process with the help of additional semantic mapping rules and classifies the results according to the internal JSE domain ontology.

D. Mapping and matching

During the processing chain of the JSE, the content of the described data channels must be repeatedly transformed from one data format to another. The most important step for creating comparable and interoperable data models is the definition of mapping functions between the used concepts. Therefore, the identified data structures must be mapped based on stored mappings that have been defined in a pre-processing phase in a formal description language. For an unambiguous assignment of the models and types of a described element, the mapping functions are specified by categories. The *Mapping Core* achieves these mappings in each component of the JSE. In the "Query" module the user's query input description is mapped to the internal domain ontology that is used for further processing in the planning process. Additionally in the "Execution" module, a mapping of the results of the called external sources to the internal JSE ontology must be fulfilled. Moreover, the spectrum of results of external sources varies from simple XML or JSON structures to complex semantic data structures. In this specific case, formal mapping rules are used to allow a higher quality data type mapping on a more generic level: new instances can be created and linked to each other. Alternatively, a taxonomy of objects can be mapped according to the internal data structure.

E. Service repository

The *Semantic Service Repository* provides access to different types of information sources like Semantic Web Services that cover information stored in external database management systems or Semantic Repositories. In the development of concepts and prototypical implementation [46] detailed service descriptions in OWL-S [47], of freely available sources of knowledge, such as DBpedia, Freebase, Flickr, were integrated. In these Web-based systems information is stored in a structured and manageable form, but can only be accessed by special query languages like SPARQL for DBpedia or MQL in the case of Freebase. The main difference of this approach, compared to conventional database management systems, is the usage of ontologies as a technology to harmonize and store semantically structured data: each concept defines and classifies information and also adds implicit knowledge characterized by its name and position in a hierarchy or taxonomy [46].

The JSE closes the gap between pure RESTful service calls and factual knowledge extracted from Semantic Web Services like Freebase or DBpedia, by mapping results and their respective annotations syntactically and semantically well-defined to a domain ontology.

F. Output presentation

The last step of the processing is done by the *Presentation Manager* which will encapsulate and transform the semantic annotations in a standardized result structure. The contents of the delivered result structures are displayed on the graphical user interface (television screen) after a parse process. Depending on the user's query, e.g., a media search, other different structured output formats (RDF, XML, JSON, etc.) can be served by the *Presentation Manager* module. This module uses filter rules and generic declarative element-based mapping techniques to create the resulting structures from the internal domain ontology and returns these structures to connected client platforms. This procedure allows a parallel distributed output: both second screen and television systems are fed with the results coming from the *Presentation Manager*. With this parallel output processing a cross-media interaction is possible.

 ISBN: 978-1-61208-353-7

VI. Conclusion and future work

With *Swoozy*, the prototypical implementation of this approach, we demonstrated that, through a seamless combination of gesture-based interaction, video information coming directly from the broadcaster and the Joint Service Engine, it is possible to provide a novel way to interact with video contents. Through this approach, television enters into a new dimension in which viewers will receive additional information and knowledge about the persons, locations, objects featured in a video or a television program.

The *Swoozy* concept is not only applicable for the sole field of television, but can also be used for other video-based systems such as interactive e-Learning systems, video casts or even online university courses, where the semantic terms would be mathematical formulas or technical concepts.

We believe that the concept of semantic television will turn television into an appealing and ludic knowledge provider and will give a brand new dimension to interactive connected television systems in the future. Moreover in addition to the input modalities (Microsoft Kinect and LeapMotion controller) used in *Swoozy*, we consider extending our gesture-based approach to SmartWatches.

References

[1] SevenOneMedia, "HbbTV macht TV clickbar," 2013.

[2] L. Aroyo, L. Nixon, and L. Miller, "NoTube: the television experience enhanced by online social and semantic data," in Consumer Electronics-Berlin (ICCE-Berlin), 2011 IEEE International Conference. IEEE, 2011, pp. 269–273.

[3] Y. B. Fernandez, J. J. Pazos Arias, M. L. Nores, A. G. Solla, and M. R. Cabrer, "AVATAR: an improved solution for personalized TV based on semantic inference," Consumer Electronics, IEEE Transactions on, vol. 52, no. 1, 2006, pp. 223–231.

[4] J. Kim and S. Kang, "An ontology-based personalized target advertisement system on interactive TV," Multimedia Tools and Applications, vol. 64, no. 3, 2013, pp. 517–534.

[5] B. Makni, S. Dietze, and J. Domingue, "Towards semantic TV services a hybrid semantic web services approach," 2010, [Retrieved: July 2014].

[6] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, 2001, [retrieved: July 2014]. [Online]. Available: http://www.jeckle.de/files/tblSW.pdf

[7] G. Klyne and J. J. Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax," W3C Recommendation, 2004. [Online]. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[8] P. Hitzler, M. Krötzsch, S. Rudolph, and Y. Sure, Semantic Web: Grundlagen. Springer Berlin Heidelberg, 2008.

[9] R. Cyganiak, D. Wood, and M. Lanthaler, "RDF 1.1 Concepts and Abstract Syntax," W3C Recommendation, 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

[10] P. F. Patel-Schneider, P. Hayes, and I. Horrocks, "OWL Web Ontology Language Semantics and Abstract Syntax," Feb. 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

[11] M. C. Surez-Figueroa, G. A. Atemezing, and O. Corcho, "The landscape of multimedia ontologies in the last decade," Multimedia Tools and Applications, vol. 62, no. 2, 2013, pp. 377–399.

[12] M. Lux, W. Klieber, and M. Granitzer, "Caliph & Emir: semantics in multimedia retrieval and annotation," in Proceedings of the 19th International CODATA Conference. Citeseer, 2004, pp. 64–75.

[13] M. Lux and M. Granitzer, "Retrieval of MPEG-7 based semantic descriptions," in In Proceedings of BTW-Workshop WebDB Meets IR, 2004.

[14] Institut für Rundfunktechnik, "Broadcast Metadata Exchange Format," BMF 2.0, 2012, [retrieved: July 2014]. [Online]. Available: http://bmf.irt.de/

[15] Adobe, "XMP - Adding intelligence to media," 2012, [retrieved: July 2014]. [Online]. Available: http://www.adobe.com/devnet/xmp.html

[16] J. Martnez, R. Koenen, and F. Pereira, "MPEG-7: the generic multimedia content description standard - part 1," Multimedia, IEEE, vol. 9, no. 2, 2002, pp. 78–87.

[17] S. Bloehdorn et al., "Semantic Annotation of Images and Videos for Multimedia Analysis," in The Semantic Web: Research and Applications, ser. Lecture Notes in Computer Science, A. Gómez-Pérez and J. Euzenat, Eds. Springer Berlin Heidelberg, 2005, vol. 3532, pp. 592–607.

[18] E. Sgarbi and D. L. Borges, "Structure in soccer videos: detecting and classifying highlights for automatic summarization," in Progress in Pattern Recognition, Image Analysis and Applications. Springer, 2005, pp. 691–700.

[19] W. Shao, G. Naghdy, and S. Phung, "Automatic image annotation for semantic image retrieval," in Advances in Visual Information Systems, ser. Lecture Notes in Computer Science, G. Qiu, C. Leung, X. Xue, and R. Laurini, Eds. Springer Berlin Heidelberg, 2007, vol. 4781, pp. 369–378.

[20] L. Ballan, M. Bertini, and G. Serra, "Video Annotation and Retrieval Using Ontologies and Rule Learning," IEEE MultiMedia, vol. 17, no. 4, 2010, pp. 80–88.

[21] U. Arslan, M. E. Dönderler, E. Saykol, O. Ulusoy, and U. Güdükbay, "A Semi-Automatic Semantic Annotation Tool for Video Databases," in Proc. of the Workshop on Multimedia Semantics (SOFSEM 2002). Milovy, Czech Republic, ser. SOFSEM-2002, 2002, pp. 1–10.

[22] G. Quénot, "TRECVID 2013 Semantic Indexing Task," 2013.

[23] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 32, no. 9, 2010, pp. 1627–1645.

[24] C. Snoek, D. Fontijne, Z. Z. Li, K. van de Sande, and A. Smeulders, "Deep Nets for Detecting, Combining, and Localizing Concepts in Video," 2013.

[25] L. J. Li, H. Su, L. Fei-Fei, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in Advances in neural information processing systems, 2010, pp. 1378–1386.

[26] T. Dajda, M. Cislak, G. Heldak, and P. Pacyna, "Design and implementation of the electronic programme guide for the MPEG-2 based DVB system," 1996.

[27] C. Peng and P. Vuorimaa, "Decoding of DVB Digital Television Subtitles," Applied Informatics Proceedings - No.3, 2002, pp. 143–148.

[28] M. Dowman, V. Tablan, H. Cunningham, C. Ursu, and B. Popov, "Semantically enhanced television news through web and video integration," in Second European Semantic Web Conference (ESWC2005). Citeseer, 2005.

[29] "Digital Video Broadcasting (DVB) Subtitling systems," European Standard ETSI EN 300 743, European Broadcasting Union, 2014.

[30] "Specification for Service Information (SI) in DVB systems," European Standard ETSI EN 300 468, European Broadcasting Union, 2014.

[31] K. Merkel, "HbbTV - Status und Ausblick," 2012, [retrieved: July 2014]. [Online]. Available: http://www.irt.de/webarchiv/showdoc.php?z=NTgwNyMxMDA1MjE1I3BkZg==

[32] R. Wang and G. Neumann, "Recognizing textual entailment using sentence similarity based on dependency tree skeletons," in Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, ser. RTE '07. Stroudsburg, PA, USA: Association for Computational Linguistics, 2007, pp. 36–41.

[33] R. Cardran, K. Wojogbe, and B. Kralyevich, "The Digital Home: Designing for the Ten-Foot User Interface," 2006, [retrieved: July 2014]. [Online]. Available: http://channel9.msdn.com/Events/MIX/MIX06/BTB029

[34] Samsung, "Design Principles for Creating Samsung Apps Content," 2013, [retrieved: July 2014]. [Online]. Available: http://www.samsungdforum.com/UxGuide/2013/01_design_principles_for_creating_samsung_apps_content.html#ux-01

 ISBN: 978-1-61208-353-7

[35] D. Loi, "Changing the TV industry through user experience design," in Design, User Experience, and Usability. Theory, Methods, Tools and Practice, ser. Lecture Notes in Computer Science, A. Marcus, Ed. Springer Berlin Heidelberg, 2011, vol. 6769, pp. 465–474.

[36] D. Porta, M. Deru, S. Bergweiler, G. Herzog, and P. Poller, "Building multimodal dialogue user interfaces in the context of the internet of services," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, pp. 149–168.

[37] D. Sonntag, M. Deru, and S. Bergweiler, "Design and implementation of combined mobile and touchscreen-based multimodal web 3.0 interfaces," in Proceedings of the International Conference on Artificial Intelligence, ser. ICAI-09, July 2009, pp. 974–979.

[38] T. Becker et al., "A unified approach for semantic-based multimodal interaction," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, pp. 135–148.

[39] S. Bergweiler, M. Deru, and D. Porta, "Integrating a Multitouch Kiosk System with Mobile Devices and Multimodal Interaction," in Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces, ser. ITS-2010, ACM. 1515 Broadway New York, New York 10036: ACM, 2010.

[40] W. Wahlster and A. Kobsa, "User models in dialog systems," in User Models in Dialog Systems, ser. Symbolic Computation, A. Kobsa and W. Wahlster, Eds. Springer Berlin Heidelberg, 1989, pp. 4–34.

[41] A. Kobsa, "Generic User Modeling Systems," User Modeling and User-Adapted Interaction, vol. 11, no. 1-2, 2001, pp. 49–63.

[42] N. Reithinger et al., "A look under the hood: design and development of the first SmartWeb system demonstrator," in Proceedings of the 7th international conference on Multimodal interfaces. ACM, 2005, pp. 159–166.

[43] F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler, "Analysis of the Accuracy and Robustness of the Leap Motion Controller," Sensors, vol. 13, no. 5, 2013, pp. 6380–6393.

[44] "SPARQL query language for RDF," W3C Recommendation, 2008, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[45] "SPARQL 1.1 query language," W3C Recommendation, 2013, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[46] S. Bergweiler, "Interactive service composition and query," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, p. 480.

[47] D. Martin et al., "OWL-S: Semantic Markup for Web Services," W3C Submission, 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/

 ISBN: 978-1-61208-353-7

Augmented Object Development using 3D Technology

An Object Redesign Process in the Ubicomp Domain

Gustavo López, Daniel Alvarado
Research Center on Information and Communication Technologies
University of Costa Rica
San José, Costa Rica
{gustavo.lopez_h,daniel.alvarado_g}@ucr.ac.cr

Luis. A Guerrero, Mariana López
School of Computer Science and Informatics
University of Costa Rica
San José, Costa Rica
{luis.guerrero, mariana.lopez}@ecci.ucr.ac.cr

***Abstract*— The construction of augmented object´s prototypes is a difficult process because most of the time these prototypes contain several sensors embedded in real life objects. In this paper, we present our experience using 3D technologies (scanning and printing) to generate prototypes of augmented objects which allow an early evaluation of them. Three different objects with embedded sensing and actuating capabilities were designed and developed. We discuss several pros and cons of applying this technology when developing augmented objects. Our results show that the use of 3D technologies enhances the quality of the final prototypes and allows a better concept validation and evaluation.**

Keywords- 3D printing; fast prototyping; augmented objects; ubiquitous computing; HCI; usability; mixed reality.

I. INTRODUCTION

An augmented object is defined as a common object, which has been provided with additional functionalities through integrated computing or software systems [1]. This definition biases people to build an augmented object from an existing one by adding other capabilities. In contrast, Waiser [2] describes Ubicomp as invisible computers in different sizes, situated to specific tasks.

Augmented Objects are normally developed to be incorporated in the Ubicomp domain. Therefore, they should be as seamless as possible depending on the goal of the objects, i.e., the capabilities added to the object. If the object to be augmented is selected from the natural context of use and additional capabilities are added, those capabilities should change it as little as possible. To do so the external look and functionalities of the object must be studied in order to add the capabilities without changing it.

In this project, we propose a strategy to develop augmented objects using 3D technologies, i.e., using a 3D scanning of the original object, redesigning it and printing a 3D version of such object with the least visible changes.

We also evaluate the proposed strategy by building versions of objects with embedded capabilities changing the object only slightly and almost imperceptibly to users.

Our main goal is to provide a strategy to allow the construction of new versions of objects as if the embedded capabilities were already there in the original design.

In a previous work, a process for the development of augmented objects (called AODeP) [3] was proposed and was later validated and improved [1].

However, the construction process was not fully addressed. This project defines some guidelines that can be followed at that stage.

Using augmented object as non-traditional interfaces for software systems can be difficult, especially when designers are reluctant to undergo a learning process. Our approach deals with that problem by designing familiar interfaces for the target users that resemble and function like the already known objects.

We add new functionalities to an object requiring the least cognitive effort possible from the user. Our approach intent to understand the object and its capabilities to assure that the new capabilities engage well with the original ones.

According to Hollinworth and Hwang [4], avoiding an increase in cognitive effort is achieved through the construction of physical objects that resemble or behave in a familiar way and context of the final user.

Implementing interfaces with familiar designs, and that provides more comfortable environments, results in better acceptance from final users. It also helps to reduce the cognitive effort and the time spent by the user to adapt to the device or system [5].

Furthermore, the use of 3D printing to make prototypes and construct real size devices in order to evaluate aesthetics, space distribution and functionality may also reduce the cost of production of the prototypes [6]. However, the use of low cost 3D printers could affect the possibility to print large objects.

Sophistication and appearance need to be addressed when creating augmented objects [7]. It is no longer enough to attach some sensors on an artifact and name it an augmented object. It is necessary to redesign the objects and really embed the new characteristics.

We use Human Computer Interaction (HCI) techniques to enhance the objects capabilities and evaluate them. HCI involves the study and design of interactions between people and computers through joint work of computer science and behavioral sciences. The main goal of HCI is to improve the interactions to create more usable interfaces [8][9].

The rest of the paper is organized as follows: Section II shows some related work. Section III explains the construction process when embedding the new capabilities in objects. Section IV shows three examples of developed augmented objects. Section V discusses our findings. Finally, Section VI shows some conclusions.

 ISBN: 978-1-61208-353-7

II. RELATED WORK

Augmented objects are also known as smart objects [10] or sentient artifacts [11]. Sánchez, Ranasinghe, Patkai and Mcfarlane [10] call smart object to a product which is capable of incorporating itself into both physical and information environments. Meanwhile, Kawsar, Fujinami and Nakajima [11] refer to sentient artifact to everyday life objects augmented with sensors to provide value added services.

Many augmented objects examples are present in the literature. However, most of these examples were all developed without following a formal design and development process and there is no special attention given to hiding the sensors or actuators.

Since the appearance of new prototyping and modeling techniques for creating custom devices, HCI researchers have been developing new forms of using this technology to improve the user experience. Due to the development of these technologies, the design processes of hardware components can now focus not only on the internal electronic components of the device but also on the external components: functionality and aesthetics [12][13]. Another sought characteristic among the researchers of these kinds of techniques is the affordability and accessibility of the components that can be obtained through custom construction.

Groups like Microsoft Research have been studying and experimenting with several physical prototyping techniques, combining different methods of generating structures like 3D printing and laser-cut wooden components [14]. These techniques provide great opportunities to integrate the design of the inner parts of a device, e.g., circuits or sensors with the design of the outer part of it, like shape, buttons and colors.

The applications and advantages of custom component fabrication go beyond customer satisfaction or market driven design. It also benefits areas like education [15] because students can build and prove their designs in a real context.

Some research groups have also been focusing in the inner parts or functionality of smart objects and tangible and more intuitive interfaces. However, authors use an approach that does not incorporate sensors. Instead of that they use conductive and capacitive interactions as a way to provide low cost prototypes with interaction capabilities [16].

Wiethoff et al.[17] propose a method to quickly prototype tangible user interfaces without having a complete design or process of construction. This is achieved by making cardboard structures for the external parts and using conductive ink to design the tangible surface of the objects.

Some research is also being made on 3D scanning quality and the influence of diverse factors, e.g., surface color, glossiness, ambient light, resolution in the scanning process and results quality [18][19].

3D Scanning and printing has been used to potentiate the transformation design practices. Stanislav and Chyon [20] demonstrated that 3D scanning and printing process enables a rediscovery of the artifacts when crafting. Authors also project the possibility of this technique used in redesign process by designers.

III. CONSTRUCTION PROCESS

The construction process detailed in this section is an in-depth description of one of the steps proposed by Guerrero, Ochoa and Horta [3], in what they called Augmented Object Development Process (AODeP).

AODeP is a method that ensures that all stages in developing augmented objects are addressed focused on the problem to solve and from an engineering perspective.

AODeP was addressed in other studies [1] and some improvements were proposed however no details on the step five were presented. AODeP proposes six main stages to develop an augmented object:

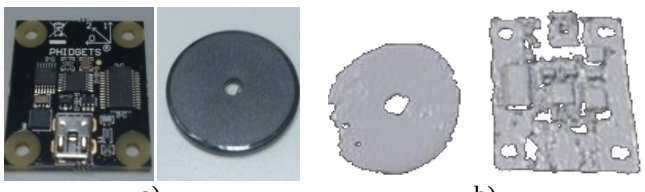
1. Problem definition
2. Context of use definition
3. Requirement definition
4. Selection of the object
5. Development
6. Testing with users

This paper addresses the fifth step of AODeP process. The development of an augmented object, assuming that a problem has been identified and requirements have been gathered; this is the next stage of the process.

Several guidelines are available to conduct a prototype development in HCI literature. However, not many people address the problem of how to augment an object. If an object has been selected for augmentation the idea is to embed the sensors or actuators in it, while changing as little as possible its external looks and of course the original functionality of such object.

We propose a Physical to Digital to Physical (P2D2P) process. Using a 3D Scanner the original object can be digitalized to be redesigned in order to make sure the new components fit the interior while changing the external looks and functionality the least possible. Once the object is redesigned, a 3D print process is used to obtain the new version of the object with the actual size of the original object with the new components.

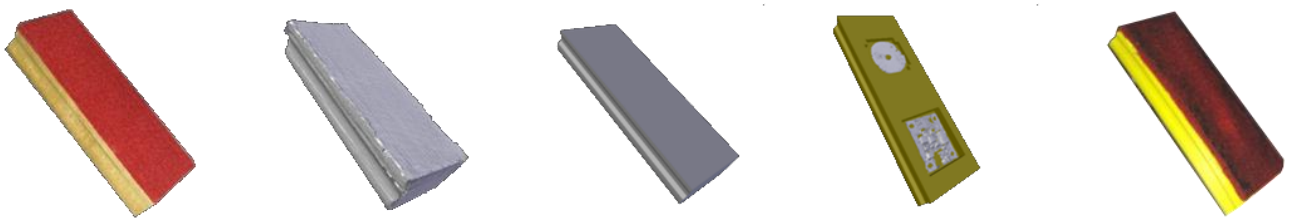
In this project, we use a NextEngine 3D Scanner HD to obtain the digitalized model. It is necessary to do at least two 360° scans to get a clear model of the original object. The software we used was NextEngine ScanStudioHD. This method creates a virtual representation of the surface geometry. If there are components inside of the original object, such components should be also scanned in order to leave the required space for them in the new version of the object. Once the object is scanned, we export it in OBJ format and import it using Blender 3D computer graphics software [21].



a) b)

Figure 1. a) Pictures of the sensors to be embed in the augmented object. b) Low quality 3D models of the sensors.

 ISBN: 978-1-61208-353-7



a) b) c) d) e)

Figure 2. a) Photo of the original object, b) Digitalized model of the original object, c) Simplified model, d) Redisigned model, e) Printed object

Using this tool we create a simplified model in order to avoid manifold edges, flipped normals or face overlapping. With the new simplified model of the object, the redesign process starts. This process needs to be addressed individually for each object since the object characteristics and the new components are specific to a given problem. For instance, we want to fit a RFID emitter and an accelerometer, both from the Phidgets kit [22], into a board eraser. Figures 1.a and 1.b show the sensors that we want to embed into the eraser. Figure 1.c shows the scanned digital representation of the sensors that would be used to measure and leave the required space inside the new object.

Once the object has been redesigned and the space for the new characteristics has been added, the replication (printing) process begins. The new model is exported in a STereoLithography (STL) format.

To process the STL file we used Replicator G [23] for Windows. Replicator G is the software that drives many computer numerical controlled machines. The model exported from the redesign process is positioned, scaled or rotated as needed to accommodate it in the print bed.

The Flashforge Creator printer is used to print the new object. It uses open source controlling software and ABS plastic filaments. It also provides an accuracy of 0.0025mm on Z and 0.011 on XY axis. The cost for using the printer is US$0.003 per printed inch of filament.

Figure 2 shows the whole process followed:

1. Digitalize the object using the 3D scanner.
2. Simplify the model reducing vertex count and avoiding problems in the model.
3. Redesign the model to fit new and old components.
4. Print and polish the final version of the new object.

Figure 2.a shows the original object to be digitalized, 2.b is the digitalized model without any modification, 2.c shows the simplified model from the digitalized one, 2.d is the redesigned model and finally, Figure 2.d shows the printed object.

IV. EXAMPLES

This section presents three examples of augmented objects we built using our P2D2P process. We redesign the objects in order to incorporate sensors and actuators to the environment changing few physical aspects of the real objects and context. All the prototypes presented in this section have a software component. However, we will briefly discuss it in this paper since our main goal is to describe the physical redesign of the objects.

A. *First augmented object: An automatic post-it note*

Email service has been developing for the past 40 years. In the past few years many companies, especially phone builders, have been developing nontraditional interfaces for email services and special notification systems to let people know they have received an important email. It does not matter if the mail is used for personal or business reasons there are always some emails that are more important than others and of course there is also spam mail.

We built an augmented object to notify a user when they have an unread important mail in their inbox. The metaphor that we decided to use was a post-it note, because normally, when you missed an important message, this kind of note is used to let you know about it.

The context of use selected was an office setting and the post-it note is displayed on top of the monitor of the computer, it could be on the keyboard or anywhere else.

The developed system consists of two software parts. One part is a Web page for configuration and the other a daemon that samples the inbox of the configured email account and determines if there is an important unread mail, if so, it sends a signal that displays the post-it note.

The case for this augmented object was printed in order to fit the needed components and so that it would disturb the least possible the context of the post-it. If an unread important mail is detected, a signal is sent and the servo contained in the case shows the post-it, otherwise it hides the note. Since we used general purpose components as sensors, the size of the prototypes are bigger than they would be if more specialized sensors were used.

The physical components of the post-it note are: (1) a servo motor, (2) a servo controller, (3) USB cables and power cable, (4) the note, (5) case and (6) case cover. Figure 3 shows the prototype set in angle to visualize the case and components.

We incorporate an HCI technique in order to validate the experience of use for this prototype. We conducted a Diary Study by asking selected participants to keep a diary of usage over 8 days.

The importance of this study was that the non-intrusive design of the object allowed the participants to keep using the original object as if a regular post-it note was there. The purpose of the notification is achieved when the participant observes the post-it note and realizes he/she got an important unread mail. The main complaint of the participants in the study was that the prototype was wired and that makes it intrusive to the environment.

 ISBN: 978-1-61208-353-7

a) b) c) d)

Figure 3. a) Front picture of the post-it displayed, b) Back picture of the post-it hidden. c) Angle picture of the case and post-it displayed, d) Angle picture of the post-it hidden.

We intent to perform one more iteration with the prototype to create a wireless version. Of course we were not able to redesign the full monitor; so, we developed this prototype trying to have a non-intrusive post-it note without affecting the monitor's functionality.

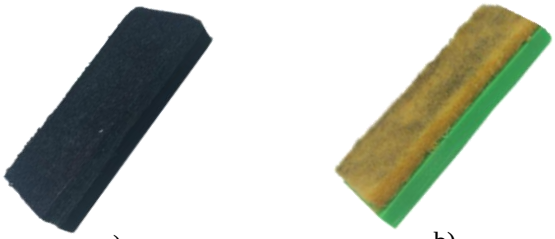
B. *Second augmented object: A whiteboard eraser*

One of the biggest problems that ubiquitous computing tries to solve is the automation of environments through what is called ambient intelligence. For instance, we designed an augmented whiteboard eraser to automate a classroom. This device is an example of overcoming the problematic of solid objects when trying to embed sensors or actuators.

Normally, when prototyping the developers attach the sensors on the outside of the objects but this may make them difficult to use, or can change the physical appearance of the object. We found this problem when trying to use a whiteboard eraser for recognition of the professor's presence in a classroom and to know if the whiteboard is being used. The context of use of this prototype is a university setting.

Common whiteboard erasers have two parts: a base and a plush. The base is normally made of wood or some kind of styrofoam and these materials are not pliable enough to insert sensors. On the other hand, the modification of the plush could jeopardize the functionality of the eraser. Therefore a redesign is needed to embed the sensors.

The physical components added to the augmented eraser were: (1) a Radio-frequency identification (RFID) transmitter; (2) a 3 axis sensor; (3) a plastic base; and (4) the plush. Figure 4 shows the original eraser and the printed version. There are no external differences although the printed version contains the sensors. An RFID receiver is on the classroom entrance and on the marker holder on the whiteboard.



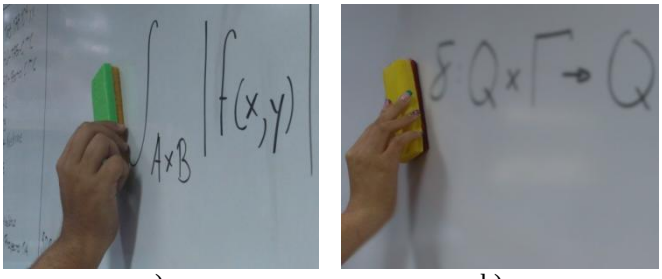
a) b)

Figure 4. a) Picture of the physical eraser to be augmented. b) Augmented Object developed using a 3D printer.

Depending on the eraser position the classroom changes settings, e.g., turning the lights on or off.

To evaluate the structure and functionality of the eraser we gave it to five faculty members of the School of Computer Science and Informatics at University of Costa Rica. They just used it as a common eraser and did not know anything about the object or the research goals. Users held it and used it in a common environment and were not able to find any difference between the regular eraser and the augmented one.

One important result of the interviews was that none of the interviewed persons were able to distinguish that there were sensors or anything different about the eraser. One issue that was derived from the questioning of the faculty members was that the erasers are commonly disposable so we should provide exchangeable plushes if we want to continue to use the base and sensors.



a) b)

Figure 5. a) Small printed eraser used in context. b) Big printed eraser used in classroom.

Figure 5 shows the use of the printed version of the eraser. We used two different models to consider external complexity. Figure 5.a shows a rectangular eraser, and figure 5.b shows a more complex 8 shaped eraser.

C. *Third augmented object: A real size classic doodler*

With the emergence of interactive video beams such as eBeam products [24], most of the interaction is being done by software and clicking in some places of the projection area. In this study, we used a Classic Doodler metaphor [25] to create a real size classic doodler that would allow the user of the interactive video beam to erase a part of the projected scene by using a little slider placed on the marker holder of the whiteboard used to project.

 ISBN: 978-1-61208-353-7

Figure 6. Front Picture of the printed doodler.

Figure 7. Picture of the Doodler slider in context

The only hardware component needed to build this prototype was the slider (Figures 6 and 7). In order to embed the slider in the marker holder we printed the holder to provide the space needed for the slider sensor. It is important to mention that this augmented whiteboard requires a connection to a computer. This connection is used in order to interact with the projector software that allows the user to write on the board. Instead of having to interact with a computer, having to erase with the mouse, or create a new blank document to be projected on the screen, the user can simply use the slider at the bottom of the board to erase all the contents of the board.

This object gives extra functionality that does not interfere with any of the other functionalities that the board already has. Finally, this augmented object does not add much cognitive effort in order to use it because of the simplicity of the device.

Figure 8.a shows a projected image with annotations using the Epson BrightLink projector software and the 8.b shows the object being used to erase the comments.

V. DISCUSSION

In this project, we found that 3D printed prototypes can not only be built to allow embedded sensors but also to enhance their capabilities. For instance, if the sensor to be embedded is a vibration sensor the case could be built to hold the sensor in a better way to enhance its functionality. This characteristic is present in the design of our email notifier. It was built so that the inner workings of the device are not visible to the user except the user sees what should be seen, which is the post-it itself. This ensures that the cognitive load for using the object is low (the physical circuits are hidden).

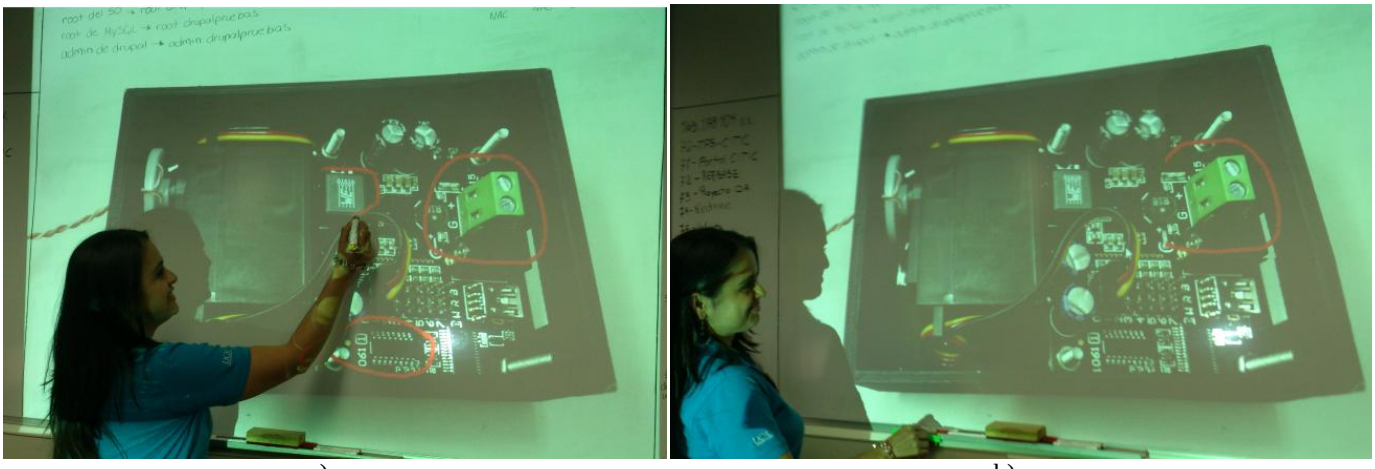
Another example of a redesigned object was the whiteboard eraser. In this prototype, we performed an interview, and no faculty member was able to distinguish that the sensors that were placed inside the redesigned object. In this way, the principle for ubiquitous computing was achieved: to embed the systems in a way that they are invisible to the users.

We found that printed prototypes can hide the new components to avoid cognitive load for the final users. Printed prototypes also avoid affecting the physical appearance of the object.

In our email notifier prototype, the augmented post-it was an actuator but we occlude it behind the monitor. Therefore, the physical appearance of the monitor remains intact except when the note is displayed.

The most valuable use of 3D printing models for rapid prototyping of augmented objects is the possibility of an internal redesign. As we told before an extensive evaluation of the object that would be redesigned is required. However, once the evaluation is performed the redesign process can help to introduce new devices to increase the object capabilities.

Our eraser validates the fact that, if a solid object is selected to be augmented, an internal redesign would allow the use of sensors without changing the physical appearance and the functionality of the object. However, the use of low cost 3D printers restricts the printing size.



a) b)

Figure 8. a) Interactive projector being used to draw some circles over an image, b) Doodler used to erase the previously drawn circles

 ISBN: 978-1-61208-353-7

The quality of the redesigned object can also be affected by the size of the sensors or actuators. With the whiteboard eraser the model was divided and printed in several parts. As an example, the printing time for each of the objects presented in this paper was approximately two and a half hours.

In our research lab, we have been developing prototypes for the past few years without using a 3D printer. We found that a printed prototype is better in terms of cost and quality. Without considering the cost of the 3D printer itself, i.e., considering only the filament and the design and printing time.

The cost for creating 3D printed prototypes is very low, especially if we can use free and open source software design tools. The cost of each of the prototypes presented in this article did not reach US$10 on total printed material cost. The sensors and actuators reach over US$250 (US$90 for the servo controller, US$12 the servo, US$6 for the RFID reader, US$2 the RFID and US$140 the main board controller). These prices correspond to the multipurpose sensors. Specific sensors can be cheaper.

We have found that 3D printing might not be the best option for small models or models that need some very precise parts, because of the coarseness and the quality of the printed object. It can be advantageous when dealing with cases or protective structures. Printing more precise models or devices, or even certain parts of a device that might need more finesse can be very counterproductive due to the lack of precision it provides.

Modeling an object following the process described in Section 3 could be facilitated by using a 3D scanner to digitalize the objects. It is interesting to mention that this 3D redesign process could be incorporated to the development process described in AODeP methodology.

VI. Conclusions And Future Work

The use of 3D printing models can improve the quality of the prototypes and allows a better and quicker evaluation of them. When the prototypes involve an actuator that cannot be easily incorporated into the object, it is useful to create a case to hide the components and reduce as much as possible the cognitive load for the user.

It is important to try not to modify the physical and aesthetic characteristics of the objects when augmenting them. This will help keep the principle of the ubiquitous computing: to make the computer invisible.

The examples presented in this paper show the use of 3D printing in prototyping. However, it is not enough to generate a systematic process of augmented object prototyping. As future work, we will design and create more printed objects and will try to generalize the creation step of the augmented object development process.

Based on the evidence presented in this paper we also present a crucial improvement to AODeP methodology by incorporating the redesign of the augmented objects with a finer level of detail.

Acknowledgment

We thank the Students Association at the School of Computer Science at Universidad de Costa Rica for letting us borrow the 3D printer and the UCR Computer Center for the 3D Scanner used in this research. This project was partially supported by CITIC-UCR (Centro de Investigaciones en Tecnologías de la Información de la Universidad de Costa Rica), grant No. 834-B4-159.

References

[1] G. López, M. López, and L. Guerrero, “Improving the Process for Developing Augmented Objects: An HCI Perspective”, Proc. Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction (UCAmI'13), Springer International Publishing, Dec. 2013, pp. 111-118, doi: 10.1007/978-3-319-03176-7_15.

[2] M. Waiser, The Computer for the 21st Century, ACM Special Interest Group on Mobility of Systems, Users, Data, and Coputing, Vol. 3, pp. 3-11, 1999, doi: 10.1145/329124.329126.

[3] L. Guerrero, H. Horta, and S. Ochoa, “Developing Augmented Objects: A Process Perspective”, Journal of Universal Computer Science, vol. 16, Aug. 2010, pp. 1612-1632, doi:10.3217/jucs-016-12-1612.

[4] N. Hollinworth and F. Hwang, “Investigating Familiar Interactions to Help Older Adults Learn Computer Applications More Easily”, Proc. 25th BCS Conference on Human-Computer Interaction (BCS-HCI '11), British Computer Society, Jul. 2011, pp. 473-478, doi:

[5] C. Leonardi, C. Mennecozzi, E. Not, F. Pianesi, and M. Zancanaro, “Designing a familiar technology for elderly people”, Gerontechnology, vol. 7, May. 2008, pp. 151, doi: 10.4017/gt.2008.07.02.088.00.

[6] N. Jarvis, D. Cameron, and A. Boucher, “Attention to Detail: Annotations of a Design Process”, Proc. 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design (NordiCHI '12), ACM, Oct. 2012, pp. 11-20, doi: 10.1145/2399016.2399019.

[7] S. Hodges, S.Taylor, N. Villar, J. Scott, and J. Helmes, “Exploring physical prototyping techniques for functional devices using .NET gadgeteer”, Proc. Seventh International Conference on Tangible, Embedded, and Embodied Interaction (TEI'13), Feb. 2013, pp. 271-274, doi:10.1145/2460625.2460670.

[8] S. Card, T. Moran, and A. Newell, "The keystroke-level model for user performance time with interactive systems", Communications of the ACM 23 (CACM), ACM, Jul. 1980, pp. 369-410, doi:10.1145/358886.358895.

[9] J. Carlisle, "Evaluating the impact of office automation on top management communication".Proceedings of the June. National Computer Conference and Exposition. (AFIPS '76), ACM, June. 1976, pp. 611–616, doi:10.1145/1499799.1499885.

[10] T. Sánchez, D. Ranasinghe, B. Patkai, and D. Mcfarlane, “Taxonomy, Technology and Applications of Smart Objects”, Information Systems Frontiers, Vol. 13, April. 2011, pp. 281-300, doi: 10.1007/s10796-009-9218-4.

[11] F. Kawsar, K. Fujinami, and T. Nakajima, “Augmenting Everyday Life with Sentient Artefacts”, Proc. 2005 Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies (sOc-EUSAI '05), ACM, Oct. 2005, pp. 141-146, doi: 10.1145/1107548.1107587.

 ISBN: 978-1-61208-353-7

[12] D. Mellis and L. Buechley, “Collaboration in Open-source Hardware: Third-party Variations on the Arduino Duemilanove”, Proc.ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12), ACM, Feb. 2012, pp. 1175-1178, doi: 10.1145/2145204.2145377.

[13] D. Mellis and L. Buechley,” Case Studies in the Personal Fabrication of Electronic Products”, Proc. Designing Interactive Systems Conference (DIS '12), ACM, June.2012, pp. 268-277, doi: 10.1145/2317956.2317998.

[14] S. Hodges, N. VillarJ. Scott, and A. Schmidt, “A New Era for Ubicomp Development”, IEEE Pervasive Computing, vol. 11, Jan. 2012, pp. 5-9, doi: 10.1109/MPRV.2012.1

[15] M. Eisenberg, and L. Buechley, “Pervasive Fabrication: Making Construction Ubiquitous in Education”, Journal of Software, vol.3, March .2008, pp. 62-68, doi: 10.4304/jsw.3.4.62-68.

[16] D. López-de-Ipiña, J. Vázquez, D. García, J. Fernández, I. García, D. Sainz, and A. Almeida, “EMI 2 lets: A Reflective Framework for Enabling AmI”, Journal of Universal Computer Science, Vol. 12, Mar. 2006, pp. 297-314, doi: 10.3217/jucs-012-03-0297.

[17] A. Wiethoff, H. Schneider, M. Rohs, A. Butz, and S. Greenberg, “Sketch-a-TUI: Low Cost Prototyping of Tangible Interactions Using Cardboard and Conductive Ink”, Proc. Sixth International Conference on Tangible, Embedded and Embodied Interaction (TEI '12), ACM, Feb. 2012, pp. 309-312, doi: 10.1145/2148131.2148196.

[18] N. Zaimovic-Uzunovic and S. Lemes, “Influences of surface parameters on laser 3D scanning”, Proc. International symposium on measurement and quality control (ISMQC2010), IMEKO, Sept. 2010, pp. 408-4011.

[19] S. Lemes and N. Zaimovic-Uzunovic, “Study of ambient light influence on laser 3D scanning”, Proc. 7th international conference on industrial tools and material processing technologies (ICIT & MPT), Slovenian tool and die developpment centre, Oct. 2009, pp. 327-330.

[20] J. Sadar and G. Chyon, “3D Scanning and Printing As a New Medium for Creativity in Product Desig”, Proc. Second Conference on Creativity and Innovation in Design (DESIRE '11), ACM, Oct. 2011, pp. 15-20, doi: 10.1145/2079216.2079218.

[21] Blender, http://www.blender.org/, [retrieved: June, 2014].

[22] Phidgets, http://www.phidgets.com/, [retrieved: June, 2014].

[23] Replicator G, http://replicat.org/, [retrieved: June, 2014].

[24] eBeam, http://www.e-beam.com/, [retrieved: June, 2014].

[25] Doodle Pro® Classic Doodler, http://www.fisher-price.com/, [retrieved: June, 2014].

 ISBN: 978-1-61208-353-7

Reversible Watermarking Based on Histogram Shifting of Difference Image between Original and Predicted images

Su-Yeon Shin, Hyang-Mi Yoo, Jae-Won Suh
School of Electrical and Computer Engineering
Chungbuk National University
Cheongju, Korea
Email: ssy6061@naver.com, hmyoo82@cbnu.ac.kr, sjwon@cbnu.ac.kr

***Abstract*—Reversible watermarking is a technique that can recover an undistorted original image from a watermarked image. The proposed watermark embedding algorithm uses histogram shifting of the difference image between a modified original image and its predicted one. In the proposed algorithm, the predicted image that works well increases the embedding capacity, so that the reference pixels for prediction are adaptively selected and filtered and the other predicted pixels are directionally interpolated with the reference pixels. The simulation results demonstrate that the proposed algorithm generates good performances in the peak signal-to-noise ratio (PSNR) values and the embedding capacity.**

Keywords–Reversible watermarking; Histogram shifting; Predicted image; Reference pixel; Directional interpolation

I. INTRODUCTION

Illegal copies of digitized image can be easily and widely distributed through various communication channels and storage devices and be a serious problem for content owners. Watermarking technique can be a good solution to prevent the use of illegal contents. Watermarking technique can be categorized into three classes by the purpose: robust, fragile, and reversible watermarking. In the robust watermarking, the watermarked message must survive the various attacks such as resizing, cropping, filtering. For the fragile watermarking, the embedded watermark should be easily broken from the attacks. Reversible watermarking means that the original image and the watermark message can be completely recovered from the watermarked image without any distortion.

Reversible watermarking algorithms are studied many ways. The difference expansion scheme proposed by Tian [1] selected some expandable difference values of neighboring two pixels and embedded one bit into each of them. Ni *et al.* [2] found the maximum and the minimum pixel levels of the image histogram and shifted the histogram to embed the secrete data. Luo *et al.* [3] utilized the interpolation error, which is the difference between the interpolated pixel value and the corresponding pixel value. However, although these reversible watermarking algorithms based on histogram shifting make sufficient space for the watermark embedding, they suffer from the overflow and underflow problems because of the wrap around pixel levels caused by histogram shifting. Hong *et al.* [4] has extended Luo's work by generalizing the distribution of the reference pixels. However, if the distance between reference pixels become longer, the performance of Hong's algorithm get worse and worse. In this paper, we propose a new reversible watermarking algorithm tho overcome these problems.

The rest of this paper is organized as follows. Section II and Section III describe the watermark embedding and extracting procedures of the proposed algorithm, respectively. In Section IV, we demonstrate the effectiveness of the proposed algorithm. Finally, conclusions are drawn in Section V.

II. WATERMARK EMBEDDING PROCEDURE

The proposed reversible watermarking algorithm uses histogram shifting of the difference image between a modified original and its predicted images. Fig. 1 shows the proposed watermarking embedding procedure. A full explanation of each procedure is given below.

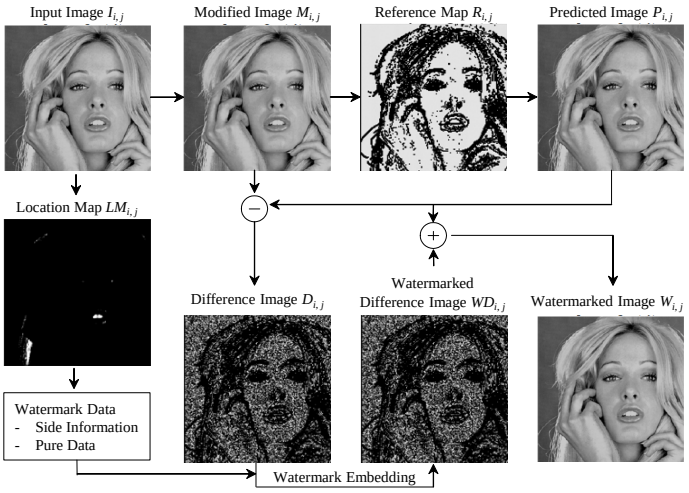


Figure 1. Watermark embedding procedure

A. Location Map and Modified Image

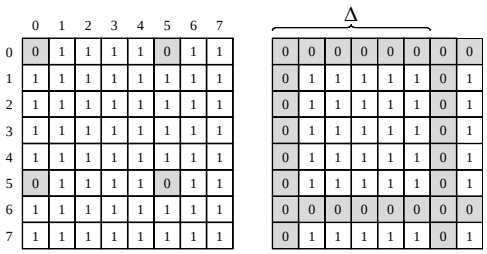
To overcome the wrap around problem, we need to monitor the lower bound pixel value "0" and the upper bound pixel value "255". Regarding the original image $I_{i,j}$ for $0 \leq i < M$ and $0 \leq i < N$, if the pixel value is equal to the lower bound pixel value or the upper bound pixel, we assign a "1" into the corresponding pixel location. Consequently, we obtain

 ISBN: 978-1-61208-353-7

an $M \times N$ binary image called as a location map. Next, we make a modified image $M_{i,j}$ by changing "0" into "1" and "255" into "254". Finally, the location map is losslessly compressed by using the joint bi-level image experts group (JBIG) compression algorithm and is inserted into some part of the embedded watermark data.

B. Predicted Image

The embedding capacity is proportional to the number of the most frequent pixel value at the difference image $D_{i,j}$, which can be obtained by obtaining a well predicted image. To do so, the predicted image $P_{i,j}$ is obtained by directional interpolation based on the proposed reference map $R_{i,j}$.



(a) If $\Delta \leq 5$ ($\Delta = 5$) (b) If $\Delta > 5$ ($\Delta = 6$)

Figure 2. Reference map

1) Reference Map: As shown in Fig. 2, we define two different types of reference map $R_{i,j}$, where Δ is a pre-defined integer and the reference pixels are differently defined by Δ along the vertical and horizontal directions. If the spatial mesh interval for Δ gradually becoming wider, the prediction performance goes from bad to worse. To prevent the decrease of the prediction performance, if Δ is greater than 5, the reference pixels are located in a line. The positions associated with the reference pixels are notified by "0" and the others are "1", which is expressed in (1)

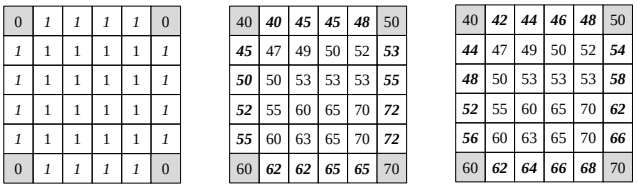
$$\begin{aligned} if \Delta \leq 5: \quad R_{i,j} &= \begin{cases} 0 & \text{if } i\%\Delta = 0 \text{ and } j\%\Delta = 0 \\ 1 & \text{otherwise} \end{cases} \\ if \Delta > 5: \quad R_{i,j} &= \begin{cases} 0 & \text{if } i\%\Delta = 0 \text{ or } j\%\Delta = 0 \\ 1 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

The reference pixels once selected should be preserved and the other pixels surrounded by the reference pixels are interpolated.

Next, we find the complex area and skip the prediction by using the reference map. It is very important because the complex area does not affect the embedding capacity and only causes the visual quality degradation of the watermarked image. To determine the complex area, we use the range function in (2). It returns the absolute difference between the maximum and minimum values of the given values.

$$\begin{aligned} Range(x_1,&x_2,x_3,x_4) = \\ &|Max(x_1,x_2,x_3,x_4) - Min(x_1,x_2,x_3,x_4)| \end{aligned} \tag{2}$$

If the given values are the equally spaced four corner reference pixel values and the return value of range function is greater than a pre-defined threshold T_0, the area surrounded by the four reference pixels are considered as a complex area. In this case, all pixels in the complex area are marked as reference pixels to prevent them from being interpolated in the process of making predicted image [4].



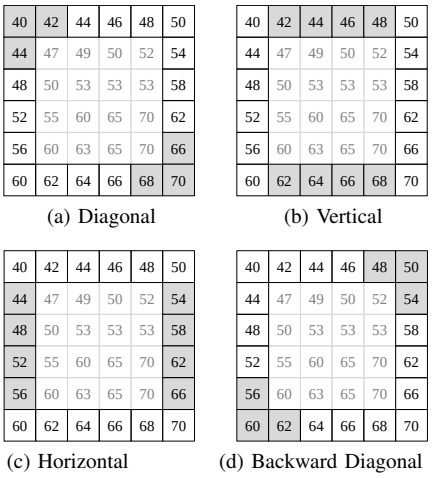
(a) $R_{i,j}$ for $\Delta \leq 5$ (b) Corresponding $M_{i,j}$ (c) Interpolation result

Figure 3. Interpolation of boundary pixels

2) Pre-Processing of Boundary Pixels: As shown in Fig. 3, we calculate an imaginary boundary pixel values for smooth area by applying linear interpolation or low pass filtering to increase the prediction accuracy for the predicted image. In case of $\Delta \leq 5$, the boundary pixel values between two reference pixels are linearly interpolated as shown in Fig. 3. In case of $\Delta > 5$, low pass filtering is applied to the reference pixels. The low pass filtered reference pixels are calculated by

$$MR_{i,j} = \begin{cases} 1/4(M_{i-1,j} + 2 \times M_{i,j} + M_{i+1,j}) & \text{if } j\%\Delta = 0 \\ 1/4(M_{i,j-1} + 2 \times M_{i,j} + M_{i,j+1}) & \text{if } i\%\Delta = 0 \end{cases} \tag{3}$$

The interpolated or low pass filtered boundary pixels are treated as reference pixels but are not reference pixels. They are only used for prediction mode decision and directional interpolation to make a predicted image.



(a) Diagonal (b) Vertical (c) Horizontal (d) Backward Diagonal

Figure 4. Prediction mode

3) Prediction Mode and Directional Interpolation: As shown in Fig. 4, the pre-processed boundary pixels are used for prediction mode decision. There are five prediction modes: diagonal mode, vertical mode, horizontal mode, backward diagonal mode, and plane mode. The mode having the smallest mean sum of absolute difference (MSAD) between far away shaded pixels is determined as the prediction mode of the

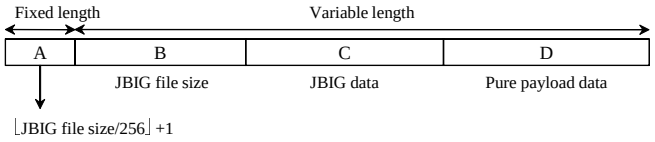
current block. The MSAD is given by (4).

$$MSAD = \frac{1}{n}\sum_{l=1}^{n}|x_l - y_l| \quad (4)$$

For examples of the above Fig. 4, the MSADs for diagonal mode, vertical mode, horizontal mode, and backward diagonal mode are 26, 20, 10, and 9, respectively. Therefore, the prediction mode of the current block is the backward diagonal mode. Therefore, the pixels of the current block are directionally interpolated by the backward diagonal mode. However, if the MSAD is larger than pre-defined threshold T_1, the block is decided as a plane mode and bilinear interpolated.

C. Data Structure of the Watermark Message

In order to be able to recover the original image and the watermark message from a watermarked image, the embedded watermark data have to be designed considering the extraction rule. The data structure of the embedded message is shown in Fig. 5 [5].



- "A" field: An eight bit integer indicating how many next bytes are used for notifying the length of the compressed location map.
- "B" field: Representing the real file size of the compressed location map by JBIG.
- "C" field: Compressed bitstream of the location map.
- "D" field: Pure watermark data.

Figure 5. Data structure of the watermark message

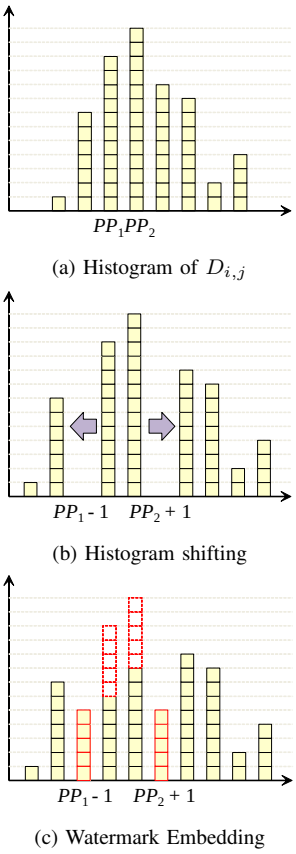
D. Watermark Embedding by Histogram Shifting

The first thing for watermark embedding is making a histogram of difference image $D_{i,j}$. The difference image $D_{i,j}$ is obtained by subtracting the predicted image $P_{i,j}$ from the modified image $M_{i,j}$ for $0 \le i < M$ and $0 \le j < N$. Fig. 6(a) shows the histogram of the difference image.

Secondly, the histogram shifting is performed around the first and the second maximum peak point (PP) pixel values in order to make the data embedding space. "PP_1" and "PP_2" are the pixel values that have the first and second highest number of pixels in ascending order. As shown in Fig. 6(b), we empty the "$PP_1 - 1$" and "$PP_2 + 1$" levels using bi-directional histogram shifting. It means that if the pixel value of $D_{i,j}$ is less than or equal to "$PP_1 - 1$", the corresponding pixel value is decremented by "1" throughout the whole image. If the pixel value of $D_{i,j}$ is greater than or equal to "PP_2+1", the corresponding pixel value is incremented by "1".

Therefore, the shifted difference image $SD_{i,j}$ can be expressed as

$$SD_{i,j} = \begin{cases} D_{i,j} - 1 & \text{if } D_{i,j} \le PP_1 - 1 \\ D_{i,j} + 1 & \text{if } D_{i,j} \ge PP_2 + 1 \end{cases} \quad (5)$$



(a) Histogram of $D_{i,j}$

(b) Histogram shifting

(c) Watermark Embedding

Figure 6. Structure of the watermark data

Next, the watermark message $W(k)$ is embedded into the two PP values. The shifted difference image is scanned again using its raster scan order. Once the "PP_1" are "PP_2" values are encountered, we sequentially check the watermark message $W(k)$. If the checked bit is a "1", the pixel value "PP_1" or "PP_2" is changed into "$PP_1 - 1$" or "$PP_2 + 1$", respectively. If the checked bit is a "0", there is no change as shown in Fig. 6(c). As a result, we can get the watermarked difference image $WD_{i,j}$ and finally obtain the watermarked image $W_{i,j}$ by adding with the predicted image $P_{i,j}$.

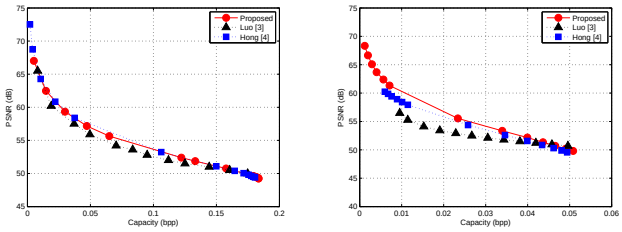
III. PROPOSED WATERMARK EXTRACTION

If the receiver has the watermarked image and the key information for extraction, the watermarked image can be separated into the original image and the watermarked data. The key information for watermark extraction are pre-determined T_0 and T_1, Δ, PP_1 and PP_2.

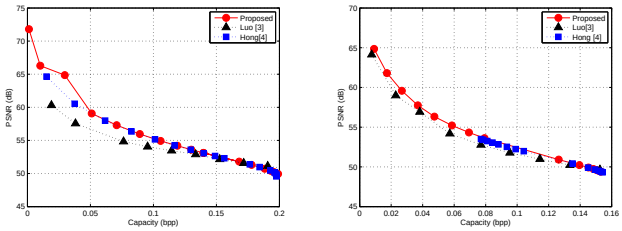
First, we make the predicted image $P_{i,j}$ by using Δ, T_0 and T_1. Because the reference pixels are not changed during the embedding process, the predicted image is exactly the same as that obtained in the embedding procedure.

Second, we generate difference image $D_{i,j}$ between the watermarked image $W_{i,j}$ and the predicted image $P_{i,j}$. During the scanning the difference image $D_{i,j}$, we can extract the embedded watermark data $W(k)$ using the PP_1 and PP_2.

 ISBN: 978-1-61208-353-7

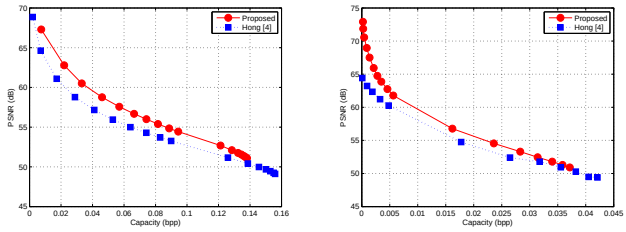


(a) LENA (b) BABOON

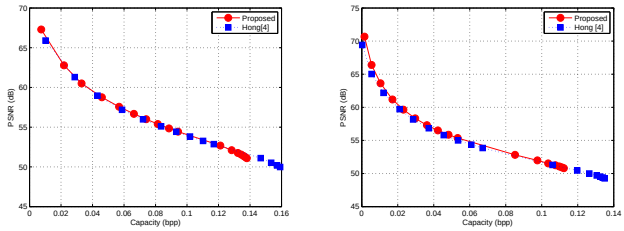


(c) AIRPLANE (d) PEPPER

Figure 7. Results for $\Delta = 3$



(a) LENA (b) BABOON



(c) AIRPLANE (d) PEPPER

Figure 8. Results for $\Delta = 6$

Whenever the corresponding pixel level is equal to "PP_1" or "PP_2", the extracted bit is "0". If the corresponding value is "$PP_1 - 1$" or "$PP_2 + 1$", the extracted bit "1".

Third, we re-scan the entire difference image $D_{i,j}$ and recover the histogram of the difference image. If the corresponding value is less than or equal to "$PP_1 - 1$", "1" is added and if the corresponding value is greater than or equal to "PP_2+1", "1" is subtracted. The returned difference image generates the modified image $M_{i,j}$ by adding the predicted image $P_{i,j}$.

Next, we parse the compressed location map information from the extracted watermark message $W(k)$. Finally, we recover the original image $I_{i,j}$ by using the modified image $M_{i,j}$ and the decompressed location map data.

IV. Experimental Results

In order to evaluate the performance of the proposed reversible watermarking algorithm, we performed computer simulations on typical 512×512 8-bit images: LENA, BABOON, AIRPLANE, and PEPPER. The performance of the proposed reversible watermarking algorithm has been compared to those presented by Luo *et al.* [3] and Hong *et al.* [4] in terms of the embedding capacity versus the PSNR of the watermarked image. In Fig. 7 and Fig. 8, we just showed two simulation results: $\Delta = 3$ and $\Delta = 6$. The simulation conditions are like this: various T_1 (1, 2, ..., 10, 20, ...80) and fixed $T_2 = 10$.

In both results, we can find the common property that the small T_1 generates high PSNR value but low embedding capacity. In Fig. 7, the proposed algorithm achieves slightly better embedding efficiency than that of the Hong's algorithm. The superiority is from the low pass filtering of the reference pixels and directional interpolation according to the prediction mode in the proposed algorithm. The excellence of the proposed algorithm is clearly shown in the case of $\Delta = 6$. By adapting the new structure of the reference pixels for prediction, the prediction errors are reduced and then it increases the embedding capacity.

V. Conclusions

In this paper, we have proposed a reversible watermarking algorithm based on the histogram shifting of the difference image between a original image and a predicted image. In order to solve the underflow and overflow problems, a location map is generated, compressed, and embedded as a part of the watermark message. In previous works, the reference pixels are distributed by equally spaced interval and the pixels surrounded the reference pixels are just bi-linearly interpolated. To enlarge the embedding capacity while keeping the visual quality of the watermarked image, we have suggested the alternative way to make reference pixel map and predicted the smooth area by directional interpolation. From the simulation results, we can conclude that the the proposed algorithm generates good watermarked image quality and embedding capacity.

References

[1] J. Tian, "Reversible Watermarking by Difference Expansion," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 8, Aug. 2003, pp. 890-896.

[2] Z. Ni, Y.Q. Shi, N. Ansari, and W. Su, "Reversible Data Hiding," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 3, Mar. 2006, pp. 354-362.

[3] L. Luo, Z. Chen, M. Chen, and X. Zeng, "Reversible Image Watermarking Using Interpolation Technique," *IEEE Trans. on Information Forensics and Security*, vol. 5, no. 1, Mar. 2010, pp. 187-193.

[4] W. Hong, and T. Chen, "Reversible Data Embedding for High Quality Images Using Interpolation and Reference Pixel Distribution Mechanism," *Journal of Visual Communication and Image Representation*, vol. 22, no. 2, Feb. 2011, pp. 131-140, 2011.

[5] H.M. Yoo, S.K. Lee, and J.W. Suh, "High capacity reversible data hiding using the histogram modification of block image," in *Proc. ISCAS*, May 2010, pp. 1137-1140.

 ISBN: 978-1-61208-353-7

Using the White Space for Digital Inclusion

Abdelnasser Abdelal and Aysha Al-Hinai
Department of Information Technology
College of Applied Sciences- Ibri
Ibri- Sultanate of Oman
Email: {abdelnasser.ibr, ayshah.ibr}@cas.edu.om

***Abstract*-People who do not have Internet access in the world are about 66 percent. These individuals are lagging behind, in the digital sense, due to their remote location, lack of economies of scale, quasi-nomadic nature, and/or low-income. WiFi-based municipal, commercial, and/or Community Wireless Networks have emerged as solutions to provide shared and affordable wireless Internet access to such digitally isolated communities. However, the spectrum used by WiFi (e.g., 2.4 Ghz) is becoming crowded. In addition, the spectrum used by WiMax is regulated in most countries. Therefore, the white space spectrum has emerged as another solution for affordable and shared connectivity solutions. White space refers to spectrum allocated for broadcasting services but not used. This paper provides a brief overview of the white space spectrum, compares it with the WiFi spectrum, outlines key existing projects, suggests a research agenda, and concludes with some policy implications and future work. We hope that this paper brings the attention of the IT community, policymakers, and community activists to the capabilities of white space.**

Keywords- white space; digital inclusion; IEEE 802.11af; broadband policy.

I. INTRODUCTION

Pervasive computing and ubiquitous communications are increasingly becoming essential to conduct our daily life affairs. The Internet, in particular, has grown to be a superhighway for accessing tremendous social economic, social, entertaining and personal services and opportunities. The International Telecommunication Union (ITU) estimates that 3.8 billion people do not have affordable mobile broadband and 2.6 billion cannot afford it [5]. According to the Internet World Stats [33], the percentage of people who do not have affordable high speed Internet is 85 in Africa; 73 in Asia, 37 in Europe; 22 in North America. In addition, 66 % of the world populations do not have access to affordable and reliable Internet services. These people are lagging behind, in the digital sense, because they are not part of the information society. Such communities usually lack viable commercial incentives to attract telecommunication companies. This could be due to their remote location, harsh terrain, high costs of deploying and maintaining infrastructures, low income and willingness to pay, and insufficient population density and/or limited capacity [29][17]. As a result, the market mechanisms have failed to achieve digital inclusion of the society at large. A study by the World Bank estimates that an increase of 10 percent in broadband penetration can boost the economic growth in developing countries by 1.38 percent [8].

Wireless communications can provide the much needed high-speed Internet access to any community in any location either through terrestrial telecommunication infrastructures (e.g., WiFi, and WiMax) or satellite backbones. Wireless communications are particularly beneficial to a wide range of populations [16]. For example, those who by their nature are quasi-nomadic (e.g., the healthcare practitioners, real estate brokers, municipal employees, nomadic herders, the mobile business persons, etc.) would find these emerging infrastructures to be of great benefit. Another group that would benefit from these emerging technologies is those who live in rural and underserved areas. Such areas usually do not have sufficient economies of scale due to lack of the population density, harsh geographical location, and/or low income of their residents. In other words, these areas do not have the necessary commercial incentives to attract telecommunication companies. Such social settings, we believe, require innovative and customized solutions for the digital inequality problem.

Wireless standards (e.g., Wi-Fi and WiMax) have gained the capabilities to provide a wide range of customized connectivity solutions that suit different social settings. They enable individuals to use laptops, Wi-Fi phones, Personal Digital Assistants (PDAs) security cameras and other portable communications devices. In addition to providing mobile and flexible real-time communications, these emerging communication technologies achieve significant time, money and effort savings to their users. As a result, wireless communications have the potential to provide ubiquitous and affordable Internet access and assist all communities to become and remain full participants in the emerging Internet-based "Information Age." Therefore, numerous societies have built autonomous Community Wireless Networks (CWNs) with their own resources, taking advantage of the free 2.4 GHz spectrum and available open source [4][32].

 ISBN: 978-1-61208-353-7

However, the 2.4 Ghz band has become too crowded because many communication technologies use it. These include Bluetooth devices, WiFi enabled devices, cordless phones, etc. Frequency regulation bodies in some countries have unregulated chunks of white space frequencies. For example, the switching to digital television has deregulated spectrum frequencies between 50 MHz and 700 MHz. In addition, WiMax spectrum is regulated in most countries.

White space has emerged as an alternative, or complement, for the licensed spectrum or the unlicensed one which is becoming more crowded. The term "White space" refers to the unused or sparsely occupied spectrum frequencies allocated to broadcasting services or guard channels of cellular networks in particular areas of a country [6][7][14][15]. In every town, there is a number of TV channels remained unused or have been abandoned after transferring to digital TV.

Innovators, researchers, and technology vendors, regulation authorities are working on devices, software, database, and policies that focus on reusing these frequency bands. They believe that reusing this spectrum can enhance the wireless landscape by offering the potential for substantial bandwidth and long transmission ranges [14]. It would also benefit rural and underserved communities which cannot afford high Internet fees of traditional Internet Service Providers (ISPs) as it improve market competition.

Therefore, there are two types of white space [7][15]:

i. Spectrum allocated for TV channels but not used locally;

ii. Spectrum allocated between radio bands to prevent channel interference. For instance, UK has allocated chunks of spectrum and left it open as buffering gaps between the high-powered transmissions carrying broadcast TV in order to avoid interference.

The remainder of this paper is structured as follows. We first discusses in Section II the importance of white space. Section III discusses the research issues related to white space. Then, in Section IV, we compare the white space spectrum with WiFi. In Section V and VI, we discuss the regulations and highlight the existing white space key projects. Finally, in Section VII, our designed project is described.

II. THE IMPORTANCE OF WHITE SPACE

It is believed that opening up this white space to lower-powered devices can provide more wireless spectrum for data transmission to support a large range of devices and services. This spectrum is important because:

i. Having high-speed Internet access has become crucial for people, organizations, and governments to conduct their business. White space-based communications would play a key role in supporting the exponential growth of mobile data communications. This is because the current spectrum allocated for wireless networks (e.g., cellular and Wi-Fi, WiMax) would not be enough. In particular, it is estimated that the number of Internet-connected devices would exceed 50bn worldwide by 2020, according to data from Cisco [7]. According to Hengeveld [5], the Internet of Things (IoT) will connect more a large number of devices which makes reusing this spectrum a necessity.

ii. It can cover the shortage of available spectrum and thus enabling communication to devices that are not well served with the previously allocated spectrums. For example, white space could be used to provide wireless broadband Internet access, similar to Wi-Fi, but over much longer distances. It can also support mobile devices like tablets and smart phones. In addition, it could be used as an extension of fixed-line broadband to reach places that are not connected via traditional cables. Moreover, white space has the capabilities to penetrate walls and underserved areas because it does not require line-of-sight technology, unlike satellite communications and microwave broadband. Reusing spectrum or using unused one increases the amount of available spectrum.

III. RESEARCH ISSUES

Hengeveld [5] suggests a research agenda for the white space-based communications. According to Hengeveld [5], research on white space could include proving broadband to rural communities, building campus networks, conducting basic research, and initiating lab trials. In addition, regulatory authorities may conduct pilot trials, studies for technical and do economic feasibility, develop prototype devices, support field test, and design measurements. With respect to the commercial sector, Hengeveld suggests opportunities including developing volume devices, adopt projects for rural broadband networks, implement campus networks, build smaller form factors, and develop standards-based devices. In addition, they could also work on building databases for available white space, provide certifications, implement use case experimentation, and adopt vertical integration.

White space devices allow secondary users to use this portion of spectrum which is not used by the primary user. Therefore, there is a major concern regarding interference between the primary signal and secondary signal. Researchers should find solutions that protect the secondary signal from interfering with the primary one. Another research area is finding new methods of access and suitable business models for this new connectivity solution. Table 1 shows key database providers, hardware providers, and other important technology vendors.

 ISBN: 978-1-61208-353-7

TABLE I. THE WHITE SPACE ECOSYSTEM [5]

Database Providers (United States)	**Hardware Providers (today)**	**Other Potential Players (growing interest)**
Frequency Finder, Inc.	Adaptrum	Atheros (Qualcomm)
Google	Airspan	ARM
Comsearch	6Harmonics	Alcatel-Lucent
Key Bridge	Carlson	Broadcom
LS Telcom AG	KTS	CSR
Microsoft	Lyrtech	Dell
Neustar	MLED	Hewlett Packard
Spectrum Bridge (Approved)	Neul	Intel
Telcordia Technologies (Ericsson)	Shared Spectrum	LG Electronics Marvell Semiconductor Nokia, Inc. Research in Motion Samsung

British Telecom plans to use white space for the purpose of providing affordable Internet access to about 500,000 households who currently lack Internet access [5]. According to its experimental tests, white space devices can provide speed up to 4 to 8 Mbps. In addition, the signal range could research 6 KM transmitter.

The key findings of the Cambridge project is that white space devices is their ability to successfully co-exist with broadcasters and other regulated data communications [5]. In addition, scientists suggest that white space could be used for a wide spectrum of applications. This trial project also evidences a growing industrial interest and readiness. The results and recommendations of such trial projects will assist regulatory authority to make the right decisions concerning white space in their countries. Similarly, we are working on a campus based network to be used for basic research, bring about awareness, and provide connectivity to our university and neighboring communities. Table 2 provides a brief description of selected projects.

IV. COMPARISON WITH OTHER SPECTRUMS

In this section, we need to compare white space with Wi-Fi standards in terms of signal range, data rate, and regulation issues. White space can be a novel choice for Wireless Internet Service Providers (WISPs) or public wireless access with greater broadband capabilities as compared to WiMAX or Wi-Fi. White space spectrum can outperform other wireless technologies (e.g., Wifi and WiMax) in terms of:

1. Providing superior communication range and it is able to penetrate solid obstacles such as trees and buildings. Unlike Wi-Fi, which has a relatively limited range, and can be blocked by obstacles, a network utilizing white-space technology can cover greater ranges than Wi-Fi while requiring less equipment. For instance, Super-WiFi [13] which is a project based on white space spectrum could provide a maximum transmission range of 250m.
2. No Line-of-Sight (LOS) is required between the points being connected because it operates in low-frequency and thus it can penetrate obstacles without the need for towers and additional infrastructure needed to prevent interference.

The long range characteristics of white space could support cellular offloading, rural broadband backhaul, Wide-coverage hotspots, bridges among small networks, sensor network, and wireless surveillance system [5].

V. WHITE SPACE REGULATIONS

The fixed spectrum allocation scheme used nowadays leads to immense underutilization of the scarce spectrum space. For instance, Shared Spectrum Company conducted a research aiming to quantize the white space spectrum in Washington DC. The results of that research detected 62% of white space even in the most crowded areas of the city [20].

Thus, a critical change is happening in the spectrum regulations. This is introducing and enabling spectrum sharing between primary or licensed users and license-exempt or secondary users. The sharing is toll free for the secondary users with a condition that they must not cause any disturbance to the primary users of the spectrum. The first instance of sharing the spectrum was sharing the unused UTF digital TV spectrum bands. Regulatory bodies, such as FCC from the USA and Ofcom from UK, stated that other spectrum should follow the digital TV white spectrum to share the unused spectrums [18]. The FCC has enabled the digital TV white space sharing in 2008 [19]. Cognitive radio is an example of a technology that relies on TV white space for communication and this technology is now running in the real-word environment.

 ISBN: 978-1-61208-353-7

TABLE II. REGULATORY LANDSCAPE OF TV WHITE SPACES.

Country/ Region	Law	Regulation	Policy	Trials/Pilots
United States	Done	Done	Done	Trials complete/Pilots ongoing
UK	Pending	Pending	Done	Trials complete/Pilots planned
Finland	Done	Done		Trials complete/Pilots planned
Canada	Pending	Pending		Trials ongoing
Singapore		Pending		Trials complete/Pilots planned
South Korea				Trials planned
European Commission		Pending	Done	Trials ongoing
China				Trials ongoing
Japan				Trials planned
Brazil				Trials planned

VI. WHITE SPACE-BASED PROJECTS

A number of white-space based wireless networks have emerged all over the world. For instance, the local government in Wilmington, North Carolina, has built the "smart city" network to extend monitoring and managing capabilities to areas that have been unreachable except by physical visits [11][22]. This experimental network is used for traffic monitoring (using wireless cameras), providing free Wi-Fi in city parks and unserved areas, and monitoring quality of water in remote wetlands. A future plan involves providing e-healthcare and offering broadband service to local schools using this white space-based solution.

Another whitespace project was held in Cambridge, England on June 29, 2011 and it was a commercial trail on whitespace Wi-Fi. It was conducted by Microsoft. In the demonstration, whitespace system successfully provided a broadband IP connectivity allowing an Xbox to stream live HD videos from the Internet.

The Blue Ridge Mountain terrain in Claudville has made Internet access hard to come by. Its citizens earn $15,574 per capita, and hence, the big ISPs haven't rushed to Claudville. However, the white space provided cheaper high-speed internet connectivity to some rural areas of Claudville [2].

The Smart Grid network uses white space for smart grid technologies in California, USA [1]. In particular, the purpose of the network is to provide more efficient, and greener and lower cost utilities wireless network. It is built by a consortium of Plumas-Sierra Rural Electric Cooperative & Telecommunications, Plumas County, and Spectrum Bridge Inc. and Google. The experimental trails have proven the white space is an effective option to deal with difficult terrain and offer another medium for affordable wireless connectivity. It has proven to have good propagation characteristics, the ability to penetrate foliage and no need for line of sight. In other words, whit space has the capability to overcome major technical challenges in difficult train areas.

West Virginia University in July 2013, became the first university in the USA to use available broadcast TV channels in order to provide wireless broadband on campus and nearby [21]. It has partnered with the Advanced Internet Regions consortium to build a wireless network for its campus and surrounding area using the TV white space [21]. These frequencies were left empty after TV stations moved to digital broadcasting. Another purpose is to study the viability of delivering connectivity to rural areas and small towns for the purpose of sustaining economic development, improving quality of life, and improving their competitive advantages in the knowledge economy. Another focus of our project is to measure the impact of our prospective network on building human capital and social capital for its remote community. Table 3 provides a brief description of selected projects.

VII. RESEARCH DESIGN

We are planning to establish a pilot TVWS based network in our university campus. The network would utilize TV white space and solar-powered base station. This new network will coexist with the current WiFi network we have. It will be used to provide broadband access in case the college's wired or wireless networks went down and to enable cheap Internet access to the surrounding society. We used a tool called "show my white space" version 2.6, which is developed by Spectrum Bridge. The tool is used to search and locate available TV white space channels in a certain area. The tool could not detect any available white space channels neither in our town nor in the capital. However, it detected several available channels in Cambridge area, as shown in Figures 1 and 2.

 ISBN: 978-1-61208-353-7

TABLE III. DESCRIPTION OF KEY WHITE SPACE-BASED PROJECTS.

Project	Region	Project Specifications
Smart City	Wilmington, North Carolina, USA	Using white space spectrum to connect the city's infrastructure and public services. It also provides public Wi-Fi to some previously underserved communities [11, 12]
Cambridge Project	Cambridge, UK	White space implementation trial
West Virginia University Project	Virginia, USA	Provide wireless broadband on campus and the surround**ings**.
Claudville	Claudville, Virginia, USA	High-speed internet connectivity to rural areas of Claudville [2]
Smart Grid	Plumas, California, USA	Real-time broadband connectivity to remote substations and switchgear

Figure 1. Cambridge area.

Figure 2.Available white space in the Cambridge area.

The reason could be that all the channels are already occupied by primary users or the white space concept is still not applied in practice. This means that the spectrum space in the area is underutilized and wasted. So, we recommend that the unused spectrum should be released, or deregulated, as it can provide free Internet access to the public in rural areas. It would also enhance related basic research.

VIII. CONCLUSION

TV band white spaces are unused spectrum left between broadcast channels. They exist in different places on different channels. More specifically, it is a spectrum band that is licensed to primary users, the part of spectrum that is unused by the primary user at specific locations and sometimes at specific time. Indeed, the use of white space will provide a new source of bandwidth and thus invaluable connectivity, while not having to rely upon traditional mobile phone networks. It can provide connectivity to both mobile and fixed devices and the internet where Wi-Fi cannot reach. Currently, we are building a pilot network in our campus to bring about the awareness, suggest relevant broadband policy, conduct basic research, and measure its social and economic effects on rural areas.

REFERENCES

[1] M. Atiyeh, (2014) Spectrum Bridge, retrieved June 2014, from http://spectrumbridge.com/Libraries/Press_Releases/Nation_s_First_Smart_Grid_White_Spaces_Network_Trial_June_23_2010.sflb.ashx

[2] N. Anderson, (Feb 21 2014) First white space broadband deployment in small Virginia town, are technical. Retrieved June 2014, from http://arstechnica.com/tech-policy/2009/10/first-white-space-broadband-deployment-in-small-virginia-town/

[3] Cambridge (2011), retrieved February 2014, from http://www.theguardian.com/technology/2013/oct/02/white-space-broadband-microsoft-google-wireless-rural

[4] J. Damsgaard, M.A. Parikh, and B. Rao, (2006) Wireless Commons: Perils in the Common Good, Communications of the ACM, 2006, vol. 49, no. 2.

[5] P. Hengeveld, (2012) Increasing Role of PPPs in ICT Ecosystem, ITU Regional Experts Group Meeting for Europe, Geneva, Switzerland 14-15 November 2012 retrieved June 2014, from http://www.itu.int/ITU-D/eur/rf/ppp/agenda_presentations.htm

[6] S. Geirhofer, L. Tong, and B.M. Sadler, "Cognitive Radios For Dynamic Spectrum Access - Dynamic Spectrum Access in the Time Domain: Modeling and Exploiting White Space," Communications Magazine, IEEE, vol.45, no.5, May 2007, pp.66, 72,

[7] S. Gibbs, (2013), White space broadband: your questions answered. Retrieved June 2014, Retrieved from: http://www.theguardian.com/technology/2013/oct/02/white-space-broadband-microsoft-google-wireless-rural

[8] The Broadband Commission (2012) The State of Broadband 2012: Achieving Digital Inclusion for All, accessed on 5th of March, from http://www.broadbandcommission.org/documents/bb-annualreport2012.pdf

[9] West Virginia University (2011), Retrieved June 2014, Retrieved from http://arstechnica.com/information-technology/2013/07/white-space-internet-may-finally-spread-through-us

[10] TV White Space (2014) Breakthrough Technology. Carlson Wireless Technologies, Retrieved June 2014, from http://www.carlsonwireless.com/solutions/tv-white-space-rural-broadband.html

 ISBN: 978-1-61208-353-7

[11] N. Patel, February 26th, 2010, White space 'Smart City' network goes up in Wilmington, North Carolina, Engadget, and Retrieved July 2014 from http://www.engadget.com/2010/02/26/white-space-smart-city-network-goes-up-in-wilmington-north-ca/

[12] J. Van De Beek, J. Riihijarvi, A. Achtzehn, and P. Mahonen, (2011, May). UHF white space in Europe—a quantitative study into the potential of the 470–790 MHz band. In New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2011 IEEE Symposium on (pp. 1-9). IEEE.

[13] J. W. Mwangoka, P. Marques, and J. Rodriguez, (2011, May). Exploiting TV white spaces in Europe: The COGEU approach. In New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2011 IEEE Symposium on (pp. 608-612). IEEE.

[14] P. Bahl, R. Chandra, T. Moscibroda, R. Murty, and M.Welsh. (2009) White space networking with wi-fi like connectivity, In Proceedings of the ACM SIGCOMM 2009 conference on Data communication (SIGCOMM) ACM, New York, NY, USA, 27-38.

[15] B. Ray, (2011). How to Build a National Cellular Wireless Network for £50m, the Register. Retrieved June 2014.

[16] Cisco (2007), Municipalities Adopt Successful Business Models for Outdoor Wireless Networks, accesses February 28. From : http://www.cisco.com/en/US/netsol/ns621/networking_solutions_white_paper0900aecd80564fa3.shtml

[17] S.O. Siochrú, and B. Girard, (2005) Community-based Networks and Innovative Technologies: New models to serve and empower the poor, series of Making ICTs Work for the Poor, the United Nations Development program, accessed on October 5th, 2008, from http://propoor-ict.net

[18] M. Fitch, M. Nekovee, S. Kawade, K. Briggs, and R. MacKenzie, (2011) Wireless service provision in TV white space with cognitive radio technology: A telecom operator's perspective and experience," Communications Magazine, IEEE , vol.49, no. 3, pp. 64,73.

[19] K. Harrison, S.M. Mishra, and A. Saha, "How Much White-Space Capacity Is There?," New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on , April 2010, pp.1,10, 6-9

[20] W. Wang, and X. Liu (2005, September). List-coloring based channel allocation for open-spectrum wireless networks. In IEEE Vehicular Technology Conference vol. 62, no. 1, p. 690.

[21] West Virginia University (2013) Nation's first campus 'Super Wi-Fi' network, retrieved March 2014, from http://wvutoday.wvu.edu/n/2013/07/09/nation-s-first-campus-super-wi-fi-network-launches-at-west-virginia-university

[22] K. S. Nanavati, (2012). Channel bonding/loading for TV white spaces in IEEE 802.11 af.

[29] C. Middleton, G. Longford, A. Clement, and A.B. Potter, (2006) ICT Infrastructure as Public Infrastructure: Exploring the Benefits of Public Wireless Networks, the proceedings of the 34th Research Conference on Communication, Information and Internet Policy.

[30] Adaptrum Demonstrates TV Whitespace Solution at Cambridge TV Whitespace Trial Launch Event. From http://www.prnewswire.com/news-releases/adaptrum-demonstrates-tv-whitespace-solution-at-cambridge-tv-whitespace-trial-launch-event-125001269.html. Retrieved June 2014.

[31] (July 9th, 2013). Nation's first campus 'Super Wi-Fi' network launches at West Virginia University, From http://wvutoday.wvu.edu/n/2013/07/09/nation-s-first-campus-super-wi-fi-network-launches-at-west-virginia-university. Retrieved March 2014.

[32] E. Vos, (2005) Reports on Municipal Wireless and Broadband Projects, March 2005 Report, from www.muniwireless.com

[33] The Internet World Usage Statistics (2014), accessed February 28, 2014, from: http://www.internetworldstats.com/stats.htm

 ISBN: 978-1-61208-353-7

Compact Three-dimensional Vision for Ubiquitous Sensing

Kumiko Yoshida
Interdisciplinary Graduate School of Agriculture and Engineering
University of Miyazaki
Miyazaki, Japan
E-mail: nc13004@student.miyazaki-u.ac.jp

Kikuhito Kawasue
Department of Environmental Robotics
University of Miyazaki
Miyazaki, Japan
E-mail: kawasue@cc.miyazaki-u.ac.jp

***Abstract*—We herein propose two computer vision systems that make use of the Microsoft KINECT sensor for ubiquitous sensing in raising stock and in industrial fields. The first system is a three-dimensional (3D) thermo-sensing system that detects 3D shape data and 3D temperature data simultaneously. These data are automatically combined, and the 3D shape and temperature distribution are reconstructed on the computer. The second system is a handheld 3D measurement system that uses a slit-ray projector in conjunction with the KINECT sensor. The 3D shape of the target is reconstructed on a computer using the detected data. These two systems are sufficiently compact and the measurement can be performed via online processing. As such, these systems will be useful in ubiquitous data acquisition systems in various fields. Typical applications of the proposed systems include environment sampling, health monitoring of animals, monitoring of facilities in raising stock, and industrial fields. The experimental results of the present study demonstrate the feasibility of the proposed system.**

Keywords-computer vision; KINECT; thermo-sensing; three-dimensional sensing.

I. INTRODUCTION

In realizing a ubiquitous society, the development of a compact system that can detect various data at a site is effective. Once they have been detected digitally, the data can be distributed through a network and can be used effectively. Recently, Charge coupled device (CCD) cameras have been made more compact and have been incorporated into a number of mobile phones. Indeed, mobile phones containing CCD cameras are considered to be the most familiar data acquisition system. Although CCD cameras capture primarily two-dimensional image data, three-dimensional (3D) data (point cloud) are required for various applications.

A point cloud is a set of vertices in a 3D coordinate system. Point clouds are used in Computer-Aided Design (CAD) data and robot vision systems. In recent years, inexpensive devices, such as the Microsoft KINECT sensor [1]-[4], which detects 3D point cloud data, have become available. The KINECT sensor is composed of a random dot projector and an Infrared (IR) camera. Random dots are projected from the laser projector, and the reflected light from the surface of objects is recorded by the IR camera. The dots recorded by the IR camera are triangulated in order to calculate the 3D position of an object based on the configuration of the laser projector and the IR camera. Such devices are useful for motion capture or modeling systems, which do not require high accuracy. Recently, The KINECT sensor has also been used in ubiquitous computing [5]-[8].

In using the KINECT sensor, there is no limitation on the measurement area size because individual 3D point cloud data sets recorded from different positions can be combined. The Iterative Closest Point (ICP) algorithm [9]-[13] is often used to combine data sets. This algorithm automatically determines the overlapping area between 3D point cloud data sets and constructs a single 3D image. The KINECT sensor can obtain thousands of point cloud data sets in real time and so is a very attractive sensor. We herein introduce two applications using the KINECT sensor.

In recent years, thermal imaging measurement technology has developed rapidly. Infrared thermography has been used in animal sciences. Non-destructive evaluation in raising stock or in animal research is one example of such an approach. However, since images used in such evaluations are two-dimensional thermal images, quantitative information such as the area, the 3D shape, and the roughness of the heat source, cannot be obtained. In order to obtain the quantitative information, we herein propose a measurement system that uses the KINECT sensor in conjunction with thermography to produce 3D shapes and 3D thermo-grams. Periodically evaluating the condition of Japanese black cattle during the growth process is important [14]. In addition to the weight and size of cattle, the body temperature should be measured as primary evaluation criteria. Quantitative measurement of 3D temperature and shape can be established using the proposed system. The proposed system can be a useful tool for monitoring the condition of cattle in breeding farms.

Another application of the proposed system is the measurement of facilities in industrial fields, such as the chemical industry. However, the original data obtained by the KINECT sensor is not sufficiently accurate for industrial applications. On the other hand, the slit-ray projection method [15]-[17] (i.e., shape from structured light), which has high measurement accuracy, is widely used in industrial applications and robot vision systems. In this method, a laser slit is projected onto the surface of the target object and the laser streak generated on the surface is detected by a CCD

 ISBN: 978-1-61208-353-7

camera. The 3D position data of the laser slit is estimated by triangulating the orientation of the laser projector and the CCD camera [18][19]. In the present study, we introduce a point cloud data acquisition system that uses slit ray projection and a KINECT sensor. The proposed system is sufficiently compact to allow its use as a hand-held device. During measurement, the user directs the laser slit ray at the target. The KINECT sensor then detects point cloud data while the CCD camera simultaneously detects the laser streak generated on the surface of the target. The user manually scans the system by directing the laser slit ray along the measurement pipe. The point cloud data obtained using the KINECT sensor is used to determine the movement of the system by adjusting overlapping data in consecutive frames using the ICP algorithm. The data obtained by the laser slit-ray projection method are more accurate than the original point cloud data obtained by the KINECT sensor because the data are obtained using a high-resolution camera. The pipe cross section is estimated based on data obtained by the slit-ray projection method. The 3D shape of the pipe is constructed on a computer from cross sections obtained from different positions. The proposed system enables onsite measurement in chemical plants and can be used in obtaining information on the current condition inside the plant, which can be used in plant evaluation and maintenance.

Two systems that use the KINECT sensor are introduced in the present paper to use as effective tools for realizing a ubiquitous society. The remainder of the present paper is organized as follows. Section 2 describes the proposed 3D thermo-sensing system using the KINECT sensor, and Section 3 describes the application of the proposed system. Section 4 describes a pipe measurement system that uses the slit-ray projection method in conjunction with the KINECT sensor. Section 5 describes the pipe reconstruction results obtained using the proposed system. Finally, the paper is concluded in Section 6.

II. THERMO-SENSING SYSTEM USING THE KINECT SENSOR

The measurement system is shown in Figure 1. The 3D thermal-sensing system is established using a 3D measurement sensor (KINECT sensor) and a thermograph (NEC, 320×240, 60 Hz). A flowchart of the measurement procedure is shown in Figure 2. The 3D shape measurement is performed using the KINECT sensor, while the thermograph simultaneously captures a thermal image. The corresponding temperature data obtained by thermography is then allocated to the reconstructed surface shape of the object on a computer. This allocation is established using a pre-determined conversion matrix, which is determined during the calibration process.

Calibration is important in determining the coordinate conversion between global coordinates and thermography coordinates. Generally, the calibration process in a 3D measurement system is complicated and not universal. In the proposed system, the thermography coordinates have to be determined precisely in the calibration process because they influence the temperature mapping accuracy onto the surface of the object. A suitable calibration method for thermography is proposed.

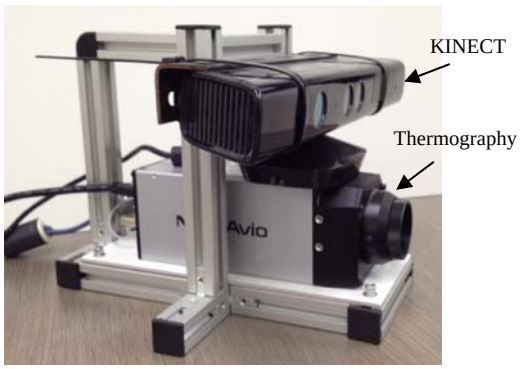


Figure 1. Measurement system.

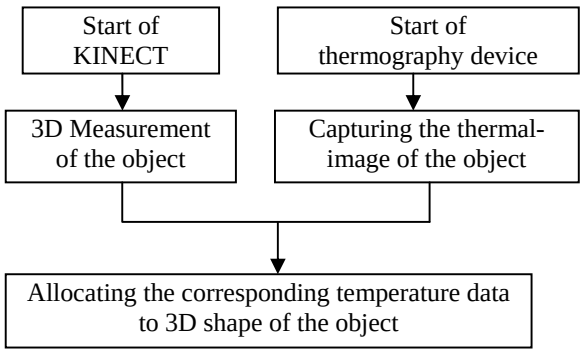


Figure 2. Measurement procedure.

A standard cube is used to match the coordinates among the global coordinate system and the thermography coordinate system, because a thermal image can be recorded by thermography. The calibration setup is shown in Figure 3. The 3D shape of the standard cube is measured using the KINECT sensor, and the thermal image of the standard cube must be recorded simultaneously by thermography because the recorded thermal image is used to determine the calibration parameters.

The relationship between the thermography coordinates (u, v) and the global coordinates (x, y, z) of a point is as follows:

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_{11} & k_{12} & k_{13} & k_{14} \\ k_{21} & k_{22} & k_{23} & k_{24} \\ k_{31} & k_{32} & k_{33} & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \quad (1)$$

where k_{11} through k_{33} are parameters that consider the rotation, scale, and displacement between the thermography coordinates and the global coordinates. These parameters are the elements of conversion matrices and are determined by inputting some corresponding positions between the thermography coordinates and the global coordinates. Here, k_{11} through k_{33} can be determined by inputting no less than

 ISBN: 978-1-61208-353-7

six corresponding points into Eq. (1).

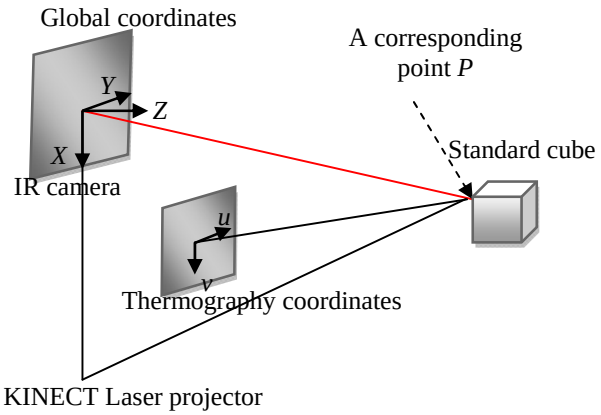


Figure 3. Calibration setup.

Equation (1) can be rewritten as follows:

$$\begin{cases} u = (k_{11}x + k_{12}y + k_{13}z + k_{14}) / (k_{31}x + k_{32}y + k_{33}z + 1) \\ v = (k_{21}x + k_{22}y + k_{23}z + k_{24}) / (k_{31}x + k_{32}y + k_{33}z + 1) \end{cases} \quad (2)$$

Once k_{11} to k_{33} are determined, the thermography coordinates can be calculated using Eq. (2), and the corresponding temperatures can also be mapped on the reconstructed surfaces of the object.

Based on the presented mathematical model, it is possible to determine the parameters that contain the position and posture of the IR camera. Once the 3D shape of the object is measured, the corresponding temperature is mapped from the thermal-image using Eq. (2).

III. APPLICATION TO MEASURE THE TEMPERATURE IN CATTLE

Periodically evaluating the condition of Japanese black cattle during the growth process is important. The weight, size, posture, body shape, and body temperature are measured as the primary evaluation criteria.

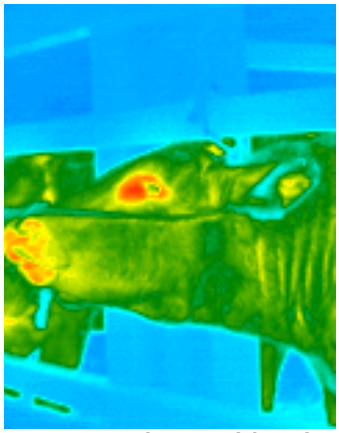


Figure 4. Two-dimensional thermal image.

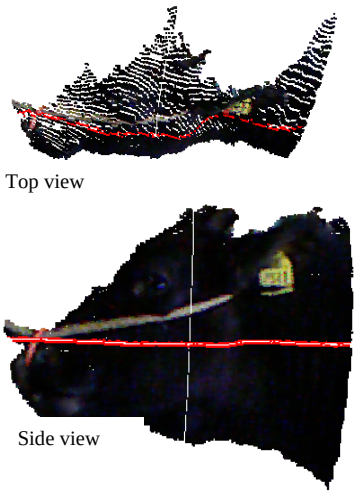


Figure 5. Point cloud data of the face of a cow.

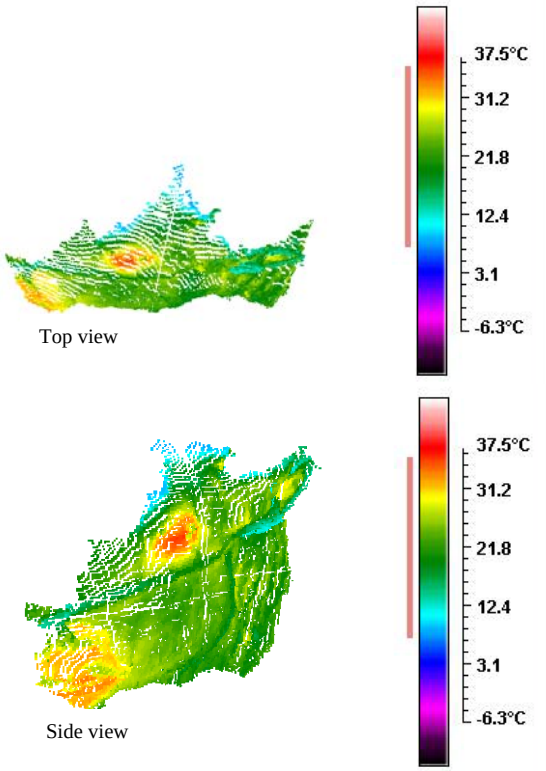


Figure 6. Constructed 3D thermal image.

A computer vision device can be a useful tool for evaluating the condition of cattle. Therefore, we used 3D thermo-sensing to measure the body shape and temperature of cattle.

Figure 4 shows a two-dimensional thermal image recorded by thermography. The image in this figure contains the temperature distribution of a cow's head but not quantitative information such as the dimensions of the cow's head. Figure 5 shows the point cloud data detected by the KINECT sensor. The temperature data contained in the

 ISBN: 978-1-61208-353-7

thermal image is mapped to point cloud data. Figure 6 shows the constructed 3D thermal image that was calculated using the proposed method. Since this result contains the 3D quantitative shape of the cattle along with temperature information, the information is useful for animal science, for example, in budget calculation.

IV. APPLICATION FOR PIPE MEASUREMENT

In the chemical plant, the shapes and the arrangement of existing pipes should be investigated before the replacement or the construction of new pipes. Generally, measurement is conducted manually using a metal tape measure and is a cumbersome task. Recently, A number of laser scanners have been introduced for use in the measurement of pipes. The laser scanner is a very attractive option because thousands of sets of point cloud data can be obtained in a short time. However, the volume of data obtained is very large, because the data includes unnecessary information such as ground data, wall data, and data related to other equipment. Extra work is required in order to extract the appropriate information from the huge volume of point cloud data for the design of new pipes. Therefore, the development of an effective method is required in order to extract the desired information from the point cloud data. In addition, certain pipes that are located at high positions that the laser cannot reach are not measured, because the laser scanner is fixed on stable ground.

Figure 7 shows the measurement system. A high-resolution CCD camera (resolution: 3,488×2,616) is attached to a Microsoft KINECT sensor. A laser slit projector (20 mW), which must be held within 300 to 500 mm from the CCD camera, is also attached to the measurement system. During the piping measurement, the user directs the laser slit ray at the target. The KINECT sensor then detects point cloud data, while the CCD camera simultaneously detects the laser streak generated on the pipe surface. The cross-sectional shape is estimated from the image of the laser streak obtained by applying the slit-ray projection method (i.e., shape obtained from structured light). The user manually scans this system by directing the laser slit ray along the pipe. The movement data (i.e., the amount and direction of movement) are estimated from the point cloud data detected by the KINECT sensor. The 3D pipe shape can be constructed on the computer from two or more sets of cross-sectional shape data obtained from the laser streak. Figure 8 shows a flowchart of the measurement procedure. The measurement system is first directed at area 1 on the target. The KINECT sensor and the laser slit measurement system are synchronized so that they are not affected by the movement of the system. The user can change the position of the measurement system by directing the system at the second measurement area (area 2). The KINECT sensor and the laser slit measurement system then detect data from this area. The point cloud data must include data from the overlap between areas 1 and 2 in order to estimate the movement of the system. Movement data (i.e., the amount of movement and the orientation of the measurement system) are estimated using the ICP algorithm [5]-[9]. The point cloud data is combined and a single pipe is constructed on the computer. The orientation of the cross section measured using the laser slit can also be determined by allocating the cross-sectional data of the constructed pipe. Thus, the shape of the entire pipe can be estimated from two or more cross-sectional data sets.

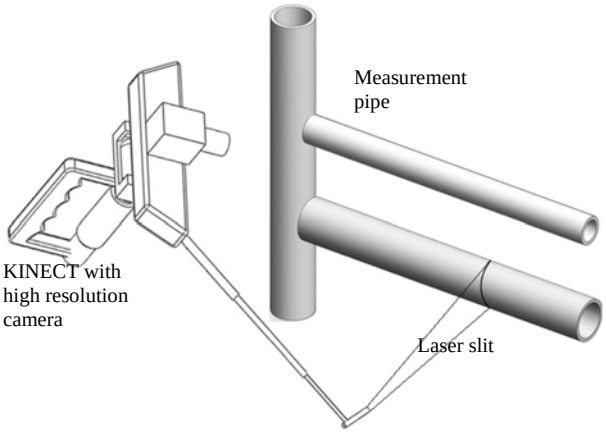


Figure 7. System for piping measurement.

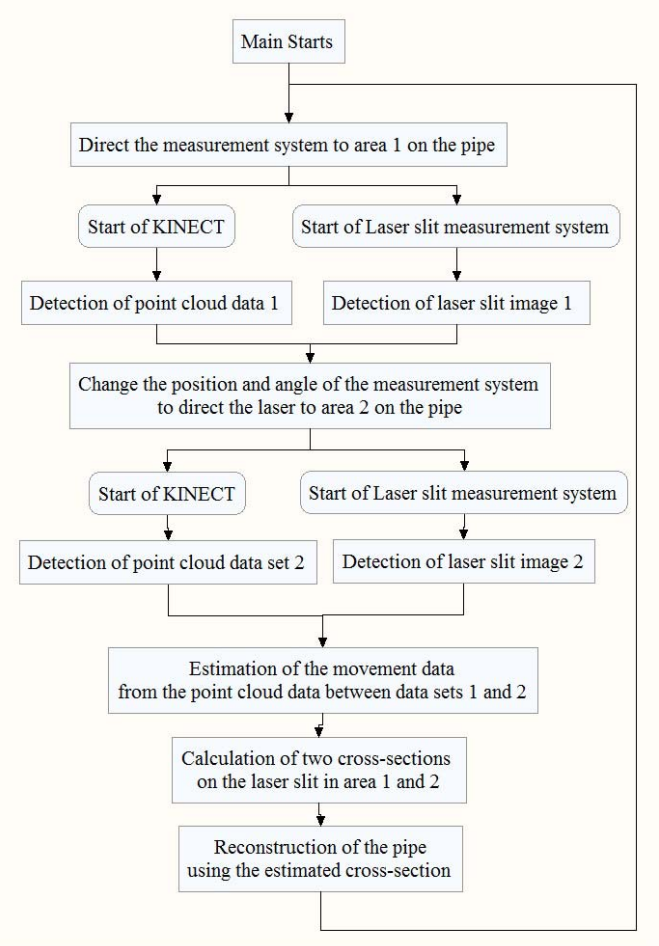


Figure 8. Flowchart of the measurement procedure.

 ISBN: 978-1-61208-353-7

Figure 9 shows the calibration setup of the high-resolution CCD camera. A calibration board is installed on the laser slit plane. An image of the scale on the calibration board is obtained by the CCD camera and is used to determine the calibration parameters. The relationship between the coordinates (u,v) of the CCD camera and the global coordinates (x,y) of the scale board is as follows:

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (3)$$

where k_{11} through k_{32} are parameters that express the rotation, scale, and displacement between the camera coordinates and global coordinates. The global coordinate system is based on the measurement system and moves with the system. These parameters are determined by inputting corresponding positions between the camera coordinates and the global coordinates. These parameters are input using a mouse for the camera coordinates and a keyboard for the corresponding global coordinates.

Parameters k_{11} through k_{32} are determined by inputting four or more corresponding points into Eq. (3). The function for converting camera coordinates to global coordinates is given as follows:

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} k_{31}u - k_{11} & k_{32}u - k_{12} \\ k_{31}v - k_{21} & k_{32}v - k_{22} \end{bmatrix}^{-1} \begin{bmatrix} k_{13} - u \\ k_{23} - v \end{bmatrix} \quad (4)$$

All points on the laser streak are converted to global coordinates, and the cross-sectional shape of the pipe is estimated.

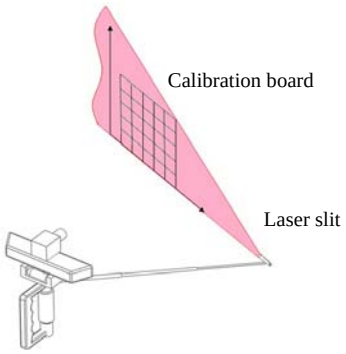


Figure 9. Calibration setup.

Figures 10(a) and 10(b) show images of the laser streak on the pipe captured by the CCD camera and the KINECT sensor, respectively. The laser streak detected by the CCD camera is clearer and more stable than the point cloud data obtained using the KINECT sensor. The resolution of the KINECT camera is 640×480, whereas the resolution of the CCD used in the present is 1,280×1,024.

The difference in the resolutions influences the measurement accuracy. The laser slit data in the point cloud data is replaced by the data of the slit-ray projection method.

(a) CCD camera image (b) KINECT image

Figure 10. Point cloud data obtained by the CCD camera and the KINECT sensor.

Figure 11. Estimated pipe cross section.

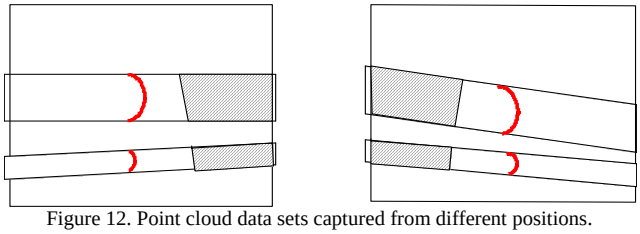


Figure 12. Point cloud data sets captured from different positions.

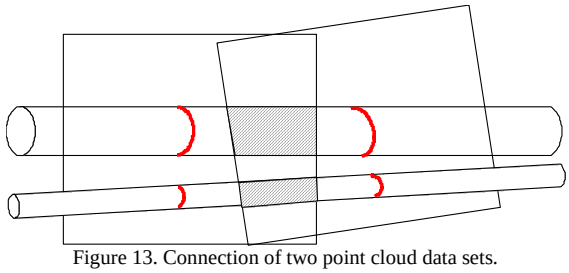


Figure 13. Connection of two point cloud data sets.

The two-dimensional coordinates (x,y) of Eq. (4) are mapped to 3D coordinates (x,y,z) by this replacement because the point cloud data have 3D coordinates.

The cross-sectional shape (ellipse) can be determined by applying the least-squares method to the data in Figure 10(a), as shown in Figure 11.

More than two point cloud data sets were obtained at different positions, as is shown in Figure 12. Each data set contains data that overlaps with data from another data set, as indicated by the shaded regions in Figure 12. These point cloud data sets are connected using the ICP algorithm as shown in Figure 13. The ICP algorithm of the Point Cloud Library (PCL) open-source framework is used in the proposed system.

 ISBN: 978-1-61208-353-7

V. RESULTS OF PIPE RECONSTRUCTION

The position and orientation of the estimated cross-section between images are then determined, and the pipe is constructed by the computer, as shown in Figure 14. In other words, the coordinates for each measurement position are converted to a common coordinate system (global coordinate system) by this connection. The proposed method can thus construct the shape of the entire pipe in a common coordinate system from point cloud data from different portions on the pipe.

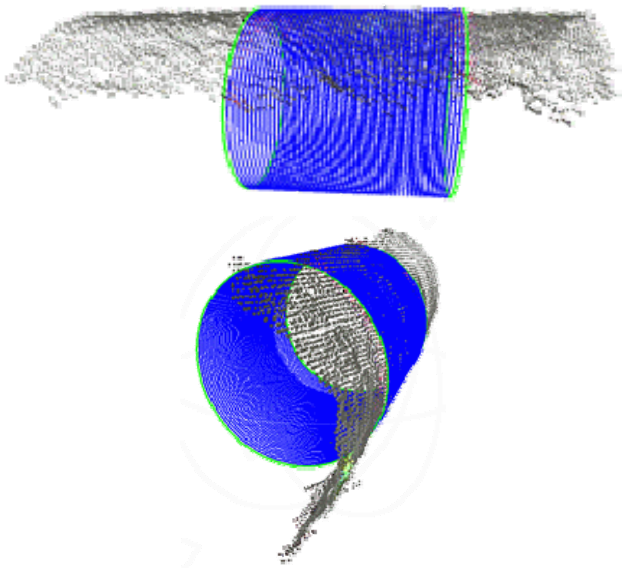


Figure 14. Pipe reconstructed on a computer.

Figure 15 shows a photograph of the pipe measured in the present experiment. The original point cloud data contains a great deal of redundant data. Six images were captured from different positions and were used to construct point cloud data on the computer. Each image includes a single slit-ray streak.

Figure 15. Photograph of the measured pipe.

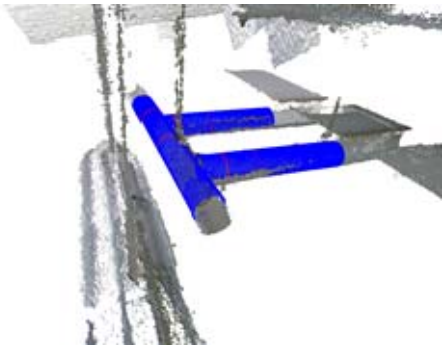


Figure 16. Pipe reconstructed on the computer.

Three pipes are reconstructed in Figure 16. Several sections are used to reconstruct each pipe. The measurement error in measuring the pipe size is less than 2 mm for distances under 2,000 mm and 0.3% over distances over 2,000 mm.

VI. CONCLUSION

Two computer vision systems that used the KINECT sensor for ubiquitous sensing in raising stock and in industrial fields were introduced. One of these systems is a 3D thermo-sensing system that detects 3D shape data and 3D temperature data simultaneously. Measurement of the 3D temperature distribution was realized by mapping thermal data obtained by thermography to the 3D position obtained by the KINECT sensor. The 3D temperature distribution of the cow head is measured for one application of the proposed system.

The other system is a handheld 3D measurement apparatus that uses a slit ray projector in conjunction with the KINECT sensor. The 3D shape of the target is reconstructed using the detected data on the computer. The measurement of pipe equipment is introduced in the present paper.

These two systems are sufficiently compact, and measurement can be performed via online processing. As such, these systems can be useful as tools for ubiquitous data acquisition systems in various fields. The experimental results obtained in the present study demonstrate the feasibility of the proposed system in raising stock and in industrial applications.

REFERENCES

[1] J. Mahoney, "Testing the goods: Xbox KINECT," 2010

[2] A. Bigdelou, T. Benz, L. Schwarz, and N. Navab, "Simultaneous categorical and spatio- temporal 3d gestures using kinect," Proc. 3D User Interface, 2012, pp. 53-60.

[3] C. Mutto, P. Zanuttigh, and G. Cortelazzo, "Time-of-Flight Cameras and Microsoft KINECT," Springer, 2012.

[4] Z. Zhang, "Microsoft kinect sensor and its effect," IEEE multimedia, vol. 19, 2012, pp. 4-10.

[5] D. Webster and O. Celik, "Experimental evaluation of Microsoft Kinect's accuracy and capture rate for stroke rehabilitation applications," Proc. Haptics Symposium IEEE, Feb. 2014, pp. 455-460.

 ISBN: 978-1-61208-353-7

[6] R. Nagayama, T. Kazuma, T. Endo, and A. He, "A basic study of human face direction estimation using depth sensor," Proc. International Joint Conference on Awareness science and technology and Ubi-Media Computing, Nov. 2013, pp. 644-648.

[7] I. Lysenkov and V. Rabaud, "Pose estimation of rigid transparent objects in transparent clutter," Proc. IEEE international conference on Robotics and Automation, May. 2013, pp. 162-169.

[8] D. Surie, S. Partonia, and H. Lindgren, "Human sensing using computer vision for personalized smart spaces," Proc. IEEE International Conference on Ubiquitous Intelligence and Computing, Nov. 2013, pp. 487-494.

[9] C. Kapoutsis, C. Vavoulidis, and I. Pitas, "Morphological techniques in the iterative closest point algorithm," IEEE Trans. on Image Processing, vol. 8, 1998, pp. 808-812.

[10] G. Sharp, S. Lee, and D. Wehe, "Icp registration using invariant features," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24 , 2002, pp. 90-102.

[11] J. Feldmar, J. Declerck, G. Malandain, and N. Ayache, "Extension of the icp algorithm to nonrigid intensity-based registration of 3d volumes," Computer Vision and Image Understanding, vol. 66, 1997, pp. 193-206.

[12] B. Lee, C. Kim, and R. Park, "An orientation reliability matrix for the iterative closest point algorithm," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 22, 2000, pp. 1205-1208.

[13] S. Gupta, K. Sengupta, and A. Kassim, "Compression of dynamic 3d geometry data using iterative closest point algorithm," Computer Vision and Image Understanding, vol. 87, 2002, pp. 116-130.

[14] K. Kawasue, T. Ikeda, T. Tokunaga, and H. Harada, "Three-dimensional shape measurement system for black cattle using KINECT sensor," International journal of circuit and Signal processing, Issue 4, vol. 7, 2013, pp. 222-230.

[15] Y. Wang, "Characterizing three-dimensional surface structure from visual images," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, 1991, pp. 52-60.

[16] D. Bhatnagar, A. Pujari, and P. Seetharamulu, "Static scene analysis using structured light," Image and Vision Computing, vol. 9 , 1991, pp. 82-87.

[17] J. Aldon and O. Strauss, "Shape decomposition using structured light vision," Visual Form, 1992, pp. 11-20.

[18] K. Kawasue, G. Uezono, Y. Gejima, and M. Nagata, “Three-dimensional Measurement by Free Scanning of a CCD Camera and Laser,” WSEAS transactions on System, vol. 3, 2004, pp. 143-147.

[19] K. Kawasue, T. Komatsu, and K. Yoshida, "Handheld three-dimensional pipe measurement system with a slit-ray projector," Proc. of SPIE vol. 8768, 2012.12, pp. 1-5.

 ISBN: 978-1-61208-353-7

A Flexible Framework for Adaptive Knowledge Retrieval and Fusion for Kiosk Systems and Mobile Clients

Simon Bergweiler

German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
Email: simon.bergweiler@dfki.de

***Abstract*—This approach describes a centralized framework that offers a flexible integration and access to different external knowledge sources via a single query interface. The advantage of this approach is to use the potential of a combination of different services to focus information and knowledge. Different SOAP or REST-based Web services can be requested in parallel and their results are integrated, analyzed and harmonized to one result structure that is sent to the user's client application. Various knowledge sources, can be integrated in the framework, without the need to know a query language of the Semantic Web like SPARQL. A special point in this approach is the discovery of services and the harmonization of the results of involved services using matching and mapping rules.**

Keywords–Semantic Web Services; OWL-S; WADL.

I. INTRODUCTION

The current paradigm of service-oriented architectures, that are highly modular, adaptable, distributed, and thus scalable leads to an increasing distribution and multi-disciplinarity of systems in today's Internet [1]. Nowadays, factual knowledge, such as pictures and videos are made accessible from anywhere through so-called cloud services. However, the distributed nature and the wide range of these services prove to be a problem. The variety of services and their different technical configurations through non-standardized custom Web interfaces (APIs) cannot be used easily from the consumer's point of view. Industry and manufacturers also recognized this and provide client applications, but only in a very static form and closely tight to their own published database, and usually in the form of mobile client apps for tablets and smartphones. However, this is precisely the problem: these client applications are customized and adapted to only fit to one specific requirement of a central source of knowledge with its interfaces. Unfortunately, with such a solution, it is not possible to formulate a comprehensive knowledge query to multiple knowledge bases across the specific domain. This is exactly the key point where the solution presented in the discussed approach will play a major role.

This paper gives an overview of the established service framework approach published by Bergweiler [2] and describes an extension and further development of the planning process and the advanced approach of the service discovery process of information sources. The objective of the approach is the design and development of a middleware that integrates information derived from various sources of knowledge. All obtained information is adapted for the particular requesting client. With this solution, new sources can be integrated, removed or modified without stringent dependencies on specific providers of information and their interfaces. Multiple heterogeneous Web services are composed and extracted data sets are analyzed and harmonized to a result set, that is returned to the client. The here presented framework combines the two classes of services: it closes the gap between classical Remote Procedure Calls (RPC) and pure Representational State Transfer (REST)-ful services calls, that binds factual knowledge extracted from Semantic Web Services like Freebase [3] or DBpedia [4], by mapping results and their respective annotations syntactically and semantically well-defined to a domain ontology.

A system prototype was developed, that answers combined requests, such as *"Give me the bordering countries of Germany and the population and some images of their capitals."*. The system decomposes the combined input and maps the respective query parts to connected matching services. After the execution of the service, the results of each query part are integrated and harmonized by means of an internal domain ontology and returned as a combined result to the respective client. In the result structure it is still indicated, from which source the relevant information comes.

The article is structured as follows: Section II gives an overview on related work. Section III describes the approach for the development of this framework for the discovery of services, the information extraction and the harmonization and fusion of all extracted information. Section IV shows how the approach can be adapted to a special domain and Section V gives a conclusion and an outlook on future work.

II. RELATED WORK

For the sake of a better understanding of Semantic Web technologies and their integration in the functioning of the presented framework, these technologies will be described below, along with references to their prototypical implementations.

Web services are defined as Web-based and self-contained software components that allow an interoperable machine-to-machine interaction and communication over networks. There exist two classes of Web services RPC and REST(ful) services. Most of the RPC-style Web services use the Simple Object Access Protocol (SOAP) [5] and the Web Service Description Language (WSDL) to describe the interface in a machine processable format. The data transfer model to which REST is based on, is present in its basic design in the Hypertext Transfer Protocol (HTTP) since the early days of the Internet. REST is used as a lightweight approach for resource-oriented loosely coupled self-contained software components - RESTful services. REST is an architectural style for distributed systems and has a syntactic description called Web Application Description Language (WADL) [6][7].

 ISBN: 978-1-61208-353-7

A. Web Service Description Language

The XML-based Web Service Description Language (WSDL) was originally designed and developed by the companies IBM, Microsoft, and Ariba to provide a standard mechanism for describing Web services in their SOAP toolkits. After the conception a proposed language draft was submitted to the World Wide Web Consortium (W3C) to define a standardized interface definition language (IDL) that defines the communicational aspect of Web services. A specific Web service endpoint with its operations and methods can be described abstractly and how the communication has to be achieved [8]. Here, WSDL strongly binds to SOAP [9] or plain HTTP [10]. Unfortunately, WSDL addresses the technical mechanisms and aspects of Web services, but does not reflect the functionality of a service and it has some strict limitations with its fixed orientation on SOAP. Furthermore, WSDL is used to generate modular source code automatically, such as stubs and skeletons for the service call. Thus, minimal interface changes mean that all parts of the program have to be regenerated until a part of the program can be used again.

The framework discussed here uses WSDL to describe classical Web services, but the new lightweight service architectures rely on another architecture model. For RESTful services this approach needs an additional service description language, the Web Application Description Language.

B. Web Application Description Language

For the description of resource-oriented loosely coupled Web services that follow the REST paradigm, first mentioned by Fielding [11], an XML-based language called WADL was developed by Sun Microsystems [7]. The aim of the development of WADL is the unified description of REST-based services in machine-readable form, in order to make processing easier and simplify the development of tools in the context of Web 2.0. Thus, WADL is the syntactic description of RESTful Web services as WSDL for SOAP-based RPC-services.

There are also efforts to attach annotations to RESTful Web services to provide automatic mediation or composition of services to so-called Web 2.0 mashups. The most common description format is SA-REST [12], Semantic annotation of Web Resources. However, this approach does not rely on WADL, but follows the basic ideas of SAWSDL. REST-based services are described in this approach by (X)HTML. A specific (X)HTML service description can be added to different metadata models, such as taxonomies or ontologies in order to describe a service semantically.

In this approach, we use WADL as service description language for the grounding of the connected RESTful sources. This is done according to the work described in [13], whereas the concrete service description for the integrated sources of knowledge of the addressed domain must be extended accordingly. See a detailed description in Section III-D *Semantic Service Repository*.

C. Ontologies and Semantic Web Technologies

A prominent and leading role in the development of Semantic Web technologies plays the W3C with the technical evaluation and specifications of its recommendations, the so-called standards. In line with the extension of the current Web to the “Semantic Web”, in which information is given well-defined meaning, to enable a better work and cooperation of people and computers [14], one of the most important data formats, the Resource Description Framework (RDF) [15], has been developed. RDF formalizes the syntax for the description of statements in the specified order of subject, predicate and object. Such triples define resources, which use a prefixed uniform resource identifier (URI) to unique characterize the type. This allows a combination of data and resources from various sources. Unfortunately, it is very difficult to model complex interrelationships in these triples. For the modeling of complex contexts and the representation of concrete domain knowledge, ontologies are the appropriate method of choice.

An ontology is often defined as a *formal, explicit specification of a shared conceptualization, with the intended purpose of enabling knowledge sharing and reuse*, whereas the conceptualization is an abstract simplified view of the world that we are trying to represent [16].

In knowledge organization the term “domain” refers to the scope for which facts of knowledge are formalized. Ontology defines a formal system, which represents knowledge by means of a fixed relevant descriptive terminology or vocabulary, relations, hierarchies and logical attributes. With this logical representation and the taxonomic structure, and the specific evaluation of these structures and interrelations, knowledge can be derived by inference mechanisms. The W3C recommends the Web Ontology Language (OWL) [17], a standardized representation language. It is currently the most advanced ontological approach, which also finds global popularity and use. The design of OWL also relies on RDF. Because of its complexity OWL currently exists in three expansion stages or versions: OWL-Lite, OWL-DL and OWL-Full. The so-called lightweight or “lite” version offers the simplest variant on the use of OWL ontologies, whereby the user of OWL-Full can exploit the full expressive power [18].

Ontologies are modeled for a specific application domain and in this approach the created ontologies define terminologies for relevant context-adaptive properties and related entities. In detail, the ontology contains a variety of customizable parameters, which can be adapted according to a specific domain, described in Section III-A *Representation structures*.

D. Semantic Web Services

The Web service ontology language OWL-S [19] is based on OWL and extends this base to a set of constructs that relate to properties, specificities and dependencies of the Web service level and is also machine readable and processable. A concrete service description in OWL-S is divided into three parts: service profile, service model and service grounding. Primarily the service profile is used for service discovery and describes what the service does. It contains information about the organization that provides the service, the preconditions, input and output values, preconditions and effects, as well as the features and benefits of the service. Once a service has been selected for use, the service profile is no longer used. For the concrete process of involving and execution of a service the description defined in the service model is used. Figure 1

 ISBN: 978-1-61208-353-7

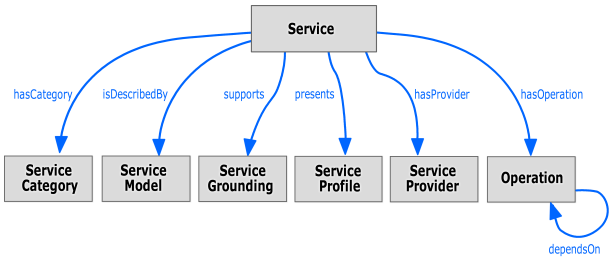


Figure 1. Main concepts of the Web service ontology language OWL-S.

shows the main concepts and relations of a service description in OWL-S.

The *Service Model* describes the actual execution process of a service. Here, this process description consists of simple atomic processes or complex composite processes that are sometimes abstract and not executable. The process describes the individual use of the service by clients by specifying input and output data, preconditions and effects.

The *Service Grounding* provides detailed technical communication information on protocols and formats as well as addressing details. Furthermore the grounding model provides a direct link or mapping between the service model and the technical service execution level. For example implementation details like input and output messages of the service model are translated into corresponding elements of the service description language. The W3C recommends and specifically describes in its member submissions WSDL, but other groundings are also possible. For a better understanding of this recommendation, it is important to know that W3C member submissions serve as input to the standards process. These descriptions contain concrete information for the service implementation and realization by enabling a direct link between the grounding and the WSDL elements. In their research articles Sirin et al. describe their prototypical implementation to directly combine OWL-S with actual executable invocations of WSDL [20][21].

This approach is based on the preliminary work in the area of semantic Web services modeling with grounding in WSDL and expands the approach to lightweight REST-based interfaces with their service descriptions in WADL [13][22].

III. Concept of the Framework

With this framework, users easily get access to Semantic Web Services without the need of special skills of RDF(S) or specific database query languages like SPARQL Protocol and RDF Query Language (SPARQL) [23]. For non-specialists the entry barrier in the world of Semantic Web technologies, to handle RDF triples, or to formulate a SPARQL query, is very high. These query languages are primarily designed to exploit the full power of the Semantic Web and make it possible to navigate in big semantic annotated data bases and specifically define restrictions to extract the specific information. As mentioned at the beginning of the article this framework uses a combination of Semantic Web technologies to create fine-grained matching answers to a given combined user query.

The composition of heterogeneous Web services and Semantic Web Services - which are playing the role of knowledge sources - poses a special challenge, because Web service interfaces or APIs might change over time. The functions are defined according to the principle of input-processing and output (IPO), as depicted in Figure 2. The interaction system (multimodal dialog system with its tablet or smartphone clients) formulates a query and the service framework generates a corresponding output after involving adequate services.

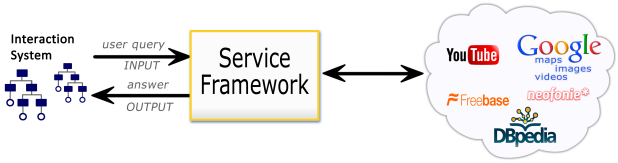


Figure 2. Process workflow of the service framework.

The focus of this paper is to provide an overview of the concept of the developed framework, to describe the discovery of services and the corresponding retrieval of semantic content. The processing chain contains many open questions in the field of:

- Query analysis and evaluation with associated mapping and matching of data structures from one format to the next. (*Query Processing*)
- Service discovery according to a complex combined query structure and the matchmaking process. (*Semantic Service Repository*)
- Composition of services to interoperable complex services. (*Planning and Execution*)
- The output presentation for each client application and the detection of duplicates. (*Output Presentation*)

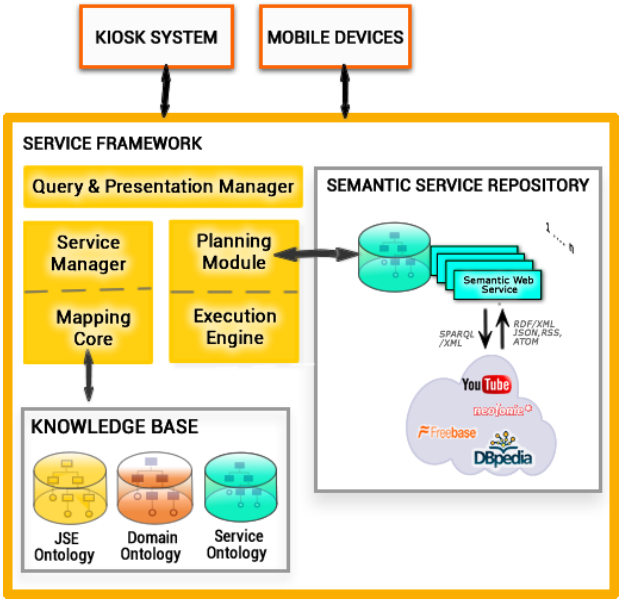


Figure 3. Architecture of the *Service Framework*.

A specific coarse-grained architecture design of the service framework is shown in Figure 3. The specific core components the (*Query Module*), the (*Planning Module and Execution Engine*) in combination with the (*Semantic Service Repository*), and finally the (*Output Presentation Manager*) form the

 ISBN: 978-1-61208-353-7

processing chain to retrieve factual knowledge out of connected knowledge sources based on semantic technologies.

A. Representation structures

The key feature of this component is the fusion of information, but with the ulterior motive or intention to integrate information and keep information interpretable and assessable. Therefore, the component harmonizes the contents on the basis of an ontology that reflects the vocabulary and domain knowledge. This solution is used in distributed application architectures as backend query point, which is the direct link in a heterogeneous world of services. The necessary knowledge contents (images, facts and videos) are introduced into the system in matching input and output formats.

The used ontologies represent and define all found metadata semantically:

- Discourse Ontology (eTFS)[24]
- Service-Framework Ontology (OWL)
- Service Ontology (OWL-S)

The *Discourse Ontology* (for modeling details see Section III-B *Query processing*) defines the vocabulary of the connected client system and characterizes the request and expected response structures. These structures must be analyzed and aligned to concepts with the similar meaning out of the *Service-Framework Ontology*. These alignments are done by using mapping data structures, defined in the *Mapping Core*. The *Service-Framework Ontology* is modeled in OWL [18]. The *Service Ontology* based on OWL-S [19] is used to describe the services in the *Semantic Service Repository* semantically.

B. Query Processing

The processing chain starts with the request of a client application to the service framework. The interface of the *Query module* distinguishes between different input formats. As mentioned at the beginning of this paper, complex natural language formulations require complex modeling languages to express and represent all dependencies and other linguistic nuances. The developed multimodal dialogue system [25] uses extended Typed Feature Structures (eTFS) [24] to formulate *complex natural language queries*. These structures comprise a hierarchy of types and typed semantic objects, which are useful to formulate semantic queries, such as *"Give me the bordering countries of Germany and the population and some images of their capitals"*. In a special case, concrete instances of pre-defined concepts of the discourse ontology of the dialog system can be adapted to formulate a specific query. However, the query is intentionally abstract and does not specify which services in particular should be used and how the result should be obtained. Figure 4 shows how to define such a composed query with focus on bordering countries and main capitals. The abstract eTFS query specifies two attributes ("slot" types):

- `CONTENT`: This slot contains a semantic object and represents input data.
- `FOCUS`: This slot points to the part what is requested and may appear more than once to address and define the abstract search options.

```
<object type="http://dfki.de/dm#AttributeRetrievalTask">
 <slot name="http://dfki.de/dm#hasTrigger">
  <object type="http://dfki.de/ont/odp#InfoRequest">
   <slot name="http://dfki.de/ont/odp#hasContent">
    <object type="http://dfki.de/dm#Country">
     <slot name="http://dfki.de/dm#countryCode">
      <value type="String"><![CDATA[DE]]></value>
     </slot>
    </object>
   </slot>
   <slot name="http://dfki.de/ont/odp#hasFocus">
    <object type="http://dfki.de/ont/odp#Focus">
     <slot name="http://dfki.de/ont/odp#hasContent">
      <object type="http://dfki.de/dm#Country">
       <slot name="http://dfki.de/dm#hasBorderingPlace"/>
      </object>
     </slot>
    </object>
   </slot>
   <slot name="http://dfki.de/ont/odp#hasFocus">
    <object type="http://dfki.de/ont/odp#Focus">
     <slot name="http://dfki.de/ont/odp#hasContent">
      <object type="http://dfki.de/dm#Capital" />
     </slot>
    </object>
   </slot>
  </object>
 </slot>
</object>
```

Figure 4. Complex combined query formulated in eTFS-Format including search topics and properties as input parameters.

In parallel to complex formulated queries, in formats like eTFS, a second simple format, where queries are assembled in XML [26], can also be used. The interpretation and analysis of simple XML queries to align the requested content with the framework's ontology is carried out in accordance to an underlying schema and taxonomy. There exist predefined mapping rules that bundle the concepts or objects of decomposed query parts of the input structures to domain concepts. For this procedure element-based matching techniques are used, according to each decomposed query part, concepts or objects are mapped to a local meta-representation, the internal framework's ontology. The query component extracts knowledge concepts and adds them to predefined ontological structures. The outcome of this are individuals, basic ontological components, and relations to internal concepts that represent the query in a processable, framework-readable form. It specifies the input type, such as a city or a country, specified by properties (complete name, keywords, etc.), implicit relations and search topics (area, population, etc.). A detailed description on query processing and decomposition of this framework in combination with a multimodal dialog system is discussed in [2].

C. Planning and Execution

Due to the high amount of services, planning has become a major task. Moreover, the goal of the system is also to carry out the planning and composition of these services automatically. Pre- and post-conditions must clearly characterize a problem in the planning process, to allow a compositional process. Due to the interoperability of services, it is easier to transfer functionality, without having to integrate the services completely. The task is to find Web services and to either decide to compose them or call them directly. Here, the composition workflow can be quite complex, under consideration and evaluation of input and output parameters. In most cases the

 ISBN: 978-1-61208-353-7

solution to a complex query can be solved by a combination of services or the sequential invocation of multiple services. This consideration in turn raises the issue of how services are interconnected and to find out if a static execution sequence is needed. An execution sequence defines the composition or concatenation of services, whereby a service at its execution time ($S1_{(t)}$) might need for its execution the result structures of another service ($S2_{(t-1)}$). This management of the supply chain opens up new challenges - services need to work together to compose complex services [27].

This paper does not set the focus on planning and composition: to discuss the algorithm in detail is beyond the scope of this paper. It is intended to give an overview of the developed framework and the discovery of services. The here presented framework uses Business Process Model and Notation (BPMN) a high-level notation that specifies semantics for composition, to orchestrate services and provide aggregated results [28]. The advantage of this approach is to use the potential of a combination of different services to focus information and knowledge. The planning component works with predefined plans, that define preconditions that must be matched by adequate modeled input query structures. A plan contains the description on how to proceed in the discovery and execution process, which abstract types of services are needed, the domain that is addressed, in which order the services have to be executed, and all essential parameters for the matchmaking process in the connected *Semantic Service Repository*, that opens up access to services that encapsulate knowledge sources of different domains.

The *Execution Engine* provides connectors and encapsulates the calls to the REST or API interfaces, by reformulating and using specific query formats like XML or languages like SPARQL and Metaweb Query Language (MQL) [3]. Once all results of different called services are received by the *Execution Engine* an internal mapping process starts a review and reasoning process and maps them with the help of additional semantic mapping rules and classifies them according to the internal framework's domain ontology.

D. Semantic Service Repository

The discovery and matchmaking of services is a major part in this processing chain. The *Semantic Service Repository* provides and manages semantic descriptions of various pre-annotated information sources. One of the advantages of this component is the functionality to add, modify, replace, and delete these stored descriptions of sources without hard programmatic dependencies and without stringent dependencies on specific providers of information and their interfaces (APIs). Certainly the component benefits from the interoperability of services. The *Semantic Service Repository* provides access to different types of information sources like Semantic Web Services that cover information stored in external database management systems or *Semantic Repositories*, such as DBpedia [4] and Freebase. These systems store information in a structured and manageable form, but can be only accessed by special query languages like SPARQL for DBpedia or MQL for Freebase. The main difference of this approach compared to conventional database management systems is the usage of ontologies as a technology to harmonize and store semantically structured data - concepts define and classify information and contain implicit knowledge characterized by name and position in the hierarchy. In this approach, all Web services are represented in OWL-S with a grounding declaration in WSDL or WADL [21][22][29].

Figure 5 shows the internal discovery process of the *Semantic Services Repository*. An XML query is sent by the *Planning and Execution* component to the *Broker* module. On the basis of a single XML schema, the *Query Handler* interprets the input, that characterizes abstract types of services, and maps them into an internal interpretable data collection, that groups multiple input elements into a single unit. Based on this input data collection a rule engine evaluates the input. A consideration of the *Rule Engine* relies on a set of predefined rules that are stored in a local rule base. When the *Broker* module is involved, the rule base is filled with these predefined rules and a working memory is generated and selected application data and objects are provided. The analysis of the planners' request is done exactly by these rules. Therefore, the rules must express in their condition part the terminology of the problem domain. When a rule matches within the current context the rule's consequence part refers to a particular SPARQL query. This means that the defined conditions are attuned and adapted to filter options of the SPARQL query in the consequence part of the rule. In line with [30], SPARQL is used to define filters that are precisely tailored to the *Service Ontology*. With these filter options the search for matching services can be arbitrarily fine-grained and started very restrictive. If no results are available, the parameters and concept types can be adjusted according to the hierarchy of the *Service Ontology*. This query is forwarded to the *Matcher* module, that connects the service repository and executes the received SPARQL query, which relates to adequate and concrete Semantic Web Service representations of deposited services. When the query matches, a list of semantic services descriptions is returned, which describe ontologically, how to interpret and execute the service. All parameters, that are required for service execution, are taken from the ontological description and added to the assembled concrete request structure for the external service call. The resulting data structure is returned to the *Planning and Execution* component that triggers the process of integration, analysis and harmonization.

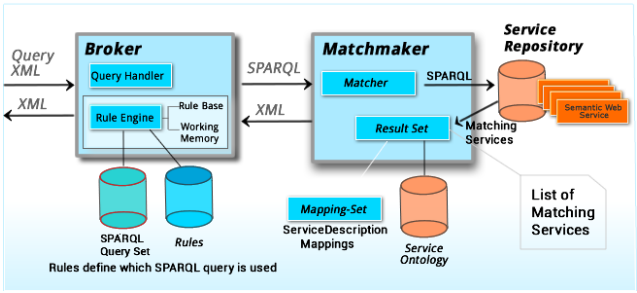


Figure 5. Service discovery process.

In this approach, the grounding description with technical concepts of WADL and WSDL, that describe the communication of external services, has been extended by parameters or properties that refer to corresponding pattern generated SPARQL queries, which are used to call sources like DBpedia or other triple stores. Furthermore, each SPARQL query defines the parameters that are used in the output structures, and which must be mapped to the internal ontology. Therefore, another reference is attached to the grounding description,

 ISBN: 978-1-61208-353-7

which describes how the results of this SPARQL call must be mapped to the internal framework's ontology.

E. Mapping and matching

In the processing chain of the framework content of heterogeneous data channels must be repeatedly aligned: this is called mapping. In this paper, the concepts of mapping and matching are used interchangeably. A distinction of these concepts is just possible in a concrete application scenario, where element transformations occur, but not out of this abstract view. The key aspect of mappings is to combine one representation structure with another. Whereas the major challenge is to create comparable and interoperable mapping functions between the used terminologies, the mappings of the result structures of the stored service descriptions have been defined in a pre-processing phase in a formal description language. For the unambiguous assignment of the models and types of an element the mapping functions are precisely specified by categories.

The *Mapping Core* realizes these mappings at each part of the framework and forms the central interface between the representation structures of each knowledge domain and can be replaced accordingly. Figure 6 illustrates the internal communication and interaction workflow of the individual core components in detail. An incoming request is evaluated by the *Query Module*, as described in Section III-B, and is mapped to the internal domain ontology that is used for further processing in the planning process. Additionally, in the *Planning Module* and the *Execution Engine* the service framework involves knowledge sources that are encapsulated or wrapped by Web service interfaces.

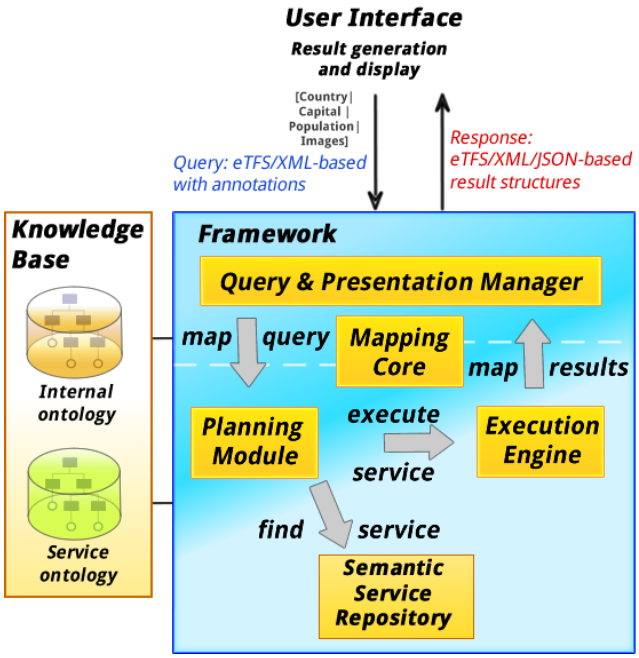


Figure 6. Internal communication workflow - query processing and data type mapping.

As an example, Figure 7 shows the procedure of mapping result structures of a remote relational data base by formal declaration of mappings to the internal framework's ontology. A mapping of the results of the called external sources, such as complex external semantic data structures or simple XML structures to the internal framework's ontology must be fulfilled and with the help of these mapping and transformation rules the process of data harmonization can be controlled.

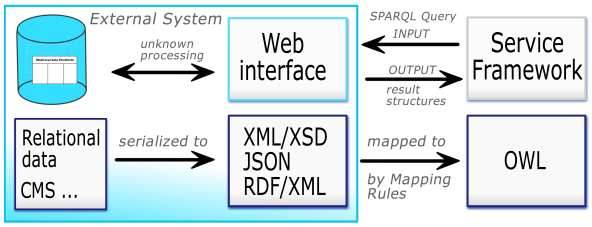


Figure 7. Dataflow of mapping external result structures.

In the case of mapping complex external semantic data structures, formal mapping rules are used that allow a higher quality data type mapping on a more abstract level: new individuals are created and linked to each other or a taxonomy of objects must be mapped to the internal data structure. For the mapping of simple XML result structures, an adaption of Extensible Stylesheet Language Transformations (XSLT) [31] can be used to create a meta XML format that is mapped to individuals of the framework's ontology. Figure 8 shows the mapping of an external data structure that contains the geo-coordinates of a city. The parameters `longitude` and `latitude` are mapped to a new individual of type `GeoLocation`. This new individual is bound by the reference `hasGeoLocation` to the existing individual with the concept type `CityTownVillage`.

The translation of data models of different domains with their vocabulary and definitions is the first step towards semantic harmonization of incoming results of connected services and the most important step in creating comparable content. These translations based on the extracted semantic information are necessary to create correct translations with the aim of identification of equality, similarity and heterogeneity. But particular problems (misspelling and typing errors) for the harmonization emerge, which need to be fixed separately by domain specialists. The analysis searches all generated individuals in the framework's ontology to find structures with the same formal and semantic criteria. Duplicates could just be filtered at the instance level, because the information with their semantics is here clearly and comparable. Found images results

```
CONDITION[
 INSTANCE[jse:CityTownVillage]
  REFERENCE[
    INSTANCE[jse:GeoLocation]
    INSTANCE_XPATH[//geo:results/geo:result]
    SET_RELATION(jse:CityTownVillage#hasGeoLocation
                    ->jse:GeoLocation)
      if(XPATH[//geo:binding/@name]==longitude) {
        XPATH[//geo:binding/geo:literal]->longitude
      }
      if(XPATH[//geo:binding/@name]==latitude) {
        XPATH[//geo:binding/geo:literal]->latitude
      }
]]]
```

Figure 8. Mapping of the geo-coordinates of a city into the framework's ontology.

 ISBN: 978-1-61208-353-7

can only be compared on the basis of textual comparison of title, captions, and descriptions.

F. Output Presentation

The creation of the output structure is the last step in the processing chain. This is done by the *Presentation Manager* which coordinates, transforms and merges the created semantic individuals in a standardized result structure. Depending on the user's query and the type of client application (smartphone, tablet or desktop) the output format is specified (RDF, XML, JSON, eTFS, etc.). The component uses the *Mapping Core* for the filtering and extraction of individuals out of the framework's ontology and the declarative element-based mapping of concepts and properties to data collections of the resulting structures. These prepared contents are delivered to the appropriate connected client applications.

The component works statelessly in terms of domain-specific parameters, the client application defines what processing has to look like and sets the processing context. After delivery of the result structure oblivion continues and the local retrieved data structures are deleted, to avoid a mixture of contents and interpretation errors when the next query is received.

IV. Scenarios

For many people mobile devices, such as tablets or smartphones, have become indispensable in today's modern world. These devices offer the possibility of full mobility, anywhere, at any time information is provided and people make use of these offers. Accordingly the behavior of the users changed and reinforces the focus on Web services, the infrastructure components of the Internet. Companies have recognized the increasing demand of that market where Web services provide dynamic factual knowledge on request, and make commercially use by bundling functionality of their Web services in cloud solutions [32].

With the presented service framework a service-oriented information kiosk system for public places, like museums or hotel lobbies, has been developed, to force and support multi-user collaboration [33]; see Figure 9. Users can connect their smartphones by app (Android, iOS) to a large terminal and share interesting facts, pictures and videos via gesture interaction.

The advantage of this multimedia platform is the seamless combination of a touch-based kiosk system and the access of heterogeneous information sources via the described service framework. It is also possible to interact and control semantic interaction elements (SIE) [34] the kiosks' applications by speech. The in-built smartphone's microphone is used to get the user's speech input (utterance). Figure 10 shows the prototypical implementation of the developed service framework in combination with multi-modal mobile access to external information by a dialog system [24].

Furthermore, a system was developed to interrogate geological facts, such as finding rivers, their length and right or left tributaries, also in combination with a multimodal dialog system [24]. The complete integration of factual knowledge out of standardized *Web Feature Services* [35] that are involved to access geological data is achieved by the service framework, using an adapted domain ontology.

Figure 9. The service-oriented Calisto kiosk system.

Figure 10. Interaction with the Calisto system prototype.

V. Conclusion and Future Work

This paper presented an approach that describes the flexible integration of heterogeneous knowledge sources and their parallel execution and analysis as well as the harmonization of the results by means of an internal framework's ontology. The solution addresses heterogeneous data sources and integrates information in spite of the different query languages and different enriched domain ontologies. Therefore, for better understanding an overview of the configuration and structure of the core components of the developed frameworks was given. Furthermore, the alignment and matching of data structures along with a detailed description of the service discovery process in semantic service repositories were presented. Moreover, it was shown how implemented prototypes of the discussed framework were used in different scenarios by using factual knowledge of their respective domains like geography or touristic information. In this context the framework acts as an important factor to connect and integrate structured information.

Future work will address the point of switching the domains. In order to make it easier for the domain experts a workbench is needed to generate the mapping structures with graphical support, because manual mapping is time consuming and error prone.

Another key aspect is the optimization of result structures. Better filtering mechanisms are tested in order to detect duplicates or merge similar content. Furthermore, depending on

 ISBN: 978-1-61208-353-7

the domain a visual pattern recognition can be used to eliminate duplicate images.

REFERENCES

[1] J. Bih, "Service Oriented Architecture (SOA) a New Paradigm to Implement Dynamic e-Business Solutions," Ubiquity, no. August, 2006, pp. 4:1–4:1.

[2] S. Bergweiler, "Interactive service composition and query," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, pp. 169–184.

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in Proceedings of the 2008 ACM SIGMOD international conference on Management of data. New York, NY, USA: ACM, 2008, pp. 1247–1250.

[4] S. Auer et al., "DBpedia: A Nucleus for a Web of Open Data," in The Semantic Web, ser. Lecture Notes in Computer Science, K. Aberer et al., Eds. Springer Berlin Heidelberg, 2007, vol. 4825, pp. 722–735.

[5] "SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)," W3C Recommendation 27 April 2007, April 2007, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/soap12-part1/

[6] L. Richardson and S. Ruby, RESTful Web Services. O'Reilly Media, 2007.

[7] M. Hadley, "Web Application Description Language (WADL): W3C Member Submission," August 2009, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/Submission/wadl/

[8] E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana, "Web Services Description Language (WSDL) 1.1," W3C, W3C Note, March 2001, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/wsdl

[9] M. Gudgin et al., "Soap version 1.2 part 1: Messaging framework (second edition)," W3C Recommendation REC-soap12-part1-20070427, April 2007.

[10] R. T. Fielding et al., "Rfc 2616, hypertext transfer protocol – http/1.1," 1999, [retrieved: July 2014]. [Online]. Available: http://www.rfc.net/rfc2616.html

[11] R. T. Fielding, "Architectural styles and the design of network-based software architectures," PhD Thesis, University of California, Irvine, 2000.

[12] K. Gomadam, A. Ranabahu, and A. Sheth, "SA-REST: Semantic Annotation of Web Resources: W3C Member Submission," April 2010, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/Submission/SA-REST/

[13] O. F. F. Filho and M. A. G. V. Ferreira, "Semantic Web Services: A RESTful Approach," in IADIS International Conference WWW/Internet 2009. IADIS, 2009, pp. 169–180, [retrieved: July 2014]. [Online]. Available: http://fullsemanticweb.com/paper/ICWI.pdf

[14] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, 2001, [retrieved: July 2014]. [Online]. Available: http://www.jeckle.de/files/tblSW.pdf

[15] G. Klyne and J. J. Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax," W3C Recommendation, 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[16] T. R. Gruber, "A translation approach to portable ontology specifications," Knowledge Acquisition, vol. 5, 1993, pp. 199–220.

[17] D. L. McGuinness and F. van Harmelen, "OWL Web Ontology Language Overview," W3C, W3C Recommendation, 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/owl-features

[18] P. F. Patel-Schneider, P. Hayes, and I. Horrocks, OWL Web Ontology Language Semantics and Abstract Syntax, W3C W3C Recommendation, Feb. 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/

[19] D. Martin et al., "OWL-S: Semantic Markup for Web Services," 2004, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/Submission/2004/SUBM-OWL-S-20041122/

[20] E. Sirin, B. Parsia, and J. Hendler, "Filtering and selecting semantic web services with interactive composition techniques," IEEE Intelligent Systems, vol. 19, no. 4, 2004, pp. 42–49.

[21] E. Sirin, "Combining description logic reasoning with ai planning for composition of web services," dissertation, University of Maryland, College Park, 2006.

[22] D. Lambert and J. Domingue, "Grounding semantic web services with rules," in Proceedings of the 5th Workshop on Semantic Web Applications and Perspectives (SWAP), ser. CEUR Workshop Proceedings, A. Gangemi, J. Keizer, V. Presutti, and H. Stoermer, Eds., vol. 426. CEUR-WS.org, 2008, [retrieved: July 2014]. [Online]. Available: http://ceur-ws.org/Vol-426/swap2008_submission_8.pdf

[23] "SPARQL 1.1 Query Language - W3C Recommendation," W3C Recommendation, 2013, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[24] T. Becker et al., "A unified approach for semantic-based multimodal interaction," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, pp. 135–148.

[25] D. Porta, M. Deru, S. Bergweiler, G. Herzog, and P. Poller, "Building multimodal dialogue user interfaces in the context of the internet of services," in Towards the Internet of Services: The Theseus Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer Berlin Heidelberg, 2014, pp. 149–168.

[26] K. G. Clark, "Extensible Markup Language (XML) 1.0 (Fifth Edition): W3C Recommendation," W3C Recommendation, November 2008, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/xml/

[27] M. Ghallab, D. S. Nau, and P. Traverso, Automated Planning - Theory and Practice. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2004.

[28] P. Y. Wong, "Compositional development of bpmn," in Software Composition, ser. Lecture Notes in Computer Science, W. Binder, E. Bodden, and W. Löwe, Eds. Springer Berlin Heidelberg, 2013, vol. 8088, pp. 97–112.

[29] M. Paolucci, T. Kawamura, T. R. Payne, and K. Sycara, "Semantic matching of web services capabilities," in The Semantic Web - ISWC 2002: First International Semantic Web Conference Sardinia, Italy, 2002, pp. 333–347.

[30] J. M. Garcia, D. Ruiz, and A. Ruiz-Cortes, "Improving Semantic Web Services Discovery Using SPARQL-Based Repository Filtering," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 17, 2012, pp. 1–15.

[31] "Xsl transformations (xslt) version 1.0," W3C Recommendation, 11 1999, [retrieved: July 2014]. [Online]. Available: http://www.w3.org/TR/xslt

[32] T. Barton, "Cloud computing," in E-Business mit Cloud Computing. Springer Fachmedien Wiesbaden, 2014, pp. 41–52.

[33] S. Bergweiler, M. Deru, and D. Porta, "Integrating a multitouch kiosk system with mobile devices and multimodal interaction," in ACM International Conference on Interactive Tabletops and Surfaces, ser. ITS '10, ACM. New York, NY, USA: ACM, 2010, pp. 245–246.

[34] D. Sonntag, M. Deru, and S. Bergweiler, "Design and implementation of combined mobile and touchscreen-based multimodal web 3.0 interfaces," in Proceedings of the International Conference on Artificial Intelligence, ser. ICAI-09, July 2009, pp. 974–979.

[35] "Web feature service implementation specification," OGC Document 04-094, Version 1.1.0, Standard, May 2005, [retrieved: July 2014]. [Online]. Available: http://portal.opengeospatial.org/files/?artifact_id=8339

 ISBN: 978-1-61208-353-7

Ubiquitous Smart Home Control on a Raspberry Pi Embedded System

Jan Gebhardt, Michael Massoth, Stefan Weber and Torsten Wiens
Department of Computer Sciences
Hochschule Darmstadt - University of Applied Science
Darmstadt, Germany
{jan-michael.gebhardt | stefan.b.weber}@stud.h-da.de
{michael.massoth | torsten.wiens}@h-da.de

***Abstract*—This paper describes an approach to use the embedded system Raspberry Pi to serve as a communication gateway between mobile devices and Konnex-Bus (KNX) home automation systems. The Session Initiation Protocol (SIP) and the Presence Service are used to build a system concept on open source and standardized software services. The concept focuses on the communication, access control and security of that gateway. This paper manifests the possible components and potential purposes of the concept. It is shown that small embedded systems like the Pi can provide a simple and cheap solution to enable ubiquitous Smart Home Control using existing infrastructures.**

Keywords-KNX; SIP; Smart Home; Raspberry Pi.

I. INTRODUCTION

The Konnex Bus (KNX) standard has been the de facto standard in home automation systems for many years. The system is widely used for new installations and has been retrofitted into many existing buildings. The standard is open, internationally accepted and standardized in several countries [1]. KNX is probably the most used building automation system on the market. The Internet brought new technologies for communication between people. The underlying technologies and protocols can also be used to communicate with machines. By merging these technologies, we get an intelligent or ”smart” home, which reflects a current trend in information technology. A smart home shall enable interaction with its owner, including the ability to monitor the status and control of home appliances and devices remotely from anywhere in the world. Such devices may consist of alarm systems, keyless access control, smoke detectors, light, heat, water or other energy management systems, medical devices, and all types of sensors, e.g., room-, door-, window- or security surveillance, monitoring and control, statistics and remote metering to every automated system and appliance in the home. With the increasing availability of smartphones and access to the Internet at any time, it is reasonable to use these devices to remotely control our smart home. The research of this project was focused on the development of a KNX-to-SIP proxy, to interconnect the home automation system with new communication protocols. The software should run on an embedded system, such as the Raspberry Pi [2], to ensure low resource consumption and to be cost-effective. Furthermore, the whole system should be compliant with open standards.

A. Purpose and Relevance

The purpose of this paper is to show that small embedded systems, in our case the Raspberry Pi, can be used to run the smart home software, developed within this project. Furthermore, a new communication model is introduced, which aims to improve the resource-consumption within the Session Initiation Protocol (SIP) [3]. More precisely, the actor and sensor information was stored within separate SIP profiles, which can be improved. The fact, that the presence information is visible to all SIP users, generates a need for a detailed security concept. This design includes a draft to implement access control, as well as communication security into the home automation remote control framework.

B. Structure of the Paper

Following this introduction, Section II describes related work and other interesting projects suitable for this concept. In Section III, an overview of the general approach is given. The components of a possible system design are discussed in Section IV. After that, we introduce two use-cases to create a basis for our concept, which will be described in Section VI, leading to a system design. Next, an overview over the possible communication security layers is given in Section VII. Sections IX and X conclude the paper and give an outlook on future work.

II. RELATED WORK

The concept of intelligent or smart home control has been well-established in IT. Many companies and institutions are working on solutions or even released their individual software. Some of these systems also come with their own hardware sensors and actors to create a Smart Home. Most of these software solutions are proprietary and not compatible with other home automation systems. With the approach developed by this project, it is possible to use the existing home automation system KNX to enable secure remote control. Whereas other projects like MavHome [4] aim to create the intelligence of a smart home, we use the existing intelligence, provided by the home automation system itself. The communication with remote clients is enabled by using existing infrastructure and open standards, such as SIP [3]. Henning Schulzrinne et al. presented how ubiquitous computing could be integrated into home networks with SIP [5]. Also, an IETF Working Group

 ISBN: 978-1-61208-353-7

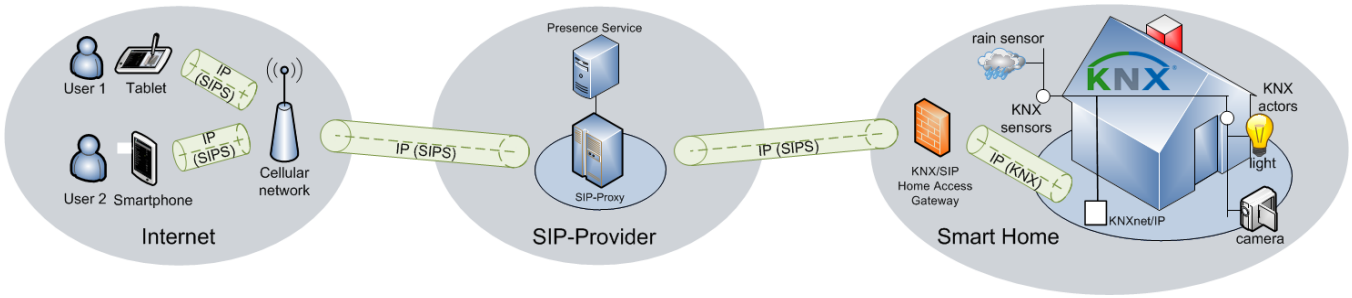


Figure 1. System architecture and components of our Smart Home.

described that future buildings will probably be equipped with a full featured IP network. They also chose SIP as the communication protocol [6]. In previous steps of the project, a proof-of-concept has been implemented [7]. The information describing the Smart Home components' status is stored within the Presence Information of a SIP user, and other SIP users can subscribe to this information via the Presence service. The existing implementation uses a separate server system as a host for the SIP proxy and the KNX/SIP Bridge. An Android tablet with a self-developed application is used to communicate with the setup. The HomeSip project described the use of SIP as a communication middleware to support home automation applications [8]. The concept is similar to our approach, where a SIP proxy is used as a gateway to enable communication between the devices. Another way to connect to the home automation network over the Internet would be to establish a Virtual Private Network (VPN) connection. In this case, the external host is virtually placed into the internal network and is able to directly communicate with the KNX/IP-bus. This solution requires a continual connection to the network. With our approach, there is no need for a continual connection. If a state change occurs, the client receives a push notification specified within the standard SIP protocol. This significantly reduces the connections in comparison to a traditional pull/continual configuration. Furthermore, a VPN connection by default enables the client to get access to all internal network devices. With our approach, we introduce access control lists to the home automation gateway, which eliminate this lack of security.

III. Approach

The new approach aims to use the standardized SIP server technology to show that our implementation does successfully interact with standard SIP proxy in the intended way. The approach will eliminate the need for a separate SIP server within the Smart Home. It is possible to use an external SIP account from any service provider. Therefore, the new implementation uses FreeSwitch as a SIP proxy with presence service. For efficiency and resource-optimization, we now use a new way to store the whole information about all sensors and actors inside the smart home within the presence state of one single SIP account. By using the Raspberry Pi as an embedded system, the separate server system is obsolete. In addition to these changes, we introduce access control, as well as general security to the project.

IV. Components

The following section details all components of our approach and sketches a general overview over the used technologies. The whole architecture of the project is shown in Figure 1.

A. Mobile Clients

Mobile Clients are used to connect to the Smart Home. Figure 1 illustrates, that several Mobile Clients can connect to the Smart Home simultaneously over various access networks. With the self-developed client, it is possible to connect from anywhere in the world to the Smart Home. The software uses SIP/SIPS to communicate with the presence service of the SIP provider to gather the information about the Smart Home. Based on that information, the KNX-sensors and -actors are displayed to the user, and it is possible to interact with actors.

B. SIP-Provider

The basic advantage of our concept is the usage of the presence service implemented as an extension to the SIP protocol. It enables event-based notifications in near real-time. The sensor and actor states are stored inside the presence information of the corresponding Smart Home SIP user. Push messages are sent whenever a sensor or actor changes its state. SIP providers already maintain the infrastructure, basically consisting of a SIP proxy and the Presence Service. The KNX/SIP Home Access Gateway sends all information about the Smart Home to the Presence Service.

C. KNX/SIP Home Access Gateway

The main task of the device is to function as a gateway between the two technologies KNX and SIP. On one side, the KNX/IP-Bus is monitored for state changes. It is also possible to write on the bus, for example to switch the light on. On the other side, the relevant information is published to the SIP Presence Service. This device also enforces the security concept detailed in section VII. The embedded system Raspberry Pi is used as platform for the KNX/SIP Home Access Gateway. The Raspberry Pi was developed by Raspberry Pi Foundation from the UK. It is a small embedded system, operating at 700Mhz, with a graphic chipset able to render up to 1080p [9]. It is a cheap but also fully functional computer system, whose performance is sufficient to run the Smart Home software. Furthermore, it is capable to be extended with other features by using the built-in GPIO-Pins [9]. As an additional feature,

 ISBN: 978-1-61208-353-7

the chipset used in this unit is equivalent to a chipset used in cellphones, which does not need additional cooling. This is a benefit for our concept, because it is silent and can be put into the electric cabinet of the Smart Home. In our concept, the Raspberry Pi is acting as a gateway between the KNX- and the SIP-protocol.

V. USE CASES

Our concept, especially our security concept as improvement to known techniques, is based on the following use cases. These scenarios focus on the access to the Smart Homes, particularly regarding the security aspects.

A. Facility Manager

As one example, we selected a facility manager, who needs to control a lot of facilities, for example Smart Homes. When we transfer this to our university, every one of its facilities be assumed a unique Smart Home, which is controlled by a facility manager. The buildings are open from 8:00 to 19:00 o'clock. If there is a lecture before or after that time, a special building has to be opened by someone. Because of this, the facility manager has to open these buildings with his user account. On the other hand, he has to monitor the state of all doors or lights to close them or turn them off at the end of the day to save energy. Besides the facility manager, every department has to be able to control their own buildings to manage the lecture halls. Because of this, there is a need of splitting actors and sensors into groups or merging Smart Homes to a bundle under the consideration of access rules. In Section VII, we will introduce such a feature.

B. Guestroom

As another example, we selected a scenario, where we have a guest at our Smart Home. Displayed in Figure 1, we assume that User1 is the Smart Home owner and User2 is our guest. The Smart Home is able to separate each room from another. Our guest should be able to control the guestroom during his visit, so he can set up the radiator to be warm, when he comes home. On the other hand, it is intended to limit his access only to relevant parts of the Smart Home. This motivates a configuration mechanism, allowing granting or denying access to specific actors and sensors, so the guest only is capable of controlling the guestroom. This leads us to the need of a technology to set up access conditions to actors and sensors, which will be introduced in Section VII.

VI. CONCEPT

Our concept aims to combine an independent Smart Home with next generation network techniques to reach a stable and secure connection over the Internet. The basic idea is to use systems already in place without any modification, so that additional implementation work is only necessary at the communication endpoints. Based on preliminary work by Massoth et al., called *"Ubiquitous Home Control based on SIP and Presence Service"* [7], we realized an enhanced example of combining a Smart Home with NGN-Technology. The main difference between this research and the work at hand is the exchange of data through the presence service. The past research needs one SIP profile for each sensor or actor,

```
<?xml version="1.0" encoding="UTF-8"?>
<presence xmlns="urn:ietf:params:...">
  <dm:person id="p1">
    <dm:note>Available</dm:note>
  </dm:person>
</presence>
```

Figure 2. Simple PIDF-Extract.

so the server-side effort is very high and not compatible to current implementations of providers. Within the new concept, all sensor data is firstly merged and then transferred to the presence service. This accumulation will be done by using a standardized XML-Scheme named Presence Information Data Format (PDIF), which will be described in the following section. Therefore this concept does not depend on cooperation with SIP providers, because it mainly adds Smart Home control functionality on top of the status information.

A. FreeSwitch as SIP Proxy

To implement our concept, we chose FreeSwitch as a SIP proxy. FreeSwitch is one of the most commonly used SIP servers on the market. It provides a very high level of standard conformity, which is useful for thorough interconnection testing as intended. It is licensed under the Mozilla Public License (MPL). In comparison to the second leading VoIP-Daemon, FreeSwitch proves better stability on a higher scale of client-usage. FreeSwitch supports a bundle of modules to achieve communication through different protocols, e.g., SIP or XMPP. It also provides the presence service, which is actually needed by this concept.

B. Construction of the Communications Protocol

The Protocol is based on the PIDF, which is introduced in the next subsection. Basically, it is a standardized XML-Scheme to exchange status information over the presence service. Our concept is to use this protocol and further enhance it by embedding additional data into that scheme.

1) Presence Information Data Format (PIDF): PIDF as standardized XML-Scheme is used as previously explained to exchange status information through the presence service. The scheme is able to divide real persons from simple or complex devices, like an answering machine or a fax. Figure 2 depicts an example of a person's current availability status. Everyone who is subscribed to him will get this information.

2) JavaScript Object Notation (JSON): JSON offers a smart and compact way to store sensor data into a PIDF-Scheme. This scheme is highly compressed and furthermore can be interpreted by JavaScript as well. Our end-users directly benefit from this behavior, because not only JavaScript or a simple Webpage is able to interpret this scheme, but also very complex software in Java. They are both able to interpret it by default. The scheme is shown in Figure 3. It can encapsulate as many arrays as exist in a database. Variable names and values are separated by a colon, whereby variables are separated by a comma. Every value can thereby encapsulate an extended set of variables. Within this structure, every actor and sensor is represented by a variable with a set of extended variables. So, the *temperature-sensor* is saved as variable with two additional

 ISBN: 978-1-61208-353-7

```
{temperatur-sensor: {
            type: Double,
            value: 22.3}}
```

Figure 3. Example of JSON encoded sensor-data.

variables. These additional variables define the type and value of the *temperature-sensor*-data. Like *double*, we also defined simple data types like: *integer, string* and *boolean*.

3) Changing actor status: In comparison to the status monitor for sensor data, the process to change the status of the data does not need the presence service. To set a new status to an actor there is only the need to send a simple message to it. These messages contain the same JSON structure, as the presence information of sensors explained above. To compare this information, we can assume *radiator-control* instead of *temperature-sensor* in Figure 3.

C. Smart Home

As already mentioned, the Smart Home is connected through a gateway, which transcodes KNX bus data into SIP and our enhanced presence exchange protocol. Every sensor pushes its current status to the KNX bus, where the bridge reads and also pushes it to the presence service. Thereby, the bridge accumulates a group of sensors and converts their data into the JSON structure explained above, which then can be sent to the presence service. On arrival of that data, the presence service sends it to all subscribers, so the end-users are able to read the current statuses of their sensors at home. Figure 4 shows an example on how the information is exchanged. Furthermore, as a extension of that behavior, the user is able

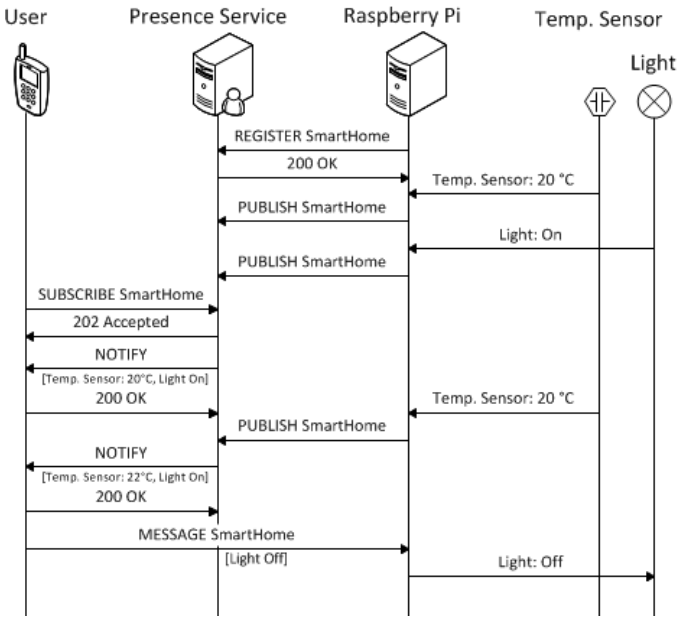


Figure 4. Smart Home information exchange.

to define these groups, like a set of all sensors and actors of one room, or to define an access control list for it. This is done by applying security layers as described hereafter.

VII. SECURITY

Nowadays, it is very important to prioritize the use of security techniques to achieve the key security concepts, like confidentiality or authenticity. In our example, we do not want anybody to control our Smart Home. To achieve this, a security concept is needed, which allows to grant access to specific persons or to prevent read and write access from third parties. It is also important to refrain from using known techniques, like a VPN tunnel, to control a Smart Home, because they only cover layer 1 of the following layers. With a VPN tunnel, the person who has access to it, has also the access to the whole Smart Home without any restrictions.

Besides a simple encryption of the connection to our Smart Home, we introduced two use-cases in Section V, which are motivating the need for an user-controlled access control list (ACL). This ACL helps us to define users who have access to specific sensors and actors. To achieve this, we elaborated a security layer concept as it is known from the OSI layer model [10].

Every layer needs to be applied on top of the lower layers, so the highest security is only reached by applying all layers of this concept. Figure 5 shows three examples on how to combine these layers. As can be seen, all layers are different

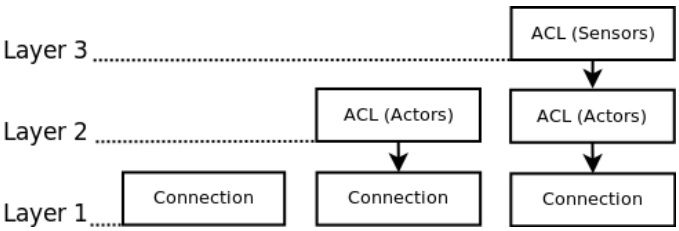


Figure 5. Security Concept - layer architecture

from each other and are needed to achieve different security goals. The exact function of each layer will be described in the next subsections. The main benefit of using such a layer concept is the ease of applying each layer on top of the others. When using this concept, one is not forced to apply all layers, but only the layers, which are needed in a specific scenario, to not overload and complicate the application. Another reason for choosing this layer concept is the separation of adaptions of the connection endpoints. Mostly, the first layer already exists in SIP applications and is supported by SIP providers. The second layer needs to be adjusted to the Smart Home, to allow the use of an ACL. At the end, the third layer requires an additional adaption at the Smart Home and the users control endpoint, because an encryption extension has to be added. In the following subsections, we are taking a closer look at these layers.

A. First Layer

The first security layer is done by the using the underlying protocol secure SIP (SIPS). SIPS enables basic encryption with the Transport Layer Security (TLS) protocol. As mentioned before, it provides the integrity and confidentiality of the communication between server and client. An attacker needs to gain access to an endpoint of the connection to decrypt or

 ISBN: 978-1-61208-353-7

manipulate the transferred data. This is the basic and recommended security layer, which does not need any adaptions on the communication protocol, because it is already implemented as a transport layer in almost every operation system, since they are implementing the OSI-Layer model [10].

B. Second Layer

To achieve access control to the actors, the second layer creates an access control list, which lists all users with access capabilities. Only users who are listed in that ACL are able to set a new state of an actor. For example, the user *owner@myhouse* is able to turn off all lights inside his house. Every command send from another user will be dropped. The whole ACL stays by the end user, so only the end user respectively the Smart Home is able to modify the ACL. So, the owner can add or remove users from it to grant or deny their access.

With this approach, the provider of SIP proxy cannot modify the status of any actor of the Smart Home, which is a security benefit. If we look at the second use-case in Section V, we have a guest at our Smart Home. That shows us the need of granting access to specific actors, like the guestroom's light, to other persons. With such a simple ACL, we are able to grant that access to almost everyone who has a valid SIP account at any provider.

A problem occurs when we look at sensor states. Everyone in the system is able to see anybody's presence state by default. This might be problematic, because everybody would be able to monitor any house if they know their SIP address.

C. Third Layer

The problem explained above can be avoided by using encryption of the sensor data. To do so, the encryption key can be requested from the Smart Home before decrypting the presence information. This is achieved by using the technique of the second layer, where only users listed in the ACL are allowed to communicate with specific actors at the Smart Home. The end user's application sends a message, requesting the encryption key, to the Smart Home. After that, the Smart Home sends back the key, also as a simple message. The whole transfer is encrypted by using the first layer (SIPS). At this point, one can see the need of such a layer-concept, because without using the other layers below, the security of the third layer is useless. For example, without layer one, an attacker is able to gain the encryption key by simply listening or without the third layer, an attacker can simply request that key and will not be rejected by the Smart Home.

VIII. Discussion

To discuss this concept, we take a deeper look at the compatibility to other products and the performance of the Raspberry Pi. With this procedure, we show the result of this research. To do so, we discuss the following aspects:

1) Power consumption
2) Total costs

A. Compatibility

By using standardized software, we achieve a high level of compatibility with other products. Only the bridge component between KNX-Bus and the SIP Service had to be implemented, so it just uses the new protocol. A comparably small effort is necessary to connect different devices to each other by only creating the interfaces. Furthermore, our protocol does not influence other services, which are using a presence-based exchange of status information, because the enhancement is based on the standardized PIDF-Scheme.

Also, the three security layers are designed on top of the normal protocol in order to not influence other services. In comparison to existing Smart Home communication solutions, which are using a standard VPN connection, the new security design achieves additional security specifications to control the access to selected actors and sensors. A VPN connection grants access to the whole Smart Home network and thereby to all sensors and actor without any policing.

Through these achievements, this concept offers a stable foundation to develop home automation systems or further software with higher complexity, e.g., monitoring or control over a Smart Grid.

B. Usage of Raspberry Pi

In this concept, communication is not the only focus. The second important aspect is the usage of resource-poor implementation. Because of this, we selected the Raspberry Pi as our hardware platform. It is especially characterized by its low power consumption and its low costs. With that solution, we save a lot of space in comparison to a normal desktop computer.

1) Power consumption: The Raspberry Pi consumes around 750 mA at 5 V, resulting in 3.75 W per hour. At a workload of about 100 %, it consumes up to 1 A, which corresponds to 5 W [9]. This heavy workload was never measured during our tests. This brings us to an average consumption of about 3 W per hour, which results in a total power consumption of 17 kW hours a year.

2) Costs: The costs of power consumption amounts to about 5 EUR a year, referring to the German electricity prices of 2012 [11], which is very low in comparison to a normal destktop computer with a power consumption of about 100 W per hour. Furthermore, the cost of purchase is very small as well. The Raspberry Pi only costs around 60 EUR with all needed peripheral equipment [9]. That leads us to a unique cost of 60 EUR and permanent costs of about 5 EUR a year. In case of a defective device, there is mostly only the need of replacing the Raspberry Pi and to replug the SD-Card, which holds the whole software and configurations. So even in the case of a faulty device, the cost is minimal.

C. Usage of FreeSwitch

The preparation of external SIP server usage has been chosen because of existing infrastructure. Our current setup is a development environment of an equivalent one of a SIP provider. With such an environment, we are able to simulate all possible situations in a high scale network, so we can evaluate this concept. To do so, we analyzed selected performance data in the following subsection.

 ISBN: 978-1-61208-353-7

1) Power consumption: Currently, the FreeSwitch is running besides the KNX-to-SIP Bridge on the Raspberry Pi. In the future, the SIP-Server of an existing VoIP-Provider will be used, so our power consumption is effectively zero. That is because we do not need an own device, as we are using an existing infrastructure.

2) Costs: Like the power consumption, our cost is nearly zero. Nearly because we need at least one valid user account. Normally, a lot of end users still have an existing user account for a VoIP-Provider, because their telephone already runs the SIP-Protocol through VoIP.

IX. Conclusion

In this paper, an updated communication and security design for the Smart Home project at University of Applied Sciences Darmstadt was presented. It is shown that only one SIP profile can store the information of all actors and sensors within a Smart Home, instead of using separate profiles. Also, it is shown that the Raspberry Pi can be used for a home access gateway as an embedded system solution. With the implementation of the Raspberry Pi the general performance could be improved as well as the energy-consumption could be reduced, compared to a standard desktop computer. This approach makes use of well-known and open source information technology standards, instead of developing new commands for SIP or any proprietary application. This ensures future compatibility and makes the approach adoptable to other home automation systems.

X. Future Work

In future work, one step would be to extend the client to comply with the new communication concept or to develop a new client, which is platform independent, so it could be used with every mobile platform. Therefore, the aim should be a web-based client using WebSocket technology.

A. Cooperation with SIP-Providers

The communication concept detailed in this paper is also extendable to fit future needs. Furthermore, sensor information may be evaluated by smart grid- or weather stations. This can be done by splitting the sensor and actor data into several groups, which then needs the cooperation with the SIP-Providers to get access to a bundle of valid SIP-Accounts to provide access to each group through these accounts. With such an extension, the network load will be lowered and the communication may be differentiated in a more efficient way.

B. Security

A disadvantage of this approach is that the server is able to read the transferred encryption key. To solve this problem, there is an ability to generate a temporary encryption key by using a Diffie-Hellman [12] key exchange and encrypt the original encryption key with that temporary one.

Another disadvantage is, that all users who are capable of calling for the password are able to see all the data that is transferred. A solution for this is to define groups of sensors and actors, as well as different passwords for each group. With such a proceeding, a permitted user is only allowed to read the data, which is required for him. An additional need is to renew all passwords inside a timeframe, so all revocations of monitoring rights are successfully deployed.

These three modifications should make up layer four of our communication architecture in the next step of our research. Another possibility is to implement a public key infrastructure to control the access to sensors and actors. This could replace layer two to four in one step.

References

[1] KNX Association, Standardisation, http://www.knx.org/knx-en/knx/technology/standardisation, [retrieved: Apr. 2014]

[2] Raspberry Pi Foundation. Raspberry Pi, http://www.raspberrypi.org/, [retrieved: June 2014]

[3] J. Rosenberg, et al., SIP: session initiation protocol, IETF, RFC 3261, Jun. 2002.

[4] D. J. Cook et al., MavHome: an agent-based smart home, Proceedings Of The First IEEE International Conference On Pervasive Computing And Communications, pp. 521-524, March 2003.

[5] H. Schulzrinne, X. Wu, S. Sidiroglou, and S. Berger, Ubiquitous Computing in Home Networks, IEEE Communications Magazine, pp. 128-135, Nov. 2003.

[6] S. Moyer, D. Marples, S. Tsang, J, Katz, P. Gurung, T.Cheng, et al., Framework Draft for Networked Appliances using the Session Initiation Protocol. IETF Internet Draft, May 2001.

[7] R. Acker, S. Brandt, N. Buchmann, T. Fugmann, and M. Massoth, Ubiquitous Home Control based on SIP and Presence Service. Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services, pp. 759-762, Nov. 2010.

[8] B. Bertran, C. Consel, P. Kadionik and B. Lamer, A SIP-basedhome automation platform: an experimental study, Proceedings of the 13th International Conference on Intelligence in Next Generation Networks, pp. 1-6, Oct. 2009.

[9] Raspberry Pi Foundation, FAQs, http://www.raspberrypi.org/help/faqs, [retrieved: Apr. 2014]

[10] H. Zimmermann, IEEE Transactions on Communications, vol. 28, no. 4, pp. 425432, Apr. 1980.

[11] Eurostat. Electricity prices for domestic consumers from 2007 onwards, http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=nrg_pc_204, [retrieved: Apr. 2014]

[12] IETF. RFC 2631: Diffie-Hellman Key Agreement Method, http://tools.ietf.org/html/rfc2631, [retrieved: June 2014]

 ISBN: 978-1-61208-353-7

Designing a Low-Cost Web-Controlled Mobile Robot for Home Monitoring

David Espes, Yvon Autret, Jean Vareille and Philippe Le Parc
Université Européenne de Bretagne, France
Université de Brest
Laboratoire en Sciences et Techniques de l'Information (LabSticc UMR CNRS 6285)
20 av. Victor Le Gorgeu, BP 809, F-29285 Brest
E-mail: {david.espes, yvon.autret, jean.vareille, philippe.le-parc}@univ-brest.fr

***Abstract*—In this paper, we focus on a web-controlled mobile robot for home monitoring. The key point is low-cost. The robot is built from standard components to reduce the cost of the hardware. A large part of the system is deported to the cloud to minimize the required software on the robot. A minimal positioning system is provided to make the robot usable. The result is a small robot wich can be used from the outside or the inside the house.**

Keywords–Mobile robot; Home monitoring; Web control; WebSocket.

I. INTRODUCTION

Web-controlled mobile devices are more and more used in ubiquitous environments [1][2][3][4]. Small monitoring robots such as the Rovio WowWee can be used [5]. Web control is not really new, but recent improvements of network performance has led to the emergence of Service Robotics [6]. Services Oriented Architectures (SOA) [7] start to be used to control physical devices [8].

Our aim is to use these approaches to control mobile home robots designed for Human Ambient Living (HAL) environments. For us, a typical application is helping elderly people who live in their houses and sometimes have difficulties to move. A mobile home robot carrying a camera could help them monitoring their house either indoors or outdoors, for example to watch the dog or to see what is going on. The mobile home robot could also be used by care helpers or relatives, as a moving phone to communicate with the inhabitants of a house from the outside.

In such an HAL environment, the total cost of the mobile home robot is the first key point. It must be kept as low as possible especially if it is a HAL environment for elderly people who often have tight budgets. This means that the mobile home robot must be built by using low-cost commercial components. Moreover, we always keep in mind that mechanical failures are unavoidable and reliability is a major key point as much as ease to repair. The basic mobile home robot is nothing but than a mobile robot which carries a camera. More sophisticated sensors such as positioning sensors are optional.

The second main key point is software and network configurations. The mobile home robot should be plug an play. This means that software and network configurations should be reduced as much as possible. Deporting a part of the system to the cloud can be a solution if it helps to get a reliable plug and play system.

The third key point is security and access control. A Web-controlled mobile home robot can be used from anywhere in the world, but the interior of a house must not be seen by unauthorized users. It is necessary to avoid any intrusive access. In case of network failure, the mobile home robot should also be able to properly stop its current action and wait for a new order.

The fourth key point is the autonomy of the battery. The robot should have an autonomy close to one hour when moving, and automatically come back to a charging dock when the battery is low.

In this paper, the second part presents a mobile home robot solution based on a commercial low-cost robot and we discuss the advantages and the disadvantages. This lead us to the design of a mobile home robot built from commercial components such as a low-cost robotic platform and a smartphone to control it. In the third part, we present the cloud control system and its performance. In the fourth part, we add video monitoring capabilities to the robot, and in the last part, we present a low-cost positioning system.

II. DESIGNING A HOME ROBOT

A. Commercial home robots

Several commercial robots such as Miabot [9] are available. The Miabot robot is rather small (about 10cm long) and fast (3.5 m/s). It has a built-in bluetooth connection and must be connected to a local central computer to be web-controlled. Even if it was not really designed for that, it can carry a small camera or other sensors.

A better robot from our point of view is the WoWee Rovio [5]. It includes a mobile base, a mobile camera and a Wi-Fi connection. Its size is 30 x 35 x 33 cm. It can be remotely controlled from anywhere in the world. When the battery is low, it automatically comes back to its charging dock. The almost 300 euros cost is acceptable.

The WoWee Rovio is an interesting robot for a HAL environment, but a weak point is the reliability and the ease to repair [10]. The WoWee Rovio can not be considered reliable. For example, sunlight may interfere with the infrared bean of the WoWee Rovio and prevent it returning to its charging dock. In case of failure, the WoWee Rovio is difficult to repair. We have used several WoWee Rovio. One of them had and infrared led problem and all of them had battery problems after one year use. This was a real problem because we had no easy solution to replace the failing components.

 ISBN: 978-1-61208-353-7

B. Using a Smartphone, an Arduino, and a basic robotic platform

Using a smartphone may help simplifying the building of a home robot because it usually includes a webcam, Wi-Fi and Bluetooth. When connected to an Arduino micro-controller [11], a smartphone can also be used to control the motors of a mobile home robot.

We use an open robotic platform which includes two tracks. It is a 4WD Rover 5 from RobotBase. The size is close to that of the WoWee Rovio. When powered, it can move forward or backward and turn. The maximum speed is 1km/h. The Rover 5 is strong enough to carry up to two kilograms.

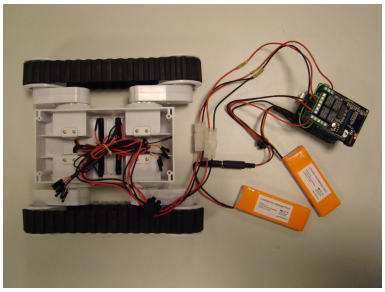


Figure 1. Components of the Web-controlled home robot.

The robot is controlled by the Arduino. Several Arduino shields are available to monitor the working speed and direction of the motors. We can use either a relay shield including four relays, or a motor shield based on a voltage regulator such as 78M05. An additional Arduino shield is required to allow Bluetooth communication between the Arduino and the smartphone.

The main advantage our solution is its simplicity. The home robot only includes six commercial components (Fig. 1 and Fig. 2):

- A mobile Rover 5 robot used as robotic platform (60 euros)
- An Android Smartphone (less than 80 euros)
- An Arduino UNO (20 euros)
- A Bluetooth shield (20 euros)
- A motor command shield (20 euros)
- Two batteries (one for the Arduino, one for the motors)

The total cost, smartphone included, is comparable to that of a WoWee Rovio. We can also use an old smartphone which has became useless. When used as a mobile home robot controller, a recycled smartphone significantly reduces the total cost.

The reliability of our mobile home robot is significantly higher than that of a Rovio. In case of failure, we only need to replace one component. Moreover, the diagnosis is very easy because each component can be individually tested.

When using 2000mAh lithium batteries, we have a 30mn autonomy when the robot is continuously moving. We have several hours of battery life when the robot is waiting for commands. Automatic battery charging is not available on our prototype.

Additional sensors can be added on the robot, but as it is a non autonomous Web-controlled robot, they are not essential. Moreover, it would increase the total cost.

III. A CONTROL SYSTEM IN THE CLOUD

We propose to deport a large part of the robot control system to the Cloud to reduce home configurations and installations. A user interface running on a standard Web Browser should make the robot usable without any special installation.

Using HTTP (Hypertext Transfer Protocol) is a solution to communicate with a distant server in the Cloud. Efficient HTTP Web servers such as Apache or Apache Tomcat are available. If the standard HTTP protocol lets easily handle problems such as client identification, it has severe limits when used for real-time monitoring.

A. The HTTP limitations

The HTTP protocol is a stateless protocol which was originally designed to get access to static HTML pages. Later, some web applications have implemented server-side sessions by using HTTP cookies. A Web server implementing sessions receives an HTTP request, establishes a connection with the server, executes the request, sends an HTTP response back, may keep a track of the HTTP request, and finally, releases the connection.

If a Web server is running on the robot, an identification sequence which gives the right to monitor the robot through the Web server can be easily implemented. The communication can be secured by using the HTTPS protocol. The main problem is the execution time of a command sent to the robot. Let us take the example of an HTTP request which should make the robot move for several seconds. As soon as the HTTP request is received on the server, the robot starts moving. If the robot moves for more than a few seconds, the HTTP response must be sent back before the robot has finished moving. In this case, the robot can get out of control.

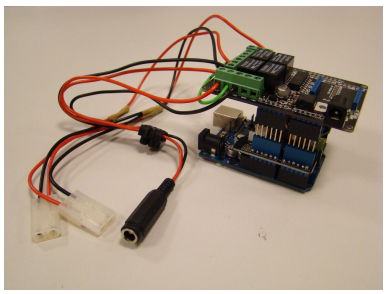


Figure 2. The Arduino command module.

This is a major problem because we must monitor a robot by using commands which execution lasts about one second. A one meter trip would require sending at least three commands to a Rover 5 moving at 1km/h. Touring a house would require hundreds of commands. When a command is sent to a distant robot, a permanent connection is required. A moving robot left

 ISBN: 978-1-61208-353-7

unsupervised just a few seconds can be dangerous. Presence and obstacle detectors working on the robot are never 100% reliable. This means that anyone who is monitoring from the outside or inside the house must have a permanent full control of the robot through the network. Moreover, the robot should be able to detect the smallest network failure, and to automatically adapt its behaviour, for example by reducing its speed.

This means that sending HTTP requests to a Web server running on the robot is not a good solution. We must continuously send HTTP requests to the robot to be able to detect network failures. That is a misuse of HTTP. Second, establishing a new connection from outside can be time consuming and sometimes takes several seconds. That is a risk we can not take. That is why we have chosen the WebSocket solution.

B. Using a WebSocket server

The WebSocket protocol was standardized in 2011 [12]. The communications are established by HTTP servers, and the communications may use TCP port 80 (or 443 when using secured communications). The client is responsible for making the connection by using an URL, consisting of a protocol, host, port, path, and optionally one or more additional parameters.

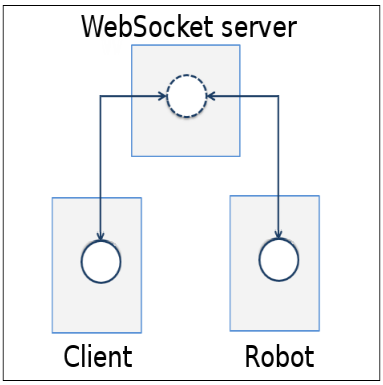


Figure 3. The WebSocket servlet.

The main advantage of WebSockets for our purpose is the fast responses coming from the server. That is due to the single connection that is established at the beginning of the communication. As soon as a connection is set, a bi-directional communication remains available. Full duplex communication over a single socket allows true real-time communication.

A standard Web browser can be used to monitor a robot through WebSockets. Most Web browsers now support WebSockets. Both the client and the robot send and receive information to and from the Web server through WebSockets. When a command is sent from the client to the Web server by using WebSocket, as soon as it is received on the server, it can be forwarded to the robot and executed. During the execution of the command on the robot, WebSockets are still used to send periodic aknowledges from the robot to the client, and from the client to the robot.

Thus, if the robot does not receive any acknowledge, or receive them too late, it can modify its state. For example, it can reduce its speed if the network is too slow. If the network is no more working, the robot can stop properly, and remain waiting until the network is working again.

C. A WebSocket server in the cloud

A WebSocket server in the clould greatly simplifies the installation of a Web-controlled home robot. The home robot just have to connect to the WebSocket server (Fig. 3). This does not require any special home configuration. An ordinary Wi-Fi connection can be used.

The well known Apache Tomcat Webserver now implements WebSockets. This means that we can use both the advantages of a standard Web server and those of WebSockets. A standard Tomcat application manages client and robot identification. The client uses an HTML form to ask for a robot. As soon as identification is successful on the server, a WebSocket communication becomes available between the client and the robot.

On the Tomcat server, we have a servlet to manage identification and robot allocation. We have also a WebSocketServlet to manage communication between the client and the robot.

The main elements of the WebSocket Servlet are shown in Fig. 4.

```
public class RobotWebSocketServlet
        extends WebSocketServlet {

  protected StreamInbound
    createWebSocketInbound(String subProtocol,
          HttpServletRequest request) {
      Manager manager = ...
      return new ClientRobot(manager);
    }
}

public class ClientRobot
     extends MessageInbound {
  Manager manager = null;
  private RobotCommunication
       (Manager manager)
       { ... }
  protected void onTextMessage
       (CharBuffer message)
       throws IOException
       { ... }
}
```

Figure 4. The WebSocket Servlet.

The ”manager” object is instantiated by the WebSocket server. From the robot point of view, it contains information about the client which is using the robot. From the client point of view, it contains information about the robot to control. The manager is stored as a Tomcat session object. It is a persistant object whose life duration is that of a session. A ”manager” object is instantiated during the identification phase, when the client asks for a robot. Another ”manager” object is instantiated when the robot connects to the WebSocket server When the WebSocket communications are set, the ”manager” objects can be retrieved and modified to help clients and robots communicate. One client is allowed to send messages to one robot, and one robot is allowed to send messages to one client.

Both the client and the robot exchange messages by sending lines of text. For example, the client sends a line containing ”forward” to make the robot move forward. Parameters can

 ISBN: 978-1-61208-353-7

also be added in the line, for example to make the robot move forward for n seconds.

D. WebSockets on the robot

As seen above, the robot is controlled by the Arduino and the Arduino is controlled by an Android smartphone using a Bluetooth communication. We use the Tyrus API to connect the smartphone to the WebSocket server. The main Java elements of the Android WebSocket connection are shown in Fig. 5.

```
final ClientManager client =
    ClientManager.createClient();
client.connectToServer(
  new Endpoint() {
    public void onOpen(Session session,
        EndpointConfig EndpointConfig) {
      session.addMessageHandler(
        new MessageHandler.Whole<String>()
          public void onMessage
              (String message) {
            ...
          }
        },
        ClientEndpointConfig.Builder.
           create().build(),
           URI.create("ws://.../robot"));
    }
  });
```

Figure 5. The Android WebSocket.

We use the Tyrus ”ClientManager” class to set a connection between the robot and the WebSocket server. When messages come from the client, the ”onMessage” method is triggered. The message is decoded and forwarded to the Arduino. During the execution of the command by the Arduino, the client and the smartphone periodically exchange messages to stop or slow down the robot in case of network failure. This program has been tested on Android 2.3 and Android 4.

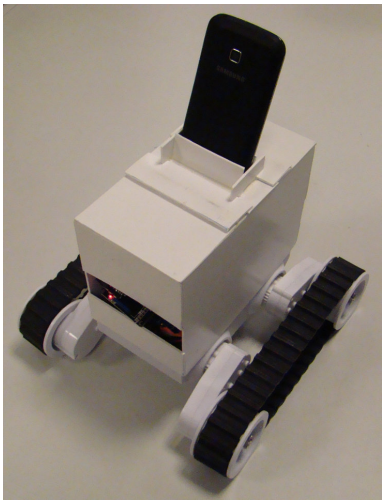


Figure 6. The Web-controlled home robot.

E. WebSockets on the client

A WebSocket connection from the client to the server is only possible if the identification phase and robot selection has been successful. This is taken into account by the standard Apache Tomcat Webserver. As soon as a client is successfully registered on the distant Web server, a WebSocket connection is established. The client uses a Web page as user interface. The only thing required to use the user interface is a WebSocket compatible Browser. The user interface is managed by the distant Web server. The main JavaScript elements of the WebSocket connection are shown in Fig. 7.

```
var client = {};
client.connect = (function(host) {
  client.socket = new WebSocket
     ("ws:// ... /client");
  client.socket.onmessage =
    function (message) {
      ...
    };
});
```

Figure 7. The client WebSocket.

The Javascript ”onmessage” function is triggered when a message comes from the WebSocket server. A widget such as a button in the user interface can trigger the ”sendMessage” function and send commands to the robot.

F. Performance

In this section, we present some experiments that illustrate the capabilities of our system. The server is connected to the local network of the laboratory, i.e., gigabit Ethernet network. It is hosted to a public address so any user are able to access it from anywhere using just a web browser. Beside the server, one robot is available. The robot is equipped with an arduino board, a bluetooth shield and a smartphone. The bluetooth shield is fully qualified to respect the Bluetooth version 2.0. Hence, the data rate is up to 2Mbps. The smartphone is connected to the local network through a WiFi connection. The wifi card on the smartphone is compliant to the IEEE 802.11g standard. Hence, the data rate is up to 54Mbps.

In order to show the performance of the system, we define the following performance metrics:

- the *Round-trip time between components* is the time to receive a response after sending a request without counting the delay due to other components. By example, if the arduino board sends a request to the smartphone, the round-trip time between these both components is the delay to receive a response without counting the delays imposed by smartphone-server connection and server-user connection.
- the *End-to-end round-trip time* is the time that the user receives a response after sending a response, i.e., it is the sum of the round-trip time between the whole components of the system. The increase of the end-to-end round-trip time degrades significantly the QoS of applications.

In order to test different scenarios, the user accesses to the robot from two different locations: our laboratory and

 ISBN: 978-1-61208-353-7

TABLE I. ROUND-TRIP TIME RELATED TO ENTITY CONNECTIONS

Entity connection	Round-trip time	
	Local (inside laboratory)	Distant (Romania)
User - Server	15 ms (±5 ms)	40 ms (±15 ms)
Server - Phone	35 ms (±10 ms)	35 ms (±10 ms)
Phone - Robot	125 ms (±40 ms)	125 ms (±40 ms)

the Military Technical Academy of Bucharest in Romania (about 2500 km from the laboratory). The user accesses to the robot through the LAN or Internet into the laboratory or the Academy of Bucharest respectively. In all the scenarios considered the server is inside our laboratory. However, due to the flexibility of our architecture, the server could be hosted in the cloud.

In Table I, we present the round-trip time related to component connections when the user accesses to the robot from different locations (laboratory and Romania). All times are expressed once the websocket connection is established.

It is interesting to see that the most important delay is added by the bluetooth connection between the arduino board and the smartphone. Indeed, the data rate of the bluetooth shield is quite low (2Mbps). The time to transmit the data from the robot to the phone, or inversely, is proportional to the data rate. This is the principal factor to this delay. Moreover, bluetooth system is a contention based system. Bluetooth systems are based on a combination of frequency-hopping and CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance) [13] methods to access to the medium. The medium is shared between all the nodes belonging to the same user and other systems such as WiFi. The delay to access to a free medium or the retransmissions due to collisions increase the round-trip time significantly. However, it is interesting to use a bluetooth connection between the smartphone and the robot due to its low consumption. The mobile robot's operational time is limited before exhausting its battery power. Indeed, Bluetooth is much more power efficient than WiFi. As mentioned by Pering et al. [14], the power consumption of Bluetooth is 10 times lower than WiFi.

The round-trip time between the smartphone and the server is quite low. Unlike Bluetooth systems, WiFi systems have a high throughput. The transmission time is significantly reduced. The round-trip time induced by WiFi is roughly 4 times lower than the one obtained with Bluetooth.

The Internet delay, i.e., when the user is located in Romania, is almost negligible as compared with local access. The university of Brest, respectively Academy of Bucharest, has a guaranteed bandwidth of 3Gbps, respectively 1Gbps, on its national network. Hence, the difference in time is particularly due to the propagation time. Let us assume a propagation speed of 200,000 km/s, the round-trip propagation time is about 25 ms.

In Table II, the end-to-end round-trip time is analyzed under

TABLE II. END-TO-END ROUND-TRIP TIME

Protocol	End-to-end round-trip time	
	Local (inside laboratory)	Distant (Romania)
HTTP	600 ms (±120 ms)	730 ms (±100 ms)
Web Sockets	205 ms (±75 ms)	250 ms (±50 ms)

two different locations (local and Romania) and two protocols (HTTP and websocket). The end-to-end round-trip time is an important parameter because it is the main criteria to determine if real-time control is possible. To control a distant robot with an acceptable quality of experience, it is commonly accepted that the delay never exceeds 400 milliseconds. We can see the HTTP protocol cannot guarantee the delay bound. Indeed, the time to establish the connection, to send a request and receive a response significantly exceeds the delay bound. In case a system requires the establishment of a TCP connection for each transaction, the real-time control of the mobile robot is not possible. The websocket protocol is more suitable for real-time control. Being designed to work well in the Web infrastructure, the protocol specifies that the websocket connection starts its life as a HTTP connection, offering backwards compatibility with no-websocket systems. The handshake of the websocket protocol has slightly the same time than the HTTP protocol. Once the connection is established, control frames are periodically sent to maintain the connection. Hence, the time is significantly reduced as compared with the HTTP protocol. For all scenarios, the end-to-end round-trip time does not exceed 300 milliseconds which is quite acceptable to transmit QoS traffic.

IV. VIDEO MONITORING

If a video stream is sent from the robot to the client, the loss of some images is not critical. Thus, videos can be obtained from a standard video Web server running on the smartphone. We have used the IP Webcam application which works on Android 1.6 and up and broadcasts video and sound. The smarphone is placed on top of the robot (Fig. 6). It also communicate with the Arduino and the cloud server as seen above.

If security is required, videos can be sent to a distant Web server through a securized channel, and forwarded to the client.

V. TOWARD A LOW-COST POSITIONING SYSTEM

We also design a low-cost localization platform for 2D-positioning. Even if the robot is not an autonomous one, as it is web-controlled, information on the position of the robot is very useful to the user of the robot. Let us assume the robot only has to monitor flat floor, i.e., the relative z-coordinate is always equal to 0. In cases where different floors have to be monitored, a robot may be on each floor. They can communicate between them in order to extend the control in the whole habitation.

The positioning system involves 4 TelosB wireless devices. The 3 auxiliary sensors have a fixed position, being place in

 ISBN: 978-1-61208-353-7

strategic places of the room, in the corners for example. The places must be chosen in such way that the robot which will have the Main Sensor attached to be in permanent Line of Sight with this sensors. This way, the communication would be done with very little interference.

Fig. 8 shows the whole system and the interaction between the components. The auxiliary sensors send a message periodically. The main sensors do not know their position. After receiving a message from an auxiliary sensor, they gather information, such as receiver's Received Signal Strength Indicator (RSSI) and the identity of the sender. In order to optimize the energy consumption, the processing of the RSSI values is skipped in this moment, being the duty of the server application to make the necessary computations from which will result the distance approximation. Once the Main Sensor acquires a message from each of the 3 fixed sensors, it will create a data packet which contains the 3 pairs of ID - RSSI value for each sender, and will sent it through the USB interface to the arduino board. The arduino board forwards this message to the server that converts the raw values into physical distance, measured in meters. At this point, the server knows the distance between the main sensor and each auxiliary sensor.

In two-dimensional geometry, the trilateration technique uses three reference nodes to calculate the position of the target node. To be localized the target node should locate at the intersection of three spheres centered at each reference position. When the signal received from the reference nodes is noisy, the system is non-linear and cannot be solved. An estimation method has to be used. To get a satisfying approximation position of the mobile robot, we use the Newton-Raphson method [15]. This method attempts to find a solution in the non-linear least squares sense. The Newton-Raphson' main idea is to use multiple iterations to find a final position based on an initial guess (by example the center of the room), that would fit into a specific margin of error.

The first results show that RSSI values are not constant due to multipath components. Hence, the precision of our system is about 2 meters. Such a precision is sufficient to know approximately where the robot is. The localization will reduce the complexity of the control interface dedicated to the distant user. The web interface will contain the cartography of the house, and the robot will be able to reach a destination, only by clicking on the map.

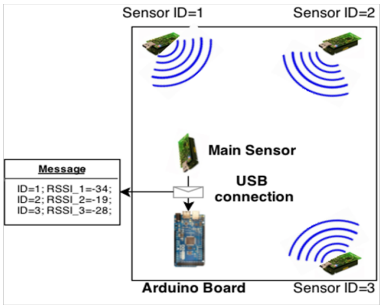


Figure 8. The positioning system.

VI. Conclusion

The smartphone and the Arduino micro-controller are the two main devices of the proposed home robot. The smartphone includes several important features such as network connection and webcam. It reduces the total cost and increases global reliability. By combining a standard Web server and WebSockets, we can deport a large part of the system to the cloud, and installing the home robot in a HAL environment becomes simple and cheap. The lack of sensors on the robot is not very important because it is not an autonomous robot. Moreover, it helps keep the price down. The positioning system is the weakest point of the system. The lack of accuracy make the robot more difficult to use. It is an important element which must be improved in the near future without increasing the total cost.

References

[1] A. Chibani, Y. Amirat, S. Mohammed, E. Matson, N. Hagita and M. Barreto, "Ubiquitous robotics: Recent challenges and future trends", *Robotics and Autonomous Systems*, Volume 61, Issue 11, November 2013, pp. 1162-1172, ISSN: 0921-8890.

[2] S. Nurmaini, "Robotics Current Issues and Trends", *Computer Engineering and Applications*", Vol. 2, No. 1, March 2013, pp. 119-122, ISSN: 2252-4274.

[3] A. Touil, J. Vareille, F. L'Herminier and P. Le Parc. "*Modeling and Analysing Ubiquitous Systems Using MDE Approach*". The Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies. Florence, Italy. October 2010.

[4] P. Le Parc, J. Vareille and L. Marce. "*Web remote control of machine-tools the whole world within less than one half-second*". ISR 2004: International Symposium on Robotics, Paris, France, March 2004.

[5] *WoWee Rovio, a Wi-Fi enabled mobile webcam*. http://www.wowwee.com/en/products/tech/telepresence/rovio/rovio. Online; accessed Apr. 15, 2014.

[6] *Robots With Their Heads in the Clouds*. IEEE Spectrum. Mars 2011.

[7] *Service-Oriented Architecture (SOA) and Cloud Computing*. http://www.service-architecture.com/articles/cloud-computing/service-oriented_architecture_soa_and_cloud_computing.html. Online; accessed Apr. 15, 2014.

[8] Y. Chen, Z. Du and M. Garcia-Acosta *Robot as a Service in Cloud Computing*. Fifth IEEE International Symposium on Service Oriented System Engineering. Nanjing, China, June 2010, pp. 151-158.

[9] Introduction to the Miabots & Robot Soccer, URL: http://eprints2.utem.edu.my/5831/1/Merlin_Miabot_Pro_Robot_Soccer_%282_Wheels%29_24_Pages.pdf. Online; accessed Apr. 15, 2014.

[10] *2009-01-Rovio-insecurity - Insufficient Access Controls - Covert Audio/Video Snooping Possible*. http://www.simplicity.net/vuln/2009-01-Rovio-insecurity.html. Online; accessed Apr. 15, 2014.

[11] *The Arduino micro-controller*. http://arduino.cc/. Online; accessed Apr. 15, 2014.

[12] *The WebSocket Protocol. Internet Engineering Task Force (IETF)*. http://tools.ietf.org/html/rfc6455. Online; accessed Apr. 15, 2014.

[13] M. Oliver and A. Escudero, "*Study of different CSMA/CA IEEE 802.11-based implementations*". In EUNICE, 1999, pp. 1-3.

[14] T. Pering, Y. Agarwal, R. Gupta and C. Power, "*Coolspots: Reducing the power consumption of wireless mobile devices with multiple radio interfaces*". In Proc. ACM MobiSys, 2006, pp. 220-232.

[15] *The Newton-Raphson Method*. http://www.math.ubc.ca/ anstee/math104/newtonmethod.pdf. Online; accessed Apr. 15, 2014.

 ISBN: 978-1-61208-353-7

A Real-time Color-matching Method Based on SmartPhones For Color-blind People

Myoungbeom Chung, Hyunseung Choo
College of Information and Communication Engineering
Sungkyunkwan University
Suwon, Republic of Korea
e-mail: nzin@ssu.ac.kr, choo@skku.edu

***Abstract*—In this paper, we propose a real-time color-matching method based on smart phones for color-blind people. Most people have no trouble seeing color, but about 8% of males and less than 1% of females have faulty color perception from birth. Recently, some applications based on mobile phones have become available for color-blind people. However, most cannot compare colors in real time but just provide color values and names based on the capture images of mobile cameras. However, the proposed method can be used to match colors in real time, as it can report similar colors based on comparisons between capture images and real-time camera images on smart phones. To evaluate the efficacy of the proposed method, we conducted a color-matching experiment, and the matching success rate was 99%. Therefore, the proposed method will be a useful technique for color-blind people.**

Keywords-color-matching application; smartphone; real-time color matching; color-blind people.

I. INTRODUCTION

Color is very important in our lives. Different aspects of color, such as its warmth, coldness, or softness, make different impressions on us. Combining various colors can change our impressions of and feelings about things. Most people have no trouble seeing color, but about 8% of males and less than 1% of females have faulty color perception from birth. Although this is not a serious issue, it can be inconvenient.

To solve the problems related to color blindness, Tomoyuki proposed real color expression technology based on color correction using a Charge-Coupled Device (CCD), a computer, and a Head-Mounted Display (HMD) [1][2]. Koji devised a color information technique whereby the color value of a specific location shows on a color wheel using the camera feature of a phone [3]. Simon proposed a color confirmation application, simulating a deutan defect and identifying color using the color wheel from the camera image capture function of a smart phone [4]. In their recent research, Manaf and Harwahyu traced the finger-pointing position from a real-time camera image on a smartphone and determined the color value of the pointing location through speech using augmented reality [5][6]. However, there are some problems with the existing studies. They show general color value or determine colors based only on an analysis of the input image. They cannot identify undefined colors at the existing method or match colors. For example, if a color-blind user employs the existing color-matching method to find a matching sock in a pile of variously colored socks, he has to take many photos to match all the socks and must remember many names of colors in order to find the same color socks. Moreover, the existing application, which applies augmented reality using finger detection on a smart phone, is uncomfortable because one hand has to hold the smart phone and the other hand has to point out something.

To solve the problems associated with the existing method, we propose a real-time color-matching method based on smart phones for color-blind people. In the proposed method, takes a photo to finding color using camera of smart phone; the photo is located on the left side of the smart phone screen. The real-time input image from the smart phone camera is shown on the right side of the screen, and an analysis of the real-time image is conducted. Therefore, the method can be used to easily match the same color in real time because the user can see the capture image and the real-time input image on the same screen at the same time. The proposed method uses Red-Green-Blue (RGB) color value and the Hue–Saturation–Intensity (HSI) color model for color matching. The method also uses a fixed center range of the input image to rapidly process color analysis of the real-time image. For a color comparison, the user can look at the capture image on the left side of the screen and the real-time input image on the right side of the screen and move the match range using a pointer. The similarity of the pointed range of the capture image and the center range of the real-time image are compared using cosine similarity, and the user is notified if the colors match. Thus, the proposed method can be very useful for color comparisons. To evaluate the performance of the proposed method, we created a real-time color matching application based on the iOS mobile operating system. We conducted a color-matching test using a printed color image and the application and achieved a 99% success rate. Therefore, the proposed method would be a useful technique for color-blind people to match colors.

This paper is organized as follows. In section 2, we explain the conversion method used to change the RGB of an image to HSI color space and the color match method, which uses cosine similarity. In section 3, we explain the screen

 ISBN: 978-1-61208-353-7

composition of the proposed application and the color-matching algorithm that supports color information for color-blind people. In section 4, we report on the color matching test results of the proposed application. Section 5 is the conclusion.

II. RELATED WORK

In this section, we explain the conversion method used to change the RGB of an image to HSI color space to match color values and the color value matching method, which uses cosine similarity. The RGB value of an image ranges from 0 to 255, according to each color. The color value of one pixel is determined based on the combination of colors in (1).

$$Color\ value\ of\ one\ pixel = R + G + B \quad (1)$$

In (1), **R** is red, **G** is green, and **B** is blue. The color value of one pixel on a computer or smart phone uses a specific amount of memory, as seen in Fig. 1, the 24 bit format of RGB color [7].

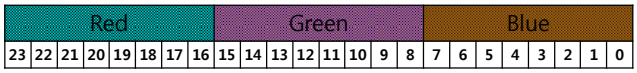


Figure 1. Memory use of smart-phone to show RGB color.

In Fig. 1, Red, Green, and Blue each have 8 bits. The range of Red is from 16 to 23, the range of Green is from 8 to 15, and the range of Blue is from 0 to 7. Thus, the RGB color of one pixel is calculated as in (2).

$$RGB\ Color = (R \times 65536) + (G \times 256) + B \quad (2)$$

In (2), Red memory space is multiplied red value by 65536 (256×256) for the 16–23 position, and Green memory space is multiplied green value by 256 for the 8–15 position. These values can be shown as 16 digit notations, according to each color. The RGB color calculated from the image can be changed to HSI color space according to (3) [8][9].

$$H = \cos^{-1}\left[0.5 \times \frac{\{(R-G)+(R-B)\}}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right] \quad (3)$$
$$S = \frac{1-\min(R,G,B)}{I}, I = \frac{(R+G+B)}{3}$$

The reason the RGB color value is changed to HSI color space is that we want to compare each color using only the hue value of each pixel. Hue of HSI color space is location of color at the visible domain of electromagnetic spectrum of light and it describes the color itself in the form of an angle between [0, 360] degree in Fig. 2. 0 degree mean red, 120 means green 240 means blue. 60 degree is yellow, 300 degrees is magenta. The saturation component signals how much the color is polluted with white color. The range of the S component is [0, 1]. And the intensity range is between [0, 1] and 0 means black, 1 means white. We can divide the HSI value by the RGB color value and can compare that with each pixel, according to hue value.

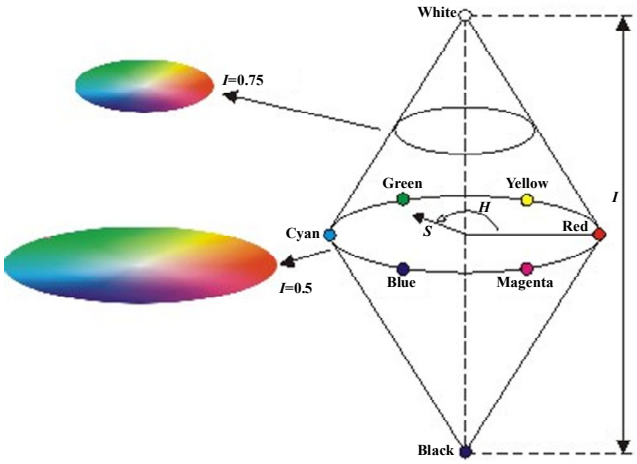


Figure 2. How the HSI color space represents colors

The color-matching method using cosine similarity is a distance calculation method that considers the direction of the color value of each pixel that needs to be compared. Normally, we use Euclidean distance to calculate the distance of comparison items. However, this method does not consider the direction of items [10]. On the other hand, the cosine similarity method considers the direction of each item and is often used in the item search field, in item similarity measurement, in vector space models, etc. Equation (4) is the calculation method for cosine similarity.

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{\|\vec{i}\|\|\vec{j}\|} = \frac{\sum_{k=1}^{n} i_k \cdot j_k}{\sum_{k=1}^{n} i_k^2 \sum_{k=1}^{n} j_k^2} \quad (4)$$

In (4), ***sim(i, j)*** is the similarity value of items ***i*** and ***j***. $\vec{i}$ and $\vec{j}$ are the vector values of each item. This is useful when finding the similarity between two items. A perfect correlation will have a score of 1 (or an angle of 0) and no correlation will have a score of 0 (or an angle of 90 degrees). According to (4), if the value of ***sim(i, j)*** is lower, the similarity of each item is bigger. On the contrary, if the value of ***sim(i, j)*** is higher, the similarity is smaller.

III. THE PROPOSED APPLICATION AND REAL-TIME COLOR-MATCHING ALGORITHM

In this section, we explain the screen composition of the proposed application and the color- matching algorithm that supports color information for color-blind people in real time. The screen composition of the proposed application is as in Fig. 3. In Fig. 3, the capture image for color matching is located on the left side of the smart phone screen, and the

 ISBN: 978-1-61208-353-7

real-time input image from the smart phone camera is located on the right side. The user can select the capture image on the left side from the images in the photo library or can display a capture image selected from the photos taken using the smart phone camera.

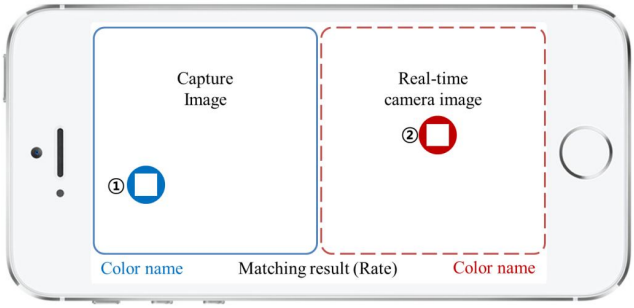


Figure 3. Screen composition of the proposed application.

The point icon (① in Fig. 3) of the capture image can be moved with a touch of the user when he/she wants to change the color matching position. The real-time camera image on the right side is the inputted image from the smart phone camera, and the center point icon (② in Fig. 3) is the color-matching range in real time. However, this icon cannot be moved by the user's touch. The pixel range for color-value matching is N×N, and each pixel range on the left and right sides calculates the RGB color value and changes it to HSI color space. The RGB color value of each pixel is stored as an array type, as in (5).

$$LR\,[a(n-1)+b] = Red\ (a, b)\ of\ capture\ image$$

$$RR[a(n-1)+b] = Red\ (a, b)\ of\ real\text{-}time\ image \quad (5)$$

In (5), $\boldsymbol{a}$ and $\boldsymbol{b}$ are the $\boldsymbol{x}$ and $\boldsymbol{y}$ coordinates of the related pixel, and n is the row number of the selected range. $\boldsymbol{L_R}$ and $\boldsymbol{R_R}$ are the red value arrays for the pixel range of the left and right images, and $\boldsymbol{L_G}$, $\boldsymbol{R_G}$, $\boldsymbol{L_B}$, and $\boldsymbol{R_B}$ are stored as in (5). The HSI color value, which is changed by (3), is stored as an array type, as in (6); we use this in the color-matching algorithm using cosine similarity.

$$LH[a(n-1)+b] = Hue\ (a, b)\ of\ capture\ image$$

$$RH[a(n-1)+b] = Hue\ (a, b)\ of\ real\text{-}time\ image \quad (6)$$

In (6), $\boldsymbol{L_H}$ and $\boldsymbol{R_H}$ are the hue value arrays for the pixel range, and $\boldsymbol{L_S}$, $\boldsymbol{R_S}$, $\boldsymbol{L_I}$, and $\boldsymbol{R_S}$ are stored as in (6).

The color-matching algorithm that supports color information in real time is shown in Fig. 4. The algorithm calculates the color-matching value using cosine similarity with each RGB and hue color value. It returns "Excellent", "Good", or "Bad" as the matching result, according to similar rates between each color. It also provides familiar color names for the user from the HSI value table using HSI color values. In Fig. 4, to obtain color-matching values, the proposed algorithm uses two types of cosine similarity. One is a similar rate using the average value of each color array; we call it the "average similar rate". The other is a similar rate using the average value of the similar rate between each pixel; we call it the "each pixel similar rate". Thus, the color-matching value is determined using the average similar rate and the each pixel similar rate, according to the pseudocode in Fig. 5.

```
If (Average similar rate is over T1) Then
  If (Each pixel similar rate is over T1) Then
    Return Excellent;
  Else If (Each pixel similar rate is over T2) Then
    Return Good;
  End If
Else If (Average similar rate is over T2) Then
  If (Each pixel similar rate is over T1) Then
    Return Good;
  END If
Else
Return Bad;
```

Figure 5. Pseudocode for color-matching results

In Fig. 5, $\boldsymbol{T_1}$ and $\boldsymbol{T_2}$ are each threshold values for color-matching results. The pseudocode returns the correct match when both the average similar rate and the each pixel similar rate is over $\boldsymbol{T_2}$. If the average similar rate or the each pixel similar rate is not over $\boldsymbol{T_2}$, the matching result is "Bad", and the color on the left side does not match that on the right side. Then, the color name is given to the user as supporting

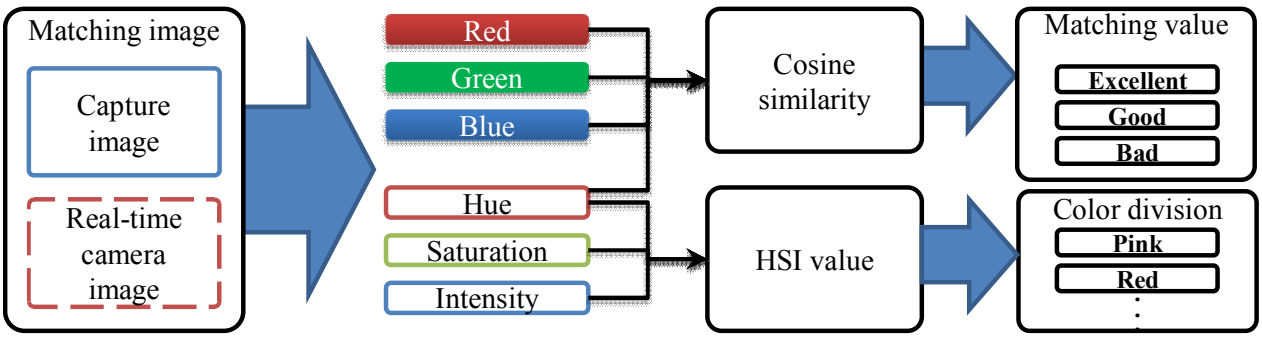


Figure 4. The flow of the color-matching algorithm

information use each array value of HSI. The color is returned one color of 16 kinds color name according to the color names defined by Harwahyu [6].

IV. EXPERIMENTS AND EVALUATION

In this section, we introduce the real-time color-matching application based on smart phones for color-blind people and explain the experiments and results to evaluate the performance of the proposed method. The application can work on iOS 6; we created it using Xcode 5. We used iPhone 5 of Apple Inc. as test device and screen size of iPhone 5 is 4 inch widescreen. And then, we set the brightness level of screen to maximum. To analyze the input image from the smart phone camera in real time, we used the AVCaptureVideoDataOutputSampleBufferDelegate protocol, which is supported by Apple and has two methods [11]. This protocol works on background of application and can converts endless captured big image of the smart phone camera to small input image. Figure 6 shows the captured image for the developed application, according to the proposed method.

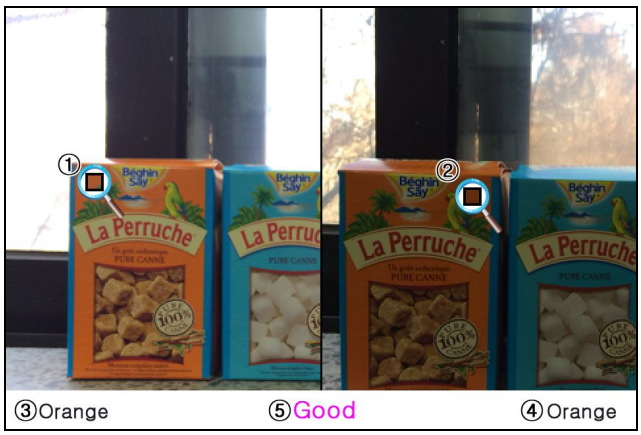


Figure 6. Capture of real-time color-matching application

① in Fig. 6 is a point icon that can be moved by user touch for color matching, and ② in Fig. 6 is the center pixel range of the real-time input image from the smart phone camera. ③ and ④ in Fig. 6 are the color name results of the pixel range of each image according to the defined color name, and ⑤ in Fig. 6 shows the color-matching results of the pixel range for both sides.

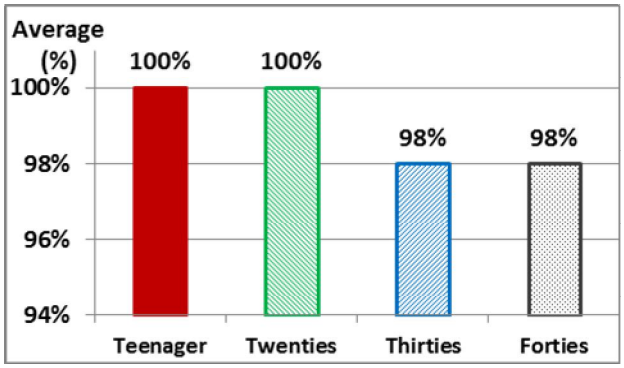
Next, we evaluated the performance of the proposed method. Twenty participants, five in each of four age groups (teenagers, twenties, thirties, and forties) were included in the experiment. To prevent the participants seeing the colors, we used the simulation eyeglasses for achromatopsia test that Tomoyuki used. The color-matching target used in evaluating the performance is a 3×3 table consisting of nine random defined color names. After participants take a capture image of the 3×3 table using the application, they find a matching color in a re-ordered 3×3 table in which the color positions are changed, even though it has the same nine colors within 5 seconds for color. In this experiment, we use 20×20 pixels as the pixel range. The values of T_1 and T_2 in the color-matching result are 90% and 80%, respectively. Figure 7 is a capture image of the smart phone screen of a participant. In Fig. 7, the capture image on the left side has green, yellow, light blue, pink, dark red, light green, red, white, and magenta from the top left to the bottom right. Each participant tries to find the matching color by moving the point icon to the yellow location.

Figure 8 shows the result of the color-matching experiment using the proposed application. Figure 8(a) shows the average results according to age group, and Figure 8(b) shows the average time it tool for participants to find the color matches. In Fig. 8(a), we see that the participants in the teenager and twenties groups could find the matching colors within the time limit. However, the participants in the thirties and forties groups could not find the matching colors within the time limit. Yet, they did not find a matching color on time only, and they found matching colors after the time limit was up.

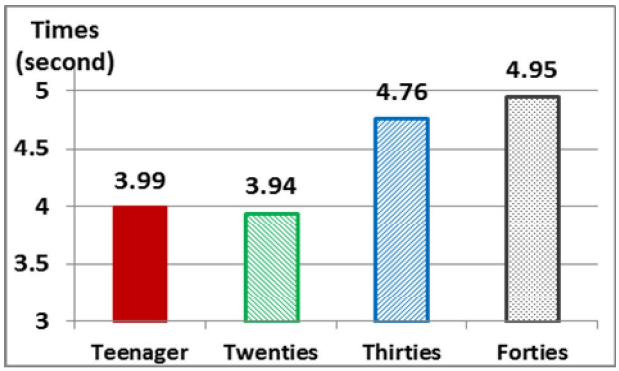
Figure 7. Screen capture of proposed application during the color-match experiment

 ISBN: 978-1-61208-353-7

Thus, the average rate for matching the colors is 99% within the time limit. If there was no time limit, it would be 100%. In Fig. 8(b), the participants in the teenager and twenties groups took about four seconds to match the colors, which is one second before the time limit was up. The participants in the thirties and forties groups took about the same amount of time as the time limit. The reason the participants in the teenager and twenties groups were faster than those in the thirties and forties groups is that they consider match color experiments like "Find the correct picture", and can concentrate better. Based on the results, the proposed method and application can successfully compare and match colors, no matter the user's age group. Therefore, the proposed method would be an effective real-time color-match method for color-blind people.



(a)



(b)

Figure 8. The result of color matching: (a)Average results according to age group, (b) Average time the method took for participants

V. CONCLUSION

The proposed method based on smart phones can be used to compare and match colors in real time, and it provides familiar color names. Thus, it is very useful for color-blind people. The proposed method had a 99% color-matching success rate over all age groups. Because it takes about 4.5 seconds for color matching, we think the proposed method could be implemented to help color-blind people.

In the future, we will determine how to improve the color similarity matching algorithm and the user interface (UI)/user experience (UX) to make the proposed method more useful and comfortable for color-blind people. We will conduct experiments with color-blind participants using the proposed application and various colors found in everyday life.

ACKNOWLEDGMENT

This research was supported in part by MSIP and MOE, Korean government, under IT R&D Program[10041244, SmartTV 2.0 Software Platform] through KEIT, Basic Science Research Program(NRF-2013R1A1A2061478) and PRCP(NRF-2010-0020210) through NRF, respectively.

REFERENCES

[1] T. Ohkubo,and K. Kobayashi, "A Color Compensation Vision System for Color-blind," In SICE Annual Conference 2008, Aug. 2008, pp. 1286-1289. doi:10.1109/SICE.2008.4654855

[2] T. Ohkubo, K. Kobayashi, K. Watanabe, Y. Kurihara, "Development of a Time-sharing-based Color-assisted Vision System for Persons with Color-vision Deficiency," In SICE Annual Conference 2010, Aug. 2010, pp. 2499-2503. ISBN:978-1-4244-7642-8

[3] V. K. Y. V. S. Kondo, and V. Y. Tsuchiya, "Development of Color-Distinguishing Application 'ColorAttendant'," FUJITSU Sci. Tech. J, vol.45, no.2, Apr. 2009, pp. 247-253.

[4] S. Schmitt, S. Stein, F. Hampe, D. Paulus, "Mobile Services Supporting Color Vision Deficiency," 13th International Conference on Optimization of Electrical and Electronic Equipment (OPTIM 2012), May. 2012, pp. 1413-1420. doi:10.1109/OPTIM.2012.6231860

[5] A. S. Manaf, and R. F. Sari, "Color Recognition System with Augmented Reality Concept and Finger Interaction: Case Study for Color Blind Aid System," 9th International Conference on ICT and Knowledge Engineering (ICT & Knowledge Engineering 2011), Jan. 2012, pp. 118-123. doi:10.1109/ICTKE.2012.6152389

[6] R. Harwahyu, A. Sheffildi Manaf, B. Sri Ananto, B. Adi Wicaksana, R. Fitri Sari, "Implementation of Color-blind Aid System," Journal of Computer Science, vol.9, no.6, Jun. 2013, pp. 794-810. doi:10.3844/jcssp.2013.794.810

[7] RGB Color Codes Chart [Online], Available from: http://www.rapidtables.com/web/color/RGB_Color.htm. 2014. 04.10

[8] M. Chung, and I. Ko, "Intelligent Copyright Protection System using a Matching Video Retrieval Algorithm," Multimedia Tools and Applications, vol.59, no.1, Jul. 2012, pp. 383-401. doi:10.1007/s11042-011-0743-z

[9] S. M. Dominguez, T. Keaton, A. H. Sayed, "Robust Finger Tracking for Wearable Computer Interfacing," In Proceedings of the 2001 workshop on Perceptive user interfaces(ACM. 2001), Nov. 2001, pp. 1-5. doi:10.1145/971478.971516

[10] G. Qian, S. Sural, Y. Gu, S. Pramanik, "Similarity between Euclidean and Cosine Angle Distance for Nearest Neighbor Queries," In Proceedings of the 2004 ACM symposium on Applied computing (ACM. 2004), Mar. 2004, pp. 1232-1237. doi:10.1145/967900.968151

[11] Apple Inc. [Online], Available from: https://developer.apple.com/library/mac/documentation/AVFoundation/Reference/AVCaptureVideoDataOutputSampleBufferDelegate_Protocol/Reference/Reference.html. 2013.09.17

 ISBN: 978-1-61208-353-7

Co-creation of Sustainable Smart Cities

Users, Participation and Service Design

Virpi Oksman, Antti Väätänen, Mari Ylikauppila
VTT Technical Research Centre of Finland
Tampere, Finland
e-mails: {virpi.oksman, antti.vaatanen, mari.ylikauppila}@vtt.fi

***Abstract*—The starting point of this paper is to develop new ubiquitous and participative services for sustainable urban planning using mixed reality and other relevant technologies. To support communication between stakeholders, we have created a demo service that can be situated in public places such as interactive displays, designed for community content sharing, close to people flows in cities. It can also be used in public planning events as well as anywhere with personal devices. The service combines visualizations and virtual environments by mixing panoramic imaging and architectural drawings of future urban plans, and includes user-centred interactions such as questionnaires and commenting tools. In this paper, we focus on social issues and the implications of this kind of services, especially trying to understand user values, needs and preferences in participative urban planning service. Political decision makers, city officials and citizens participated in this research to clarify how they perceive the new digital concepts, and how these digital services should be designed and offered for users to support public participation and collaboration in future urban planning projects. Important changes in urban planning would be to increase information and communication and present more alternatives at the early stages of projects. According to the different stakeholders involved in this research, informing the public how their feedback has been taken into consideration and developing real-time feedback channels would enhance participatory urban planning.**

Keywords-participatory design; mobile mixed reality; ubiquitous services; urban planning; user experiences.

I. INTRODUCTION

Urban planning has traditionally been perceived as a complex process. It includes different stages and multiple stakeholders, and it is not very understandable for lay people. Law defines urban planning process and it consists of certain phases, which require acceptance of stakeholders and decision makers. In Finland, statutory status of urban planning and land use ensures involvement of all citizens and other stakeholders who live in the target area. It guarantees any stakeholder the right to see the materials and to leave feedback before any decisions. According to the Land Use and Building Act, citizens are able also challenge the decisions if necessary. However, cities and communities have recently started to pay more attention to making the urban planning processes more participative. Negotiation between different stakeholders and considering citizens' needs and preferences has become more important. In Finland, for instance, besides the official process the stakeholders can also be involved at earlier phase of planning process in order to get deeper understanding of target area. The participatory or co-creative approach which engages citizens and other stakeholders is beneficial in many ways. Co-creation and co-design benefits have been associated to improving processes of idea generation and decision-making and promoting co-operation and creativity. In addition, they have impact on improving users' satisfaction and building trust or loyalty over the long-term [1].

Especially when urban planning ideas are presented and tested already in the early stages, the projects are more likely to proceed smoothly, in a good spirit and are not in danger of being delayed or halted as a result of political or social resistance. When possible problems in the planning can be detected already at the early stages, the result can be qualitatively better in many ways. There is also a possibility of minimizing economic risks when there is no need to make costly changes afterwards, when it is noticed that something went wrong in planning [1].

There is a wide range of different kinds of participatory urban planning tools, methods and technologies in the practice in different countries and cultural contexts [2][3].

Increasing participation demands developing easy-to- use services that are situated in places where people notice them and can be used anytime and anywhere users want to use them. In this paper, we focus on exploring qualitatively how different stakeholders, especially citizens and political decision makes perceive ubiquitous, mixed reality technologies as a part of future participatory urban planning. We are interested in how different technology concepts using any device and any location could open up and make the urban planning process more visible, easy-to-access and understand. Moreover, we study what kind of devices, applications, locations and situations would support users to participate in urban planning. Consequently, we are interested in how to find urban planning solutions that enhance co-operation and take into consideration user values such as maintaining the quality of living environments, clean nature and the protection of historic buildings.

To understand different stakeholders' views, we have conducted 13 interviews among political decision makers responsible for urban planning in local government and city officials. These interviews shed light on different stakeholders' expectations towards participatory urban planning service and contributed to our first urban planning service demo, which is presented in this paper. In addition, we have conducted user studies in a small local community,

 ISBN: 978-1-61208-353-7

where several environmental urban planning projects are taking place. These projects include, for instance, supplementary construction, green design and planning of noise barriers. Through interviews, demos and case pilots we aimed to gain an understanding of how users perceived this kind of participatory mixed reality services in real urban planning projects and how to develop the service further for large-scale participatory projects.

This paper is structured as follows: Section II describes how the possibilities of mixed reality technologies have been experienced in urban planning. Section III describes the ubiquitous mixed reality concepts we discussed with different stakeholders to involve them to participative urban planning service design. Section IV describes the political decision makers' and city officials' views on participatory urban planning concepts. Section V addresses the environment project and goes into detail about the citizens' feedback on participatory urban planning service demo. Finally, in Section, VI we present conclusions and our future work.

II. THE ROLE OF MIXED REALITY SERVICES IN URBAN PLANNING: SOME EXPERIENCES

Different technological approaches such as virtual reality, mirror worlds and mobile augmented reality have been experimented on for aiding public participation in urban planning [2][3][4]. Immersive visualisation tools help users to understand what is being proposed and planned as many non-experts have difficulties to understand maps and plans [5][6]. Technological barriers to participatory urban planning and e-government are coming down, particularly at the local municipal level, and there are new opportunities for public engagement based on local needs and capacities [8]. Design of interactive systems can affect citizen participation in local governance and urban planning. The interactive systems should be flexible and versatile and enable participatory design approach, which goes beyond professional design projects and allows users to suggest further adaptations. New types of user interaction and technology design solutions should be considered to encourage citizens and other stakeholders to participate [9]. There are several recent examples of this. A prototype of a mixed reality application supporting a range of devices for a collaborative multimodal interaction was developed by Wagner et al. to enable group of participants to create a vision of urban projects. The stakeholders and users involved in the urban planning project had various backgrounds ranging from local urban planning specialists to other stakeholders such as members of local commerce. Mixed reality visualizations proved useful in enriching the available repertoire of representations and enhancing stakeholders' understanding of urban situations. 3D visualizations, videos and sounds helped to express and co-construct their ideas. Sound was perceived also an important element in urban planning, but a more complex medium in the participatory process [7].

Web-based solutions provide good support for the traditional methods used in the participatory urban planning process. These applications are especially suitable for acquiring local knowledge; they are an easy and inexpensive way to reach large and diverse groups of respondents. Noujua et al. noticed that there are also challenges related to the web-based participation; for instance that it may produce shallow information and the participation may be quite random [11].

A smartphone augmented reality system for urban planning was tested with 18 members of the public. The objective was to test if a smartphone augmented reality system would increase the willingness of the public to participate in urban planning events and if the participation was actually increased. The aim was also study qualitatively how the public perceived the smartphone augmented reality system in urban planning. The results of the study show, as expected, that the younger members (the 18–25 age group) of the test groups were more familiar with smartphone technology and saw the system as easy-to-use. Only the youngest age group showed an increase in willingness participate in urban events if the smartphone augmented reality system was used in the events. However, the study could not show any evidence that the actual participation in urban planning events would have increased because of the use of the smartphone-augmented system [4].

Increasing participation requires time and resources from all stakeholders. From city officials' and other planning professionals' point of view seeking citizen involvement via web based and mixed reality services does not necessarily decrease the workload, and the professionals need to be strategically prepared to manage new flows and ideas coming from citizens [4].

III. UBIQUITOUS MIXED REALITY CONCEPTS FOR URBAN PLANNING

In the beginning of the interviews with political decision makers, city officials and companies, we introduced four ubiquitous mixed reality concepts for urban planning. The examples helped in figuring out the idea of new visual approaches to community planning and aimed to facilitate feedback and ideas related to the different approaches.

A. Mixed reality mobile tools

We described possibilities of visualizing urban planning solutions with smartphones and tablet devices. The idea is for users to be able to move around the surroundings under development and see merged virtual 3D objects and a camera view on a handheld device (Fig. 1). Moving in real environment and utilising mobile augmented reality (AR) solutions, which integrate virtual information into physical environment, help users to perceive scales and sizes related to the existing buildings and surroundings when virtual building objects will be located in their intended locations. The demonstrated mobile mixed reality tool for architectural sites has been described and evaluated in earlier studies [8][9][10]. Current trend in mobile devices supports the idea of utilising on-site AR solutions. Smartphones have advanced processors and cameras, they utilise large displays and tablet devices are already commonly used.

 ISBN: 978-1-61208-353-7

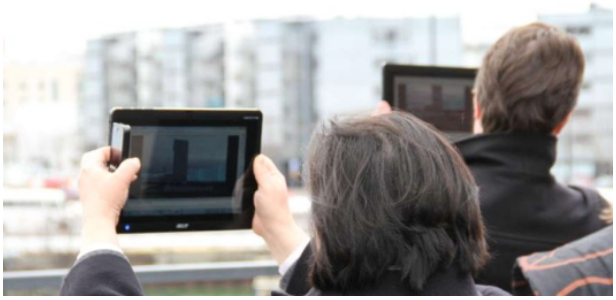


Figure 1. On-site augmented reality solution

B. Interactive public screens

The other approach presented was interactive public screens with mixed reality and advanced control methods features (Fig. 2). The screen shows areas under development, and new digital visualizations are embedded into the views. One or several users can manipulate the views and community plan options using their gestures or the touch screen input method. Other users can watch and discuss the views at the same time. Gesture recognition would be implemented with the help of depth camera sensors or handheld control devices. On the other hand, utilisation of personal smartphones as a second screen device could also be applicable method in the public screen approach. This kind of public screens can be located next to the area under development, in shopping centres or in municipal office buildings.

Figure 2. Interactive public screen in urban planning.

C. Multiuser design tables

Thirdly, the users can explore and co-create urban planning solutions using interactive and multiuser design tables. The tables can be a combination of tangible objects or 3D printed building models, projected information and camera recognition systems. The users are able to browse different urban planning options or manipulate objects on a table, and they can receive more information using, e.g., pointing, touching or gestures.

The user moves and indicates building options using AR markers on the table. A tangible 3D table-top system in which physical objects on a table can be recognized has been developed [10]. The same kind of table-top systems in urban planning includes the *Spatial Design Table* and the *Bionicle Table* [11][12]. They both enable 3D visualizations showing how different buildings look in their environments.

D. Web-based service with panoramic visions

One concept developed for participatory urban planning is web-based open service, which can easily be used with personal devices as they run on web browsers of different devices such as tablet devices and PCs. The user-centric service mixes panoramic imaging and architectural drawings of future urban plans, and includes user-centred interactions such as questionnaires and commenting tools so that citizens can participate and comment on timely issues, such as future urban planning, construction of green walls and urban gardening, and sustainable energy solutions. Fig. 3 is a screenshot of the web-based urban planning service, which visualizes sound barrier plans between a highway and a field. Left side includes map information of the area and how the sound barrier is located. The right side visualizes the sound barrier plans and its future construction phases embedded in panoramic images of surroundings.

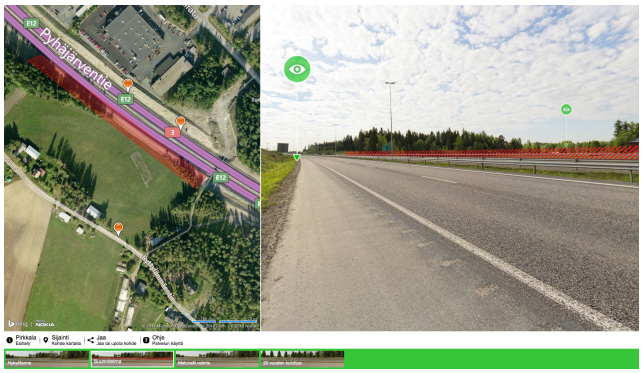


Figure 3. Web-based service with mixing panoramic imaging and architectural drawings of future urban plans and including user-centric feedback tools

IV. POLITICAL DECISION MAKERS' AND CITY OFFICIALS' VIEWS ON THE PARTICIPATORY URBAN PLANNING CONCEPTS

In the interviews with political decision makers we discussed how to increase the awareness of citizens' opportunities to participate in planning of their living environment and how to increase real interaction with other stakeholders. The political decision-makers were selected from all the parties presented at these boards. The participating political decision-makers were both experienced and new decision-makers who were in their first term on the Board of Governors. They represented both genders and as members of environmental or technical boards they were in a central role in organizing services related to urban planning. The political decision-makers were selected from all the parties presented at these boards.

A common concern was that people are not aware of the general process, they lack of information on the channels for interaction and participation and at which phase they have an opportunity of commenting and providing feedback. Active citizens who acquire knowledge are well informed about the future urban projects of their communities and their

 ISBN: 978-1-61208-353-7

possibilities to contribute, but for the average citizen the statutory planning process is unfamiliar. In many cases, citizens do not realise that anything is going until before actual building begins in neighbourhoods and at that point it is too late to contribute. In addition it became clear, that recent participatory methods in urban planning are not suitable for all. As digital services may be unfamiliar to older generations, public urban planning events are tied to a certain time and place and in Finnish context situation is such that especially busy working generations and young people are easily left out. Encouraging young people and working families to participate in urban planning events is challenging.

By providing better accessible information, easier ways for participation and open communication awareness and a general understanding of the nature of the planning process could be increased.

Another concern was the state of planning when involvement of different stakeholders usually happens. Too mature, detailed plans and quality of materials were seen as the main problems from both the decision-makers' and citizens' point of view. Decision makers would like to have information on new projects at an earlier stage when planning has not yet reached a high level of maturity. They noted that they often face a situation where they can only accept or reject a proposal. There is a demand for alternative design solutions, a discussion of the impacts of different solutions and for a better illustration of the overall picture. Materials were also complained of as being too complicated and difficult to understand, not only from the decision makers' point of view, but even more by average citizens. Also, the amount of material is often huge and that makes it almost impossible to go through all the information to get an overview and a clear understanding of the wider impacts of the project. The situation is the same with the citizen. That easily leads to uncertainty and resistance among the citizens, which may result in an increased number of complaints and a longer time for processing.

Open communication between different stakeholders was seen as a way to enhance real co-operation and participation, which would help the discussion to resolve the problematic issues already at an early stage of planning. The numbers of complaints are assumed to decrease if different viewpoints can be taken into account as early as possible.

Different concepts of participatory mixed reality tools (described in Section III) were presented for the interviewees. Participants were asked for feedback and to evaluate the possible impacts of the use of such tools. All the concepts were seen as interesting and the possible impact for urban planning process was seen positive. Use of public tools would enable more flexible and diverse ways for stakeholders to participate in commenting on the plans. It would make communication more effective when the information can be brought among people flows. Scalability of tools was also seen as positive.

Of the presented technology approaches, the decision-makers and city officials prioritized lightweight, web-based mobile solutions. Other presented solutions, such as the interactive design table and public screens, were also seen as viable in the long run. They were seen as suitable for large urban planning projects and as tools for both decision-makers and citizens. Interactive public screens were seen as effective attention grabbers and information channels: they were considered a good way of spreading knowledge of urban planning projects. However, screens were seen as less suitable for collecting feedback and ideas from the general public. It was assumed that people would be hesitant to use a technical device that was for public use. The actual participation and feedback would happen via a personal mobile or other personal device, or in a more closed facility organized by the city or community. User interfaces that recognize gestures were seen as better suited to public spaces than touch screens. Touch screens in public use were perceived as uncomfortable especially because of hygienic reasons.

Even if the need for easy-to-use, light and adaptive visualisation tools was recognised, a concern was how the tools will be adopted. Ability and willingness to take new technical solutions into use were seen as challenging.

Another concern that was raised in discussions with decision makers was that, due to new tools, the amount of data is likely to be increased and for that reason new tools and methods are also needed to handle and analyse all that data effectively and to produce readable reports. Even now the amount of information and material is often great, especially in the case of larger planning projects, and a lot of effort is needed to go through all that material.

Considering the earlier experiences of mixed reality technologies in urban planning and the feedback from political decision makers and city officials, we developed a service demo, which is an open urban planning service. It can easily be used with personal devices as they run on web browsers of different devices such as tablet devices and PCs. The service mixes panoramic imaging and architectural drawings of future urban plans, and includes questionnaires to acquire local knowledge. In the next Section, we describe how the demo was used in a local environment project and how citizens perceived use of this kind of service in urban planning. The interviews with political decision makers helped us to clarify what kind of technologies and what kind of user features would suit best for participatory urban planning. We have used this information in designing of our participatory urban planning demo. In the next Section, we will describe how we used this demo as a part of an environment project and how users responded to it.

V. THE ENVIRONMENT PROJECT

We conducted user studies and participatory urban planning pilots related to real environment projects with our service demo in Western Finland. In this region, there are several large future urban planning projects planned related to public traffic and development of city and community centres. For instance, the international airport area is under lively development. Local media reports frequently on new urban plans, and in local government there are debates for and against different urban planning and environment projects. We conducted our first user demo in a small local community. We wanted to ascertain how to support citizens

 ISBN: 978-1-61208-353-7

and other stakeholders in involving them planning of the sustainability and quality of their living environments through digital services. We wanted to find out how our demo service suited this purpose, and how to develop it further, especially trying to understand user values, needs and preferences in participative urban planning. We first conducted a user study in a small village near the highway where a new noise barrier is planned to protect inhabitants from noise pollution. There are only town houses in this area, and residents of the village consisted mainly of families with children and older people. Our aim was to reach residents living near the noise barrier to respond to our inquiry, so we published an online questionnaire link in a municipal community web portal, community Facebook site and in a local newspaper. The query was available over a period of a few weeks in March and in April 2014.

A. *Citizens' feedback on participatory urban planning demo and devices*

In all 25 respondents (12 males, 13 females) completed the web-based questionnaire, which included both multiple choice and qualitative open-ended questions. Most of them belonged to the age group from 35 to 44 year olds. They were quite highly educated: 10 of them had a bachelor's degree, 6 of them had a Master's degree and two of them had the level of doctorate.

The survey included basic background information questions, and focused on topics such as clarifying requirements for a future community planning, perceptions on visualisation and participation services, and most preferred places and information channels and devices for utilising a future participatory urban planning service. Users were also asked to try out the web-based pilot service which mixed panoramic imaging and architectural drawings of the planned noise barrier near their homes. The demo illustrated noise barrier building stages and the area five and ten years later.

Fig. 4 shows how users would like to have access to participatory urban planning service. From the options given, users would most likely to use the service from municipal web pages. They also preferred mobile devices as a convenient way of using the services in the local environments. However, users reflected that they would quite unlikely to use it from a municipal service point. Also municipal public events and notices in public transport were quite uncertain or unlikely places to access and use these services. In open-ended questions, users reflected that it would be problematic to give their opinions in such public places, if they wanted to maintain their privacy. One respondent pondered that it would be most convenient to participate with a personal device on a couch at home and it would be more likely to increase the possibility of participating in a public planning event also, if the plans are easy to access with personal devices and they are related to one's own neighbourhood.

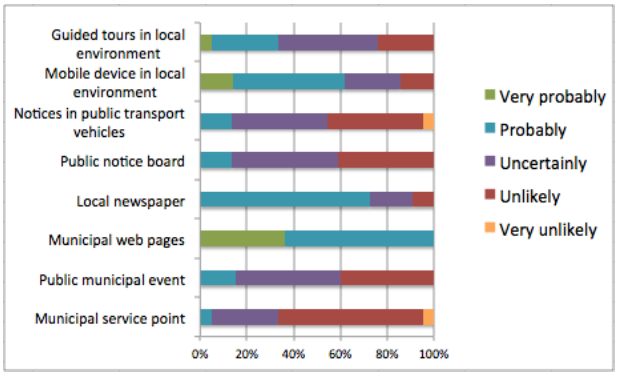


Figure 4. Where and with what devices users would you like to have an access to participatory urban planning service?

B. *Citizens' views on using participatory service in environment and sustainability related projects*

Fig. 5 provides an overview of the questionnaire responses to the question of presenting information related to municipalities' community planning and construction projects. Half of the respondents stated that impacts on the environment and showing alternative plans are very important. Also, explicit information on timetables and upcoming phases are at least as important to more than four fifth of the respondents. Only the need for visualising influences of seasons of the year was not seen as very important.

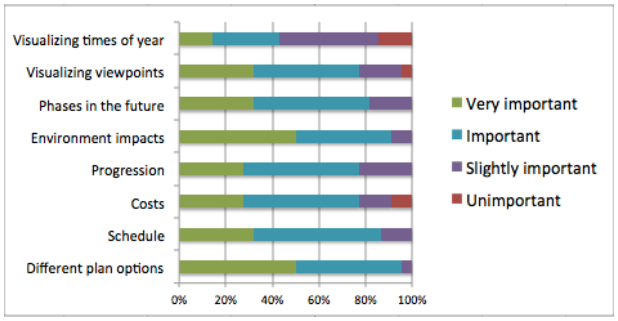


Figure 5. Importance of visualising and sharing different information on community planning projects through the participatory service.

Fig. 6 indicates respondents' feedback related to how well future visualization and participation tools are applicable for municipalities' environmental development domains. Their attitudes towards environment-related development activities were mainly very positive. Only playgrounds had one negative feedback, but on the other hand nine users stated that playgrounds fit excellently with future community planning services.

 ISBN: 978-1-61208-353-7

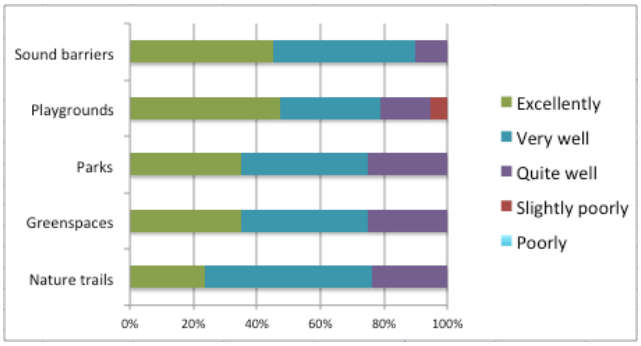


Figure 6. How well does the service suit different environment and sustainability projects?

The results of the survey were very much in line with the issues discussed with decision makers. In open-ended questions, urban planning information was complained of as being difficult to find, and the participation process is perceived as being too complex. Opportunities to interact and be heard were claimed to be challenging. Respondents requested involvement at an earlier phase of the planning process, more alternative solutions to be compared, clear timetables and information on how the process is progressing.

Respondents were mainly interested in the projects that are linked to their neighbourhood area, somehow reflect their everyday lives or projects that are supposed to have large, revolutionary influences not only geographical but also at the societal level. At present, the information on ongoing projects is sought from municipal websites and from the local newspaper, which are both listed as municipal official communication channels.

Participants were asked how the visualisation service succeeded in visualising the example case. In general, the service was found to be interesting, useful, and easy to use. The way the service visualises planning material was seen to be beneficial compared to traditional methods. Especially the possibility of viewing the target area from different viewpoints in a real environment was seen to be valuable. This feature also helps to locate plans, e.g., new buildings in the current surroundings and to illustrate the effects on the landscape.

VI. CONCLUSION AND FUTURE WORK

The design of participatory urban planning services can have a great impact in developing smart and sustainable environments. To have optimal user-satisfaction, services should be flexible and adaptive and provide access to plans at any time and with any devices users prefer. Citizens in general are interested in commenting on and participating in urban planning projects, which are related to their everyday lives and their own neighbourhood. Urban places such as public interactive screens at transportation or municipal service points are good options in informing about future plans, however, users may be more hesitant to use them to give comments or feedback than to use their own, personal devices. Young people have been a little more active in responding to on-line surveys, and interested in new technological approaches such as smartphone augmented reality, but it is still challenging to find a method to activate young people to participate and influence their living environments.

Moreover, in order to become effective, co-creative and influential, the service should be either specially designed for each relevant user segment and research theme or activate large numbers of users to comment and share their ideas. Important changes in urban planning would be to increase information, communication, collaboration and present more alternatives at the early stages of projects. Also, informing the public how their feedback has been taken into consideration in urban planning is important. Up-to-date information should be easy to find for instance under the same service.

In the next steps of our participatory urban planning service development project, we will pilot the demo in a large urban planning project to demonstrate green design and urban farming in a city centre. Furthermore, another important question to study further is how to consider and use this user feedback in decision making so as to plan and co-create user-centric smart cities and sustainable living environments.

REFERENCES

[1] M. Steen, M. Manschot, and N. De Koning, “Benefits of co-design in service design projects.” International Journal of Design, vol. 5, 2011, pp. 53-60.

[2] G. Rambaldi, P. A Kwaku Kyem, M. McCall, and D. Weiner “Participatory Spatial Information Management and Communication in Developing Counties.” The Electronic Journal on Information Systems in Developing Countries, vol. 25, 2006, pp. 1-90.

[3] M. Majale “Employment Creation through Participatory Urban Planning and Slum Upgrading: The Case of Kitale, Kenya.”, Habitant International, vol. 32, 2008, pp. 270-282.

[4] M. Allen, H. Regenbrecht, and M. Abbott, “Smart-Phone Augmented Reality for Public Partcipation in Urban Planning” Proceedings of the 23rd Australian Computer-Human Interaction Conference. ACM, New York, NY, USA, 2011, pp. 11-20.

[5] J. D. Salter, C. Campbell, M. Journey, and S.R.J. Sheppard “The Digital Workshop: Exploring the Use of Interactive and Immersive Visualisation Tools in Participatory Planning.” Journal of Environmental Management, vol. 90, 2009, pp. 2090-2101.

[6] S. R.J. Sheppard “Landscape Visualisation and Climate Change: the Potential for Influencing Perceptions and Behaviour.” Environmental Science & Policy, vol. 8, 2005, pp. 637-654.

[7] I. Wagner, M. Basile, L. Ehrenstrasser, V. Maquil, J-J. Terrin, and M. Wagner, “Supporting Community Engagement in the City: Urban Planning in the MR-Tent.” C&T '09 Proceedings of the fourth international conference on Communities and technologies, ACM New York, NY, USA, 2009, pp. 185-194.

[8] C. Skelton, M. Koplin, and V. Cipolla, “Massively participatory urban planning and design tools and process: the Betaville Project.” Proceedings of the 12th Annual international digital government research conference: Digital government innovation in challenging times. ACM New York, NY, USA, 2011, pp. 355-358.

 ISBN: 978-1-61208-353-7

[9] J. Saad-Sulonen, A. Botero, and K. Kuutti, "A long-term strategy for designing (in) the wild: lessons from the Urban Mediator and traffic planning in Helsinki" Proceedings of Designing Interactive Systems (DIS'12). ACM New York, NY, USA, 2012, pp. 166-175.

[10] V. Oksman, A. Väätänen, and M. Ylikauppila, "Future Illustrative and Participative Urban Planning." Proceedings of CONTENT 2014, The Sixth International Conference on Creative Content Technologies, 2014, pp. 22-29.

[11] J. Noujua and A. Jussila, "Exploring Web-based Participation Methods," Proceedings of the Tenth Anniversary Conference on Participatory Design (PDC '08), Indiana University Indianapolis, IN, USA, 2008, pp. 274-277.

[12] E. Seltzer and D. Mahmoudi, "Citizen Participation, Open Innovation, and Crowsourcing: Challenges and Opprtunities for Planning", Journal of Planning Literature. Sage Publications, vol. 28, 2013, pp. 3–18.

[13] T. Olsson, A. Savisalo, M. Hakkarainen and C. Woodward, "User evaluation of mobile augmented reality in Architecture", Engineering and Construction, 2012, pp. 733–740.

[14] C. Woodward and M. Hakkarainen, "Mobile mixed reality system for architectural and construction site visualization". In augmented reality –Some Emerging Application Areas", Andrew Yeh Ching Nee (ed.), InTech, 2011, pp. 115-130.

[15] P. Dalsgaard and K. Halskov, "Tangible 3D tabletops: combining tabletop interaction and 3D projection". Proceedings of NordiCHI 2012, ACM Press, 2012, pp. 109–118.

[16] H. Ishii, et al., "Augmented urban planning workbench: Overlaying drawings, physical models and digitalsimulation." Proceedings of the 1st International Symposium on Mixed and Augmented Reality. IEEE Computer Society, 2002, pp. 203-211.

[17] R. Nielsen, J. Fritsch, J. K. Halskov and M. Brynskov, "Out of the box – Exploring the richness of children's use of an interactive table". Proceedings of the 8th International Conference on Interaction Design and Children (IDC '09), ACM New York, NY, USA , 2009, pp. 61-69.

 ISBN: 978-1-61208-353-7

A Case Study on Understanding 2nd Screen Usage during a Live Broadcast

A Qualitative Multi-Method Approach

Mari Ainasoja, Juhani Linna
School of Information Sciences
University of Tampere
Tampere, Finland
firstname.lastname@uta.fi

Päivi Heikkilä, Hanna Lammi, Virpi Oksman
VTT Technical Research Centre of Finland
Tampere, Finland
firstname.lastname@vtt.fi

***Abstract*— Media multitasking with different 2nd screen devices – i.e., with tablets, smartphones and laptop computers that are used simultaneously with viewing a television broadcast – has rapidly become a common user behavior. While this behavior can be measured quantitatively in several ways, fairly little is known about the reasons and motives behind it. To compose a better understanding of the role that 2nd screens have in the viewing experience of the user, we conducted a case study with 12 participants by using a combination of qualitative data collection methods. Through thematic analysis we combined four ideal types, Commentator, Analyzer, Home Gamer and Active Follower, which exemplify the different meanings that 2nd screen usage has for the viewer.**

Keywords-media multitasking; 2nd screen; case study; user types.

I. INTRODUCTION

The emergence and evolution of mobile technology and mobile services have changed the media landscape irrevocably. Media users can now "go online" virtually in and from anyplace and have nearly unlimited access to media content and services from myriads of providers. Media businesses – from newspaper companies and TV broadcasters to digital content providers – are trying to re-invent their products and business models to suit the evolving needs of the mobile user. This re-invention needs empirically supported knowledge from all aspects of mobile device use.

Using mobile devices simultaneously with other media content is a change in user behavior that has emerged rather recently. An estimation from Nielsen [1] suggests that 69 % of US users aged 13 and above use tablets while watching TV at least several times per week. Google's survey [2] suggests that 81 % of smart phone users used it simultaneously with TV. These figures may be among the highest estimations, but it is nonetheless safe to say that this type of multitasking has become common very rapidly.

We call tablets, smartphones and laptop computers that are used simultaneously with TV viewing as 2nd screens. While media multitasking is an old phenomenon on a non-specific level (e.g. reading a magazine while listening radio), the fact that 2nd screens have the capability to be connected to the same media experience in a personalized way creates a whole new design paradigm for both media business and media research.

Currently, users' behavior with 2nd screen services is tracked through a variety of quantitative variables – e.g. amounts of downloads, clicks, traffic sources and use flows. These variables are mainly used for measuring service performance and the business impact of a specific service. Researchers, whose goals tend to be wider and more of a theoretical nature, have used mainly surveys and laboratory experiments to study the use of 2nd screens.

While quantitative methods such as tracking digital footprints or conducting online surveys may be the only practical way to study media use among the masses, our case study takes a qualitative perspective on 2nd screens and thus contributes to understanding reasons and motives behind the usage figures. In order to design 2nd screen services that support or enhance users' ways of using media, we need to understand why 2nd screens are used and how they could potentially change the media experience. These issues are particularly interesting in the context of live broadcasts, as they make synchronized information flow between the broadcaster and the 2nd screen possible. Consequently, our research question is defined as follows:

What kind of role does the 2nd screen have in the users' viewing experience during a live broadcast?

An entertainment program called the Voice of Finland (VoF) served as a practical setting for the research. We selected a sample of 12 adults aged from 20 to 38 years who considered themselves "followers" of the show and studied their 2nd screen use relating to the show. We used a mixture of data collection methods – media diaries, theme interviews and on-site observations – to provide a broad perspective on the research problem.

The article proceeds as follows: the next Section summarizes earlier research on 2nd screen usage, and Section 3 elaborates the qualitative case study approach and the combination of data collection methods utilized in this study. Empirical results are described in Section 4 and discussed further in Section 5. In Section 6, implications and limitations are considered.

II. STUDIES ABOUT THE USE OF 2ND SCREENS

A. *Media multitasking as a general phenomenon*

In research, media multitasking has been the most common viewpoint used to describe multiple media use. This concept covers a broad phenomenon including different

 ISBN: 978-1-61208-353-7

media channels and both non-media [3] and media [4] multitasking activities: In this article, we concentrate on the latter: "the practice of participating in multiple exposures to several media forms simultaneously" [4].

Media multitasking has been on the research agenda for around 10 to 15 years; for example, Pilotta et al. [5] found out already in 2004 that only 16 percent of the US media population did not engage in simultaneous media usage. Popular viewpoints in prior research on media multitasking include the ability to process information and perform tasks in multitasking environments, gaze distribution between media and age-based differences in multitasking behavior [6].

Our case study focuses on multiple media use that is related to a specific, live TV program. In a more general laboratory study on concurrent TV and laptop use, Brasel et al. [6] found out that participants switched their attention between a television and a laptop at an extremely high rate. The computer dominated the television for visual attention, and the gazes captured by the television were shorter than gazes captured by the laptop. However, the time was split between more web pages than channels.

B. *From media multitasking to program-related 2nd screens*

Most of the media multitasking around TV – say, email checking or online shopping – is unrelated to TV watching and only part of it enriches the actual program content and watching experience [7][8]. The simultaneous, program-related use of other devices during a TV broadcast has gained less attention in academic research than the wider phenomenon of media multitasking. To date, studies taking the user perspective have been conducted mainly at the level of industry reports using the terms 2nd screen or social TV [9], and academic studies are still scarce.

One stream of research has focused on developing and designing 2nd screen devices and solutions both for controlling TV programs/services and for enriching TV programs with interactive features such as quizzes and voting [10]. Cesar et al. [7] summarize the roles of a program-related 2nd screen through a taxonomy of three activities: content control, content enrichment and content sharing.

Another stream of academic studies on 2nd screens has focused on social media (e.g. Twitter, Facebook or instant messaging) usage during TV broadcasts. For example, Han et al. [9] used a qualitative convenience sample consisting mainly of students and found five motivational categories for the complementary use of text-based media (mainly instant messaging and social media) during live TV broadcasts: communicating about impressions of a broadcast, information sharing and seeking, feelings about co-viewing, curiosity about others' opinions and program recommendations.

In addition to established social media channels, some research has also been done on social TV applications such as GetGlue, Intonow and Miso in the US. Basabur et al. [11] emphasized in their field trial that program-related use of social TV applications is characterized by the innate need for social validation: a place to show off the knowledge, compete on who makes the funniest comment, and validate the feeling that friends think alike and belong together.

More recent academic research has focused on 2nd screen applications that are offered by TV companies to enrich the viewing experience of one specific program (like the VoF and the HomeCoach application in our case study). The use of these applications is in line with more general media multitasking: interest in 2nd screen applications and the intention to use them is higher among people who are used to using other media during television viewing [12]. However, some studies have pointed out a critical view on the potential of program-related 2nd screens, emphasizing the strengths of TV as a low-effort medium and reporting fairly low usage rates of interactive features unique to specific TV programs [10][13]. Also in the study of Courtois et al. [12], respondents were only slightly interested in using program-related 2nd screen applications and preferred using established social media channels instead of specially designed applications. On the other hand, in the study of Basabur et al. [11], participants appreciated the idea of getting different aspects of the 2nd screen in the same application if the integration with established social media was smooth.

C. *2nd screen and entertainment programs*

The program genre plays an important role in, for example, the use of TV in general, and more particularly in the use of social TV [11][14][15]. For example, the genre of television content had stronger effects on gaze duration distributions than individual psychological differences in the study by Hawkins et al. [16].

Social features have been pointed out to be especially suitable for sports, reality TV, quizzes and home decorating shows [14][15]. A study by Geerts et al. [15] relates this to the plot structure of a show claiming that people do not talk while engaged to a plot. Also Basabur et al. [11] concluded in their field trial that the program genre affected how much effort users were ready to put on making and reading comments on the 2nd screen. If the program required a lot of attention, like dramatic shows, users experienced creating links and informational posts to a 2nd screen as distracting; however, simple commenting was still accepted and fans of shows knew the patterns of the shows and became skillful in knowing when they can take their eyes off from primary TV content.

In a qualitative study by Han et al. [9], it was reported that communication about one's impressions of a broadcast was the strongest motivation for using instant messaging and/or social media during entertainment program broadcasts. This motivation means exchanging mutual thoughts or opinions, developing a bond of sympathy, using the content of a broadcast as a topic of conversation and talking about persons on air. The second most frequent motivations with entertainment programs were 'information sharing and seeking' and 'feelings of co-viewing'.

III. QUALITATIVE CASE STUDY APPROACH

Earlier research has used mainly surveys and laboratory experiments to study the use and users of 2nd screens. Our

 ISBN: 978-1-61208-353-7

goal was to compose a deeper and more coherent understanding of the role that 2nd screens can have in the personal viewing experience – not to generate statistically generalizable results from a specific form of data or evaluate the performance of the 2nd screen application itself. A qualitative case study approach was found to best suit the purpose of our research: Case studies focus on understanding the dynamics that are present in single settings [17], and qualitative data offer insight into complex processes that cannot be reached with quantitative data [18].

The advantages of qualitative data collection methods are known, but so are their disadvantages. Reflective methods like interviews, where the respondent reflects his or her media experiences, are known to have reliability issues [6]. Respondents do not remember their doings correctly, and nor are they aware of everything they do [19]. On the other hand, direct observational methods do not reveal the motivational background or users' own interpretations. Further, data collection that is conducted in laboratories or in controlled settings fails to address the natural complexity of the media experience in an interpersonal context [20], for example.

Based on the rationale above, we designed a research framework that had its groundings in contextual inquiry. Contextual inquiry is an ethnographic research method that aims to find naturally used roles, attitudes and behaviors to support design work [19]. We collected the data with multiple qualitative methods, both observational and self-reflective methods, to overcome the disadvantages that a single method may have had. All prerequisites and requirements relating to 2nd screen devices or their use were discarded once the participants were chosen. In other words, we encouraged natural use of 2nd screens – in contrast to Brasel et al. [6] and Tsekelevs et al. [10], for example.

A. *Case: The Voice of Finland*

The VoF offered versatile possibilities for 2nd screen interaction and thus also a rich practical setting for the research. The show is based on a Dutch concept, the Voice of Holland, developed by Talpa Holding NV in 2010 and now franchised to over 20 countries. In the show, a group of amateur singers compete against each other under the guidance of professional artists. In Finland and in several other countries it is one of the most popular shows on TV. The VoF's 2nd screen channels included active and actively promoted Twitter and Facebook channels, partly simultaneous broadcasting with different content through the program's web page, and a specifically developed, free and interactive 2nd screen application – HomeCoach – that was synchronized with the live broadcast (Figure 1). The application (Kotivalmentaja in Finnish) was freely available for Android and iOS smartphones during the season and it offered different functionalities in different phases of the VoF, from guessing the course of the show to evaluating the performances and cheering the contestants. In a sentence, at the time of the study the VoF was the most comprehensive attempt to bring together TV broadcast and 2nd screens into one rich media experience.

B. *User participants*

The user participants for the study were chosen from a participant pool which was formed on the basis of a recruitment questionnaire. The questionnaire was advertised

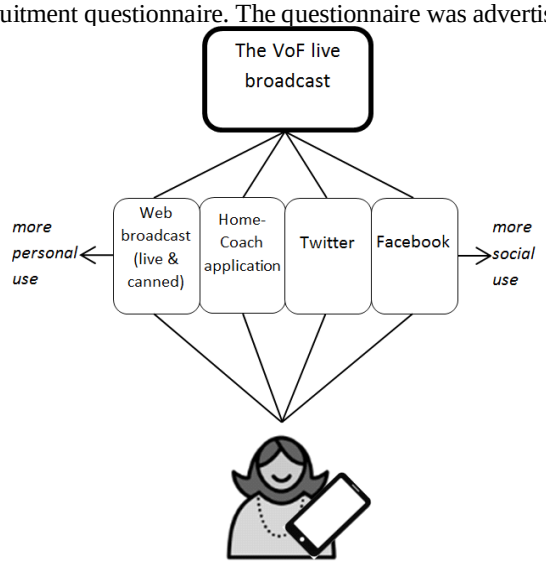


Figure 1. The VoF and the official 2nd screen channels

on the VoF Facebook page. Besides demographics, we collected information about general 2nd screen use and watching habits. From the 261 people that filled in the questionnaire, we picked 12 adults who watched the show in different social settings and whom we regarded as early adopters or early majority (as defined by Rogers [21]) in broadcasting-related 2nd screen application adoption. The 12 participants who resided in Southern Finland and were aged from 20 to 38 are summarized in Table 1.

C. *Media Diaries and Contextual Interviews*

All participants were asked to fill in a qualitative and mostly open-ended media diary for seven consecutive days, and they were interviewed both in the beginning and at the end of the diary period.

TABLE I. THE PARTICIPANTS OF THE STUDY INCLUDED NINE FEMALES (F) AND THREE MALES (M).

	Age	**Household type; living**	**Described typical VoF watching situation**	**Home Coach**
F1*	20	With a friend	With a friend	Yes
F2*	20	With a spouse	With a spouse	Yes
F3*	24	Alone	With a group of friends	Yes
M1	28	Alone	Alone or with a friend	Yes
F4*	32	With a spouse	Alone	Yes
F5	34	Alone	Alone or with a friend	Yes
M2	35	With a spouse and 2 children	With family	No
F6*	35	Alone	With a friend	Yes
F7	35	With a spouse and 2 children	With family	Yes
F8	36	With a spouse and 2 children	With family or with a spouse	No
F9*	37	With a spouse and 3 children	With family	No
M3	38	With a spouse and 3 children	With family or with a spouse	Yes

* Participated in the on-site observation

 ISBN: 978-1-61208-353-7

The first interview happened in all but one occasion in the participant's home, i.e., in the context where she or he viewed the VoF. In the interview we were interested in the personal 2nd screen usage motives and habits on one hand, and in the social context (both virtual and physical) in which this usage happens on the other. We encouraged the participants to provide their own interpretations, e.g. by asking "*How would you describe your use*" instead of "*How do you use (the 2nd screen device in a certain situation)*". All interviews were recorded and transcribed for further analysis.

The main purpose of the pre-structured media diaries was to provide a longer temporal viewpoint on the research question: While the weekly live broadcasts were in the central focus of the study, show-related 2nd screen channels were active also between the shows. The content of the media diaries was walked through with the participants to ensure that they understood what was required from them. In the diaries we asked the participants to make notes on their 2nd screen use before, during and after a chosen TV viewing event, and the tasks they did relating to VoF each day. We also asked the participants to describe the social setting where the viewing happened, and even the emotions and feelings that were attached to the situation. With the question setting we also aimed to provide rich stimuli for the second interview, which was conducted on the basis of the filled-in diary.

D. *On-Site Participant Observations*

In addition to media diaries and contextual interviews, we conducted six participant observations in the participants' homes – i.e., in their typical watching environment – during the VoF live broadcasts on Friday nights.

Two researchers were present in each observation. The observation sessions started with an explanation of the process (i.e., research methods and use of data) and were followed by an interview, as described in the previous section. The participants and other people present were free to ask questions at any time. These discussions that preceded the observation also served as a way of building trust – or rapport, as described by Guest et al. [22] – with the participants. This trust-building session was an essential part of the participant observation, since natural behavior was encouraged. The main challenge of participant observation, namely, the possible impact of the observer on the behavior of participants, was recognized. However, it was justified to use this method because asking clarifying questions immediately on-site enabled us to study the motivational background and users' own interpretations, which are central to our research question. Additionally, in this multi-method approach, the results do not build solely on observation, but the diaries and interviews complemented the picture.

During the VoF live broadcast, the researchers took notes on the behavior of the participant, from time to time asking clarifying questions; for example, when it was unclear what the participant did with the 2nd screen. It was agreed beforehand that the other researcher would focus especially on the 2nd screen use, while the other observed the social context of the viewing event. All the observations were videotaped, with the permission of everyone involved.

After the broadcast, the participants were briefly interviewed on the basis of the observations. This was done to asses both the typicality of the viewing event and the validity of the researchers' interpretations.

E. *Thematic Analysis*

Thematic analysis is a flexible analysis method that allows the use of different types of qualitative data. There is no explicit way of practicing this widely used method, but as a general definition it is used for identifying, analyzing, and reporting patterns (i.e., themes) within the data [23].

In this study the collected data – i.e., the transcribed interviews, media diaries and observation reports – were first initially coded separately by the researcher who interacted with the participant in question. The results were then shared and discussed within the team to build mutual understanding between the researchers. These discussions were especially important since the research team was a multidisciplinary one with backgrounds in psychology, marketing and technology.

From the initially coded data, we identified themes that would most clearly define the role of the 2nd screen use in the viewing experience. From these themes we combined four ideal viewer types and named them as Commentator, Analyzer, Home Gamer and Active Follower. The purpose of these ideal types is to illustrate the different roles that the 2nd screen serves in a TV viewing experience. Ideal types are common mental constructs in social sciences; they do not conform to reality in detail, but rather approximately [24]. Each of these types has a different use motivation and a different way of using 2nd screens.

IV. RESULTS: DIFFERENT VIEWER TYPES

A. *Commentator*

A Commentator mainly uses the 2nd screen because of two reasons: it provides social amusement and gives novelty value. Commentators like to watch interesting TV shows with others and they enjoy observing and commenting on various aspects of the show spontaneously as they pop up. In the case of the VoF, they are not only interested in the music or the competition but also in the styling, musicians, audience and the speakers. Some Commentators searched these details online while watching. The show provides a forum for spotting interesting people and trends and the 2nd screen can support this. As the 2nd screen solution of the VoF did not support this kind of trend spotting, some commentators felt that it is irrelevant or disturbing, as it interferes with observing and commenting. For this user group, information on their interests, such as brands of clothing or members of the house band, could make the 2nd screen more attractive.

For Commentators, watching TV and using 2nd screens is often a social event. Watching the show together with others gives a possibility to spontaneously exchange opinions of anything seen in the show. Commentators are also willing to try out new things, and 2nd screen applications are one opportunity that they want to check out with their friends or family. 2nd screens can be used together or at least shown to

 ISBN: 978-1-61208-353-7

others once in a while. For those commentators who watch TV alone, Facebook, instant messaging and Twitter can serve as the social aspects. Commentators are an optimal target group for social 2nd screen solutions or social media which give interesting information on trends and various aspects of the show. However, the 2nd screen should not require intense concentration, as there may be interesting issues on TV to observe any time. Some commentators fluently use several 2nd screens, social media and 2nd screen applications during one TV show, while some prefer to concentrate on the main screen – in order not to miss any important moments.

The following observational note demonstrates how Commentators constantly comment on the different aspects of a program and how the use of a 2nd screen together with a friend blends in with this commenting:

> One of the competitors, Eve Hotti, has started to sing. *"I would like to win a prize that includes a stylist that would style us like that, looking so nice."* Annikki says. Annikki and Mira compliment Eve's make-up and clothes. Mira sings along with the song and comments on the name of the performer: *"It would be nice to be Mira Hotti."* [The last name refers to "hot" in Finnish]. Annikki evaluates the performance with the HomeCoach application as good and tells this evaluation to Mira. When one of the judges gives the feedback in the show, Annikki comments: *"He [the judge] has become more masculine. That hair fits him well."*
>
> (Observational note, Annikki 35 years)

B. *Analyzer*

For an Analyzer, the 2nd screen serves as a tool for obtaining relevant additional information related to the content of the TV program. In the case of the VoF, this information helps them in analyzing the potential success of contestants. Receiving real-time updates of other viewers' reactions to the performances gives them a means to anticipate the results of the show and compare the general reaction of the audience to their own opinions. It is typical of them to ponder on the reasons for the success of certain performers – whether it is due to song selections, feedback from coaches or lack of similar contestants.

Analyzers are not only interested in seeing other TV viewers' responses, but they also want to express their own opinions and thus have an impact on the statistics of the show. As they value the accuracy of the information, it is also characteristic of them to evaluate performances precisely and even fine-tune their evaluations several times during a performance. Their commenting has a wider time frame: Analyzers may even compare the program content to previous program seasons, while Commentators spontaneously comment on things as they pop up in the program. As analyzing requires concentration on the program in question, Analyzers mainly use one 2nd screen at a time and are an optimal target group for program related 2nd screen applications. The 2nd screen application related to the VoF was especially liked by Analyzers.

The following observational note illustrates how Analyzers really put their effort into expressing their opinion precisely with a 2nd screen and how comparing one's own opinions to those of other viewers' becomes a part of 2nd screen usage:

> Tiina and Kimi are listening to performances closely and Tiina evaluates each with the HomeCoach during the song. They discuss about the song choices and the competitors that they believe will continue the competition from each team. During the performance of Antti Railio, Tiina first gives a rating with the HomeCoach that the performance is excellent. While the song goes on, she fine-tunes the evaluation and finally decides to use the Wow button [the best rating in the application] and tells about it to Kimi: *"I will give him a Wow"*. Before the song ends, Tiina shows Kimi the evaluations that others have given with the application: *"See, almost everyone says that this was excellent."*
>
> (Observational note, Tiina 20 years)

C. *Home Gamer*

For a Home Gamer, the 2nd screen brings engaging extra activity besides watching the show. Home Gamers are excited about possibilities which support playful competition between the TV viewers. Although Home Gamers are naturally interested in the competition between the contestants of the show, they may be even keener on competing with other TV viewers through the 2nd screen or on their own gaming habits that they have generated around the program. When competing with other TV viewers, it is important to see one's own points and placement in the 2nd screen application, for example. Good rewards or recognition through the application may also motivate them.

As the VoF was broadcasted on Friday nights, it was also a time for starting a weekend with the family or getting together with friends. In these contexts, a well-designed 2nd screen application can give an extra spice for watching the show together and create playfulness in the audience. For example, one group of four friends organized 'VoF parties' to watch the show, eat and chat together. Each one of them used her own iPhone to guess which contestants can continue in the show. The answers were first hidden and then revealed simultaneously. Although the 2nd screen application did not actually support this kind of competition, it encouraged them to play together.

Some home gamers wished for a 2nd screen solution which would truly support them in competing with each other – not only competing with other TV viewers. In some families or friend groups, the gaming habits could reach even a bit more serious level. For example, in one family everyone's guesses were recorded to an excel sheet to calculate the points and the winner was announced after the show with excitement. In another family, children had their own competition: they guessed which contestants continue and bet on their last candies. In guessing competitions the 2nd screen and social media sometimes served as a way to receive extra information – publicly or secretly - for supporting one's guess, but it could be designed to support these playful activities in a much more focused and engaging ways.

The following observational note and interview quotation demonstrate how a 2nd screen application can enable playfulness even if the person is watching TV alone:

 ISBN: 978-1-61208-353-7

> Heli is really focused on watching the the VoF in her favorite armchair. She uses the HomeCoach application actively during the program and rates each competitor as quickly as possible at the beginning of each song. Right after the show, she checks her points and placement in the weekly competition like she does every week: *"It brings more excitement, a new twist when you can't wait to see if you guessed right. I actually won one weekly competition. I got some kind of widget but I sold it. But it feels great to be right."*
>
> (Observational note and interview quotation, Heli 32 years)

D. *Active Follower*

For an Active Follower, the primary role of the 2nd screen is to follow and walk along the journey of certain personas in the program. In the case of the VoF, these personas were competitors, coaches and/or hosts. The 2nd screen becomes meaningful to Active Followers through showing support to one's favorites and becoming an active fan. It provides an opportunity to get closer to the people in the program than by watching traditional linear broadcast without a 2nd screen.

Through 2nd screen usage, Active Followers look for additional information about the facts, backgrounds, learning, history and career development of their favorite personas. They are looking for an insider feeling, a feeling of being a part of the program and the lives of people in it. The interest in personas is not limited to the broadcast or the program season: they are also interested in getting updates about the life and career of their favorites after or outside the program. They follow contestants through different media, both traditional and digital, during the whole program season. TV broadcasts are already enriched by magazines, tabloids and websites, and the 2nd screen adds an extra spice to that.

Active Followers want the possibility to participate in live broadcast, discuss and cheer their favorite performers. Social media is an important part of 2nd screen usage through the willingness to express and share the support of persons with other viewers or friends. This need can be fulfilled by using Twitter or Facebook or by using program-related applications. For Active Followers, an ideal 2nd screen application would be one that gives an opportunity to follow your favorites during and outside the broadcasting time.

The following interview quotations illustrate the central role of following certain persons in the program and how it is reflected in the use of the 2nd screen:

> *"I think this year they have managed to promote the career of all artists slightly better than last year. Their personality has been shown and I have seen that they have got gigs." … "I did not use the application anymore during the final week because my favorite performers were not in the competition anymore" … "I tried sharing the HomeCoach results in Twitter, but I was disappointed that it did not share who I was cheering."*
>
> (Interview quotations, Anna 34 years)

V. DISCUSSION

Table 2 summarizes the key differences between the ideal user types that stem from the data. The role of the 2nd screen for each type is further elaborated by describing the key contextual aspects and motivational factors. Motivations for 2nd screen usage are categorized by following the well-established research stream of uses and gratifications. This research perspective emphasizes that the audience may use the same media in different ways and to meet different needs according to their own wants and contexts (cf. Katz et al. [25]). It has been widely used especially in assessing motivations to use media [26]-[29], and in different studies they are roughly categorized into motivations related to information benefits, entertainment benefits and social benefits [9]. A similar categorization can be also found in the literature of perceived value, in which categories are named after utilitarian, hedonic and social value, respectively [30]. Compared to the majority of earlier research in 2nd screen usage [10]-[13], our results describing the different roles, contexts and motivations that 2nd screen have in the users' viewing experience shift the emphasis from usage intention figures and interface design to understanding user experience.

The primary role of the 2nd screen in the viewing experience is different for different viewer types. Our results include all three activities from the taxonomy of Cesar et al. [7], namely content enrichment, content sharing and content control, but give more detailed description of different role these activities can take in viewing experiences. For Commentators, the 2nd screen supports first and foremost the social aspects and commenting. It can be used together in order to strengthen the social ties in the living room or alone in order to avoid the feeling of watching alone. Key contextual aspects include other people and the role of different 2nd screens. Analyzers, instead, use the 2nd screen to support the analysis and anticipation of results. In their context the program content, directing the subject of analysis, has a more central role in 2nd screen usage. In addition to reality shows, detective stories, for example, could have suitable program content for this kind of 2nd screen usage. For Home Gamers, the competition and social gatherings or parties organized around the program are essential for using the 2nd screen. Finally, Active Followers concentrate on following persons in the program, and their context for the 2nd screen is influenced by multiple media channels in addition to TV.

Viewer types can be roughly compared by concluding that Commentators emphasize social motivations, Analyzers are motivated mainly by information and Home Gamers emphasize entertainment. However, all 2nd screen user types have motivations related to all three aspects: information, entertainment and social aspects.

From the information viewpoint, Commentators are interested in spotting trends and getting conversation topics from funny details in the program. The motivation of communication about impressions of a program, which was pointed out in earlier research as a strongest motivation in case of entertainment programs [9], applies especially in this user category in our study. Analyzers are more interested in accurate information and statistics from the program, while Home Gamers appreciate information related to competitions and quizzes between viewers. Active Followers want to get to know their favorite persons in the program in detail.

 ISBN: 978-1-61208-353-7

Entertainment is especially important for Home Gamers who look for playfulness and additional activity from 2nd screens. It is also characteristic of Commentators who look for humorous comments and want to keep up-to-date. Analyzers are entertained by the feeling of acquired expertise and correct analysis.

The social motivations of Commentators and Home Gamers focus more on their own social circle, typically people in their own living room. Instead, Analyzers and especially Active Followers are keener on a wider social circle formed by the program viewers. The latter types are also interested in participating in the program through a 2nd screen but the emphasis differs: Analyzers want to contribute to results, while Active Followers want to feel that they are part of the program and live broadcast.

VI. CONCLUSION

A. *Academic Implications*

From an academic point of view, we qualitatively explored a phenomenon of 2nd screen usage to address the lack of published studies. The limited prior research has focused more on the simultaneous use of social media with TV, but our research covered a broader set of 2nd screen activity including a show-specific application with synchronized content. The described viewer types and the motivations that drive their behavior in the context of media use can serve as one viewpoint for further academic debate considering 2nd screens. In addition, our findings contribute to the body of knowledge regarding the wider phenomenon of media multitasking.

B. *Practical Implications*

From a practitioner's point of view, TV broadcasters and application developers can use our results in developing 2nd screen solutions that serve better the needs of different user types. Rich, contextual data helps designers to emphasize the role of service users and gives inspiration for enhancing the user experience of designs. Understanding the different tendencies and habits that viewer types have is necessary in designing and targeting future applications – especially since 2nd screens also provide new possibilities for advertisers. Our results are applicable especially to reality TV, but to some extent also to the entertainment program genre more generally.

C. *Research Limitations and Future Research Directions*

This case study has four limitations. Firstly, it focuses on only one TV show with a limited number of participants that were considered both as followers of the show and active 2nd screen users. The insights of this study help in understanding and targeting different 2nd screen user types, but forming statistically justified design principles or business calculations would require the results to be tested with a larger sample size and quantitative analysis in the future. The future study with more participants also could reveal new viewer categories for example from the late adopters of 2nd screen applications excluded in this study. Secondly, as the program genre, for example, has been suggested to affect the use of 2nd screens [9], the viewer types could be refined and/or compared through a study that would target different genres. Thirdly, although the limitations of data collection methods – e.g. the researchers' influence in participant observation – were narrowed down in this study by using a combination of complementary methods, the results could be validated in the future through a research with different methodological choices like video ethnography. Fourthly, and since TV and media concepts are often international, cross-cultural research would be highly desirable.

ACKNOWLEDGMENT

This study received financial support from the Finnish Funding Agency for Technology and Innovation (TEKES) as part of the Next Media research programme, initiated by Digile Ltd.

TABLE II. DIFFERENT ROLES OF 2ND SCREEN IN THE VIEWING EXPERIENCE

	Commentator	**Analyzer**	**Home gamer**	**Active follower**
Role of the 2nd screen	social amusement and commenting	support for analysis and anticipation of results	tool for playful competition	following and supporting persons in the program
Key contextual factors for the 2nd screen	other people, social media, different 2nd screens	program content, other viewers' opinions	social gathering / party around the program	different media channels around the persons
Motivation: information	spotting trends, information about various aspects of the program	accurate information and statistics that support analysis	information about points and placement in the game, inside information to support one's guess	background information of personas
Motivation: entertainment	trying out new things, keeping up-to-date, humor	expertise, joy of successful evaluation	playfulness, additional activity	getting closer to persons, feeling of being part of the program
Motivation: social	experience of watching together, using the content as a topic of conversation, using the 2nd screen together	possibility to influence and contribute, curiosity about the opinions of others, analyzing further	competing with others, spending time together, rewards and recognition	participating in live broadcast, feeling of community / fan group around the program

 ISBN: 978-1-61208-353-7

REFERENCES

[1] Nielsen, "State of the Media Spring 2012 Advertising & Audiences, Part 2: By Demographic", April 27, 2012. http://www.nielsen.com/content/dam/corporate/us/en/newswire/uploads/2012/04/nielsen-advertising-and-audiences-spring-2012.pdf

[2] Google, "The New Multi-screen World: Understanding Cross-platform User Behavior," August 2012. http://services.google.com/fh/files/misc/multiscreenworld_final.pdf

[3] S.-H. Jeong, and M. Fishbein. "Predictors of multitasking with media: Media factors and audience factors", Media Psychology, vol. 10, issue 3, pp. 364-384, 2007.

[4] F. Bardhi, A.J. Rohm, and F. Sultan, "Tuning in and tuning out: media multitasking among young users". Journal of User Behaviour, vol. 9, issue 4, pp. 316-332, July/August 2010.

[5] J.J. Pilotta, D.E. Schultz, G. Drenik, and R. Philip, R. "Simultaneous media usage: A Critical User Orientation to Media Planning". Journal of User Behaviour, vol. 3, issue 3, pp. 285-292, 2004.

[6] S.A. Brasel, and J. Gips, "Media multitasking behaviour: concurrent television and computer usage", Cyberpsychology, Behaviour, and Social Networking 14 (9), pp. 527-534, 2011.

[7] P. Cesar, D.C.A. Bulterman, and J. Jansen, "Leveraging user impact: an architecture for secondary screens usage in interactive television", Multimedia Systems, vol. 15, issue 3, pp. 127-142. 2009.

[8] V. Oksman, M. Ainasoja, J. Linna, M. Alaoja, P. Heikkilä and K. Alijoki K., "2nd screen usage while watching TV: An ethnographic study", TIVIT Next Media Deliverables, 2013. (http://www.nextmedia.fi)

[9] E. Han, and S.-W. Lee, "Motivations for the complementary use of text-based media during linear TV viewing: An exploratory study", Computers in Human Behavior vol. 32 (March), pp. 235-243, 2014.

[10] E. Tsekleves, R. Whitman, K. Kondo, and A. Hill, "Investigating media use and the television user experience in the home". Entertainment Computing, vol. 2, issue 3, pp. 151-161, 2011.

[11] S. Basabur, H. Mandalia, S. Chaysinh, Y. Lee, N. Venkitaraman, and C. Metcalf, "FANFEEDS: Evaluation of socially generated information feed on second screen as a TV show companion", In Proceedings of EuroITV'12, July 4-6 2012, Berlin, Germany, pp. 87-96, 2012.

[12] C. Courtois, and E. D'Heer, E. "Second screen applications and tablet users: Constellation, awareness, experience, and interest", Proceedings of EuroITV'12, July 4-6 2012, Berlin, Germany, pp. 153-156, 2012.

[13] L. Cruickshank, E. Tsekleves, R. Whitham, A. Hill, K. Kondo, "Making interactive TV easier to use: Interface design for a second screen approach", The Design Journal, vol. 10, issue 3, pp. 41-53, 2012.

[14] G. Harboe, N. Massey, C. Metcalf, D. Wheatley, and G. Romano, "The uses of social television". ACM Computers in Entertainment, vol. 6, issue 1, 2008.

[15] D. Geerts, P. Cesar, and D. Bulterman, "The implications of program genres for the design of social television systems", In Proceedings of the international conference on designing interactive user experiences for TV and video, pp. 71-80, 2008.

[16] R. Hawkins, S. Pingree, J. Hitchon, "What produces television attention and attention style? Genre, situation and individual differences as predictors", Human Communications Research, vol. 31, pp. 162-167, 2005.

[17] K.M. Eisenhardt, "Building theories from case study research," Academy of Management Review 14 (4), pp. 532-550, 1989.

[18] K.M. Eisenhardt and M.E. Graebner, "Theory building from cases: Opportunities and challenges", Academy of Management Journal, 50 (1), pp. 25–32, 2007.

[19] H.R. Beyer, and K. Holtzblatt, "Apprenticing with the customer," Communications of the ACM, vol. 38, issue 5, pp. 45-52, May 1995.

[20] L. Jayasinghe, and M. Ritson, "Everyday advertising context: An ethnography of advertising response in the family living room", Journal of User Research, vol. 40, issue 1, pp. 104-121, 2013.

[21] E.M. Rogers, Diffusion of Innovations, 4th ed, The Free Press, New York, 1995.

[22] G. Guest, E.E. Namey, and M.L. Mitchell, "Participant observation," in Collecting Qualitative Data: A Field Manual for Applied Research. SAGE Publications, Inc., Thousand Oaks, pp. 75-112, 2013.

[23] V. Braun, and V. Clarke, "Using thematic analysis in psychology", Qualitative Research in Psychology, vol. 3, issue 2, pp. 77-101, 2006.

[24] Ideal type. (n.d.). Encyclopedia Britannica, Inc.. Retrieved June 25, 2014, from Dictionary.com website: http://dictionary.reference.com/browse/ideal type

[25] E. Katz, J.G. Blumler, and M. Gurevitsch, Utilization of mass communication by the individual. In: Blumler, J.G. & Katz, E. (Eds.). The Uses of Mass Communications: Current Perspectives on Gratifications Research. Beverly Hills, Sage, pp.19–32, 1974.

[26] C.R. Bantz, "Exploring uses and gratifications: A comparison of reported uses of television and reported uses of favourite program type", Communication Research, vol. 9, issue 3, pp. 352-379, 1982.

[27] J.C. Conway, and A.M. Rubin, "Psychological predictors of television viewing motivations", Communication Research, vol. 18, issue 4, pp. 443-463, 1991.

[28] T.F. Stafford, M.R. Stafford, and L.L. Schkade, "Determining uses and gratifications for the internet", Decision Sciences, vol. 35, issue 2, pp. 259–288, 2004.

[29] A. Quan-Haase, and A.L. Young, "Uses and gratifications of social media: A comparison of Facebook and instant messaging", Bulletin of Science, Technology & Society, vol. 30, issue 5, pp. 350-361, 2010.

[30] B.J. Babin, W.R. Darden, and M. Griffin, "Work and/or fun: measuring hedonic and utilitarian shopping value", Journal of User Research, vol. 20, issue 4, pp. 644–656, 1994.

 ISBN: 978-1-61208-353-7

A Study on the Ka-band Satellite 4K-UHD Broadcasting Service Provisioning in Korea

Min-Su Shin, Joon-Gyu Ryu, Deock-Gil Oh

Dept. Satellite Wireless Convergence
Electronics and Telecommunications Research Institute
Daejeon, S.Korea 305-700
Email: {msshin, jgryurt, dgoh}@etri.re.kr

Yong-Goo Kim

Dept. Media Technology
Korean German Institute of Technology
Seoul, S.Korea 121-913
Email: ygkim@kgit.ac.kr

***Abstract*—Advances in ubiquitous system and service are recently made thanks to the fruition of hardware and software developments so far achieved. The key elements of the ubiquitous services might be the omnipresence of service equipment and the distribution network to provide information to users everywhere. In this sense, satellite network could be one of the most suitable candidates for ubiquitous services in mobile and fixed environment for bi-directional communication and wideband broadcasting network system, which could be the optimal choice for high quality educational information services. This paper analyzes the feasibility of new satellite UHD broadcasting service scenarios using the Ka frequency band while increasing the service availability in Korea. Some countries have started their service trials, and plan to launch commercial broadcasting through a satellite link. For these services, diverse service scenarios should be evaluated to identify the most efficient way to provide the target service on schedule. For this purpose, a rain attenuation analysis was conducted to recognize the amount of expected attenuation in the Korean territory, and the results were applied to the design of the DVB-S2 satellite link. The service scenarios were then analyzed from a variety of aspects considering as many technologies as possible that are expected to be available in the near future. Some of these service scenarios were evaluated for their service availability within the Korean territory through live experiments, the results of which showed that satellite UHDTV service in the Ka band is possible if the proper technologies are selected. This study will be helpful for determining the most reasonable way for other countries preparing similar services at the initial stage, and can contribute to a stable provisioning of UHDTV services in the consumer market.**

***Index Terms*—Satellite UHD Broadcasting; Immersive Broadcasting; Ka-band; Channel Adaptive Broadcasting.**

I. Introduction

Since its inception with a 24 hours-a-day single channel service in 2003, digital HDTV satellite broadcasting service in Korea has developed into a large market with more than 100 HDTV channels and about four-million subscribers. Currently, digital satellite broadcasting services are provided through the Ku-band transponder of the KoreaSAT-6 satellite, which was launched in December 2010, and they will be extended for next-generation broadcasting services, such as stereoscopic 3-Dimentional TV (3DTV) and Ultra HDTV (UHDTV). These new broadcasting services will demand frequency capacity more than the current saturated Ku-band, and to resolve such limitations, the 21.4 to 22.0 GHz frequency band was allocated for Broadcasting Satellite Service (BSS) in regions 1 and 3 at the World Administrative Radio Conference-92 (WARC-92) for implementation after April 1, 2007. According to this allocation, many countries in these regions have been competitively requesting frequency registration for this frequency band, and the number of registrations has increased significantly to up to 700 satellite networks. However, since propagation attenuations in this badn may place a heavy restriction on the service availability and system feasibility, mitigation techniques have been studied from diverse perspectives [1]. Many advanced countries have developed various Ultra High Definition (UHD) broadcasting technologies, opening a new horizon for the possibility of commercial UHD broadcasting service. Since it began its R&D activities for UHD broadcasting service in 1995, Japan has established a new concept for next-generation broadcasting service that fully satisfies the human perception capacity in a visual and auditory sense [2]. With overall research results in the every part of broadcasting chain [3], the Japanese government announced the launch of trial broadcasting at 4K resolution for mid-2014 and the start of test broadcasting at 8K resolution for 2016. In the case of Korea, the pay TV operators conducted experiments for UHD broadcasting service and started their trial service from the first quarther of 2014. For this end, satellite transmission tests and terrestrial broadcasting tests for 4K UHD broadcasting service were successfully conducted in 2013. During the initial stage of such service, satellites are expected to be the major medium because of its flexibility and adaptability for new services. However, since the propagation attenuation in the Ka band, particularly from rain, poses a significant challenge in service availability, it is needed to investigate the pertinence and possibility of commercial broadcasting services using this band for immersive media including UHD video. Moreover, since the weather in Korea has been changing to an increase in rainfall in recent years, worsening the conditions for satellite broadcasting [4][5], reasonable service scenarios should be carefully taken into consideration. For this purpose, a simple analysis was conducted to identify what kinds of services are possible [6]. In this paper, an analysis of previous rain attenuation models is conducted to confirm their results, and a

 ISBN: 978-1-61208-353-7

link analysis is extended to consider more detailed parameters. Through these analyses, much more diverse service scenarios are envisaged. The remainder of this paper is organized as follows. A rain attenuation analysis for UHD broadcasting using a satellite at the Ka frequency band, to identify how much rain attenuation has to be expected using the regional rain distribution of Korea, is presented in Section II. In addition, an analysis of the link margin for each code rate and the modulation method of DVB-S2 at a certain rain rate are presented to evaluate the service availability in Section III. Next, channel-adaptive UHD satellite broadcasting scenarios based on the combinations of various technical elements are provided in Section IV, and the experimental progress for UHD satellite broadcasting service conducted in Korea is described in Section V. Finally, in Section VI, some concluding remarks are offered.

II. Rain Attenuation Model Analysis in Ka-band

The quality of microwave signals propagating through the atmosphere over the satellite links is affected by complex contributions, such as the absorption and scattering caused by atmospheric gases and water droplets in the precipitation. Among these contributions, attenuation by atmospheric gases may normally be neglected when it comes to Ka band satellite communications. On the other hand, attenuation by precipitation significantly degrades the performance of the transmission link and varies greatly depending on the geographic location and climate [7][8]. It is therefore essential to precisely predict the attenuation by rain on the propagation links for the proper planning of satellite systems and the evaluation of the feasibility for commercial UHD broadcasting services over the Ka frequency band in Korea.

A. Standard rain attenuation model

The rain attenuation can be estimated using the model proposed by Olsen [9], where the rain attenuation for a satellite system can be predicted based on the specific attenuation calculated theoretically according to the scattering properties of hydrometeors, and the effective path length estimated using local rainfall data. The Olsen model and more recent studies for the different microphysical properties of hydrometeors were adopted into the international standards to predict the specific attenuation as a function of the rain rate and target frequency [10]. ITU-R adopted the rain attenuation prediction model for a frequency range of 1 to 400 GHz using the rainfall intensity distribution model for a spherical raindrop shape, and ITU-R P.838 recommended a model for calculating the attenuation from rain based on knowledge of the rain rates, and was revised twice to adopt the type of polarization and the scattering properties for non-spherical raindrops [11]. The estimation model of rain attenuation [12] currently used as the international standard is the DAH model [13]. This model uses an empirical approach for estimating the effective path length, and additionally includes a vertical adjustment factor to consider the combined effect of several propagation impairments. The rainfall rate exceeding 0.01% of an average year with an integration time of 1-min, i.e., $R_{0.01}$, is desirable to take local measurements for an accurate estimation of the rain attenuation whenever possible. Otherwise, an estimate can be used from the Recommendation ITU-R P.837-6. We found that the characteristics of precipitation recommended by the ITU-R model does not correctly represent the local precipitation intensity in Korea these days, and therefore decided to use local measurement data for calculating $R_{0.01}$, based on a conversion of the integration time from 20-min to 1-min. The local rainfall rate will be discussed again in Section II-B. Next, it is necessary to obtain the specific attenuation, γ, at a rainfall rate exceeding 0.01% of an average year, which can be calculated based on the frequency-dependent coefficients given in the recommendation [10] and $R_{0.01}$ determined in the previous step. The ITU-R model adopts adjustment factors in the horizontal and vertical directions to consider the relation between the path length affected by rain and the diameter of rain cell to estimate the effective path length, L_{eff}. Finally, the rain attenuation for 0.01% of a one-year time period, $A_{0.01}$, as well as for the other percentages of an average year, can be predicted.

B. Rainfall rates in Korea

It is necessary to use rainfall statistics that are as accurate as possible in the calculation of the specific attenuation to obtain the proper rain attenuation value for a certain area under consideration. To this end, ITU-R classified Korea and Japan as rain climate zone K, and recommended using 50.6 mm/h for the rainfall rate exceeded for 0.01% of the average year [14], which is an important parameter for predicting rain attenuation. However, the recommended rainfall rate has turned out to be quite different from the measurement data within the Korea territory.

Table I shows the rainfall rate statistics in Korea, which were measured over a ten-year period in several regional areas with a 20-min integration time, and then converted into a 1-min integration time [5]. It has been reported that the climate of Korea is becoming more like subtropical weather in terms of temperature changes, and that the mean rainfall rate for the last ten years has increased by 9.1% compared to the previous 30

TABLE I. Rainfall Rate Distribution in Korea

Time percentage [%]	Rainfall rate [%]				
	Measurements in Korea			ITU-R P.837-1	ITU-R P.837-6
	Seoul	Busan	Daejeon		
0.01	66.4	66.5	62.9	42.0	50.67
0.02	53.6	49.5	50.3	-	37.21
0.03	45.4	40.1	37.5	23.0	30.11
0.05	35.5	31.1	30.0	-	22.33
0.10	23.5	21.4	21.3	12.0	14.23
0.20	14.1	14.7	12.8	-	8.87
0.30	10.5	10.5	8.8	4.2	6.71
0.50	6.5	6.9	5.5	-	4.69
1.00	3.2	3.8	2.6	1.5	2.79

 ISBN: 978-1-61208-353-7

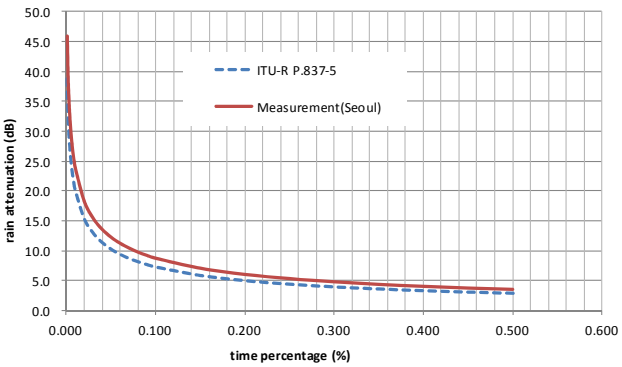


Fig. 1. Predicted rain attenuation distribution in Korea

years [4]. The difference in rainfall rate between ITU-R models and the measurements listed in Table I indicates that the rain rates obtained using the ITU-R P.837-6 model do not fully reflect these changes in Koreas local rainfall characteristics. We therefore use the measured rainfall rates to predict the various propagation parameters in this paper since the ITU-R model may give erroneous rain attenuation values on the radio links in Korea. For calculating the rain attenuation for UHD satellite broadcasting over the Ka band in Korea, the center frequency of channel 3 of the Chollian satellite was used for the operating frequency, and vertical polarization was therefore selected. To take into account domestic rain distribution trends, we take 66.4 mm/h as the rainfall rate for $R_{0.01}$, which is the measured value for Seoul with location data of 126.58E and 37.33N. With these values, the estimated attenuation from rain for the regional use of the Ka-band satellite system is obtained, and the results compared with those from the recommended rainfall rate in ITU-R P.837-5, as shown in Fig. 1. The difference between the two results becomes larger as the rainfall rate increases reaching more than 3.5 dB at 0.01% time rate, which means that the ITU-R model does not fully reflect domestic weather changes, and regional statistics should therefore be used for a better quality service provision.

III. Link Analysis for Ka-band Satellite System

The second version of Digital Video Broadcasting via Satellite (DVB-S2) standard is one of the most popular technologies in satellite broadcasting. We take the DVB-S2 technology into consideration for the transmission system of a UHD satellite broadcasting service since the system has been adopted for Ku band satellite broadcasting in Korea, and this is therefore the best way to launch a new service minimizing the burden of risk in terms of investment and service compatibility. In this section, we describe DVB-S2 system performance characteristics, which was conducted by finding the C/N required to achieve a packet error rate (PER) of less than 10^{-7} for the TS packet for each coding and modulation technique of DVB-S2. The link margin of the DVB-S2 system under the link conditions addressed in Section II is then estimated.

A. DVB-S2 system performance

The DVB-S2 standard has been specified to meet the demands for the best transmission performance, total flexibility, and reasonable receiver complexity. To achieve the best performance-complexity tradeoff, DVB-S2 benefits from more recent developments in channel coding and modulation. Moreover, the system was intended to be used for several applications including interactive point-to-point applications and professional applications, such as IP unicasting, digital TV, news gathering, and data content distribution. The unique features making these diverse applications achievable is the adoption of adaptive coding and modulation (ACM) functionality, which allows an optimization of the transmission parameters for each individual user on a frame-by-frame basis depending on the path conditions under closed-loop control through a return channel. To keep the packet error rate at less than 10^{-7} over an AWGN channel, it is known that the required C/N of the DVB-S2 system varies from -2.4 dB with QPSK 1/4 to 16 dB with 32APSK 9/10. However, since the effects of nonlinearity and synchronization loss and phase noise should be considered along with the ideal performance [15], we take 0.6/1.0/2.0/4.0 dB as an additional power losses for QPSK/8PSK/16APSK/32APSK, as reasonable estimates with typical equipment characteristics, respectively. Therefore, the C/N values required to meet 10^{-7} of PER over an AWGN channel for each modulation and coding (MODCOD) parameter of the DVB-S2 system are calculated as shown in Table II.

Table II can be used to select feasible transmission schemes for a certain service availability by comparing the required signal level with the link performance as in Table III. Moreover, this can be used to design service composition meaing how many channels and what kind of channels could be transmitted within the dedicated bandwidth because total data rate is simply calculated when the transmission schemes used are determined.

B. Link analysis of Ka band satellite broadcasting system

To recognize the possible transmittion configuration for 4K UHD satellite broadcasting over the Ka band, it is necessary

TABLE II. Required C/N of DVB-S2 system

Code rate	QPSK	8PSK	16APSK	32APSK
1/4	-2.59	-	-	-
1/3	-1.39	-	-	-
2/5	-0.19	-	-	-
1/2	0.81	-	-	-
3/5	2.11	5.71	-	-
2/3	2.91	6.81	10.21	-
3/4	3.81	8.11	11.41	15.91
4/5	4.51	-	12.21	16.81
5/6	5.01	9.61	12.81	17.61
8/9	6.11	10.91	14.11	18.91
9/10	6.61	11.21	14.31	19.31

 ISBN: 978-1-61208-353-7

to evaluate the link margin of the Chollian satellite and the available bitrates that can be transmitted through its corresponding bandwidth for each MODCOD parameter of the DVB-S2 system under a certain rain attenuation condition. The link margins of the Chollian satellite downlink, which can be used to define the possible transmission method to be used for a certain rainfall rate, are given in Table III. For a more realistic analysis, the required C/N contains non-linear power loss under the assumption of using a pre-distortion technique at an earth station, as discussed in Section III-A.

TABLE III. Down-link Performance of Chollian Satellite

D/L (20.13GHz)	Clear	Rain (0.07%)	Rain (0.1%)	Rain (0.3%)	Rain (0.5%)
Saturated EIRP	60	60	60	60	60
Free Space Loss	209.9	209.9	209.9	209.9	209.9
Rain Attenuation	0	10.44	8.73	4.78	3.5
ES G/T	15.51	13.30	13.38	13.79	14.04
Channel Bandwidth	80	80	80	80	80
C/Nd	13.2	0.55	2.35	6.7	8.23
D/L C/N[dB]	13.2	0.55	2.35	6.70	8.23

*Above listed figures are assumed that antenna diameter = 45cm(38dBi), antenna Noise Temperature = 92K, Noise Bandwidth = 100MHz and all the carrier to interference ratios (C/I) = 60 dB.

For the simple design of the uplink from the transmitter station to satellite, it is assumed that no rain loss is considered thanks to the perfect uplink power control, and the total transponder bandwidth is occupied with the minimum input and output backoff. The earth station antenna gain is calculated under the assumption that a 70 cm antenna diameter would be proper for the Ka band satellite link. As shown in Table III, it turns out that the link margin for clear sky conditions is around 13 dB, and the broadcasting link is unable to maintain its connection with the receivers at rain conditions of higher than 0.07% of the time percentage because the only possible tramission schemes are lower than QPSK 1/2 code rate which could be usable but very impractical choices.

IV. Ka-Band UHD Satellite Broadcasting Scenarios

Based on the above link analysis of the DVB-S2 satellite UHD broadcasting system using the Chollian satellite, several satellite UHD broadcasting scenarios are considered, and their service performances are analyzed. As candidate technologies to be used for the service scenario establishment, diverse technologies available at present or expected to be available in the near future are taken into account. Special regard is paid to the backward compatibility with the current HDTV services and the adaptability to channel variation for extending service availability. Channel adaptability is a technology being introduced to mitigate rain attenuation, which is a critical issue in Ka band satellite services. The related technologies in terms of transmission are being widely developed in many countries [1]. To discuss the backward compatibility, it is necessary to identify the legacy HDTV receiver and new UHDTV receiver in terms of the element technologies. Legacy HDTV receivers are defined to have the capability to deal with H.264/AVC HP@4.1 for the video codec, and DVB-S2 constant coding and modulation (CCM) mode for the transmission. On the other hand, the new UHDTV receivers have the capability to support H.264/AVC with higher than HP@5.1 and HEVC for the video codec, and DVB-S2 variable coding and modulation (VCM) mode for the transmission. Therefore, new UHDTV receivers are able to receive the frame resolution of UHD video as well as HD video and a satellite signal with multiple protection levels, which makes it possible to design much more flexible service scenarios. The service scenarios for satellite UHD broadcasting are categorized according to the perspective on the video codec, transmission mode and frequency band. In the first step, service scenarios are classified into single layered and multiple layered services according to whether the service is composed of multiple layers in terms of video quality. Single-layered service is the simplest scenario and can be applied for dedicated channel service, which only covers new subscribers for UHDTV services. However, in multiple-layered service scenarios, several combinations of technologies are considered to find a way to provide backward compatibility and mitigate the channel deterioration from rain attenuation. To meet these requirements, scalable video coding (SVC) and simulcasting schemes are considered in the scenario as well.

A. Single layered scenarios

Single layered scenarios only aim to provide UHD broadcasting service itself, and thus do not consider subscribers with legacy HDTV receivers. Since they are not configured to have multiple layers in the UHD program, a single-frequency band and DVB-S2 CCM mode are used in the scenarios. Therefore, these scenarios do not support backward compatibility because they require video coding technology higher than H.264/AVC HP@5.1 profile to deal with the 4K frame resolution of the video. These scenarios include two types of scenarios, which are classified according to which video codec is used for the UHD program compression. The single layered scenario system is expected to be the most appropriate system for the initial UHD satellite test broadcasting because it does not affect the current on-air services if the frequency band is properly selected.

B. Multiple layered scenarios

In multiple layered scenarios, each program is transmitted at different layers with different protection levels to provide adaptive service in a channel variation environment. To this end, the base layer stream and enhancement layer stream are constructed from one program. For the sake of convenience, in this paper a low-quality video stream is called a base layer stream, and a high-quality video stream is called an enhancement layer stream. The base layer stream and enhancement layer stream can be generated by either way of SVC or simulcasting scheme. In addition, each layered stream can be transmitted in diverse ways depending on the transmission mode and target frequency band.

In the first step, multiple layered scenarios are classified depending on their backward compatibility possibility. To

 ISBN: 978-1-61208-353-7

TABLE IV. Classification of Multiple Layered Service Scenarios

	Multiple layered Service Scenarios							
	Non-Backward Compatibility Supported				Backward Compatibility Supported			
	SVC		Simulcasting		SVC		Simulcasting	
	Single Codec	Multi Codec	Single Codec	Multi Codec	Single Codec	Multi Codec	Single Codec	Multi Codec
Single Band	[B]HEVC,Ka [E]HEVC,Ka [T]VCM	[B]AVC,Ka [E]HEVC,Ka [T]VCM	[B]AVC,Ka [E]AVC,Ka [T]VCM [B]HEVC,Ka [E]HEVC,Ka [T]VCM	[B]AVC,Ka [E]HEVC,Ka [T]VCM	N/A	N/A	N/A	N/A
Multiple Band	[B]HEVC,Ku [E]HEVC,Ka [T]CCM	N/A	[B]HEVC,Ku [E]HEVC,Ka [T]CCM	N/A	N/A	[B]AVC,Ku [E]HEVC,Ka [T]CCM	[B]AVC,Ku [E]AVC,Ka [T]CCM	[B]AVC,Ku [E]HEVC,Ka [T]CCM

*[B]: Base layer of UHD service, which means it is for an HD program, [E]: Enhancement layer of the UHD service, which means it is used for reproducing a UHD program, and could be an enhancement layer for SVC coding and the UHD program signal itself in a simulcasting scheme, and [T]: Transmission mode of DVB-S2 technology, where N/A denotes that the service scenario is not able to meet the corresponding conditions.

*It is assumed that SVC coding with H.264 would not support UHD resolution, and thus the enhanced layer of SVC in the scenarios is generated by only HEVC video coding. It is also assumed that the SVC with HEVC supports both H.264 and HEVC for its base layer stream.

support backward compatibility to the current HDTV service in Korea, the scenario should involve H.264/AVC and DVB-S2 CCM mode for the base layer stream. Providing backward compatibility with DVB-S2 CCM mode in a single band is not considered in the paper because of its inefficiency. Therefore, there are three scenarios that can provide backward compatibility, as shown in Table IV. The base layer stream used for HDTV service should be coded with the H.264/AVC scheme of HP@4.1, and transmitted through the Ku band frequency. In addition, the enhancement layer stream is transmitted through the Ka band frequency using either H.264/AVC of HP@5.1 or the HEVC coding schemes. Since these service scenarios use dual-band transmission, each layer stream is transmitted with DVB-S2 CCM mode, and the new UHDTV receiver requires to have Ku/Ka dual-band signal reception capability when the SVC scheme is considered. As for the non-backward compatibility scenarios, the HEVC coding scheme for the base layer stream and DVB-S2 VCM mode can be considered for the scenario analysis. Both single-band transmission and multiple-band transmission can be possible because they can utilize DVB-S2 VCM mode to transmit each layer stream with differentiating its protection level even in single-band transmission. For the single-band transmission scenarios, every combination of technology is technically possible, while multiple codec scenarios for multiple band transmission are not available because they are involved in the backward compatibility scenario. As shown in Table IV, it is assumed that only HEVC SVC usage is considered because SVC in H.264 is hard to be practically applied. H.264 SVC is currently not supported by most commercial products, and HEVC SVC technology will be used later if such functionality is thought to be needed. However, since HEVC SVC technology is under standardization, it will take time to appear on the market, and more importantly, it is not clear at the moment whether the business market will demand scalable video coding applications. Therefore, simulcasting scenarios will be more reasonable for satellite broadcasting in the near future if channel adaptability is required for the target service, even though it is necessary to submit to a sacrifice in bandwidth.

V. Ka-Band 4K-UHD Satellite Experimental Broadcasting Demonstration

4K UHD satellite experimental broadcasting services through the Ka band satellite have been conducted in past years. In this experiments, only two service scenarios with single band and single codecs with one of AVC and HEVC schemes are applied, as shown in Fig. 2. The experiments were conducted to evaluate the multi-channel service provisioning with channel adaptibility for 4K UHD satellite broadcasting. To meet the requirement, a high-speed DVB-S2 modem was developed using VCM mode support. Most of the legacy DVB-S2 modems are developed to support Ku band transponders of around 30MHz providing 80Mbps of maximum capacity. Our new modem extends this capacity upto around 300Mbps to support wideband Ka band transponders, which are required to provide multiple UHDTV channel services. In addition, this modem supports all the MODCOD listed in the DVB-S2 standard with multiple TS interfaces functionality for VCM transmission, and implemented with 0.3 dB margin in average comparing to the ideal required C/N of each MODCOD. For the channel adaptive service verification, several combinations of two MODCODs are selected that each combination can serve more than 99.7% of their service availibilty.

According to the link budget analysis in Section III and the experiment results, it turned out that this channel adaptive

 ISBN: 978-1-61208-353-7

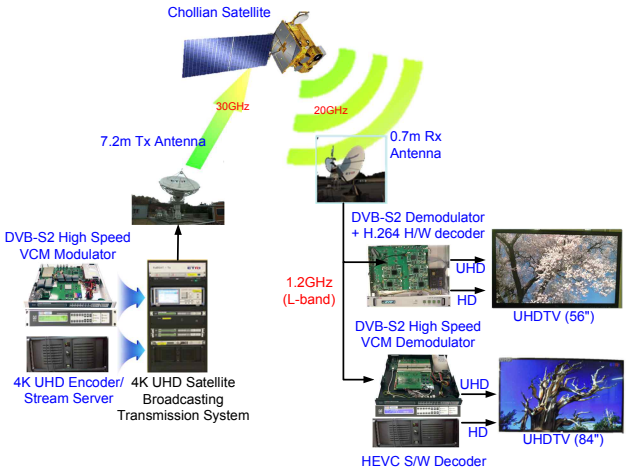


Fig. 2. Ka-band Satellite UHD Broadcasting Demonstration

functionality can be able to increase the service availability by more than 7 hours per year. It is clear that the channel adaptability is worthy of attention, especially in regions of heavy rainfall. Moreover, it would be much more important in applications in which the link should always be connected between earth stations, such as for public protection and disaster relief.

VI. Conclusion

This paper presented a variety of analyses on Ka band satellite broadcasting service for immersive media, providing the technological background for a verification of satellite UHD service commercialization. Rain attenuation modeling was performed first through the international standard method with the domestic rainfall rate statistics, since the rain attenuation is the most critical factor for satellite broadcasting in the Ka band. Next, the satellite link performance was analyzed to determine the most suitable transmission method for the service under the local environment, and it turned out that the channel adaptive functionality should be seriously considered for more stable service continuation. The paper presents diverse feasible service scenarios with the combinations of technologies which are currently available and expected to be available in the near future as well. Some of these combinations are open to discuss in terms of economics. However, most of them are worth taking into consideration since it is still vague about what will be the correct answer in every element for UHD service. Since technology developments will make rapid progress, it will reduce the cost and make it possible to bring about much more complicated scenarios in reality. Along with the service scenarios, the result of satellite experiments which were conducted for the preparation of satellite UHD broadcasting service in Korea was presented. Thanks to these gradual developments of service technologies, plans for satellite UHD broadcasting service will materialize in the very near future as announced by certain countries.

Acknowledgment

This work was supported by ICT R&D program of MSIP/IITP. [14-000-01-001, Development of Adaptive Satellite Broadcasting and Communication Transmission Technologies]

References

[1] Mitigation techniques for rain attenuation for broadcasting-satellite service systems in frequency bands between 17.3 GHz and 42.5 GHz, Rec. ITU-R BO.1659-1, Jan. 2012.
[2] E. Nakasu, "Super Hi-Vision on the Horizon: A Future TV System That Conveys an Enhanced Sense of Reality and Presence," IEEE Consum. Electron. Mag., vol. 1, no. 2, Apr. 2012, pp. 36–42.
[3] K. Oyamada, S. Okabe, K. Aoki, and Y. Suzuki, "Progress of Transmission Technologies for UDTV," Proc. Of the IEEE, vol. 101, no. 1, Jan. 2013, pp. 154–168.
[4] D. Choi, J. Pyun, S. Noh, and S. Lee, "Comparison of Measured Rain Attenuation in the 12.25GHz Band with Predictions by the ITU-R Model," International Journal of Antennas and Propagation, vol. 2012, Article ID 415398, 5 pages, 2012. doi:10.1155/2012/415398
[5] Regional Rain-rate Distribution for Korean Territory, TTAS.KO-06.0122, Dec. 2006.
[6] M. Shin, J. Ryu, D. Oh, and Y. Kim, "The Feasibility Study on the 4K-UHD Satellite Broadcasting Service in Ka-band," Proc. IEEE ICCE, Jan. 2013, pp. 486–487.
[7] R. K. Crane, Electromagnetic Wave Propagation Through Rain, John Wiley & Sons, 1996.
[8] A. Kumar and I. S. Hudiara, "Measurement of Rain-Induced Attenuation of Microwaves at 19.4 GHz," IEEE Antenna and Wireless Propagation Letters, vol. 1, no. 1, 2002, pp. 84–86.
[9] R. L. Olsen, D. V. Rogers, and D. B. Hodge, "The aRb relation in the calculation of rain attenuation," IEEE Trans. Antennas Propag., vol. 26, no. 2, 1978, pp. 318–329.
[10] Specific attenuation model for rain for use in prediction methods, Rec. ITU-R P.838-3, Mar. 2005.
[11] T. Oguchi, "Electromagnetic Wave Propagation and Scattering in Rain and Other Hydrometeors," Proc. of the IEEE, vol. 71, no. 9, Sep. 1983, pp. 1029–1078.
[12] Propagation data and prediction methods required for the design of Earth-space telecommunication systems, Rec. ITU-R P.618-11, Sep. 2013.
[13] A. Dissanayake, J. Allnutt, and F. Haidara, "A Prediction Model that Combines Rain Attenuation and Other Propagation Impairments Along Earth-Satellite Paths," IEEE Trans. Antennas Propag., vol. 45, no. 10, Oct. 1997, pp. 1546–1558.
[14] Characteristics of precipitation for propagation modeling, Rec. ITU-R P.837-6, Feb. 2012.
[15] E. Casini, R. De Gaudenzi, and A. Ginesi, "DVB-S2 modem algorithms design and performance over typical satellite channels," Int. J. Satell. Communi. Network., vol.22, no.3, 2004, pp. 281–381.
[16] X. F. Li, N. Zhou, and H. S. Liu, "Joint Source/Channel Coding Based on Two-Dimensional Optimization for Scalable H.264/AVC Video," ETRI Journal, vol. 33, no. 2, Apr. 2011, pp. 155–162.
[17] H. Kim, S. Lee, J. Lee, and Y. Lee, "Reducing Channel Capacity for Scalable Video Coding in a Distributed Network," ETRI Journal, vol. 32, no. 6, Dec. 2010, pp. 863–870.

 ISBN: 978-1-61208-353-7

SmartRoadSense: Collaborative Road Surface Condition Monitoring

G. Alessandroni, L. C. Klopfenstein, S. Delpriori, M. Dromedari, G. Luchetti
B. D. Paolini, A. Seraghiti, E. Lattanzi, V. Freschi, A. Carini, A. Bogliolo
DiSBeF – University of Urbino
email: alessandro.bogliolo@uniurb.it

***Abstract*—Monitoring of road surface conditions is a critical activity in transport infrastructure management. Many research solutions have been proposed in order to automatically control and check the quality of road surfaces. Most of them make use of expensive sensors embedded in vehicles or mainly focus on detection of specific anomalies during monitoring activity.In this paper, we describe the design of a system for collaborative monitoring of road surface quality. The overall architecture encompasses the integration of a custom mobile application, a georeferenced database system and a visualization front-end. Road surface condition is summarized through a roughness parameter computed using signal processing algorithms running on mobile devices. The roughness values computed are subsequently transmitted and stored into a back-end geographic information system enabling processing of aggregated traces and visualization of road conditions. The proposed approach introduces a thoroughly integrated system suitable for monitoring applications in a scalable, crowdsourcing collaborative setting.**

Keywords–Roughness; Accelerometer; Smartphone; Monitoring; Cloud.

I. INTRODUCTION

Nowadays, all consumer-level mobile devices (e.g., smartphones) feature a rich set of embedded instruments. The presence of triaxial accelerometers and Global Positioning System (GPS) sensors allow the device to track its position and motion states with high degree precision.

Additionally, mobile devices also enable the development of applications that can acquire data from such instruments. Thus, it is possible to access sensor data in real time, store it in memory, handle it using the processing power of the device itself and transmit the data to remote servers using the device's connectivity features.

These features, combined with the ubiquitous and pervasive nature of smartphones and to the inherent scalability of cloud based computing, make possible the design of systems aimed at fine-grained, massive distributed sensing.

In this paper, we propose and describe a system, called "SmartRoadSense", aimed at supporting collaborative monitoring of road surface roughness using mobile smart devices. To this purpose, we designed a three-tiered architecture encompassing: i) a mobile application at user level that processes raw data from the embedded accelerometers and transmits the result of the computation (i.e., a roughness index) together with geographic localization data from GPS to a server; ii) a back-end server running a geographic information system where georeferenced data are properly aggregated, organized and stored; iii) a graphical front-end based on a cloud platform service for visualization.

In order to use data from the accelerometer to study the condition of the road surface, we propose to use Linear Predictive Coding (LPC) [1]. LPC is a method that allows us to predict a particular value in an analog signal by means of a linear combination of the past values of the signal itself. This signal processing technique is used to compute the redundant information contained in the signal. In our case, it can be used to remove accelerations not attributable to irregularities in the road surface.

The mobile application designed in the SmartRoadSense architecture exploits LPC for deriving an estimate of the roughness of a road from sampled points. The values collected by this parameter are computed on board by a smartphone, and transmitted in batch to a remote dedicated server. The back-end server functionalities are in charge of collecting data, mapping traces on the geospatial database and consistently aggregating them for further processing and statistical analysis.

A. Previous Work

Starting in the late 1950s several studies of road surface have proved that its quality is the most important criteria for the evaluation of a road path and its drive comfort. The deterioration of roads leads to added vehicle operating costs, increased fuel consumption (with more emissions to the environment) and increased pavement failures, due to the added dynamic loads of the vehicle [2][3][4].

Several studies have tried to model the road elevation profile, using sine waves, step functions, or triangular waves [5], or as the sum of randomly generated sinusoidal functions with different amplitudes and phases [6]. More recently, it was shown that the spatial Power Spectral Density (PSD) of a typical road surface has a low-pass characteristic, which decreases at the increase of the spatial frequency (measured in cycles/m) [6][7]. In these studies, the road surface profile is modeled as a white Gaussian noise filtered by a first order low-pass filter. It was also shown, that the vertical acceleration of a point following the road profile depends on the horizontal velocity, i.e., the vertical acceleration is related to the car velocity.

A consolidated approach for estimating road surface condition entails the adoption of costly and sophisticated hardware equipment such, for instance, laser profilers [8], specific accelerometers and data acquisition systems [9] whose cost (also taking into account calibration and installation) can be significant.

Another trend of studies explored the feasibility of exploiting low-cost sensors, for instance those embedded in mobile devices such as smartphones. A first work towards this direction have been proposed by Eriksson et al. [10] that built

 ISBN: 978-1-61208-353-7

a system (termed the "Pothole Patrol") targeted at monitoring road anomalies. They used a set of accelerometers and GPS devices deployed in embedded computers in cars. The sampled signals, processed by a given set of filters to remove artifacts and noise, are given as input to machine learning algorithms for detection of potholes and road anomalies. Mohan et al. introduced "Nericell", a road and traffic monitoring system based on smartphones [11]. Sensing devices embedded in smartphones (namely microphones, accelerometer and GPS) are exploited for detecting potholes, bumps and also other traffic related events such as braking and honking.

Our work shares some features with these approaches while we believe it, differs in several aspects. First, while previous works mostly focus on given events for monitoring road quality such, for instance, pothole detection, we aimed at building a continuous monitoring system by assigning a numerical value to each of the samples of the sensed signals by means of LPC algorithms, resulting into a roughness index for potentially each point the monitored road. Second, we integrate the information gathered from several different users into a single consistent aggregated stream thus opening the way to statistical analysis and data fusion techniques for possible error compensation. Third, we take advantage of scaling capabilities provided by cloud computing facilities providing a suitable interface of our system to cloud based platforms, an example of which is given by SmartRoadSense visualization engine, therefore making it possible collaborative crowd-sensing.

B. Contribution and Organization

This paper introduces a system for measuring road quality based on low-cost sensors, a mathematical model to extract a quality index from sensor data, and a software architecture system which allows measurements to be collected and aggregated in an average estimate of road roughness.

The mathematical model upon which we developed our signal processing algorithm is described in section II. The design choices and system-level features of adopted and implemented software components are introduced in section III. Finally, preliminary experiments based on the current implementation are shown in section IV. Concluding remarks, open issues and future work are discussed in section V.

II. MATHEMATICAL MODEL

In this section, we describe the mathematical model used to extract information of the road surface conditions.

According to [6], the road surface profile $w(x)$ can be modeled as white Gaussian noise filtered by a first order low-pass filter. The white Gaussian noise has the following autocorrelation function $\rho_{ww} = q\delta(x)$, where q is the PSD magnitude and $\delta(x)$ the Dirac delta function. PSD is given by $S_{ww}(\Lambda) = q$, where Λ is the spatial frequency measured in cycles.

The first order low-pass filter has frequency response

$$H(\Lambda) = \frac{1}{p + j2\pi\Lambda}. \quad (1)$$

Thus, the PSD of the road elevation profile $S_{rr}(\Lambda)$ is given by

$$S_{rr}(\Lambda) = S_{ww}(\Lambda)\,|H(\Lambda)|^2 = q\left|\frac{1}{p + j2\pi\Lambda}\right|^2. \quad (2)$$

In this model, the statistical properties of the road profile are completely characterized by parameters q and p.

Let us consider an ideal point closely following the road profile and moving with constant horizontal velocity v. From equation (2), it can be proved that the vertical acceleration has a continuous time Fourier transform given by

$$A_y(f) = \frac{(j2\pi f)^2}{p + j2\pi f\frac{1}{v}}W(f), \quad (3)$$

and has the following temporal PSD

$$S_{A_yA_y}(f) = qv\left|\frac{(j2\pi f)^2}{pv + j2\pi f}\right|^2. \quad (4)$$

Thus, road parameters q and p of (2) can also be obtained by analyzing the PSD of the vertical acceleration.

The scenario of an accelerometer embedded in a mobile device, rigidly anchored inside the car cabin, is very different from that of an ideal point following the road profile. The accelerometer senses the road through tires, suspensions, and the mechanical coupling with the car cabin. In real applications, the PSD in (4) is sensed by the accelerometer filtered by an unknown transfer function modeling the effect of tires, suspensions, and mechanical coupling. The waveform detected by the accelerometer originated by the road profile is a noise signal with a large spectral content that depends on the road parameters q and p. The accelerometer samples the waveform at a given sample frequency F_s and outputs a discrete time vector signal composed by the triaxial components $a_x(n)$, $a_y(n)$, and $a_z(n)$, according to some internal axial reference. Figure 1 shows an example of the three components recorded by a Motorola G smartphone on a car following a straight road at 40 km/hour. The broadband noise behavior is apparent from the figure.

Other undesired contributions add to this waveform. Indeed, the accelerometer also senses the gravity acceleration, vehicle accelerations, centrifugal accelerations at curves, roll, pitch, and yaw accelerations due to road trend, and vibrations due to the engine. These contributions to the signal have a significant magnitude that can entirely mask the acceleration fluctuations due to the road profile. Nevertheless, some of these accelerations vary slowly and have a low spectral content, others have a periodic spectral content (e.g., vibrations caused by the engine). Thus, the undesired contributions can be removed with a prediction filter, that estimates the accelerometer current sample $a(n)$ (with $a(n) = a_x(n)$, $a_y(n)$, or $a_z(n)$) from past samples, i.e., with an LPC analysis [12], [13]

$$e(n) = a(n) + \sum_{i=1}^{N} \lambda_i\ a(n-i), \quad (5)$$

 ISBN: 978-1-61208-353-7

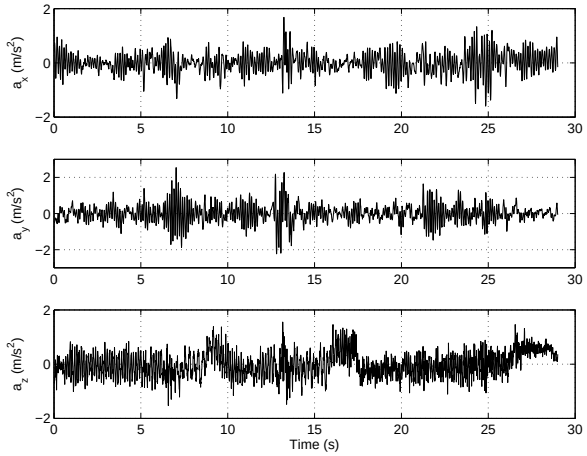


Figure 1. Triaxial accelerations components measured with a Motorola Moto G smartphone on a car at 40 km/h.

where λ_i, with $i = 1, \ldots, N$, are the LPC coefficients, N represents the prediction filter memory length, and $e(n)$ the residual prediction error.

In order to compute the prediction filter and the prediction error, a block based approach is applied: the signal $a(n)$ is split in segments of length M, with M sufficiently large to have an accurate estimate of the prediction filter and, at the same time, sufficiently small to be able to consider the signal stationary. The prediction filter is computed with the Levinson-Durbin recursion [14][15] summarized in Table I (using the Matlab notation). $R(0)$, $R(1)$, ..., $R(N)$ is the autocorrelation sequence on $a(n)$ estimated over a segment; $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_N]^T$ is the prediction filter coefficient vector.

TABLE I. PSEUDO-CODE FOR LEVINSON-DURBIN RECURSION.

```
k = R(2)/R(1);
λ = k;
E = (1 − k^2) * R(1);
for i = 2 : N
    k = (R(i + 1) − λ * R(2 : i))/E;
    λ = [k, λ − k * λ(i − 1 : −1 : 1)];
    E = (1 − k^2) * E;
end
```

The prediction error $e(n)$ maintains the information on the road parameter q (which is a proportionality parameter in the PSD) while the information on the parameter p is lost in the signal whitening produced by the prediction filter. Thus, a parameter proportional to q can be obtained estimating the power of the prediction error P_{P_E} on each segment

$$P_{P_E} = \frac{1}{M} \sum_{n=1}^{M-1} e(n)^2. \quad (6)$$

An index of the road roughness, R_I, is eventually obtained by averaging the power of the prediction error for the three axial components

$$R_I = \frac{1}{3} \left(P_{P_{E_X}} + P_{P_{E_Y}} + P_{P_{E_Z}} \right). \quad (7)$$

Figure 2 shows the behavior of the roughness index R_I and the smoothed roughness index as a function of time on a car following a straight road at 40 km/h. The smoothed roughness index has been obtained by averaging R_I with a sliding window of 11 samples.

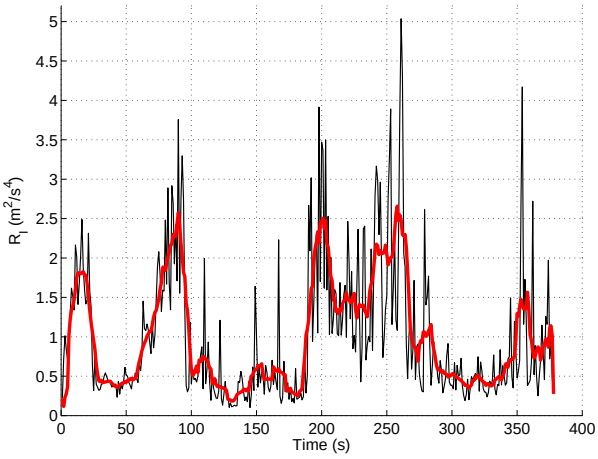


Figure 2. R_I (black curve) and smoothed roughness index (red curve) on a car at 40 km/h.

While this parameter is sensitive to the specific car, the mobile device, and guide style of the driver, the compact information contained in this single parameter can be easily collected and stored. Data from different vehicles, different devices and different drivers on the same road can be combined to achieve a meaningful metric of the road quality.

III. SYSTEM ARCHITECTURE

The mathematical process described above was provisionally implemented using Matlab. After testing on tracks of data collected using real devices, the algorithm was ported to the Java programming language.

In order to perform experiments on real data and to aggregate information collected about road conditions, the following system architecture has been devised: the Java algorithm is embedded in an Android application which runs on an Android mobile device. The application gathers accelerometer data, thus computing and storing P_{P_E} results annotated with GPS data about the device's position. Data is periodically sent to a remote server, which uses a geolocalized database to link each data point to specific roads. Results are aggregated and provided to the user as a geographical map whose roads are enriched with data about their estimated roughness.

The software architecture, graphically displayed in Figure 3, is described in more detail in the following sections.

A. Android application

The SmartRoadSense project is built around an Android Application, displaying a user-friendly interface and relying on a background service that gathers data from the device's sensor and processes it to compute P_{P_E} values in real-time. Results are stored in memory along with GPS data, bearing,

 ISBN: 978-1-61208-353-7

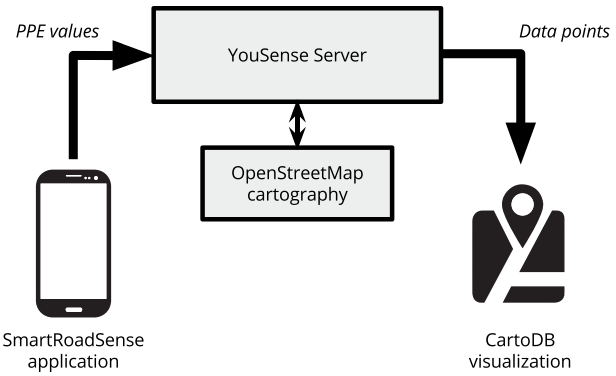


Figure 3. Software architecture of SmartRoadSense.

velocity, timestamp and other metadata. Each recording session is identified by a unique "track ID" (identifying both the device and the session).

Both sensors, triaxial accelerometers and GPS, are run at their highest possible frequency. The resulting data rate depends on the device used: usually accelerometers provide data with a frequency of 100 Hz (or higher) and GPS works at 1 Hz once fixed.

Data gathering is limited by the lower frequency (i.e., the GPS sensor), thus the application records a sample of data once per second on average. As described before, to allow real-time computation, the algorithm operates on windows of data: currently a single P_{P_E} result is extracted from a total of 100 samples, of which 25 are taken from the previous window and 25 from the next window (giving a total overlap of 50 samples). Thus, each result is computed from a total of 100 samples, extracted from an average of 100 seconds of data gathering.

The collected results are periodically transmitted to a remote server (each 15 minutes on average, when a data connection is available).

B. Data collection and aggregation

Data collection server is implemented using the ASP.NET platform on a Linux machine running Mono. A PostgreSQL database with PostGIS extensions acts as the storage back-end for all collected data.

The server application exposes a set of RESTful HTTP APIs that can be used by registered users in order to submit data. As described before, raw roughness data computed by the devices is gathered together with accessory GPS information and track ID metadata. Data points are indexed by geographical position for fast access.

A background process collects all new data points recorded and uses the new data to update information about road roughness. This process is executed periodically (at the time it runs once a day). This process works as follows: the set $P_{new} = \{P_1, P_2, \cdots, P_n\}$ of new data points registered by the server is collected; each point in P_{new} is mapped to the closest road, using a geographical database (in our implementation, we use the open road data available from OpenStreetMap). Roads are represented by a geometric path, thus the mapping of points will yield a set $R_{new} = \{R_1, R_2, \cdots, R_m\}$ of paths, representing all roads for which the database has new data points; each road in R_{new} is updated by extracting points at regular distance intervals from one end to the other (see Figure 4). Thus, each road contributes with one or more averaging points for which the roughness will be estimated; for each averaging point (shown as black dots in Figure 4) all existing data points in a given range are extracted from the database and contribute to the final average roughness value. At the moment, these values are computed as the average between all values.

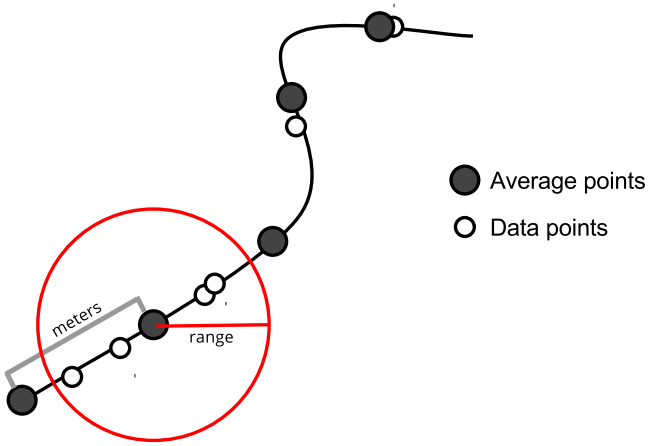


Figure 4. From single data points to average roughness points.

After this process has completed, the database has a new set of average roughness points for each road for which new raw data was sampled.

C. Data visualization

Once the average roughness data has been extracted from the raw dataset and mapped to a road, the final data is pushed to an external CartoDB server: CartoDB is an online service allowing easy visualization and handling of maps with rich data overlays. It is well suited to represent geographic maps with a great amount of data points, while allowing the user to filter and manipulate the data.

An example of the CartoDB interface (using Google Maps cartography) with the average roughness points overlay is shown in Figure 5: the road is filled with equidistant points, colored ranging from green to red. Red points mark positions where the average P_{P_E} value was relatively high, indicating a bumpy road. Green points, on the contrary, indicate low P_{P_E} values, which means that the vehicle was traveling smoothly.

On the left of Figure 5 a sample screenshot of the Android application is shown, indicating the last P_{P_E} value computed.

IV. Preliminary Experiments

Two Motorola Moto G smartphones have been equipped with the SmartRoadSense application and have been setup in order to automatically record track data when in movement and to transfer the collected data to the central server every 15 minutes.

 ISBN: 978-1-61208-353-7

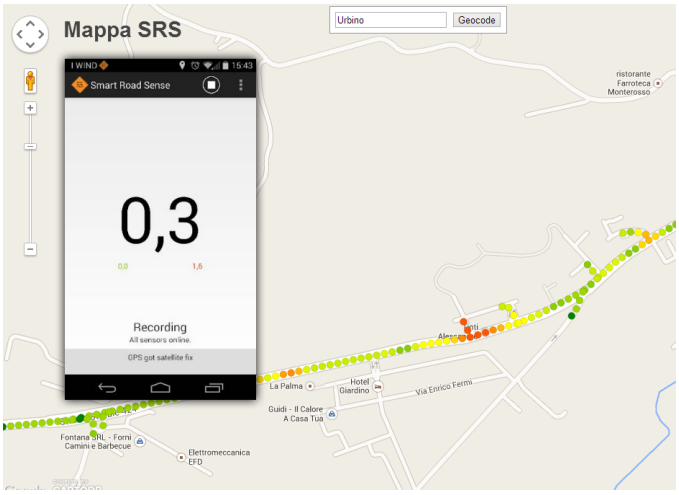


Figure 5. Application screenshot and sample data projection on the map.

The devices have been installed inside two public busses, owned by a local transportation company, using a mobile device rigidly anchored inside the bus cabin. Both busses run twice daily between the cities of Fano, Marotta and Pergola (Marche, Italy). Data was collected over the course of two weeks, from April 1st 2014 to April 16th 2014, totalling ca. 215300 data points. Those points match a total of 744 roads, according to the OpenStreetMap database used (which includes various segments of roads, crossings, etc.), which account for 275089 meters of coverage. On average, data points collected were associated to a road at 5.19 m of distance during the road matching phase of the aggregation algorithm. This gives an estimation the raw GPS data precision. Data occupation of all the collected data amounts to approximately 95 MB (including the overhead given by the PostgreSQL database and indexes).

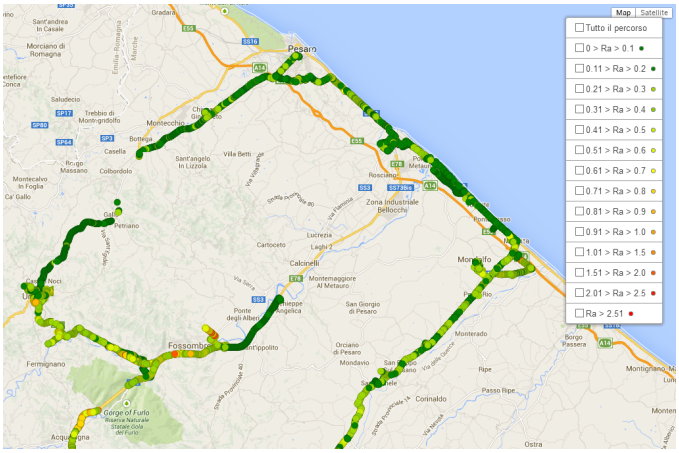


Figure 6. Map of data collected from the experiment.

The resulting data, processed by the server and averaged over equidistant aggregation points, is shown in Figure 6 as an overlay to a geographical map provided by Google Maps. Each shown point represents an aggregated roughness point, whose color varies from red to green as previously described in III-C.

V. Conclusions

In this paper, we have shown how a standard mobile device, running a custom-built application while being anchored to a vehicle in movement, is able to collect data that can be used to detect the quality and irregularities of the road surface.

Such data, collected from the triaxial accelerometer and the GPS sensor, can be appropriately processed by the computational power of a mobile device. Moreover, the dependence of the measurements on vehicle's velocity can be appropriately compensated.

As shown in section III, a data acquisition system has been built that post-processes the data collected by mobile devices and is able to compute a compound roughness index, which is reliably applied to roads marked on a geographical map. Raw data collected by inexpensive devices on personal vehicles, thus transformed into a clear overview of road quality, can be used for the benefit of institutions or drivers. For instance, it may be used by local authorities to detect the presence of critical road surface segments, thus focusing expenditure on roads showing higher maintenance needs.

A. *Future work*

The mathematical model described in section II could be improved, for example giving the possibility to analyze the characteristics of the road surface without the constraint of having the mobile device rigidly anchored to the vehicle.

Moreover, the interface of the central data server will be improved in order to expose a well documented Application Programming Interface (API), allowing client applications to submit data, manipulate it and handle registered tracks.

Future work also includes improvements to the SmartRoadSense mobile application, possibly polishing the experience for end-users and providing means of user registration, in order to enable the distribution of the application via "Google Play Store" and make it possible to collect road data from virtually any user willing to contribute to the project.

Finally, we plan further testing and improvements to the aggregation method used by the server to compute final roughness values: at the time of writing, all data points contribute to an unweighted arithmetic average. However, we can envision averaging methods which weigh contributes (for instance based on age) and/or an evaluation of the data point's source quality.

Acknowledgments

The project was partially supported by "Cassa di Risparmio di Pesaro" foundation and the NeuNet cultural association. The experiment was made possible thanks to the cooperation with "Autolinee Vitali" of Fano.

 ISBN: 978-1-61208-353-7

REFERENCES

[1] L. Deng and D. O'Shaughnessy, Speech processing: a dynamic and optimization-oriented approach. CRC Press, 2003.

[2] Society of Automotive Engineers, Ride and Vibration Data Manual (SAE J6A), ser. SAE information report. Society of Automotive Engineers, 1965.

[3] T. Dahlberg, "Optimization criteria for vehicles travelling on a randomly profiled road–a survey," Vehicle system dynamics, vol. 8, no. 4, 1979, pp. 239–352.

[4] National Quality Initiative Steering Committee, "National highway user survey," Washington, DC, 1996.

[5] J. Y. Wong, Theory of ground vehicles. Jhon Wiley and Sons, Inc, 2001.

[6] T. D. Gillespie, Fundamentals of vehicle dynamics (R-114). SAE International, March, 1992.

[7] M. Ndoye et al., "Sensing and signal processing for a distributed pavement monitoring system," in Digital Signal Processing Workshop, 12th-Signal Processing Education Workshop, 4th. IEEE, 2006, pp. 162–167.

[8] J. Laurent, M. Talbot, and M. Doucet, "Road surface inspection using laser scanners adapted for the high precision 3d measurements of large flat surfaces," in 3-D Digital Imaging and Modeling, 1997. Proceedings., International Conference on Recent Advances in. IEEE, 1997, pp. 303–310.

[9] M. Ndoye, A. M. Barker, J. V. Krogmeier, and D. M. Bullock, "A recursive multiscale correlation-averaging algorithm for an automated distributed road-condition-monitoring system," Intelligent Transportation Systems, IEEE Transactions on, vol. 12, no. 3, 2011, pp. 795–808.

[10] J. Eriksson et al., "The pothole patrol: using a mobile sensor network for road surface monitoring," in Proceedings of the 6th international conference on Mobile systems, applications, and services. ACM, 2008, pp. 29–39.

[11] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in Proceedings of the 6th ACM conference on Embedded network sensor systems. ACM, 2008, pp. 323–336.

[12] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals," in Transactions on Acoustics, Speech and Signal Processing. IEEE, Jun. 1979, pp. 247–254.

[13] L. B. Jackson, Digital filters and signal processing, 2nd ed. Boston: Kluwer Academic Publishers, 1989.

[14] N. Levinson, "The wiener RMS error criterion in filter design and prediction," J. Math. Phys., vol. 25, 1947, pp. 261–278.

[15] J. Durbin, "The fitting of time-series models," Revue de l'Institut International de Statistique, 1960, pp. 233–244.

 ISBN: 978-1-61208-353-7

Applying Flow-based Programming Methodology to Data-driven Applications Development for Smart Environments

Oleksandr Lobunets and Alexandr Krylovskiy

Fraunhofer FIT,
Sankt Augustin, Germany
emails: {oleksandr.lobunets, alexandr.krylovskiy}@fit.fraunhofer.de

***Abstract*—This paper describes initial results of applying the Flow-based Programming methodology to developing data-driven applications for smart environments. This paradigm recently gained popularity in creating concurrent data-driven applications in a wider domain of distributed systems. We investigate this approach applied to the smart environment applications domain and compare it to the Object-Oriented approach typically used in the framework of SOA-based middlewares for the Internet of Things. Our preliminary results show that the Flow-based Programming approach leads to a clear transformation of the design architecture into the software implementation, speeds up the development process, and increases code reuse and maintainability.**

Keywords–Flow-based programming; data flow; data-driven application; smart environment; software engineering.

I. INTRODUCTION

The IoT (Internet of Things) is envisioned as an Internet-like network of interconnected objects, which *allows people and things to be connected anytime, anyplace, with anything and anyone, ideally using any path/network and any service* [1]. With the growth of the IoT deployments in recent years, the data volumes generated by the IoT devices rapidly increase [2], which is recognized by the industry as a motivation for development of data-driven platforms [3]. Similarly, it poses new challenges to the applications development, requiring Big Data processing and dealing with real-time data streams [4]. Such *data-driven* applications, therefore, become one of the most important classes in the growing IoT applications domain.

While the main concern of the IoT is to provide connectivity at the network level, several related technologies facilitating the applications development emerged in recent years. Specifically, two major approaches are being adopted in the research and industry communities: the WoT (Web of Things) [5] and SOA (Service Oriented Architecture) based middlewares [6].

The WoT proposes an HTTP-like protocol to build an application layer on top of the IoT similar to the Web, reusing the widely adopted standards and knowledge of the Web applications development. It does not, however, define a specific development paradigm, leaving this choice to the application developers and the task at hand.

The SOA-based middlewares address the IoT devices heterogeneity by introducing device abstraction layers and enable the integration of the IoT infrastructures with the existing Information Systems. They leverage the developers knowledge in the enterprise software development, where OOSC (Object-Oriented Software Construction) remains by far the most popular software development paradigm.

Despite the success of the OOSC in the broader software development domain, our experience suggests that it sometimes becomes a burden: applying its principles directly to development of data-driven applications for smart environments is challenging and often results in a significant gap between the initial architecture design and the software implementation. Looking for an alternative approach capable of a better realization of the design ideas, we discovered the FBP (Flow-based Programming) [7], which is a subset of a more general Data-Flow approach to software construction [8].

In this work, we approach the task of developing a typical data-driven application for smart environments. The smart environment in our case is considered as a services and applications layer on top of a general IoT infrastructure of interconnected sensors, actuators, displays, and other various computational elements, as originally described by Mark Weiser [9]. We demonstrate the problems arising from employment of the conventional OOSC approach to this task and share our experience in applying the FBP paradigm instead. Our preliminary results show that FBP allows for a clear transformation of the design architecture into the software implementation, speeds up the development process, as well as increases the code reuse and maintainability.

The rest of the paper is structured as follows: Section II provides an overview of related work, Section III describes the application domain and the software design process using OOSC and FBP methodologies, emphasizing the benefits of the latter. Section IV describes the details of our initial FBP system implementation, and Section V provides a conclusion.

II. RELATED WORK

The WoT and SOA-based middlewares are the two main foundations for the development of IoT applications. On the one hand, mash-ups of the Web-enabled devices in the WoT can be constructed similarly to the mash-ups in the Web 2.0 to create ad hoc applications [10]. On the other hand, applications can be built using a SOA-based middleware by composing the services it provides. Such services may include both

 ISBN: 978-1-61208-353-7

WoT-like communication with devices through the middleware abstractions, as well as more complex network-wide services [6].

The WoT mash-ups use direct communication with Web-enabled devices via a Web API, allowing application developers to reuse their Web-development experience and enabling rapid prototyping. However, when building complex data-driven applications requiring communication with many devices and data processing, using the WoT mash-ups alone results in systems which are hard to scale and maintain. In such scenarios, a more systematic approach to applications development is needed, which is one of the core motivations for development of SOA-based middlewares.

In addition to providing a unified API to the IoT devices, SOA-based middlewares offer other services to simplify the applications development. The data processing functionality required for data-driven applications in such systems is typically implemented in the so-called context-awareness services [6]. However, these services need the context models to be explicitly defined, effectively limiting the application developers in the expressiveness of the supported modeling techniques. Moreover, they rarely consider real-time processing of data streams [6], which in practice means that application developers need to implement new services for the middleware itself.

With respect to the integration of the system components using a messaging middleware, various integration solutions and corresponding visual notation languages for connecting distributed components are described by Hophe and Woolf [11]. These patterns resemble approaches in the general Data Flow methodology, but the described solutions mainly rely on the existing enterprise software and protocols, which can be overwhelming for the smart environment applications outside of the enterprise domain. The visual notation employed in these solutions is useful for System Architects and Business Analysts to communicate the system design, but the actual implementation usually follows a different control flow and requires complex configuration and integration code depending on the specific case.

FBP is a form of reactive programming [12] that was initially developed at IBM in 1970s as a software development paradigm, where an application is constructed as a network of asynchronous processes exchanging data chunks and applying transformations to them [7]. It has gained a momentum again recently with the NoFlo project [13], which focuses on enabling visual programming based on the FBP methodology. Several other industrial projects exploring similar principles have appeared in recent years: Streamtools from The New York Times R&D Lab [14], NodeRed from IBM [15], etc. All these projects focus on the processing of *data flows*, which is a major requirement of the modern data-driven applications.

III. DESIGN

Without loss of generality, we assume that our application deals with the data from a large sensor network (IoT), which is used for monitoring the energy consumption of a production line. The connection to the sensor network can be established using either a specialized IoT middleware [16], a publish-subscribe bridge using a message broker [17], a Web-Socket gateway, or a RESTful API. Using any of these technologies,

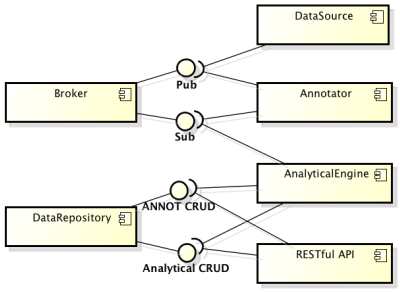


Figure 1. Static OOSC model (UML Component Diagram).

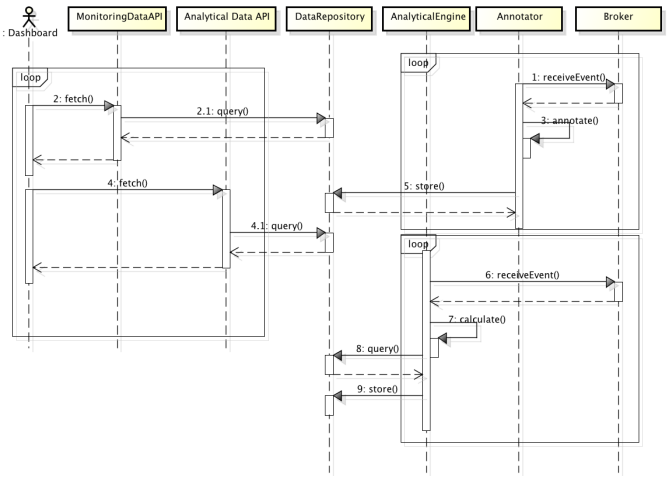


Figure 2. Dynamic OOSC model (UML Sequence Diagram).

application collects sensor data and annotates it with additional context data, e.g., information about the production process. The annotated data is then archived and used for historical and nearly real-time analytics and monitoring presented to the end-user. Our experience suggests that archival, annotation, and analytics are the typical and most often used data transformations in the data-driven applications.

Designing the system architecture for the described application, we first follow the conventional approach suggested by the OOSC paradigm, which we used to practice for years. Then, we apply the FBP principles to the same task and discuss this approach and its benefits in more details in the next sections.

A. Applying OOSC Methodology

OOSC can be considered as a conventional methodology, as it is the dominating approach in software development, although *Post Object-Orientation* methodologies (Component-based Engineering, Aspect-Oriented Development, Service-Oriented Architectures) [18] are getting more attention recently. Following the OOSC methodology, the requirements elicitation is followed by the system design phase, during which a number of static and dynamic models are created and described using a visual notation language such as UML (Unified Modeling Language).

 ISBN: 978-1-61208-353-7

The static model of our application is provided in Figure 1. This component diagram describes the following subsystems: Broker, DataRepository, DataSource, Annotator, AnalyticalEngine, RESTfulAPI. These components are interconnected using different interfaces: Pub, Sub, Annotation CRUD, and Analytical CRUD, where CRUD stands for Create-Retrieve-Update-Delete actions.

While static model describes the structural characteristics of the system, the dynamic model describes its information and control flows. A UML sequence diagram is shown in Figure 2. An additional actor Dashboard is placed at the left of the diagram to depict an interactive application that consumes the output of our system. The Dashboard is communicating with the RESTfulAPI component, which in turn queries the DataRepository and returns either monitoring (raw) or analytical (implied) data. From the right of the diagram, two looping sequences can be identified: annotating loop and analytics loop. The former is retrieving sensor events from the Broker, annotates them and puts in the DataRepository. The latter loop receives a signal from the Broker about new data, queries the DataRepository for raw data, performs calculation and puts results back into the DataRepository.

According to the OOSC approach, designing the system architecture is followed by the software implementation based on the defined models by creating classes, components, libraries, and services, which layout at the code level typically differs from the design models. For instance, in Java the application can be implemented as a single multi-threaded executable, or as a number of OSGi bundles. Independent of the implementation, however, the implementation results in a significant difference between the static and dynamic models, and the way the actual code is executed. The sequence diagram is the closest in resembling the data flow in the system, but its interpretation is already challenging, not to mention how difficult it is to read the Object-Oriented code that implements these flows.

B. Applying Flow-based Programming Methodology

The data-driven applications represent a type of problems, which are mainly concerned with data input, transformation, storage and output. This is the area where the FBP methodology promises to help by designing the application as a flow of interconnected building blocks.

In FBP, application is a network of interconnected reusable components (*black boxes*). Each component has a number of named input and output ports, which are used for receiving incoming data and sending outgoing data correspondingly. Components can be elementary and programmed in some HLL (High-Level Language) or composite (defined as an FBP network). The network execution engine or *scheduler* creates a process for each component and establishes connections between their input and output ports as *bounded buffers*. The data chunks traveling across a connection are encapsulated into IP (Information Packets), which can be grouped into streams or tree-like hierarchies. The parametrization of a process is performed using a special type of IP – IIP (Initial Information Packet), which can also be sent by the *scheduler*. A detailed description of the FBP can be found in [7] and [19].

In this work, we have used the *output-backwards* design approach, starting from visual prototypes of the reports for the end-users. While moving iteratively from the application outputs to the core of the business logic, we were revising our FBP network from a high-level composite components definition to the elementary components with a specific purpose to be programmed right away.

On the first iterations, the top-level network diagram describing our application was produced as depicted in Figure 3, which depicts FBP components as rectangles. Reading the diagram from left to right allows to follow the application business logic and figure out the resulting output. The IPs stream from the *Data Source* (a sensor network), flow to the *PubSub* component, which sends IPs from RAW[0] and RAW[1] output ports to the *Monitoring Publisher* and *Data Annotator* components correspondingly. The *Monitoring Publisher* component was dropped from the OOSC example intentionally to avoid complex cluttered and non-readable diagrams. This component transforms the data into a format suitable for the *Monitoring UI* system, depicted as a subnetwork in the diagram. The *Data Annotator* component annotates the incoming IPs and sends them via its output port to the *PubSub*'s ANNIN input port. The *PubSub* component splits the annotated IPs and sends them to output ports ANNOT[0] and ANNOT[1], which are connected to the IN ports of the *Annotated Data Archiver* and the *Analytics Calculator* components correspondingly. The latter calculates the implied data from the received IPs and sends it to the *Analytics Data Archiver* component. The *Monitoring UI* and *Analytics UI* are depicted as 3rd party systems (or other complex networks in terms of FBP). This description of the data flow is clear enough for a client or the project manager, but not for the developers. For implementation, the complex components need to be described in fine-grained details and each composite component needs to be defined as a network.

A detailed FBP network is presented in Figure 4. It preserves the original layout, but provides more detailed description of components and connections. The composite components described previously are now marked with dotted lines and consist of other elementary components. This level of detail is already sufficient for implementation and, as we will show it in the next section, the actual software implementation also reflects this design. Note that even at such level of detail, it is still possible to follow the data flow from the *Data Source* and all subnetworks down to the resulting output. There are also several repeating components: *STDOUTSUB* (passes through IPs from the *PubSub* output to the system standard output stream, *stdout*), *STDINPUB* (passes through IPs from the system standard input stream, *stdin*, to the *PubSub* input), *CONVERT* (transforms IPs to the storage format of the database), *DBWRITER* (writes formatted IPs to the database). The other new elementary components include: *DF* (transforms IPs from the *PubSub* output to the format acceptable by the monitoring system), *SOCKSEND* (sends the IPs from its input to the TCP/UDP socket defined by the IIP), *TROUTER* (parses the topic of the input IP and creates a new IP with the parsed topic and data), *CTXA* (annotates the incoming IPs with context data).

 ISBN: 978-1-61208-353-7

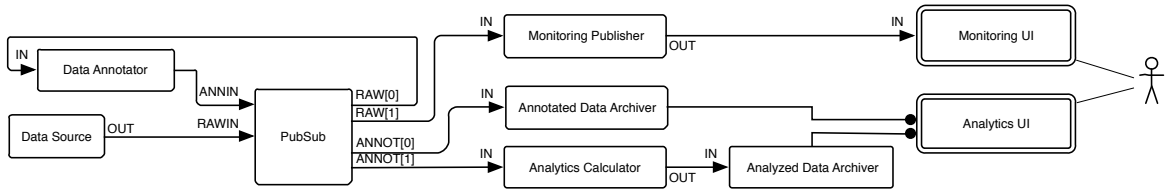


Figure 3. Top-level view of the FBP application.

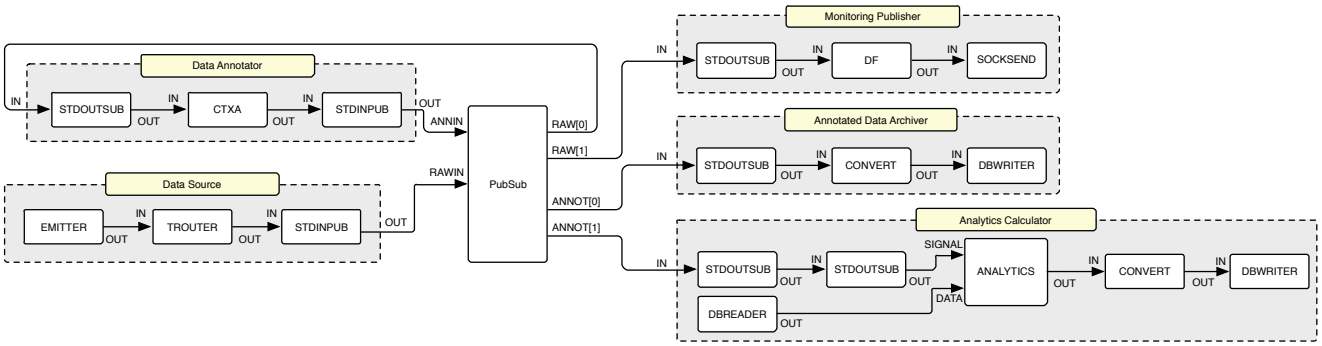


Figure 4. Detailed view of the FBP application.

IV. IMPLEMENTATION

There are different ways to proceed with the actual system implementation. There are many existing FBP development platforms and frameworks, implemented in different programming languages. In the appendix of [7][19], one can find different related concepts, forerunners, and FBP-related technologies. Due to the distributed nature of the smart environment and the absence of a suitable ready-to-use solution, we have started with a custom FBP system implementation by integrating different 3rd-party systems, as described below.

A. Components and ports

Each elementary component depicted in Figure 4 is implemented as a stand-alone executable in a language that in our opinion is the most appropriate for its logic. Some of the components are implemented in Java, some – in C/C++, others – in Go. Each component is a self-sufficing program, which can be used outside of the current application. The *EMITTER*, *TROUTER*, *STDOUTSUB*, *STDINPUB*, and *DF* programs were written in Go using its powerful share by communication concurrency programming model. The *PUBSUB* component is implemented using the Mosquitto MQTT broker [17], while *CONVERT* and *ANALYTICS* components are written in Java. Ports of the components use different protocols for interconnection: stdin/stdout, TCP/UDP sockets, and MQTT [20].

B. Coordination language and scheduler

FBP distinguishes between programming languages: a HLL is used to implement the logic of components and a coordination language is used to describe the data flow and the network structure. The coordination language should be simple and understandable to both developers and users of the system. In the first iteration, we focused on the components and ports that can be easily connected to each other. In order to accomplish this, we used a Linux shell scripting language with POSIX conventions for command line arguments of the executables and UNIX pipes for *stdin/stdout* connections. Although it is a powerful way to express the programming logic, it is not an appropriate way to describe the connections. We are currently evaluating a better DSL (Domain Specific Language) for application description. One of the possible alternatives is the NoFlo's FBP language [21].

Another important part of the FBP system is the *scheduler*, as described in Section III. In our prototype implementation, we used the Foreman Procfile-based applications manager tool [22] and the Upstart event-based process management [23] as the FBP execution engine. This combination of tools requires more efforts on manual description of the processes and dependencies in order to execute the application. In the future we are planning to implement the scheduler that would require only a DSL-based description of the flow for execution.

V. CONCLUSION

The experience described in this paper is the first step towards exploring the efficiency of the FBP methodology applied for the data-driven applications development in the domain of smart environments. The FBP approach mitigates the gap between the information flow in the system design and the control (execution) flow in its implementation. The application design depicted in Figure 4 has been mapped to a source code almost one-to-one, preserving the identical data and control flows.

An FBP application consists of reusable building blocks, and their reorganization into a new application does not require code modification or recompilation, which enables the component reuse and increases the development speed. It is also independent of the HLL used for components implementation, enabling polyglot components in the system. As we

 ISBN: 978-1-61208-353-7

demonstrated, a simple coordinating language can be used to create a flow from components written in Java, C/C++ and Go: only ports definitions need to be known, which should be obviously language-agnostic. Besides, FBP components are interchangeable: each component can be replaced with another block that has the same input and output ports. During development we replaced the sensor network with an event simulation component, and later we replaced a simple rule-based annotator with a BPMN-driven [24] annotation system without writing a single line of code.

The loose coupling between the components allows to manage the code entropy at the low level: the unified flow for data and control allows for easier identification of the faulty component in the network and make the whole debugging processes easier to follow. The unit testing is naturally applied to a FBP application: each component is tested separately by feeding fixtures to input ports and asserting the output ports data against the expected results. The unit testing of the complete application does not differ from testing a composite component (which is a network of components by definition).

VI. Future work

Despite the fact that we are referring to the FBP approach, it is not implemented in our system completely yet, as only some of its fundamental principles are applied at the moment.

Our next steps towards the further exploration of the FBP approach for smart environment applications development include the following: development of general purpose components for most required transformations on the IPs in the system; switching to a unified standard protocol for bounded connections implementation; adding support for a coordination DSL and development of a runtime system (*scheduler*) that will use it for execution of application components; integration with existing FBP visual editors, such as mentioned earlier NoFlo UI or NodeRed to create a complete FBP development environment.

Acknowledgment

The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007-2013 under Grant Agreement no. 257852

References

[1] O. Vermesan, P. Friess, P. Guillemin, S. Gusmeroli, H. Sundmaeker, and A. Bassi, "Internet of things strategic research roadmap," Internet of Things-Global Technological and Societal Trends, 2011, pp. 9–52.

[2] A. B. Zaslavsky, C. Perera, and D. Georgakopoulos, "Sensing as a service and big data," CoRR, vol. abs/1301.0159, 2013.

[3] F. V. Lingen. Data driven platforms to support iot, sdn, and cloud. [Online]. Available: http://blogs.cisco.com/perspectives/data-driven-platforms-to-support-iot-sdn-and-cloud/ (retrieved: June, 2014)

[4] D. Harris. Why the internet of things is big datas latest killer app if you do it right. [Online]. Available: http://gigaom.com/2014/03/04/why-the-internet-of-things-is-big-datas-latest-killer-app-if-you-do-it-right/ (retrieved: June, 2014)

[5] V. Trifa, "Building blocks for a participatory web of things: Devices, infrastructures, and programming frameworks," Ph.D. dissertation, ETH Zurich, 2011.

[6] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," Communications Surveys Tutorials, IEEE, vol. 16, no. 1, First 2014, pp. 414–454.

[7] J. Morrison, Flow-Based Programming, 2nd Edition: A New Approach to Application Development. CreateSpace Independent Publishing Platform, 2010. [Online]. Available: http://books.google.de/books?id=R06TSQAACAAJ

[8] M. Carkci, Dataflow and Reactive Programming Systems. Lean Publishing, 2014.

[9] M. Weiser, R. Gold, and J. S. Brown, "The origins of ubiquitous computing research at parc in the late 1980s," IBM Systems Journal, Vol. 38, No 4, 1999.

[10] D. Guinard, V. Trifa, T. Pham, and O. Liechti, "Towards physical mashups in the web of things," in Proceedings of the 6th International Conference on Networked Sensing Systems, ser. INSS'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 196–199. [Online]. Available: http://dl.acm.org/citation.cfm?id=1802340.1802386

[11] G. Hohpe and B. Woolf, Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2003.

[12] D. Harel and A. Pnueli, "Logics and models of concurrent systems," K. R. Apt, Ed. New York, NY, USA: Springer-Verlag New York, Inc., 1985, ch. On the Development of Reactive Systems, pp. 477–498. [Online]. Available: http://dl.acm.org/citation.cfm?id=101969.101990

[13] H. Bergius. Noflo - visual control flows for javascript. [Online]. Available: http://noflojs.org/ (retrieved: June, 2014)

[14] The New York Times Research & Development group. Tools for working with streams of data. [Online]. Available: https://github.com/nytlabs/streamtools (retrieved: June, 2014)

[15] IBM Emerging Technology. Node-red. [Online]. Available: http://noflojs.org/ (retrieved: June, 2014)

[16] M. Jahn, M. Eisenhauer, R. Serban, A. Salden, and A. Stam, "Towards a context control model for simulation and optimization of energy performance in buildings," in 9th European conference on product and process modeling (ECPPM 2012), 3rd Workshop on eeBDM, eeBIM. Reykjavik, Iceland, 2012.

[17] An Open Source MQTT v3.1/v3.1.1 Broker. Mosquitto. [Online]. Available: http://mosquitto.org/ (retrieved: June, 2014)

[18] A. Przybyek, "Post object-oriented paradigms in software development: a comparative analysis," in Proceedings of the International Multiconference on Computer Science and Information Technology, ser. IMCSIT'07, 2007, pp. 1009–1020. [Online]. Available: http://www.proceedings2007.imcsit.org/pliks/67.pdf

[19] J. Morrison. Flow-based programming. [retrieved: June, 2014]. [Online]. Available: http://www.jpaulmorrison.com/fbp/ (2014)

[20] MQTT.ORG. MQ Telemetry Transport. [Online]. Available: http://mqtt.org (retrieved: June, 2014)

[21] H. Bergius. Language for flow-based programming. [Online]. Available: http://noflojs.org/documentation/fbp/ (retrieved: June, 2014)

[22] D. Dollar. Foreman - manage procfile-based applications. [Online]. Available: http://ddollar.github.io/foreman/ (retrieved: June, 2014)

[23] Canonical Ltd. Upstart - event-based init daemon. [Online]. Available: http://upstart.ubuntu.com (retrieved: June, 2014)

[24] Object Management Group, Inc., "Bpmn: Business process model and notation," retrieved: June, 2014. [Online]. Available: http://www.bpmn.org

 ISBN: 978-1-61208-353-7

Exploring Social Accountability for Pervasive Fitness Apps

Yu Chen
Human Computer Interaction Group
Swiss Federal Institute of Technology,
Lausanne, Switzerland
yu.chen@epfl.ch

Jiyong Zhang
Artificial Intelligence Laboratory
Swiss Federal Institute of Technology,
Lausanne, Switzerland
jiyong.zhang@epfl.ch

Pearl Pu
Human Computer Interaction Group
Swiss Federal Institute of Technology,
Lausanne, Switzerland
pearl.pu@epfl.ch

***Abstract*—Mobile fitness applications have gained increasing popularity to help users walk and exercise more. A key component in such apps is its ability to motivate users. Traditional gamification methods have focused on competition such as leaderboard for community users, self-reflection for individual users, or a combination of the two. Motivated by recent work showing a promising effect of social capital, we have designed and developed a mobile game, HealthyTogether, based on such ideas. We are further interested in how users behave in different settings of gamification methods compared to a baseline. To this end, we have designed and conducted an in-depth user study (N=24) involving 12 dyads playing these games in 4 conditions over a period of two weeks. We report here the design of the application as well as the user study. Among the various rewarding schemes, one that uses a hybrid concept of competition and social accountability gives the most desirable outcome.**

Keywords-health; pervasive fitness applications; gamification; competition; social accountability.

I. INTRODUCTION

Wellness and lifestyle change have gained significant attention in recent years. Both research communities and commercial sectors are putting increasing effort to develop wearable sensors and mobile applications that help and "nudge" individuals to increase their physical activities, eat healthier diet, better manage their sleep and stress, and engage in social lives with family and friends.

Many of the applications use gamification -- the use of game elements in non-game context [14] -- to motivate users to exercise more. Concrete methods include competition such as leaderboard for community users [9], self-reflection such as visualization for individual users [2][5][6], or a combination of the two [7][12]. Recent work shows a promising effect of social responsibility for the sake of helping each other especially among family members, friends, and people who share same interests and goals [1]. We define this concept as social accountability, which refers to a person's awareness of another person's goal and rendering himself/herself responsible to the goal's successful fulfillment.

In this work, we are interested in how users behave in various settings of gamification methods: competition, social accountability, a hybrid model of a mixture of competition and social accountability, and a baseline non-social setting. To this end, we have developed a mobile application, HealthyTogether, which enables dyads to participate in physical activities together, send each other messages, and earn badges. We use this application as an experimental platform for an in-depth user study (N=24) to evaluate how the various reward schemes influence users' exercises and social interactions, both quantitatively and qualitatively.

The rest of the paper is organized as follows. After covering related work in Section II, we present HealthyTogether in Section III, user study design in Section IV and results in Section V. We conclude this paper in Section VI.

II. RELATED WORK

Self-reflection is considered as a self-motivation and successful strategy for pervasive health applications. Research prototypes such as Shakra [4] and Houston [13] and commercial products such as Fitbit and Nike+ all visualize users' daily activities to achieve self-reflection. A number of systems also present physical activities using metaphors. UbiFit Garden [12] visualizes users' daily steps by the growing status of plants. The more activities a user takes, the healthier his plant looks. Fish'n'Steps [7] uses the metaphor of fish tank to visualize users' step count. Recent work has employed informative art as a visualization tool, such as research prototype Spark. The above work mainly motivates users in an individual setting.

Social interaction, including peer-support, cooperation, competition and belonging to a group has been a clear motivator for wellness activities [1][8]. Commercial products have widely adopted competition to motivate user, such as Nike+ and Fitbit. Fitster [9] is a research prototype that visualizes users' steps in a social network and places users in a virtual competition environment. Kukini [16], Fish'n'Steps [7] and Life Coaching Application [11] support competition by helping users to form a team and explicitly introducing social interaction and social pressure.

Research also shows that social communication can motivate users to exercise. Consolvo et al. [13] show that message exchange can help users to increase the responsibility and give support to group members [9]. Champbell et al. [16] suggest that communication using everyday fitness games can help enhance players' social relationship and sustainability in everyday fitness.

Social accountability has been shown to be effective in helping users to achieve goals. Ahtinen et al. [1] have found out that connecting with family members and loved ones can

 ISBN: 978-1-61208-353-7

Figure 1. a) The FitBit tracker, b) FitBit in use, and c) the Samsung Galaxy.

help motivate users; connecting with people with similar wellness targets from communities within short distances can also increase motivation towards wellness activities. Stickk [15] helps users to achieve their goals by allowing them to appoint another person to monitor the progress and verify the accuracy of progress report. They can add supporters who can encourage them by commenting on their progress. Users can also put stake on the goal and specify where the stake would go if they fail in the goal. GoalSponsor [3] allows users to set up goals and sponsors whom they should be accounted for. A sponsor can be a friend, a professional in healthcare, or someone who has accomplished the goal successfully. Users are more committed in fulfilling the goals either because they do not want to let others down or because they do not want to lose reputation in front of others [3]. In the above work, the structure includes one person who has a goal to fulfill and another person who monitors the progress.

To the best of our knowledge, current fitness applications have not well studied interaction schemes in which users mutually account for each other's progress. We are motivated to investigate social accountability factor in pervasive fitness applications using an experimental platform called HealthyTogether.

III. HEALTHYTOGETHER

HealthyTogether is a mobile application that involves a pair of users to exercise together, and it is implemented on the Android platform. To measure users' physical activities, we choose the Fitbit sensor (as shown in Figure 1 a) and b)) among many off-the-shelf sensors as the activity tracker for our HealthyTogether system. Here, we describe the user interface design and the underlying rewarding mechanisms.

A. Game Rules

We designed a series of rewarding mechanisms for HealthyTogether in order to investigate the impacts of different social settings in pervasive fitness application. A user can win badges based on *Karma Points*, which are calculated as below.

$$kp(u) = \alpha \cdot steps(u) + \beta \cdot steps(u')$$

Based on different α and β values, HealthyTogether provides the following three reward settings:

- **Competition setting**, where $\alpha = 100\%$, $\beta = 0$;
- **Accountability setting**, where $\alpha = 0$, $\beta = 100\%$;
- **Hybrid setting**, where $\alpha = 80\%$, $\beta = 20\%$.

In **competition** setting, a user's Karma points are calculated purely by his or her steps. To gain more badges, a user only needs to focus on his or her own activities even if he is exercising with a buddy. Thus, we name this rule competition setting. In **accountability** setting, a user's Karma points are calculated by the steps of the buddy. Therefore, the more he encourages his buddy to exercise, the more points he earns. Thus, we name it accountability setting. On the other hand, even if a user does not move at all, he can still gain badges from the buddy's activities. In the **hybrid** setting, a user's Karma points are calculated based on both his (her) own and that of the buddy, proportionally. The idea behind this reward scheme is to encourage competition while also motivating users to cheer each other. Initially, we set $\alpha = 80\%$ and $\beta = 20\%$ based on the well-known Pareto Principle. In the future, we will also experiment different ratios of competition and social accountability, such as 50%-50% and 20%-80%.

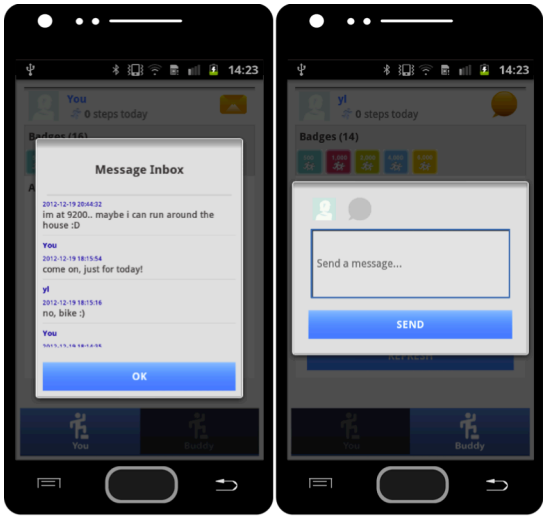
B. Badges

HealthyTogether issues badges based on *kp(u)*. The first badge is issued if *kp(u)>500*, to help users get started in a short time. This number is followed by 1,000 and 2,000 and then increases by every 2,000 points. HealthyTogether calculates Karma points in a daily basis but accumulates badges over time. For example, if a user earns 5,353 Karma points in a day, he can gain 4 badges, i.e., 500, 1000, 2000 and 4000. If a user earns 5,353 and 6,086 points in the first two days, he can gain 4 and 5 badges respectively and a total of 9 badges.

C. Interaction Design

The main interface of the HealthyTogether system is shown in Figure 1 c). It contains a 'self' tab and a 'buddy' tab. Each tab displays information about step count, active time and badges of the current day. We use a pie chart to visualize the proportion of time that a user is in various

 ISBN: 978-1-61208-353-7



a) b)

Figure 2. Screenshots of messaging components.

activity modes, i.e., sitting, lightly active, fairly active and very active.

The badge area displays the total number and the badges that a user has earned. The badges are accumulated over time. In Figure 1 c), the user has earned 6 types of badges with a total number of 16. When he/she clicks on a badge icon, a dialog box pops out explaining the details of this badge type, including how many badges the user has earned and how he/she earned the badges.

There is a messaging button on the top-right corner of each page. When it is clicked, users can either view message history (Figure 2 a)) or send messages to their buddies (Figure 2 b)). Users will receive a vibrated notification when buddies send them new messages.

IV. USER STUDY

To study different game settings in real situations, we designed an exploratory deployment study. We first conducted a user study (Study 1) that spans for six continuous working days, which was divided into a three-day control session and a three-day experimental session. After conducting the study, we were able to discover some interesting results. For example, participants suggested that we extend the study to two weeks, excluding the weekends, so that the control and the experimental sessions span over identical days of the week, thus minimizing the influence of a given day's schedule to the physical activities being monitored. For example, a user may work in the office on Mondays but conduct experiments in the laboratory on Wednesdays. We therefore conducted the second study (we name it Study 2) with duration of two weeks. We refer to the control session as Phase I and experiment session as Phase II in both studies.

A. Participants

We recruited the participants on campus via word-of-mouth. After one person signs up, we asked her to invite a buddy of her choice to join. Their ages range between 22 and 33 and they never used Fitbit before. We required that each dyad should not work in the same office or too close to each other. We offered all participants a 50CHF gift card as compensation for their time.

B. Materials

We provided users with an Android phone with 3G SIM card and a Fitbit. Three users requested to use their own Android phones because it would be more convenient for them. We checked that their phones are compatible for installing HealthyTogether.

C. Procedure

Both Study 1 and Study 2 were structured as a two-phase, within-subjects design. Phase I allowed participants to become accustomed to using Fitbit and allowed us to collect baseline fitness data. In this phase, all participants use Fitbit alone without connecting with buddies. In Phase II, participants in baseline groups (Group A1- A3) continued to use only the Fitbit while groups in social settings (Group B1- B3 in competition setting, C1- C3 in accountability setting, D1- D3 in hybrid setting) started to use Fitbit and HealthyTogether with buddies. The structure of Study 1 was the same as the Study 2, except the duration was extended to two weeks, with both phases extended from three days to five days.

At the beginning of the study, we invited each pair of participants to our laboratory and helped them to set up their Fitbit accounts. We also had a short interview with them on their experience in using fitness sensors. At the end of Phase I, we invited participants in social settings to our laboratory again to install HealthyTogether with different game rules.

Since our user study lasts for up to two weeks, we requested participants to fill in a daily experience survey related the study. At the end of each day, we sent a reminder email with the survey link to participants asking them whether they have anything to share with us about their experience using Fitbit or HealthyTogether. The survey only contains one question: "Do you have anything to share with us on your experience using Fitbit/HealthyTogether today?" Daily survey not only helps us to gain an in-depth understanding of users' experience, but also facilitates us to explain their step data with activities during that day.

At the end of the study, we organized a semi-structured interview. We invited two participants in each group to attend the session together, so that they could share their stories. We did not ask a fixed set of questions, but mediated the session with the following aspects: overall impression, experience, attitudes and aptitudes, motivation of usage, social relationship.

V. RESULTS

In this section, we report both quantitative and qualitative results collected in Study 1 and Study 2. To facilitate describing results, we encoded the two participants in each

 ISBN: 978-1-61208-353-7

dyad with 'a' and 'b' together with their group ID. For example, we encode the two participants in Group C1 as 'C1a' and 'C1b' respectively.

A. *Quantitative results*

Study I

We first investigate users' step count across the 6 days. The overall average daily step count is 7,439 (min=3,185, max = 11,490). We then compare the average daily step count between baseline group (A1) and groups who used HealthyTogether (B1—D1) to evaluate the effectiveness of social interaction incentives (see Figure 3). Results show a slight decrease of steps from Phase I to Phase II across all groups. One explanation is the novelty effect of using Fitbit in the first 1—2 days, as reflected in daily survey. One interesting finding is that in Group A1 the average step count decreased by 20.4% but in Group B1—D1 it decreased by only 10.6%. This implies that HealthyTogether with social settings could help users to persist in physical activities.

We further compare Group B1—D1 with different rules of calculating the accumulative Karma points of the three days. Results show that users in Group B1 (competition setting) have largest difference of Karma points (Δkp = 14,556) compared with Group C1 (Δkp = 839) and Group D1 (Δkp = 2,982). One explanation is that participants in Group B1 focus more on their own performance compared with other groups. In other words, it implies that the social accountability factor, applied in Group C1 and D1, could lead to more balanced performance of physical activities between buddies.

Meanwhile, users exchanged 72 messages, shared by Group B1 (N=43), Group C1 (N=27) and Group D1 (N=2). The distribution shows that Group C1 (accountability setting) and D1 (hybrid setting) interact more compared with Group B1 (competition setting). Particularly, participants in Group D1 exchanged 59.2% more messages than Group C1. This implies that hybrid setting is most useful to encourage participants to interact with each other.

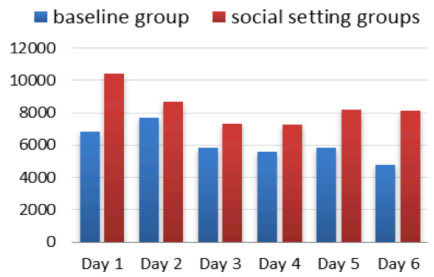


Figure. 3. Distribution of average daily steps for groups with non-social setting vs. social setting in Study 1.

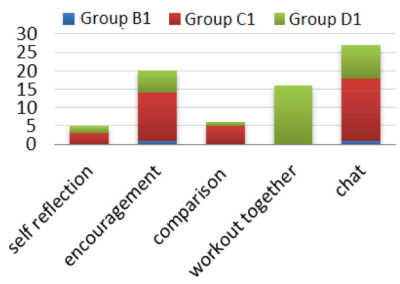


Figure 4. Topic distribution of messages exchanged using HealthyTogether in Study 1.

TABLE I. MESSAGE TOPICS AND EXAMPLES

Topics	Examples
Self-reflection	*"im at 9200.. maybe i can run more"*
Cheering	*"you should make it 8k for a new badge"*
Comparison	*"the first time i am higher than you!!!"*
Workout together	*"we should walk around the floor together to take a break:)"*
Chat	*"feeling so tired now, go to bed soon"*

Several topics emerged from the analysis of message content, and we present them with sample text in Table I. The distribution of each topic shared in Group B1—D1 is shown in Figure 4. It reveals the following phenomenon: 1) in total, there are 27 chat messages, which have the largest share; 2) encouragement is the main topic that is relevant with physical activity (20 messages); 3) Group C1 has the largest share (13 messages) in encouraging messages; 4) the major topic of Group D1 is workout together (16 messages), and it only appears in Group D1. The results imply that hybrid setting introduces most conversation in the topic of workout together.

Study 2

The deployment of Study 2 is the same as Study 1 except for the duration. We first verify whether discoveries in Study 1 still exist in Study 2. The average daily step count is 9,501 (min=3,200, max = 24,334). Figure 5 is a distribution of average daily steps between baseline groups (Group A2, A3) and social setting groups (Group B2, B3, C2, C3, D2, D3) in two weeks. The distribution shows a steady increase of average daily steps in social groups from Phase I to Phase II. Comparing social setting groups and baseline groups, we found average steps increased by 9.8% from Phase I to Phase II in social setting groups but decreased by 10.1% in baseline groups. This is consistent with implication in Study 1 that social settings could motivate users to exercise compared to when they walk alone.

We then compare groups using HealthyTogether in different social settings. Figure 6 shows each participant's average daily steps in Phase I vs. Phase II. The average daily steps in competition groups (B2 and B3) have increased from 9,747 to 10,128 (Δ=381). In accountability groups (C2 and C3), this number increased from 8,888 to 9,717 (Δ=829), and in hybrid setting (D2 and D3) from 10,762 to 12,437 (Δ=1,675). The average daily step increase of hybrid group is 51% more than that of accountability group and three times more than that of competition group. If we assume that participants have the same schedule of the same workday in different weeks, and that Phase I is a baseline for participants, then the above results suggest that hybrid setting encourage users to walk more.

 ISBN: 978-1-61208-353-7

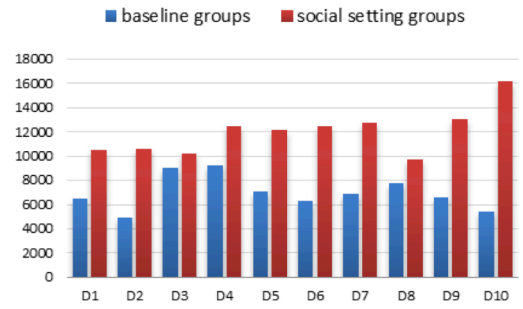


Figure 5. Distribution of average steps non-social setting vs. social setting in Study 2.

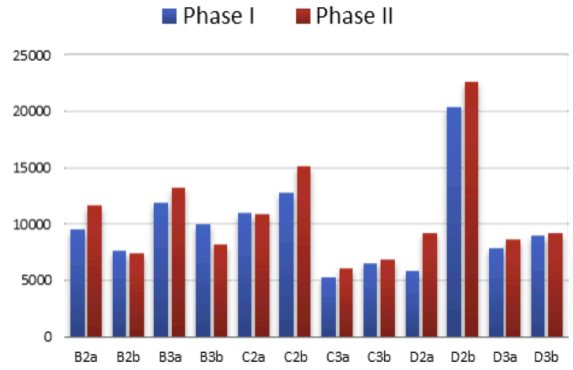


Figure 6. Average step count Phase I vs. Phase II in Study 2.

We further compare the steps between participants in a dyad. As shown in Figure 6, dyads from both accountability groups (C2 and C3) and hybrid groups (D2 and D3) have increased steps together from Phase II to Phase I. On the other hand, in competition groups, average step count of B2a and B3a have increased 22.5% and 11.0% respectively from Phase I to Phase II while this number decreased for their buddies (2.1% for B2b and 17.8% for B3b). This implies that accountability factor helps dyads to walk more steps together. In Study 1, we found participants in accountability group and hybrid group are trying to achieve a balanced number of badges. Even though we do not have the same finding in Study 2, the results concur with the implication in Study 1 that users have more balanced working performance that both users in a dyad improve together.

We then analyze the 86 messages sent between the dyads in Study 2. Participants in hybrid groups sent 58 messages, which is more than twice the number of accountability group (N=21) and seven times more than that of competition group (N=7). Figure 7 shows the distribution of message topics within groups of the three social settings. Different from Study 1, messages with topics about self-reflection and encouragement have the largest share in the total of messages (27.8% for both topics). We also discover that hybrid groups have the largest share of messages (81%) in the topic of self-reflection. The distribution accords with what we have found from Study 1 that 1) hybrid groups have most share of messages (75%) in the topic of workout together, and 2) encouragement is the major topic (54%) in accountability groups. If we consider the number of messages as one metric to evaluate social interaction, results in Study 2 further provide evidence that hybrid setting is more likely to stimulate social interactions.

B. *Qualitative analysis*

In this section, we report the results we found through the logged daily survey and post-study interviews from both Study 1 and Study 2.

During the user interview, we found some qualitative results that are related to our findings from quantitative analysis. Overall, the feedback about HealthyTogether was very positive. First, HealthyTogether has helped them to compare with each other. As B1b said in the interview, "to check her (buddy's) steps and compare with mine is most important for me". Second, they could interact with each other via HealthyTogether. B1a reported in survey: "I received message on the first day, so the next day I intentionally walked more between the buildings."

We also found some evidence that the accountability factor (applied in accountability and hybrid setting) could help users to care about each other. As C1b reported in her daily survey, "I discovered his step is twice more than mine. As his badges depend on my steps, I feel I should walk more in order not to discourage him." This supported what we found in Study 1 that the participants in C1 have more balanced performance when using HealthyTogether. Additionally, when we asked whether users about their social relationship before and after using HealthyTogether, participants in D3 revealed that they were already very close friends. Participants in C1, C2, and D1, and D2 reported that they had developed further relationship with buddies after using HealthyTogether. For example, D1b said: "Even though we are colleagues, we did not talk much. Finding something to do together rapidly brought us closer." For another example, C2a reported that she knows more about her buddy: "When I woke up, my buddy already had 3,000+ steps... She is already on campus..." However, we did not find the same report from competition groups. This suggests that social accountability could be helpful for a user to enhance social relationship with the buddy.

Participants have reported their concerns regarding competition setting. Both B1a and B1b reported competing with each other cause demotivation. "I knew I would never beat him because he needs to walk a lot from home to

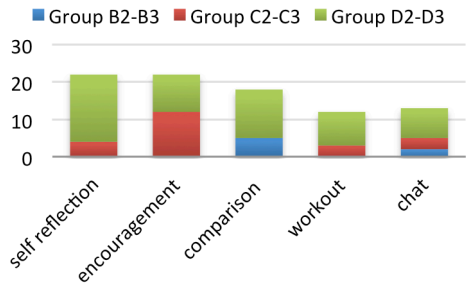


Figure 7. Topic distribution of messages in Study 2

 ISBN: 978-1-61208-353-7

work…" –B1b. "I clearly had advantage in winning the game and sometimes I'm afraid walking too much makes her pressure…" –B1a. B3a also reported that when he noticed his buddy walking less, he also became lazier. Admittedly, the demotivation effect could be caused by the unbalanced abilities in physical exercise. This suggests that choosing a suitable exercise buddy can be important in order to maximize the effect of competition.

Evidence shows that the hybrid setting is more preferred than the accountability setting. "I feel a little bit wired if my badges only depend on my buddy's steps. Is it a little bit de-motivating me?" mentioned by C3a. By comparison, instead of 'depending on' others, both D1 and D3 have reported they arranged activities together, such as walking to a farther away cafeteria to have lunch (D1), and going to Zumba course together (D3). D2a reported in his survey that his buddy gave him suggestions to increase the steps: "Without trying too hard, I almost reached 10k steps. Seeing his progress during the day motivated me to move more. It was also useful to talk to him (via message) - he gets a high step count by walking around while reading articles. I also did this walking in circles thing for about 1h, which helped my brainstorming." Even though D2b (avg=20,372) have clear advantages over D2a (avg=5,852) in Phase I, this discrepancy did not demotivate either of them. Instead, D2a's average daily step has increased by 57.3% (avg=9,207) from Phase I to Phase II and D2b has increased by 11.2% (avg=22,661). As D1a said: "Helping others to become better is a `plus' rather than `minus' to your life."

VI. CONCLUSIONS

In this work, we have developed a mobile application called HealthyTogether that allows dyads to participate in daily physical exercises as a game. We conducted an in-depth user study with 12 dyads with a period of up to 2 weeks and compared participants using HealthyTogether in 3 social settings and a baseline non-social setting. Results show that social settings, even in the competition mode, can help users to persist more in physical activities compared with baseline group. Additionally, the hybrid setting is more likely to motivate users to walk more and more actively help others. Furthermore, the number of messages sent between participants in the hybrid settings is 8 times more than those in the competition setting and twice of those in the accountability setting. Integrating social accountability factor is also promising to enhance social relationship between buddies.

In the future, we plan to conduct longitudinal studies with more users in various conditions to validate our findings with statistical analysis.

ACKNOWLEDGEMENT

We thank Swiss National Science Foundation for sponsoring this research work. We are also grateful for all anonymous reviewers for their valuable feedback.

REFERENCES

[1] A. Ahtinen, M. Isomursu, M. Mukhtar, J. Mäntyjärvi, J. Häkkilä, and J. Blom. Designing social features for mobile and ubiquitous wellness applications. The 8th international Conference on Mobile and Ubiquitous Multimedia (MUM'09), ACM. Nov. 2009.

[2] C. Fan, J. Forlizzi, and A. K. Dey. Spark Of Activity: Exploring Informative Art As Visualization For Physical Activity. Proc. of the ACM Conference on Ubiquitous Computing (Ubicomp'12), ACM. Sept. 2012, pp. 2–5.

[3] GoalSponsor. https://www.goalsponsors.com. Accessed: August 6, 2014.

[4] I. Anderson, J. Maitland, S. Sherwood, L. Barkhuus, M. Chalmers, M. Hall, and H. Muller. Shakra: tracking and sharing daily activity levels with unaugmented mobile phones. Mobile Networks and Applications, 12(2-3), 2007, pp. 185-199.

[5] I. Li, A. K. Dey, and J. Forlizzi. A stage-based model of personal informatics systems. Proc. of the SIGCHI conference on Human Factors in computing systems (CHI'10), ACM. Apr.2010, pp. 557-566.

[6] I. Li, A. K. Dey, and J. Forlizzi. Understanding my data, myself: supporting self-reflection with ubicomp technologies. Proc. of the ACM Conference on Ubiquitous Computing (Ubicomp'11), ACM. Sept. 2011, pp. 405-414.

[7] J. Lin, L. Mamykina, S. Lindtner, G. Delajoux, and H. Strub. Fish'n'Steps: Encouraging physical activity with an interactive computer game. UbiComp 2006, pp. 261-278.

[8] J. Maitland, and M. Chalmers. Designing for peer involvement in weight management. Proc. of the SIGCHI conference on Human Factors in computing systems (CHI'09), ACM Press (2011), pp. 315-324.

[9] N. Ali-Hasan, D. Gavales, A. Peterson, and M. Raw. Fitster: social fitness information visualizer. In CHI'06 Extended Abstracts on Human Factors in Computing Systems. Apr. 2006, pp. 1795-1800.

[10] P. Klasnja, S. Consolvo, and W. Pratt. How to evaluate technologies for health behavior change in HCI research. Proc. of the SIGCHI conference on Human Factors in computing systems (CHI'11), ACM. May 2011, pp. 3063-3072.

[11] R. Gasser, D. Brodbeck, M. Degen, J. Luthiger, R. Wyss, and S. Reichlin. Persuasiveness of a mobile lifestyle coaching application using social facilitation. Persuasive Technology 2006, pp. 27-38.

[12] S. Consolvo, D. W. McDonald, T. Toscos, M. Y. Chen, J. Froehlich, B. Harrison, and J. A. Landay. Activity sensing in the wild: a field trial of UbiFit garden. Proc. of the SIGCHI conference on the SIGCHI conference on Human Factors in computing systems (CHI'08), ACM. Apr. 2008, pp. 1797-1806.

[13] S. Consolvo, K. Everitt, I. Smith, and J. A. Landay. Design requirements for technologies that encourage physical activity. Proc. of the SIGCHI conference on Human Factors in computing systems (CHI'06), ACM. Apr. 2006, pp. 457-466.

[14] S. Deterding, M. Sicart, L. Nacke, K. O'Hara, and D. Dixon. Gamification: using game-design elements in non-gaming contexts. In CHI'11 Extended Abstracts on Human Factors in Computing Systems, ACM. May 2011, pp. 2425-2428.

[15] Stickk. http://www.stickk.com. Accessed: August 6, 2014. 2

[16] T. Campbell, B. Ngo, and J. Fogarty. Game design principles in everyday fitness applications. Proc. of the 2008 ACM conference on Computer supported cooperative work (CSCW'08), ACM. Nov. 2008, pp. 249-252.

 ISBN: 978-1-61208-353-7

Multimodal Task Assignment and Introspection in Distributed Agricultural Harvesting Processes

Daniel Porta*, Zeynep Tuncer†, Michael Wirth‡ and Michael Hellenschmidt§
*German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany, daniel.porta@dfki.de
†John Deere GmbH & Co. KG, Kaiserslautern, Germany, TuncerZeynep@JohnDeere.com
‡Eyeled GmbH, Saarbrücken, Germany, wirth@eyeled.de
§SAP AG, Darmstadt, Germany, michael.hellenschmidt@sap.com

***Abstract*—Harvesting processes are in fact industrial manufacturing processes that follow a tight schedule. Unexpected incidents can disturb a harvest and require a replanning of the process in order to avoid severe financial losses. When a new plan has been found, it must be communicated to the affected process participants, i.e., drivers of agricultural machinery. This paper presents a cloud-based system for orchestrating and coordinating a fleet of agricultural machinery and their drivers during an ongoing harvest in case of an unexpected incident. A management dashboard allows the real-time replanning of a harvesting process and sends updated instructions to each affected driver's mobile device. The paper focuses on the communication between the contractor and a driver in the field as well as the interaction of the driver with his mobile device. It is explained how the system accomplishes a fast, traceable, and safe communication with the drivers that may suffer from bad network conditions and a high cognitive load. In order to understand the details of his new tasks, a driver can examine them in a multimodal dialogue including speech with the system. This is beneficial in a driving situation. By interacting with the mobile client, the system also deduces if the driver correctly understood his new instructions and can intervene if not.**

Keywords–Multimodal Dialogue; Task Assignment; Task Introspection; Agriculture; Harvesting Process.

I. INTRODUCTION

Because of its complexity, agricultural production can nowadays be regarded as an industrial manufacturing process, in which a tractor in a sense represents not merely a vehicle but a complex tool that is, due to the size of the production area, mobile by means of wheels. The interests of agricultural enterprises are also comparable to other production companies. On the one hand, the entrepreneur, i.e., the farmer, wants to design process chains that achieve the highest possible efficiency. On the other hand, these production processes should to some extend be robust against possible interference. Thus, the designed processes have significant effects on the efficiency of the machines, but also on the products themselves. Additionally, a particular focus in agriculture lies upon ecology and sustainability, which are also affected by the operating procedures. Considering a harvest campaign as a complex agricultural process chain, a large number of agricultural machinery, e.g., tractors and forage harvesters, is involved in the distributed process. Their coordination and cooperation must be perfectly organized to accomplish the harvesting process in an economical manner. They are based on a complex orchestration of all participating employees and machines which is planned before the actual harvest begins. The necessary interaction between the machines require process technologies as in any conventional factories. However, external and unexpected influences, such as changing weather or traffic conditions, may affect this sensitive structure as well as a drop out of a machine or driver.

A harvest is typically planned and conducted by a contractor on behalf of the farmers. A contractor is a service provider that has a large fleet of agricultural machinery and personnel at his disposal. In case of an unexpected incident that affects an ongoing harvest, the contractor gets in trouble. He normally serves multiple customers at different locations at the same time and his schedule for the remaining harvest season often leaves no room for catching up a bigger delay. The contractor, therefore, requires tools based on Information Technology (IT) for decision making and transmission of information in order to immediately change action plans for agricultural machinery and their drivers. Subsequently, all stakeholders need to be informed about the change of plan. This is done today often via mobile phones or radios. Telephony as a synchronous form of communication requires a time-consuming, sequential approach in order to inform the large number of affected drivers (we consider a smaller number of 20 drivers in total) and also suffers in rural areas from poor network coverage. Radios, however, often do not have the required range and are just not safe to use for inexperienced seasonal workers. Thus, some driver might follow the new plan while others still follow the outdated plan.

This paper presents a cloud-based harvest and communication management system for the COordination of Agricultural Production in Real-Time (COAP-RT). It is capable of orchestrating and coordinating a fleet of agricultural machinery and their drivers during an ongoing harvest. The paper focuses especially on the drivers' multimodal mobile User Interface (UI) for communicating changed action plans as a result of an unexpected incident. This communication must be *quick* and *traceable* for the contractor and *safe* for the driver. These three factors are important for keeping track of a tight schedule. Also, these factors become even crucial considering that tractor drivers suffer from a high cognitive load in phases of high concentration, e.g., (un-) loading procedures or driving over rather small roads with tons of cargo and opposing traffic. This leaves only little room for additional attention. In such situations, multimodal dialogue interaction including speech and gestures proved to be beneficial [1].

The outline of the paper is as follows. In the next Section II, related work is discussed. Section III covers the overall system architecture and describes the implemented components. In Section IV, a complete user interaction with the mobile client is discussed. Finally, we conclude in Section V.

 ISBN: 978-1-61208-353-7

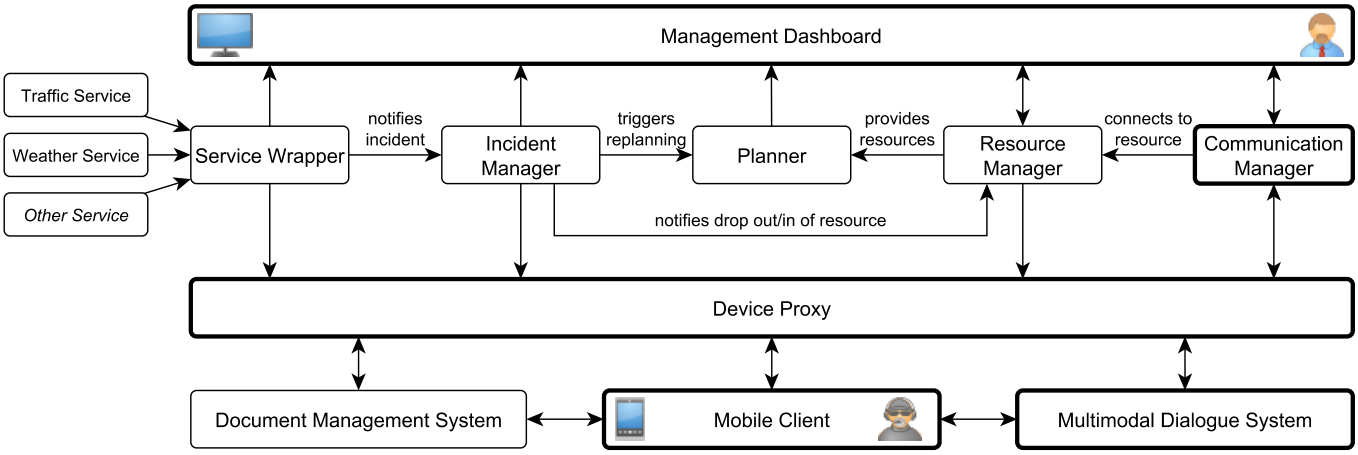


Figure 1. Overall logical system architecture. The highlighted components are relevant for the focus of this paper.

II. Related Work

There is an increasing number of IT-based research regarding precision farming or precision agriculture aiming at knowledge representation, information management, and decision support for various agricultural tasks that apply Future Internet technologies and concepts like the Internet of Services, the Internet of Things, the Semantic Web, and Industrial Internet [2][3].

The iGreen Project [4] developed an infrastructure for knowledge and service networks based on a service-oriented architecture in order to implement mobile decision support systems and tools for collecting and exchanging knowledge over organizational boundaries in the domain of agricultural production [5]. They use agroXML [6] and linked data technologies [7] in order to fit information systems to the requirements of agricultural processes. In essence, they propose the adoption of ReSTful [8] services because these are ”especially well suited to agriculture as they allow quick adaption to new conditions and reuse of data in unforeseen contexts” [9]. We also implemented a ReSTful service backend. The AGRICOLA project [10] developed an agent-based dynamic resource planning network in the agriculture domain. The planning considers weather and drop outs of personnel or machinery as dynamic disturbance factors. The planning itself relies on simulation-based dynamic coalition forming [11] in order to a achieve stable groups of cooperative agents. Our planning component is inspired by this approach. The Marion project [12] focuses on a dynamic and distributed infield planning system for harvesting. This planning considers route planning of (autonomous) vehicles within a field for optimal and smooth harvesting processes. This is a fine-grained micromanagement approach. In contrast to this, the COAP-RT system rather looks at the harvesting process from a bird’s eye view. However, both approaches would perfectly fit together and complement each other. There are also commercial software-as-a-service platforms like Farmpilot [13] or 365FarmNet [14] available. Farmpilot can be seen as a fleet management system and harvest planning tool. However, it does not support automatic replanning of a process during an ongoing harvest campaign in case of an unexpected incident and the provided mobile application is only intended for use by the contractor, not the drivers. 365FarmNet aims at an open and holistic process management and service platform for the agricultural domain ranging from sowing to harvest. In this sense, the COAP-RT system can be seen as a specialized service for replanning an ongoing harvest campaign with an advanced multimodal mobile UI for communicating changed action plans to affected drivers. The Farming 4.0 initiative [15] of Deutsche Telekom and the German agricultural machinery manufacturer CLAAS tackle similar issues like we do. However, a coordinating higher instance in case of an incident is not intended. Also, the mobile client does not provide clear instructions but only data whose interpretation is up to the driver. The initiative mainly aims at testing communication technologies, such as Long-Term Evolution (LTE) and infield navigation in rural areas.

III. Overall System Architecture

This section describes the logical system architecture of the cloud-based COAP-RT system. Figure 1 depicts the respective components of the system and their connections. The highlighted components are relevant for the focus of this paper. However, the others are briefly presented in order to get a complete picture of the system. The cloud infrastructure relies on the SAP HANA Cloud platform [16].

The *Service Wrapper* makes data from external data sources internally available. Currently, we include the ReSTful traffic service from Microsoft [17] and the ReSTful weather service from OpenWeatherMap [18] which also provide weather forecast. Additional services can easily be integrated. The Service Wrapper polls these Web services in regular short intervals and republishes the data in an internal representation on a dedicated ReSTful application programming interface (API). Thus, the Service Wrapper acts as a Meta Web Service [19]. It also generates incidents based on rules if a process interfering situation is observed. Such an incident is propagated to the *Incident Manager*, which can also receive incidents from drivers created via their mobile client or even automatically generated incidents from their vehicles. For testing and demonstration purposes, incidents can be simulated. The Incident Manager then updates the

 ISBN: 978-1-61208-353-7

remaining resources, e.g., in case a tractor dropped out, notifies the responsible contractor as well as affected drivers in the harvesting chain about the incident (that means they should prepare for forthcoming updates of their instructions) and triggers a replanning of the current harvesting process. The Incident Manager also keeps track of reported incidents. They can be retrieved from a dedicated ReSTful API. The replanning is carried out by the *Planner*. It takes available resources and other contextual constraints, such as deadlines and economical cost models, into account in order to come to an optimal solution. However, the replanning of a harvesting process is not an automatism. In the end, the contractor has to approve (or reject) the proposed replanned process. So, the contractor is always in charge and retains full control over the harvesting process, which will raise acceptance for IT-based tools in the experience-based agricultural domain. The Planner can also be accessed via a dedicated ReSTful API.

The *Management Dashboard* is the stationary UI for the contractor, similar to the Management and Monitoring Tool described in [20]. It is implemented as a browser-based mashup of available data and relevant information utilizing the ReSTful APIs of the system's back-end components. Thus, the contractor can initially enter, e.g., agricultural machinery at his disposal as well as customer records into the *Resource Manager*, a ReSTful component for managing master data. During a harvest campaign, he can get a quick overview of unexpected incidents that require his attention, affected customers, fields, and vehicles. In case a replanning becomes necessary, the contractor can revise countermeasures proposed by the Planner and, if he finally agrees, can broadcast a set of derived instructions to all affected drivers via the mobile Internet.

Communication with drivers is accomplished by the *Communication Manager*. When a driver signs in and connects to the system, this component links a virtual vehicle resource with the corresponding real-world vehicle. Afterward, the Communication Manager distributes messages addressed to the virtual resource to the actual driver's mobile client. It must be *traceable* for the contractor (1) when a message was send to a driver, (2) when it was received, (3) that a driver noticed the message and (4) that he understood the contents of the message. While the former two points tackle technical issues regarding network coverage in rural areas, the latter two points tackle important human factors because the contractor does not personally talk to the driver. When a driver has to concentrate on his actual work, he might not notice a new message or misunderstand (parts of) its content. If point (3) and/or (4) cannot be checked, it is very likely that drivers follow outdated, perhaps contradicting instructions and the replanned harvesting process is not executed as desired. Without asking the driver in such a situation, this cannot be revealed directly but may be derived later, e.g., by looking at position traces. In the next section, we will explain in depth how the COAP-RT system copes with the traceability of these four points. The Communication Manager does not communicate directly with the drivers' mobile clients but via the *Device Proxy*. Technically, both components can be merged. However, from a logical perspective, they serve different purposes. The Device Proxy is based on the lightweight Node.js framework and leverages the Session Initiation Protocol (SIP) [21] normally used by IP telephony. It implements a SIP registrar and proxy that manages a dictionary of SIP contact URIs and user names provided by the clients through SIP registrations. In this way, the Device Proxy can relay messages from the Communication Manager (received via HTTP) to the driver's actual mobile client (transmitted via SIP). This is by design of the SIP protocol efficient and to some extend robust against poor network coverage since the proxy uses the User Datagram Protocol (UDP) as SIP transport protocol. This communication is also easily traceable in a sense that the mobile client automatically acknowledges receipt of a message at the application layer.

The *Mobile Client* runs on a Google Nexus 7 with Android 4.2.2 or higher and is implemented as a cross-platform Apache Cordova application. It contains native Java plugins for SIP communication (we extended the source code of the stock Android SIP stack by instant messaging capabilities [22]), local speech synthesis, and bidirectional audio streaming in the common 8-bit A-law telephony codec. The graphical UI (GUI) as shown in Figure 2 is based on HTML5 and JavaScript, and thus, rendered in a Web browser component. It allows a driver to be always in touch with the most recent developments in the current harvest campaign. Pushing the red microphone button opens a channel for speech interaction. This function can also be triggered by pushing the call button of a connected bluetooth headset, since the client leverages vendor-specific headset events of the Android platform. The GUI consists of six interconnected views that are accessible after the driver signed in by providing his credentials and selecting a vehicle (Figure 2, screen (1)). This paper concentrates on the first two views. Most important, the *task view* (2) always explains the driver in a very clear and unmistakable fashion his current role and tasks within the harvesting chain. Tapping on a pin icon, the GUI switches to the map view where the corresponding target location is displayed. Incidents from the Indidents Manager are listed in the *incidents view* (3). The symbols indicate the source (from a vehicle, a weather, or traffic service), the severity, and the current status (unhandled, handled) of an incident. Tapping on a list entry reveals further details. A driver can report an incident by means of a dedicated form that is available via the menu button in the lower left corner of every screen. Neßelrath and Porta, [23] showed that reporting an incident can also be accomplished in a multimodal dialogue fashion. The *map view* (4) visualizes the driver's operational area and provides navigation aids to go there. The map view also visualizes the locations of relevant incidents that occurred during the harvest campaign. The *weather view* presents fine-grained current weather information and a three day forecast for the area of operation. This information comes from the Service Wrapper. The *vehicle view* displays selected vehicle specifications like dimensions and weights. This information comes from the Resource Manager and helps inexperienced drivers get acquainted more easily with the new vehicle. The *document view* allows the driver to access all manuals, regulations, and directions relevant for his current tasks as required by statutory law in PDF format. This is especially useful after a replanning occurred. In case a driver gets new instructions or is assigned to a different field which is now located in a landscape protection area or in a hillside situation, he does not need to return to the home base for picking up the required security advises in a paper-based form. These documents are provided by the context-sensitive *Document Management System*.

 ISBN: 978-1-61208-353-7

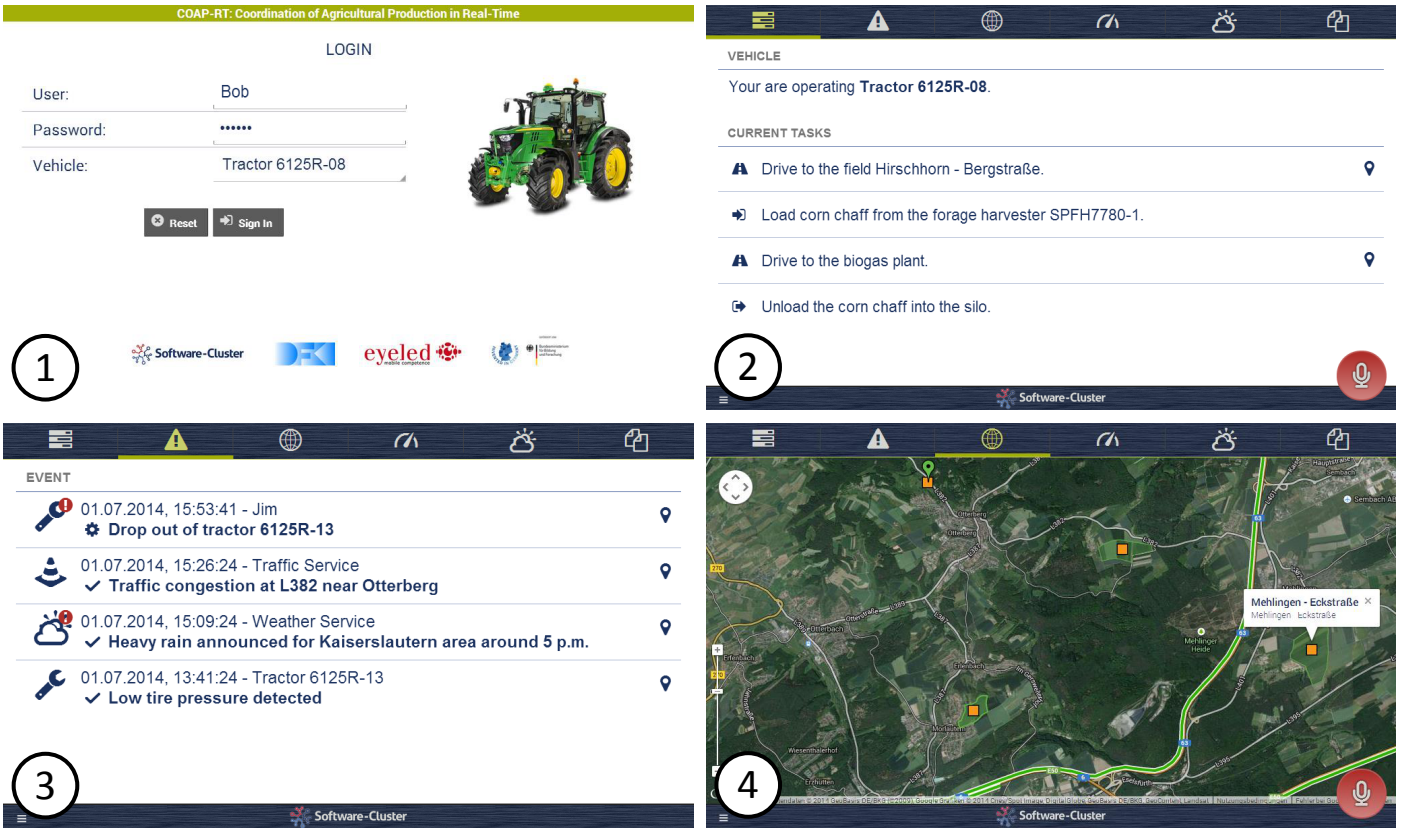


Figure 2. The GUI of the Mobile Client. Shown are the login screen (1), the task view (2), the incidents view (3), and the map view (4). The vehicle view, the weather view, and the document view are omitted for brevity. The paper focuses on the task view and the incidents view.

The *Multimodal Dialogue System* (MMDS) finally allows a driver to interact with the COAP-RT system in an advanced and non distracting fashion by using touch, speech, and other modalities. It is based on the SiAM-dp multitenant multimodal dialogue platform initially developed for the automotive domain [24]. It pursues a model-based development approach by means of the Eclipse Modeling Framework (EMF) [25] in order to build context-adaptive dialogue applications and it aims at considering the driver's cognitive load as one adaptation criterion. Consequently, SiAM-dp consists of a runtime environment and an Eclipse-based workbench for the rapid development of dialogue applications. Technically, the MMDS is also connected to the Device Proxy via SIP. When a driver signs in the COAP-RT system, the mobile client automatically creates a SIP session at the MMDS using the SIP INVITE procedure. Both components, the mobile client and the MMDS, can now communicate with each other by exchanging SIP messages. Eventually, this enables mixed-initiative conversations, i.e., either a driver or the COAP-RT system (by means of the MMDS) can start a conversation. The MMDS integrates off-the-shelf network speech recognition and synthesis solutions by means of SIP and MRCP [26]. Also, the MMDS implements a dialogue strategy, i.e., a strategy for leading a conversation with the driver, such that the contractor can be confident that the driver understood his new instructions. In this way, the MMDS can actually simulate a phone call with the driver on behalf of the contractor. Thus, the driver does not need to take his eyes from the street and can adapt to his current cognitive load.

IV. Multimodal Task Assignment and Introspection

As already mentioned in the introduction, the transmission of updated instructions from the contractor to the driver must be *quick*, *traceable*, and *safe*.

The transmission is *quick* as the MMDS can lead several conversations simultaneously. Thus, the COAP-RT system is faster than calling each affected driver individually and manually in sequence. If only one driver has to be instructed (which is unlikely in complex harvesting processes), the reaction time of the system still outperforms the contractor's reaction time, who might have to first search the correct phone number by hand. The system relies on the mobile Internet as the actual communication channel and improvements on the network infrastructure in rural areas are out of our scope. However, speech interaction requires less bandwidth than a normal phone call because the same audio codec is used for server-side speech recognition. Speech synthesis is performed directly on the mobile device. So, only the textual content needs to be transmitted from the MMDS to the client. If a driver's mobile device has no signal, he cannot be reached neither automatically by the system nor manually by the contractor. However, the system immediately recognizes when the driver is back online due to the device's renewed SIP registrations.

Regarding the technical *traceability* of the successful delivery of messages, we utilize the SIP protocol. So, the Communication Manager logs the timestamps of when a new message

 ISBN: 978-1-61208-353-7

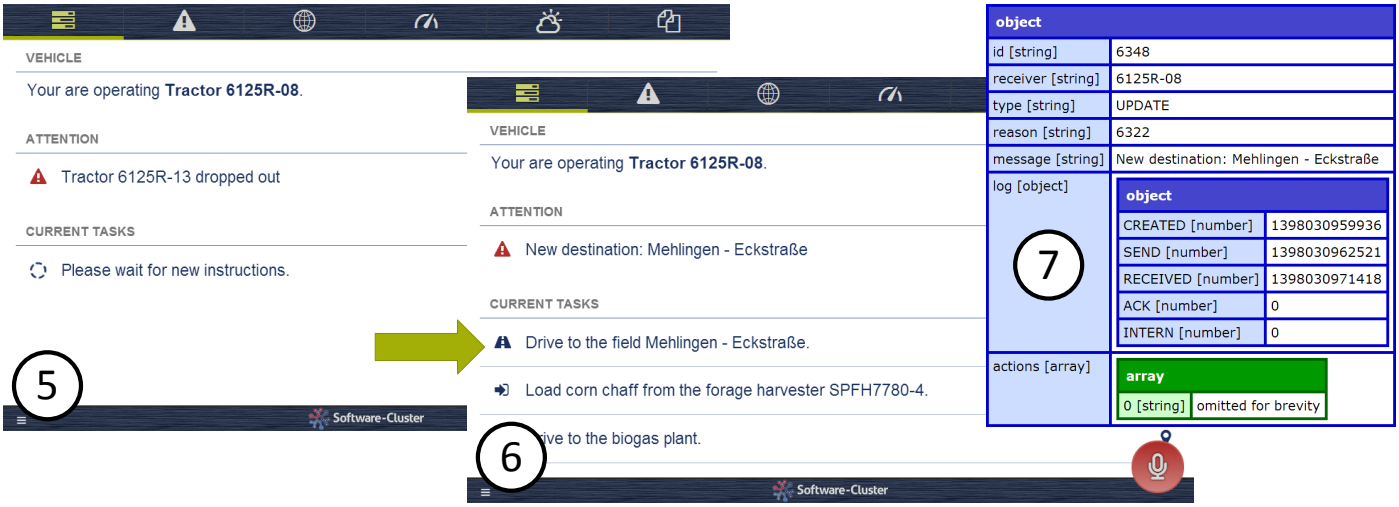


Figure 3. Screen (5) immediately informs an affected driver of an unexpected incident, here the drop out of another tractor. Screen (6) displays an updated list of tasks as reaction to the incident. Screen (7) shows the actual message behind screen (6) in a beautified JavaScript Object Notation (JSON).

was created, when it was sent out, and when it was received and attach this information to the message. This is shown in Figure 3, screen (7). This way, delivery delays, e.g., due to bad network coverage can be detected and separately handled as long as the retransmission approach of messages included in the SIP protocol does not succeed after a specified time. Also shown is that the message is not yet acknowledged and understood by the receiver although it was already successfully delivered. The missing timestamps are added to the message log by the MMDS if the driver responded (ACK) and the MMDS is finally convinced that the driver understood or better internalized the contents of the message (INTERN).

We lend the term *introspection* from psychology as the ability of self-reflection and combine it with the term *task* in order to point out that a driver, when assigned with a new task, first has to reflect if he correctly understood it. If he then comes to the conclusion that something is unclear, he is encouraged to ask the system and thus *inspect* the details in a multimodal dialogue. Currently, a task consists of a location and an action to perform at the location. The MMDS constantly observes the user interaction. Using its dialogue strategy which implements prototypical grounding mechanisms known from dialogue theory [27], the MMDS is able to conclude based on the drivers interactions whether he already internalized his new task, i.e., his new location and action, or not. If not, the MMDS should give or ask for the so far not grounded information and thus stimulating the introspection process of the driver.

Every vehicle in a harvest campaign is equipped with a mobile COAP-RT client. Bob, a tractor driver, wears a bluetooth headset that is connected with the mobile client of his tractor. So, he can benefit from the built-in noise canceling mechanisms of the headset for an improved speech recognition performance and is not required to touch the mobile device while driving. He just unloaded corn chaff at a biogas plant and wants to return to his actual field of activity when an unexpected incident occurs at a different location that affects him. In the following, two alternative examples of typical dialogues between the driver (U) and the system (S) are presented. GUI updates are explained in square brackets.

(1) S: Hello Bob! You will get new instructions due to an unexpected incident. I'll get back with you soon. [*The task view is updated (Figure 3, screen (5)) and the incident appears in the incident view (Figure 2, screen (3), first list item)*]
(2) U: Ok, why that?
(3) S: Tractor 13 operated by Jim dropped out due to a tire damage.
[*The system performs a replanning. The contractor finally agrees to the proposed replanned harvesting process and broadcasts new instructions.*]
(4) S: Attention Bob! You have new instructions. Please drive immediately to the field Mehlingen. [*The task view is again updated (Figure 3, screen (6)).*]
(5) U: Ok, what should I do at the field in Mehlingen?
(6) S: Please support the forage harvester 4 there.
(7) U: Ok, I'm on my way.
(8) S: [*The navigation system is silently updated.*]

In utterance (2), Bob acknowledges receipt of the message with "Ok" and indicates that he understood an incident has occurred with the subsequent "why" question. The answer is given by the system in (3). After replanning, Bob is initially informed about his new task in (4). In (5), Bob again acknowledges receipt of the message. He also repeats his new destination. This immediately indicates the MMDS that he understood the location part of his new task. Since he also actively asks for more information, the MMDS finally assumes in (8) that he also understood the action part his new task. As an alternative to utterance (5), the driver might be less communicative, perhaps due to a higher cognitive load. This results in the following slightly different dialogue.

(1) S: Attention Bob! You have new instructions. Please...
(2) U: Ok, I'm on my way.
(3) S: Your navigation system is updated with Mehlingen as your new destination. Please follow the instructions.
(4) U: Ok, thanks.
(5) S: Please support the forage harvester 4 there.
(6) U: Ok.
(7) U: If something is unclear, don't hesitate to ask.

 ISBN: 978-1-61208-353-7

In (2), Bob acknowledges receipt of the message. He also indicates that he is assigned to a new destination. However, the MMDS is not convinced yet that Bob correctly understood his new destination. Therefore, the MMDS gives an additional hint and repeats the destination in (3). Still, the MMDS is not sure whether Bob knows what to do at his new destination. So, it introduces this information in an anticipatory manner in (5) which is acknowledged in (6). In (7), the MMDS is sufficiently convinced to end the conversation. However, it encourages Bob to ask if questions arise. Please note that the system takes a stronger initiative here in order to achieve a sufficient confidence that Bob understood his new instructions, whereas Bob's utterances are shorter (basically, these are only confirmations) causing less additional cognitive load. Thus, the implemented dialogue strategy is also beneficial for the *safeness* of the driver while driving and inspecting the details of his recently updated instructions.

V. Conclusion and Future Work

We presented the cloud-based COAP-RT system for orchestrating and coordinating a fleet of agricultural machinery and their drivers during an ongoing harvest campaign in case of an unexpected incident. It has been shown how the system contributes to a quick, traceable and safe communication of changed action plans from the contractor to affected drivers. The overall system was successfully demonstrated at the CeBIT 2014 where we got in contact with domain experts who appreciated our approach. Currently, we simulate vehicle drop outs. So, next steps are to get access to real telemetry data, e.g., from a CAN bus or ISOBUS, and to increase the planning granularity in order to be able to conduct a field test for evaluation purposes. Regarding the communication aspect, refinements of the dialogue strategy and related artifacts will be beneficial for a more natural user experience. A user study conducted as an extended lane change test in an appropriate simulation environment as described in [28] can finally asses the driver's cognitive workload. We should also consider more fine-grained escalation mechanisms, i.e., how and when to indicate the contractor that his manual intervention is required.

Acknowledgment

The work presented in this paper was funded by the German Federal Ministry of Education and Research (BMBF) in the Software-Cluster project SINNODIUM (www.software-cluster.org) under grant number "01IC12S01D". The authors assume responsibility for the content.

References

[1] P. Bastian, K. Dagmar, D. Tanja, and S. Albrecht, "Reducing non-primary task distraction in cars through multi-modal interaction," it - Information Technology Methoden und innovative Anwendungen der Informatik und Informationstechnik, vol. 54, no. 4, Aug. 2012, pp. 179–187.

[2] R. J. Lehmann, R. Reiche, and G. Schiefer, "Review: Future internet and the agri-food sector: State-of-the-art in literature and research," Comput. Electron. Agric., vol. 89, 2012, pp. 158–174.

[3] A. Kaloxylos et al, "The use of future internet technologies in the agriculture and foodsectors: Integrating the supply chain," Procedia Technology, vol. 8, 2013, pp. 51 – 60.

[4] "The iGreen Project," Available: http://www.igreen-projekt.de/ [retrieved: Aug, 2014].

[5] "The iGreen Project - Data Services," Available: http://data.igreen-services.com/ [retrieved: Aug, 2014].

[6] "argoXML," Available: http://www.agroxml.de/ [retrieved: Aug, 2014].

[7] "Linked Data," Available: http://linkeddata.org/ [retrieved: Aug, 2014].

[8] R. T. Fielding, "REST: Architectural styles and the design of network-based software architectures," Doctoral dissertation, University of California, Irvine, 2000.

[9] D. Martini, M. Schmitz, R. Kullick, and M. Kunisch, "Fitting information systems to the requirements of agricultural processes: A flexible approach using agroxml and linked data technologies," in Proc. of the Int. Conf. on Agricultural Engineering (AgEng 2010), 2010, pp. 157–163.

[10] A. Gerber and M. Klusch, "AGRICOLA - Agenten für mobile Planungsdienste in der Landwirtschaft (in English: AGRICOLA - Agents for Mobile Planning Services in Agriculture)," KI, vol. 18, no. 1, 2004, pp. 38–42.

[11] M. Klusch and A. Gerber, "Dynamic coalition formation among rational agents," IEEE Intelligent Systems, vol. 17, no. 3, May 2002, pp. 42–47.

[12] M. Reinecke, H.-P. Grothaus, G. Hembach, S. Scheuren, and R. Hartanto, "Dynamic and distributed infield-planning system for harvesting," in Proc. of American Society of Agricultural and Biological Engineers Annual International Meeting (ASABE-2013), vol. 1. Curran Associates, Inc., 2013, pp. 156–160.

[13] "Farmpilot," Available: http://www.farmpilot.de/com/ [retrieved: Aug, 2014].

[14] "356FarmNet," Available: http://www.365farmnet.com/en/ [retrieved: Aug, 2014].

[15] "Farming 4.0," Available: http://www.laboratories.telekom.com/public/english/innovation/exponate/pages/industrie-4.0.aspx [retrieved: Aug, 2014].

[16] "SAP HANA Cloud Platform," Available: http://www.sap.com/pc/tech/cloud/software/hana-cloud-platform-as-a-service/index.html [retrieved: Aug, 2014].

[17] "Bing Traffic API," Available: http://msdn.microsoft.com/en-us/library/hh441725.aspx [retrieved: Aug, 2014].

[18] "OpenWeatherMap," Available: http://openweathermap.org/ [retrieved: Aug, 2014].

[19] D. Sonntag, D. Porta, and J. Setz, "HTTP/REST-based meta web services in mobile application frameworks," in Proc. of the 4nd Int. Conf. on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM '10). IARIA, 2010, pp. 170–175.

[20] D. Porta, "A novel, community-enabled mobile information system for hikers," in Proc. of the 2nd Int. Conf. on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM '08). IEEE Computer Society, 2008, pp. 438–444.

[21] "SIP: Session Initiation Protocol," Available: http://tools.ietf.org/html/rfc3261 [retrieved: Aug, 2014].

[22] "Session Initiation Protocol (SIP) Extension for Instant Messaging," Available: http://tools.ietf.org/html/rfc3428 [retrieved: Aug, 2014].

[23] R. Neßelrath and D. Porta, "Rapid development of multimodal dialogue applications with semantic models," in Proc. of the 7th IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems (KRPDS '11), 2011, pp. 37–47.

[24] R. Neßelrath and M. Feld, "SiAM-dp: A platform for the model-based development of context-aware multimodal dialogue applications," in Proc. of the 10th Int. Conf. on Intelligent Environments (IE '14). IEEE Computer Society, 2014, in press.

[25] "Eclipse Modeling Framework Project (EMF)," Available: http://www.eclipse.org/modeling/emf/ [retrieved: Aug, 2014].

[26] "Media Resource Control Protocol Version 2 (MRCPv2)," Available: http://tools.ietf.org/html/rfc6787 [retrieved: Aug, 2014].

[27] H. H. Clark and E. F. Schaefer, "Contributing to discourse," Cognitive Science, vol. 13, no. 2, 1989, pp. 259–294.

[28] A. Castronovo, M. Feld, M. Moniri, and R. Math, "The ConTRe (continuous tracking and reaction) task: A flexible approach for assessing driver cognitive workload with high sensitivity," in Adjunct Proc. of the 4th Int. Conf. on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '12). ACM, 2012, pp. 88–91.

 ISBN: 978-1-61208-353-7

A Location Management System for Destination Prediction from Smartphone Sensors

Sun-You Kim, Sung-Bae Cho
Dept. of Computer Science
Yonsei University
Seoul, Korea
{sykim@sclab.yonsei.ac.kr, sbcho@yonsei.ac.kr}

***Abstract*— Several applications based on smartphones have been developed for user's requirements. Among them, the location based services (LBS) are demanding to many people. As a result, location management systems become more important to manage the locations and acquire new information in places because location information can be obtained in diverse sensors. This paper proposes a location management system which manages the information of location and predicts a future location and moving path from the current sensor values. The proposed system consists of five modules. Three modules perform to manage the information of location and user's context, and two modules predict future information about the locations. In order to show the feasibility of the proposed system, we conducted the evaluation on each module with a real dataset collected from mobile devices.**

Keywords-Location management; location-based service; destination prediction; hidden Markov model

I. INTRODUCTION

As many people use smartphones, such as Android phones, the smartphone market grows rapidly [1]. Smartphone is easy to develop the applications using device sensors because various mobile OSs provide open platform. It enables developers to access easily user's locations and sensor data.

In many cases, the applications of smartphones such as Location-Based Service (LBS) use location information in mobile device [2][3][4]. With the increase of LBS, many researchers investigate on the locations using smartphone. As a result, it is necessary to manage the data of locations and forecast user's future locations based on the sensor information.

In a mobile phone, the system which manages location information and predicts future places should include the following functions. Each function is classified into two categories. First category is management of location information. It manages user's Point Of Interests (POI) and finds new locations which are meaningful for user. Second category is a service using location information. It provides new information such as user's current location, moving time and destination. This system, which includes two categories, is composed of the following functions.

1) Data collection: It collects sensor data and user's information. Using data collection, the system generates new information.

2) Location extraction: It extacts location, which is a frequently visited place or meaningful place for user. By extracting meaningful location, location management system can recommend to register symbolic location for user.

3) Location recognition: It means to classify where user's current location is. The system is able to offer a suitable service by finding user's location exactly.

4) Prediction of departure time: It forecasts when user departs. The new information required for future can be allowed to the user at the appropriate time by using it.

5) Prediction of destination: It predicts the user's place and path of desitination not yet reached. Destination prediction should use all information of location manager system because it requires many data.

In location management system for destination prediction, user's mobile phone gathers data using sensors. And it extracts the POI, which is meaningful for user, using collected data. Extracted locations are managed by this system. Also, it classifies user's current location and predicts user's departure time in present location. Based on this information, this system predicts user's future location.

Some researchers have studied on the prediction of destination using various smartphone sensors. Do *et al.* proposed a location prediction method using linear regression, logistic regression and random forest [5]. It uses the information of mobile device, such as GPS, Bluetooth, call log, application history and proximity. Lu *et al.* developed a forecasting method based on Support Vector Machine (SVM) using GPS, acceleration, Bluetooth, Wi-Fi and call log in mobile device [6]. Kim *et al.* used the Bayesian network for destination prediction [7]. Its input values are location information, visiting time, staying time and user's gender.

Gambs *et al.* proposed a method for prediction of destination using mobility Markov chains with Point Of Interest (POI) sequence [8]. Liao *et al.* developed a destination forecasting system based on hierarchical dynamic Bayesian network with GPS [9]. Simmons *et al.* proposed a destination prediction framework to use hidden Markov model using GPS sensor and map database in mobile environment [10]. However, previous studies did not build a system to manage the whole information related with locations. This paper proposes a system to manage several sensor data related with location and predict the future location.

The rest of this paper is organized as follows. Section II describes each module that constructs the proposed system, which aims at the management and service. Section III addresses the result of experiments on each component.

 ISBN: 978-1-61208-353-7

Section IV summarizes the location management system for the destination prediction and draws a conclusion.

II. THE PROPOSED SYSTEM

The proposed system consists of two functions, which are the management of location and the service with location. The location management is conducted by data collection and location extraction. The location service is provided by transportation recognition, location recognition, moving time prediction and destination prediction.

To manage locations in this system, first, this system collects sensor data and user's input information in data collection module. Then, it identifies user's meaningful locations using location extraction module. Extracted locations are managed in symbolic locations.

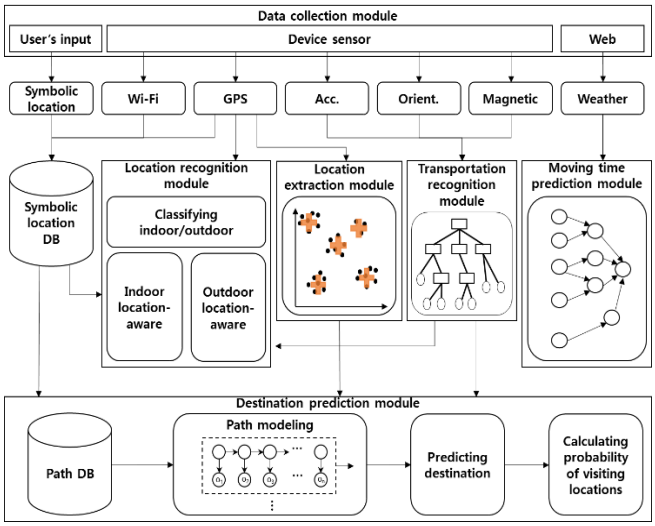


Figure 1. System overview

This system provides three service functions. It offers user's present location using location recognition module, and predicts departure time in future with moving time prediction module. Finally, it gives information about user's future location and moving trajectory using destination prediction module. Figure 1 illustrates an overview of the system.

A. Sensor data collection

We collected raw data such as acceleration, orientation, magnetic field, GPS information and Wi-Fi information by using Android phone API. GPS information includes latitude, longitude, accuracy, number of satellites, and SNR (Signal to Noise Ratio). Wi-Fi information contains SSID (Service Set Identifier), mac address, and RSSI (Received Signal Strength Indication). Most of data are collected twice per one second. GPS information is collected whenever the state of GPS sensor is changed. When user registers a location, Wi-Fi information is obtained.

Also, we collected weather information using Yahoo weather API. The weather data express temperature and weather state of 47 types. We transform the weather state into 7 types. The data specification and frequencies are summarized in Table I.

TABLE I. SENSOR DATA FOR DESTINATION PREDICTION

Sensor Type	Frequency	Description
Acceleration	Two times for one minute	3 axis acceleration (-2g~2g)
Orientation	Two times for one minute	Orientation, pitch, roll
Magnetic field	Two times for one minute	3 axis magnetic field (uT)
GPS	When GPS state is changed	Latitude, longitude, accuracy, SNR, number of satellites
Wi-Fi	When user registers a location	SSID, mac address, RSSI
Weather	Once for five minutes	Temperature (℃), Weather state (7 types)
Time	Two times for one minute	Current time

Acceleration, orientation and magnetic field can be used to check user's transportation mode. GPS is the information necessary to perform the location extraction and the location recognition. Wi-Fi is used to recognize indoor location, and weather is input of prediction of moving time.

This system stores the names of symbolic location, which are entered by user. Each location name is connected with Wi-Fi information and GPS information such as latitude and longitude.

B. Location extraction

For inducing to register user's meaningful locations, this system extracts candidate locations, which can be meaningful. Because it is impossible to use all GPS data, which are very big size in mobile device, the locations extracted are transformed into symbolic locations.

Previous studies about location extraction used *k*-means clustering, which is density-based algorithm [11]. However, *k*-means clustering should determine the number of '*k*'. In the location extraction, '*k*' means the number of locations. We do not know the number of locations extracted in advance. User's meaningful locations follow a Gaussian distribution [12][13]. However, because the criteria of density in *k*-means clustering are ambiguous, *k*-means clustering is not suitable in the location extraction problem.

Instead of it, we use G-means clustering method [14]. This is a clustering method to test each cluster in Gaussian distribution through statistical verification and repeat the *k*-means clustering until all clusters follow the Gaussian distribution. The statistical verification is performed by Anderson-Daring test, which is represented by the following equation:

$$A^2(Z) = \frac{-1}{n}\sum_{i=1}^{n}(2i-1)[\log(z_i) + \log(1 - z_{n+l-i})] - n \quad (1)$$

Here, x_i is transformed into a value of average 0 and variance 1. When $x_{(i)}$ is the i-th value, we define $Z_i = F(x_{(i)})$. In this equation, F is $N(0,1)$ Cumulative Distribution Function (CDF). Using the latitudes and longitudes, which are obtained from GPS sensor, this system performs G-means clustering.

After the clustering, user's key locations are extracted. The extracted locations contain the information of latitude and longitude. Figure 2 is an example of the result of location extraction.

Extracted locations is used by two objects. Extracted locaitons of stop state are user's meaningful locations, and all the extracted locations are used for constructing a path. A path has many points of location. Therefore, it is nessesary to reduce locations, which express a path. We use the extracted locations to make a path.

Figure 2. An example of location extraction

C. Transportation type recognition

The transportation recognition needs to judge whether to perform the location recognition. It is necessary as an input for the destination prediction. To classify the moving state, we transformed sensor data such as acceleration, orientation, and magnetic field using decision tree algorithm. In some cases, decision tree shows better performance than alternative algorithms to process time series data such as acceleration [15].

For using sensor data in mobile device for decision tree, we extract some features such as the difference between previous and current sensor values, average sensor value for a specific period, and standard deviation of the sensor value for a specific period. Following (2)~(4) are the equations for pre-processing in decision tree.

$$sum_X = \sum_{i=1}^{N} \sqrt{(x_i - x_{i-1})^2} \quad (2)$$

$$mean_X = (\sum_{i=1}^{N} \sqrt{(x_i - x_{i-1})^2})/N \quad (3)$$

$$std_X = \sqrt{\sum \sqrt{((x_i - x_{i-1})^2 - mean_x)^2}} \quad (4)$$

In the equations, X means specific sensor, and x_i represents the i-th sensor value. Equation (2) is the summation of the difference between previous value x_{i-1} and current value x_i from a sensor X. Equation (3) represents the mean of the difference value which is calculated by equation (2). Equation (4) denotes the standard deviation of the difference between previous and current sensor data. Acceleration, orientation and magnetic field exist in 3-axis. This method uses average and standard deviation as the feature of decision tree. Each sensor is calculated by three averages and three standard deviations [16].

Using these features, decision tree is generated, which classifies the transportation types about input values, which are acceleration, orientation and magnetic field. Transportation types are staying, walking, running, and in vehicle.

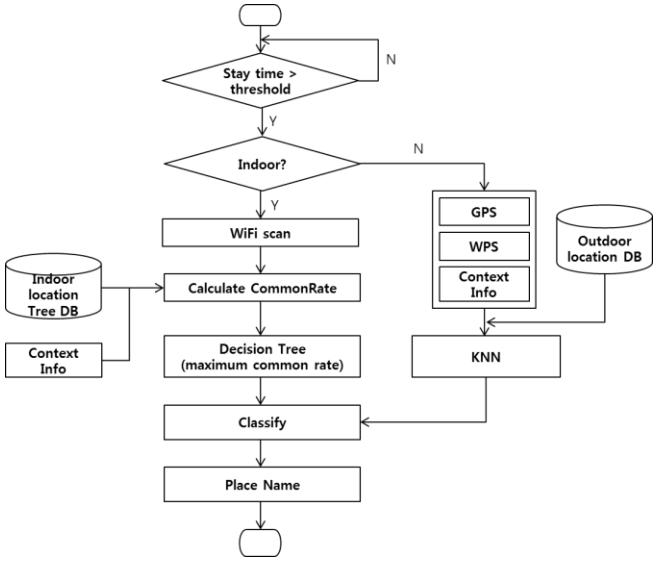


Figure 3. Process of location recognition

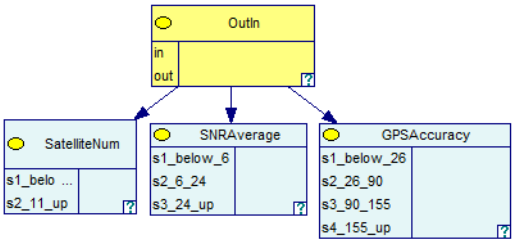


Figure 4. Model of na ïve Bayes classifier for distinction of indoor / outdoor

D. Location recognition

In the location management system, it is an important problem to identify user's location. In order to know user's place, this system includes the location recognition module. For recognizing an outdoor place, it is easy to identify user's current location by using GPS value of latitude and longitude. However, in an indoor location, the signals of GPS satellite cannot pass wall of the buildings. Therefore, we should use a different method for an indoor location. Recently, in the field of the location recognition, many researchers use Wi-Fi AP (Access Point) [17][18]. Thanks to the ubiquitous Internet, it does not need additional configuration, so that we use Wi-Fi AP information to identify user's current location in indoor. The process of the location recognition is shown in Figure 3.

For accurate recognition of user's current location, this system distinguishes between indoor and outdoor. Next, if the result is outdoor, it performs outdoor location recognition using the k-nearest neighbor. Otherwise, it executes the algorithm for indoor location recognition using decision tree.

1) Classification of indoor/outdoor: To discriminate indoor and outdoor, we use the na ïve Bayes classifier, which

is a fast and simple inference method. Naïve Bayes is a probabilistic model, which is based on Bayes rule under the strong conditional independence assumption. In the proposed method, the input of naïve Bayes includes GPS information such as number of statellite, SNR, and GPS accuracy. The number of statellite converts to 2 discrete values, which are seperated by a threshold of 11. SNR is made by 3 values, and GPS accuracy is transformed into 4 values. The preprocessing values are entered by inputs of naïve Bayes model. The naïve Bayes model used is shown in Figure 4.

2) Outdoor location recognition: *k*-NN method is used to identify outdoor place. *k*-NN classifies the data by performing majority vote with the *k* neighbors closest to the input data. In the stored symbolic locations, after selecting *k* GPS points closest to the current position, it selects a location, which is the largest number in *k*-point locations.

3) Indoor location recognition: For indoor place, recogintion method cannot use the GPS sensor because of the unavailable GPS signals. So, we adopt the Wi-Fi finger print method based on the decision tree. Decision tree is generated by Wi-Fi information such as RSSI and MAC address, which is from the previously stored symbolic location. It makes a result, which is symbolic location to use decision tree with the new input of Wi-Fi information.

E. Prediction of moving time

If the system predicts the departure time, it can offer information, which is necessary for user in advance. To predict user's departure time based on context has the disadvantage, which is high error rate because it should determine a specific time in 24 hours. So, our system calculates user's staying time and predicts departure time after inferring how long the user stays at the current place.

The system uses Bayesian network for the prediction of user's moving time. Bayesian network is a stochastic model, which has a Directed Acyclic Graph (DAG) structure and Conditional Probability Tables (CPTs). Bayesian network is used to handle the uncertainty with probability. It supports the efficient probability calculation based on conditional independence assumption. The Bayesian network modeled in this system is as shown in Figure 5.

This network is a structure, which has root nodes, intermediate nodes, and observation nodes. Observation nodes are for input of system time, current location and weather information. Intermediate nodes calculate the probability values based on evidence values of observation node. The root nodes are computed by intermediate nodes. Each root node means at least 1 hour, 1~2 hours, 2~4 hours, and 4~8 hours, which are the result values in this module.

F. Destination prediction

The destination forecasting informs the location that user has reached at the last time and calculates the probability of visiting location, which is intermediate location of the path. Destination prediction uses Hidden Markov Model (HMM) [19]. HMM is a statistical model characterized by a Markov process with unknown parameters, modeling observations to

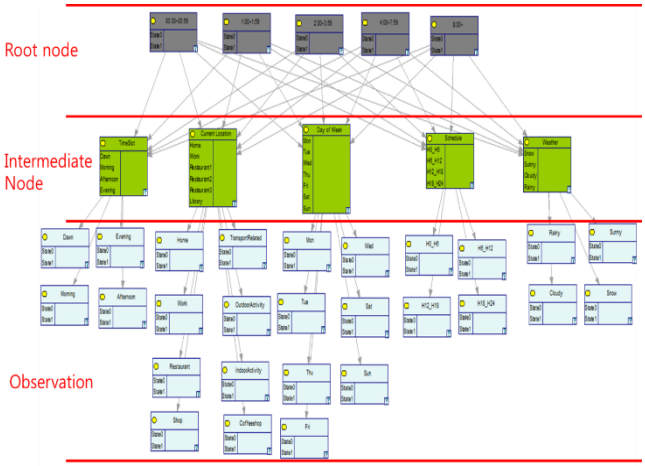


Figure 5. Model of Bayesian network to predict moving time

determine these hidden parameters. HMM is a widely used technique that stochastically models sequence data of the time series. It is mainly composed of the state transition probabilities, which select the observation value at each state. HMM composed by state transition probability A, probability distribution of observed value B, and probability distribution of initial state Π. One HMM, λ, is represented as (5).

$$\lambda = \{A, B, \Pi\} \quad (5)$$

In the proposed system, destination prediction uses extracted locations, transportation type and time. This module consists of the three parts.

1) Building path model: The HMM of path has information about the start and end points of a path. HMM is built by number of pairs of source and destination locations. Path information is made up by sequence, which contains the extracted location, the transportation type and the time that is quantized. The HMM included the path information is learned by Baum-Welch algorithm [20], which is a learning method typical to represent probabilistic information of multiple sequences.

2) Predicting destination: About the new input, which is information of departure or sub-path, this method evaluates all HMMs for finding the path of the highest similarity. Evaluation is conducted by the forward algorithm, which is basic method to check the similarity between a sequence and an HMM.

3) Calculating visiting probability: Based on destination. which is determined by optimal path, the probabilities of visiting destination and intermediate locations are calculated. First, we find out an optimal path sequence, which is the same as a departure and a destination of optimal HMM and includes current path, from the path repository. By determining a sequence of future movements of the location, it finds out the optimal state sequence from the HMM and calculates the probability of visiting locations based on the optimal state sequence. The calculation of the optimal state

 ISBN: 978-1-61208-353-7

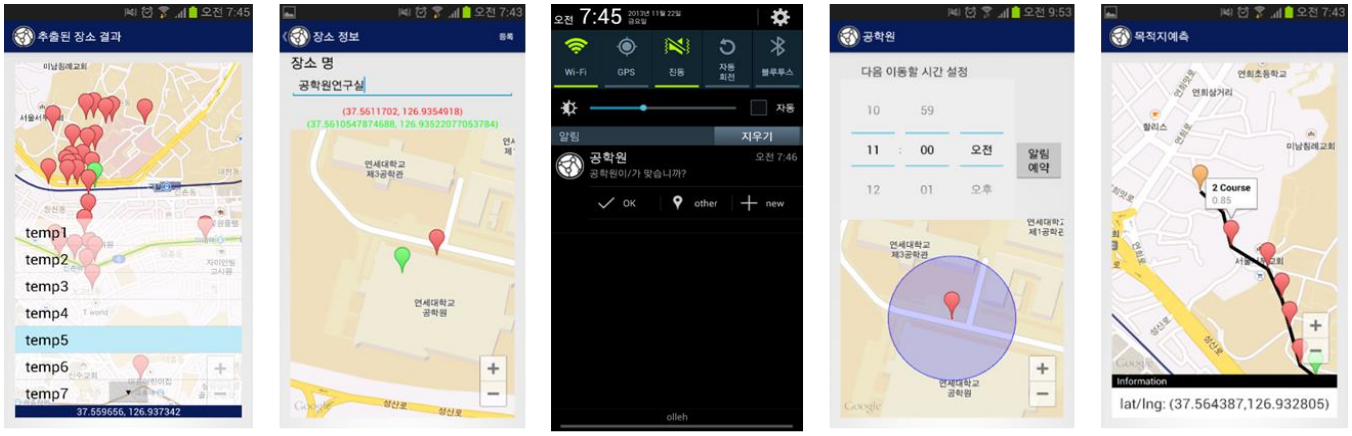


Figure 6. Interface of the proposed system

sequence is conducted by Viterbi algorithm [21]. It can determine the most probable sequence of states in optimal state sequence.

III. EXPERIMENTAL RESULTS

In order to evaluate the proposed system, we applied it to a SAMSUNG Galaxy S4 Android phone and conducted experiments. Data set is gathered for four months from ten people. The specification of data set is shown in Table II.

We implement this system using Android API. In order to speed up the operation, all core modules are implemented by using Android NDK API. The NDK is a toolset that allows to implement parts of App using native-code languages such as C and C++ [22].

TABLE II. DESCRIPTION OF DATASET

	#Location	#Path	Size of storage
User 1	16	193	2.44GB
User 2	20	268	2.62GB
User 3	32	149	1.46GB
User 4	50	288	3.41GB
User 5	42	309	3.66GB
User 6	32	233	1.34GB
User 7	28	236	1.48GB
User 8	24	294	3.21GB
User 9	36	237	2.37GB
User 10	14	189	2.08GB

The interfaces for each component are illustrated in Figure 6. Screen of (a) is an interface, which shows the result of the location extraction. The interface for entering the symbolic location in the system is shown in (b). When user stays for a certain period of time, this system notifies a place where it is now for user, such as (c). After the location-awareness, in current location, the screen predicted when the user might depart in (d). When it comes to the predicted starting point for the user, predicting a destination and a path to reach the destination are shown as in (e).

A. Location recognition

In this system, location recognition module uses the methods of classifying indoor or outdoor, outdoor location recognition and indoor location recognition. To evaluate this module, the performance of each method is measured.

Figure 7 illustrates the performance of discriminating indoor or outdoor for the ten users. This experiment is evaluated by using naïve Bayes classifier with 10-fold cross-validation. In the experiment, accuracy shows 96.6% in average.

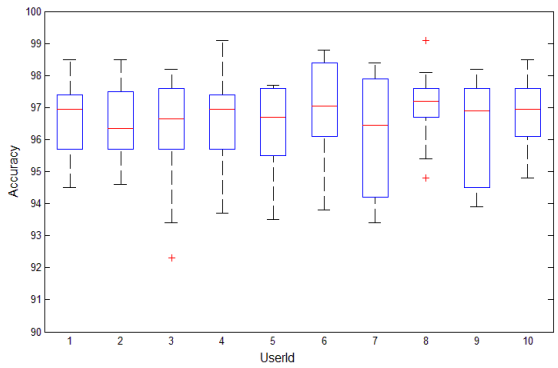


Figure 7. Performance evaluation of classifying indoor/outdoor

Figures 8 and 9 show the accuracy of outdoor and indoor location recognitions, respectively. These experiments are conducted by 10-fold cross-validation. Outdoor location recognition result of the *k*-NN method shows average accuracy of 98.96%. However, in case of indoor location recognition, we obtain 95.36% of average accuracy, which is relatively low.

B. Prediction of departure time

To identify the prediction of departure time can offer new input at appropriate time for user. In this system, we evaluate

the accuracy for the ten users. The average accuracy results in 80.6% as shown in Figure 10. This system has high usability by customizing user's own moving time.

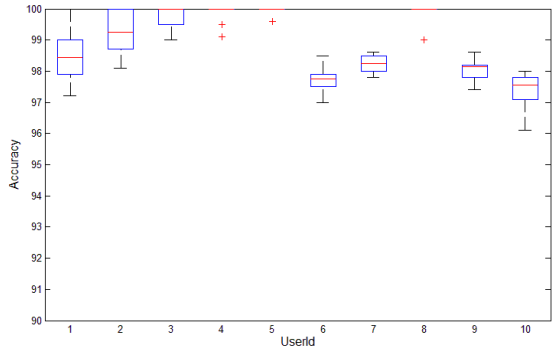


Figure 8. Performance evaluation of outdoor location recognition

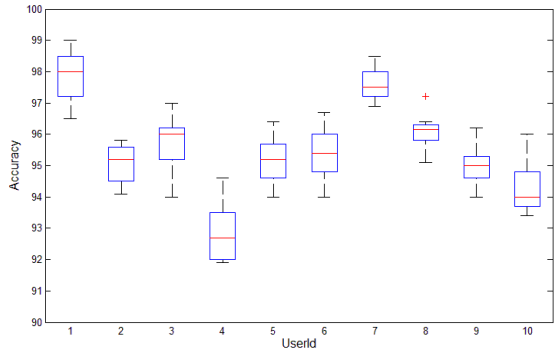


Figure 9. Performance evaluation of indoor location recognition

C. *Prediction of desitinaion*

In order to evaluate the accuracy of the proposed detination prediction method, we measured accuracy according to the path of progress. Prediction result of the advancement of the path is illustrated in Figure 11. It is a result that is preformed by 10-fold cross-validation. We trained the path models using 90% paths and measured accuraccy using 10% paths. Looking at the prediction accuracy in accordance with the progress of the path, as the path is largely moves, it can be seen that the prediction accuracy becomes higher because the information of the location movement is increased. For 0% progression of the path, which is capable of predicting only location information from the starting place, HMM showed accuracy of 57.96% in average only with the information of departure.

IV. CONCLUDING REMARKS

In this paper, we have proposed a system to manage location information and predict user's destination with outputs of modules using smartphone sensors such as GPS, Wi-Fi, acceleration, orientation, and magnetic field. The proposed system consists of location extraction, transportation type recognition, location recognition, prediction of moving time and prediction of destination. Destination prediction preforms calculating the similarity between new path and probabilistic model which contains information of the path with different modules, and finds a user's destination. Experimental results with real data collected from ten people show the usefulness of the proposed system.

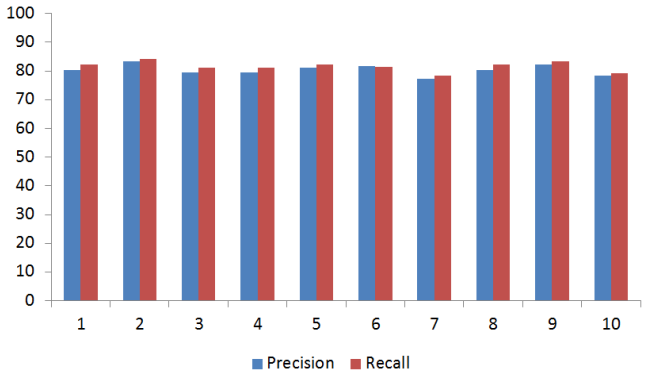


Figure 10. Performance evaluation of departure time prediction

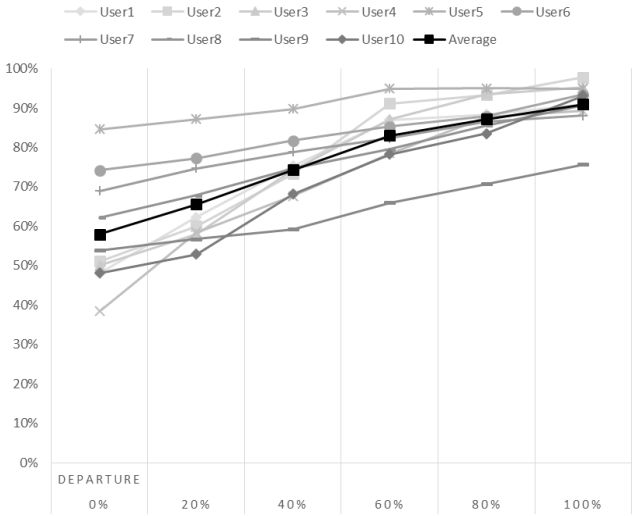


Figure 11. Performance evaluation of destination prediction

ACKNOWLEDGENMENTS

This work was supported by Samsung Electronics, Inc.

REFERENCES

[1] M. Butler, "Android: Changing the Mobile Landscape," IEEE Pervasive Computing, vol. 10, no. 1, 2011, pp. 4-7.

[2] Y. Chon and H. Cha, "LifeMap: A Smartphone based context provider for Location-based Services," IEEE Pervasive Computing, vol. 10, no. 2, 2011, pp. 58-67.

[3] S. Bell, W. R. Jung, and V. Krishnakumar, "WiFi-based enhanced positioning systems: Accuracy through mapping, calibration, and classification," Proc. of the 2nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness, 2010, pp. 3-9.

[4] N. Brouwers and M. Woehrle, "Dwelling in the canyons: Dwelling detection in urban environments using GPS, Wi-Fi, and jeolocation," Pervasive and Mobile Computing, 2012, pp. 1-16.

[5] T. M. T. Do and D. Gatica-Perez, "Contextual conditional models for smartphone-based human mobility prediction,"

 ISBN: 978-1-61208-353-7

Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 2012, pp. 163-172.

[6] Z. Lu, Y. Zhu, V. W. Zheng, and Q. Yang, , "Next Place Prediction by Learning with Multiple Models," In Proceedings of the Mobile Data Challenge Workshop, 2012.

[7] B. Kim, J. Y. Ha, S. Lee, S. Kang, and Y. Lee, "AdNext: A Visit-Pattern-Aware Mobile Advertising System for Urban Commercial Complexes," In Proceedings of the 12th Workshop on Mobile Computing Systems and Applications, 2011, pp. 7-12.

[8] S. Gambs, M. O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility Markov chains," In Proceedings of the 1st Workshop on Measurement, Privacy, and Mobility, 2012, pp. 1-6.

[9] L. Liao, D. Fox, and H. Kautz, "Learning and Inferring Transportation Routines," In Proceedings of the National Conference on Artificial Intelligence, pp. 348-353, 2004.

[10] R. Simmons, B. Browning, Y. Zhang, and V. Sadekar, "Learning to Predict Driver Route and Destination Intent," In Proceedings of the IEEE Intelligent Transportation Systems Conference, 2006, pp. 127-132.

[11] A. J. Dou *et al*, "Data clustering on a network of mobile smartphones," IEEE/IPSJ Symposium on Applications and the Internet , 2011, pp. 118-127.

[12] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," in Proc. IEEE Infocom, 2006, pp. 1-13,.

[13] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. Campbell, "Nextplace: A spatio-temporal prediction framework for pervasive systems," Pervasive Computing, 2011, pp. 152-169.

[14] G. Hamerly and C. Elkan, "Learning the k in k means," Advances in Neural Information Processing Systems, vol. 16, 2004, pp. 281.

[15] J.-K. Min and S.-B. Cho, "Mobile Human Network Management and Recommendation by Probabilistic Social Mining," IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 41, no. 3, 2011, pp. 761-771.

[16] Y.-S. Lee and S.-B. Cho, "An Efficient Energy Management System for Android Phone Using Bayesian Networks," Distributed Computing Systems Workshops (ICDCS), 2012, pp. 102-107.

[17] I. Bisio, F. Lavagetto, M. Marchese, and A. Sciarrone, "GPS/HPS-and Wi-Fi Fingerprint-Based Location Recognition for Check-In Applications Over Smartphones in Cloud-Based LBSs," IEEE Transactions on Multimedia, vol. 15, no. 4, 2013, pp. 858-869.

[18] D. H. Kim, J. Hightower, R. Govindan, and D. Estrin, "Discovering semantically meaningful places from pervasive RF-Beacons, " Proc. of Int. Conf. on Ubiquitous computing, 2009, pp. 21-30.

[19] Y. J. Kim and S.-B. Cho, "A HMM-based location prediction framework with location recognizer combining k-nearest neighbor and multiple decision trees," 8th International Conference on Hybrid Artificial Intelligent Systems, 2013, pp. 618-628.

[20] L. E. Baum, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math. Statist, vol. 41, 1970, pp.164 –171.

[21] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, 1989, pp. 257–286.

[22] http://developer.android.com/tools/sdk/ndk/index.html: July, 2014.

 ISBN: 978-1-61208-353-7

An SMT-based Accurate Algorithm for the K-Coverage Problem in Sensor Network

Weiqiang Kong, Ming Li, Long Han, and Akira Fukuda
Graduate School of IS&EE, Kyushu University, Japan.
weiqiang@qito.kyushu-u.ac.jp, {ziqiangliming, l-han}@f.ait.kyushu-u.ac.jp,
fukuda@ait.kyushu-u.ac.jp

***Abstract*—In the context of wireless sensor network (WSN), the K-Coverage problem denotes that each point in a certain network area is covered by at least K sensors at the same time so as to guarantee the quality of services provided by the WSN. In this paper, we first propose a bottom-up modeling method for the K-coverage problem. Based on this method, we investigate a set of iteratively-applicable simplification techniques for simplifying the problem. Furthermore, we propose a satisfiability modulo theory (SMT) based algorithm for computing an accurate solution to the K-coverage problem. Experimental results have shown that our proposed simplification techniques and algorithm provide sufficiently satisfiable performance with respect to both computing speed and problem size.**

Keywords-K-coverage; wireless sensor network; satisfiability modulo theory; accurate algorithm;

I. INTRODUCTION

Wireless sensor network (WSN) is an infrastructure comprised of sensing (measuring), computing, and communication elements that gives an administrator the ability to instrument, observe, and react to events and phenomena in a specified environment [1].

WSN has been developing rapidly in recent years and been applied to many fields including military, business, and agriculture, etc. Coverage problems are one of the most active research topics related to WSN. It is generally necessary to deploy multiple sensors to cover an entire WSN area so as to provide services within the area. Each sensor used in WSN has a limited sensing radius range. A point in a WSN area is said to be covered if it is within the radius range of a sensor. If a point is covered by only one sensor, then it is said to be 1-covered. Consequentially, a WSN area is said to be K-covered if every point in the area is covered by at least K sensors at the same time. Such a restriction is called the K-coverage problem. Due to costs and/or interference among multiple sensors, it is often not practically feasible to deploy an arbitrarily large number of sensors to simply fulfill the K-coverage restriction. How to deploy sensors in a reasonable way so as to decrease the number of sensors while fulfilling the K-coverage restriction, is the research topic of this paper.

The K-coverage problem is a typical combinatorial optimization problem. As the size increase of the target WSN area, the scale of solution space grows exponentially. In a large-scale K-coverage problem, it is generally difficult to compute optimal solutions. Therefore, existing algorithms usually circumvent this problem by sacrificing coverage rate or solution accuracy to improve algorithms' performance [2], [3], [4]. In the case of sacrificing coverage rate, a compromised coverage rate is used to replace the strict 100% rate; in the case of sacrificing solution accuracy, it is allowed to deploy redundant sensors.

In this paper, we investigate high performance algorithms for the K-coverage problem, which should not sacrifice coverage rate (i.e., fulfill strictly 100% rate) and should find the minimum number of sensors. Our contributions made in this paper are as follows: (1) we proposed a bottom-up modeling method for the K-coverage problem; (2) we proposed a set of iteratively-applicable simplification techniques for simplifying the K-coverage problem; (3) furthermore, we proposed an SMT-based efficient algorithm for computing accurate solutions to the K-coverage problem. As shown by our preliminary experiments, the simplification techniques as well as the algorithms provide sufficiently satisfiable performance.

The paper is organized as follows. Section II introduces the modeling method of K-coverage problem; Section III investigates a set of methods for simplifying the K-coverage problem; Section IV describes our proposed SMT-based accurate algorithms and experiment evaluation of the algorithms; Section V concludes the paper.

II. MODELING OF THE K-COVERAGE PROBLEM

Candidate-Points. In a practical environment, positions in which sensors are allowed to be deployed are usually limited. For example, in an indoor environment, they are often deployed on walls or pillars. Therefore, we abstract the positions in which sensors can be deployed and call them as "candidate points for sensor deployment" (called "candidate-points" for simplicity).

For each candidate-point, we can only deploy one sensor (or not deploy). The number and location of candidate-points depend on the physical environment of the target area. As one of the most important input-data of K-coverage algorithms, the number of candidate-points determines directly the complexity and solution space of the problem. For each candidate-point, we can make a choice to deploy a sensor there or not. Therefore, the number of choices is equal to 2^N, where N denotes the number of candidate-points.

 ISBN: 978-1-61208-353-7

Observation-Points. In addition to "candidate-points", to represent the target area under consideration, we define another type of points, which are called as "observation-points". An observation-point is said to be covered by a sensor if and only if the point is within the radius range of the sensor. With this definition, we say that an observation point is K-covered if it is covered by K sensors. If all the observation-points in the target area are covered by sensors, we say that the area is covered by sensors. The distribution of observation-points can be even, or the distribution can be customized according to actual requirements.

The density of observation-points reflects the accuracy of the model, and generally, the higher the density, the more accurate the coverage by the sensors of the target area is. However, as the increase of the number of observation-points, the complexity of the problem also increases. The K-coverage problem can be defined again based on the above two definitions. That is, to find the minimum number of sensors, the deployment of which can satisfy that all observation-points in the target area is K-covered by sensors.

A. Basic Algorithms for Computing K-Coverage

We can easily come up with a simple exhaustive algorithm to compute this problem. The pseudo-code of the algorithm is shown in Algorithm 1. (Actually, this algorithm is also mentioned in [5] and is called as the Original Combination Algorithm). We assume there are a set O of observation-points and a set N of candidate-points; We use C to denote the set of all possible combination of sensor deployment; We use function $k\text{-}covered(c)$ to evaluate if a sensor deployment $c \in C$ satisfies K-coverage, which returns `true` when c satisfies and `false` otherwise; We use function $num(c)$ to denote the number of sensors in c; We use min to hold the minimum number of sensors that satisfies K-coverage; We use $outputDeploy$ to hold the sensor deployment that satisfies K-coverage and has the minimum number of sensors.

Algorithm 1. A Simple Exhaustive Algorithm for K-Coverage

```
1. Input: The sets O, N, and C.
2. Output: The sensor deployment outputDeploy
3.
4. for each c ∈ C
5.     if k-covered(c) == true then
6.         if num(c) < min then
7.             min = num(c);
8.             outputDeploy = c;
9.         end if
10.    end if
11. end for
12. return outputDeploy
```

However, we can notice that there are a lot of performance deficiencies in Algorithm 1: let us assume that there are N candidate-points, the time complexity of the algorithm is $O(2^N)$ since the number of all possible combination of sensor-deployment is 2^N, and in the worst case, we need to check every combination of them. Even if there are some ways of improving the performance of this algorithm, e.g., by checking combinations in an ascending order of the number of sensors, it is still difficult to change the time complexity essentially. The time complexity is still O(2^N).

Therefore such kind of simple exhaustive algorithms, which are based on the combinations of sensor deployment, is not practically feasible. As the increase of the number of candidate-points, the complexity of such algorithms will grow exponentially.

In addition, there is one more reason for the low performance of Algorithm 1. In the step of computing whether a sensor deployment c satisfies K-coverage (namely the function $k\text{-}covered(c)$), there are a lot of unnecessary (and thus avoidable) repeated calculation. We explain this computation repetition in more detail below.

We use sets to denote the combination of sensor deployment, where the elements of each set are (sensor) candidate-points. As shown in Figures 1 and 2, let us assume that there are two combinations of sensor deployment $C1$ and $C2$, where $C1 = \{s1, s2, s3, s4\}$ and $C2 = \{s1, s2, s3, s5\}$; also assume that there are two observation-points $o1$ and $o2$ in the target area, and the objective coverage considered is 2-coverage. The difference of $C1$ and $C2$ only lies on $s4$ and $s5$, and $s4$ and $s5$ are in positions far away from the observation-point $o1$. In Algorithm 1, we need to check whether $C1$ and $C2$ can make observation-points $o1$ and $o2$ satisfy 2-coverage. Although $C1$ and $C2$ can both satisfy the 2-coverage for $o1$ and $o2$. But it should be noted that, compared to the sensors observable by $o2$, the sensors that can be observed by $o1$ are the same in the two deployments $C1$ and $C2$. Consequentially, if deployment $C1$ could satisfy the 2-coverage for $o1$, then $C2$ could as well. Therefore, it is actually not necessary to check the coverage for $o1$ twice. We can imagine that the performance loss here due to unnecessary repeated computation is huge.

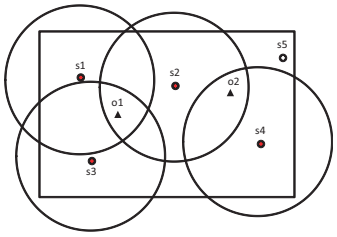


Figure 1. A Sample Combination of Sensor Deployment $C1$

In order to reduce this kind of performance loss analyzed above, we could, instead of directly computing the *combination* of desired sensor deployment, compute in advance the sets of (sensor deployment) candidate-points for each observation-point, which make the observation point satisfy K-coverage. We can see that the sets of candidate-points for

 ISBN: 978-1-61208-353-7

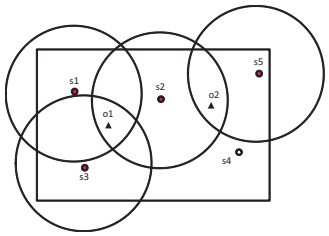


Figure 2. A Sample Combination of Sensor Deployment $C2$

$o1$ is $S11 = \{s1, s2\}$, $S12 = \{s1, s3\}$, $S13 = \{s2, s3\}$; for $o2$, there are two such sets $S21 = \{s2, s4\}$, $S22 = \{s2, s5\}$. Because our goal is to find a minimum-sized number of sensors, we can easily make an assertion that the final result (for our goal of a combination of candidate-points for 2-coverage) must contain and only need to contain one set of $S11$, $S12$, or $S13$, and must contain and only need to contain one subset of $S21$ or $S22$. That is, the final result for the best 2-coverage combination must be one of the follows: $S11 \cup S21 = \{s1, s2, s4\}$, $S12 \cup S21 = \{s1, s2, s3, s4\}$, $S13 \cup S21 = \{s2, s3, s4\}$, $S11 \cup S22 = \{s1, s2, s5\}$, $S12 \cup S22 = \{s1, s2, s3, s5\}$, $S13 \cup S22 = \{s2, s3, s5\}$. Among of them, the combinations with minimum number of sensors are $\{s1, s2, s4\}$, $\{s2, s3, s4\}$, $\{s1, s2, s5\}$, $\{s2, s3, s5\}$. The four sets of sensors are all optimal solutions. As can be seen, we have used here the bottom-up method (by focusing on observation-points) to replace the top-down method (by focusing on candidate-points), and this could avoid the repeated-checking problem mentioned above.

Another advantage of such bottom-up method is that it makes simplification of the K-coverage problem easier.

III. Simplification

We have proposed a set of simplification techniques that can be used as premises of the SMT-based algorithm (to be introduced in Section IV) for simplifying the K-coverage computation problem. These techniques can be applied iteratively for the simplification purpose. In this section, due to space limitation, we only mention the basic ideas of two such techniques. For each observation-point, we can find sets of candidate-points that satisfy the K-coverage restriction. We called such sets as "satisfiable sets" of the observation-point. One observation-point may usually have more than one "satisfiable set". We denote a single "satisfiable set" as S, and the set of all "satisfiable (sub)sets" as SS.

A. Eliminating Certain Observation-Points

Consider two observation-points $o1$ and $o2$, the sets of satisfiable set of $o1$ is $SS1 = \{S11 = \{s1, s2\}, S12 = \{s1, s3\}\}$ and the sets of satisfiable set of $o2$ is $SS2 = \{S21 = \{s1\}, S22 = \{s2\}, S23 = \{s3\}\}$. Note here that, under the restriction of K-coverage, the element number of satisfiable sets of $o1$ is 2, but the element number of $o2$ is 1. Although seemingly strange, this is possible as an intermediate result of simplification, which will be shown in the next subsection.

For the above example, we can see that $S11$ is the parent set of $S21$ and $S22$ (namely $S21, S22 \subseteq S11$); $S12$ is the parent set of $S23$. We can assert that if $o1$ is satisfied for K-coverage, then $o2$ must be satisfied as well. We can see that $o2$ can be eliminated since its satisfiability is contained by that of $o1$ and the existence of $o2$ does not have any effect on the final result. We call $o2$ as a *similar-point* of $o1$, and similar-points have transitivity but does not have commutativity.

B. Must-Be-Chosen Candidate-Points

Let us assume an observation-point $o1$, which has a *unique* satisfiable set $S1$. It can be asserted that every element $s \in S1$ must be contained in the final optimal result $outputDeploy$ (See Algorithm 1), because, otherwise, $o1$ will never be satisfied for K-coverage. We call the candidate-points contained in this unique satisfiable set as "must-be-chosen candidate-points". Generally, when the distribution of observation-points is sparse, such kind of observation-points often exist. Eliminating such observation-points can reduce the number of candidate-points, and consequentially, reduce the complexity of the problem directly.

Consider an example: there are three observation-points $o1$, $o2$, and $o3$; the set of satisfiable sets of $o1$ is $SS1 = \{S11 = \{s1, s2\}, S12 = \{s1, s3\}\}$, the set of satisfiable sets of $o2$ is $SS2 = \{S21 = \{s1, s2\}, S22 = \{s3, s4\}\}$, and the set of satisfiable sets of $o3$ is $SS3 = \{S31 = \{s4, s5\}\}$. We can see that there is only one satisfiable set $S31$ for $o3$, so all the elements in $S31$ should be in the final result. We thus mark $s4$ and $s5$ as Must-Be-Chosen candidate points by assigning $s4 = 1$ and $s5 = 1$ (we use 1 to denote that a point is chosen and 0 otherwise); remove $o3$ from the model, and further simplify $SS1$ and $SS2$. We use $SS1'$ and $SS2'$ to denote the simplified results of $SS1$ and $SS2$, respectively. $SS1' = \{S11 = \{s1, s2\}, S12 = \{s1, s3\}\}$ and $SS2' = \{S21 = \{s1, s2\}, S22 = \{s3\}\}$.

We can see that the element numbers of satisfiable sets of $SS2'$ are inconsistent (two for $S21$ and one for $S22$). As mentioned before, such situation is possible during simplification. Furthermore, observation-point $o2$ has now become a similar-point of $o1$. As introduced in Section III.A, $o2$ can be removed. Such kind of iterative simplification can greatly decrease the complexity of the problem. It is necessary to emphasize that multiple simplification may not need to be carried out in a fixed order. In the final computation algorithm for K-coverage, we will form these simplification techniques in a set and try repeatedly these techniques until no further simplification could be done.

IV. An SMT-based Accurate Algorithm

In general, SAT/SMT solving [6] is a technique to find variable-assignments to all the Boolean variables contained

 ISBN: 978-1-61208-353-7

in a logical formula so as to make the formula `true`, or to determine that there is not such variable-assignments.

In a SAT problem, the formula is a Boolean expression written using only logical operators of $\wedge$, $\vee$, $\neg$, together with Boolean variables and parentheses. Therefore, the problem here is to, through assigning `true` or `false` to each of the Boolean variables, try to find an assignment that makes the entire formula have the value `true`. If all possible variable-assignments have been tried and no such assignment exists, then the problem here is said to be unsatisfiable.

SMT solving is an extension of SAT solving technique. In SMT problems, the formulas can contain variables of other types such as Integers and Reals etc., rather than merely Boolean variables as in SAT problems. An example of SMT problem is given as follows. The formula here is $(x + 1) < 2 \wedge (y \vee z)$, where x is an Integer, y and z are Boolean variables. The formula is satisfiable since there exists at least one variable assignment, for example, $x = 0$, $y =$ `true`, $z =$ `false`. Another formula $(x + 1) = 2 \wedge (x \neq 1)$ is obviously unsatisfiable. There are a lot of well-known SMT solvers such as Z3 [7], Yices [8], and CVC3 [9]. In this paper, we use the Z3 solver since it generally performs best.

A. A SMT-based Algorithm with Z3

To utilize the Z3 SMT solver to compute the K-coverage problem, we need in the first place to encode a K-coverage problem into the input language of Z3 (namely a set of Z3-recognizable formulas). Note that we do not directly encode the original K-coverage problem, but instead, we only encode the intermediate problem after simplification, into a set of formulas written in Z3 input language.

The general idea of encoding the problem (after simplification) is as follows: (Step 1) Declare an Integer variable for each of the candidate-points and define their values as either 0 or 1, where 0 denotes `false` and 1 denotes `true` ; (Step 2) Define a set of logical formulas using these variables, which restricts the conditions that are necessary to fulfill the K-coverage restriction; (Step 3) Define a logical formula, which restricts the condition that the sum (of the values) of all the candidate-points should be smaller or equal to the minimum value (of the sum of the candidate-points) computed so far.

To make it easier to understand, we use a concrete example to explain the idea of encoding. In the example: we consider a 2-coverage problem; there are two observation-points; after applying the simplification technique introduced in section III, the sets of satisfiable sets of these two observation-points are $SS1 = \{S11 = \{s1, s2\}, S12 = \{s3, s4\}\}$ and $SS2 = \{S21 = \{s1, s4\}, S22 = \{s2, s3\}\}$. We encode the problem into the following Z3 language (essentially, we use the SMT-LIB 2.0 language [10], a standard language, formulas written in which can be accepted by most state-of-the-art SMT Solvers including Z3):

```
1.  (declare-const s1 Int)
2.  (declare-const s2 Int)
3.  (declare-const s3 Int)
4.  (declare-const s4 Int)

5.  (assert (or (= s1 0) (= s1 1)))
6.  (assert (or (= s2 0) (= s2 1)))
7.  (assert (or (= s3 0) (= s3 1)))
8.  (assert (or (= s4 0) (= s4 1)))

9.  (assert (or (and (= s1 1) (= s2 1))
                (and (= s3 1) (= s4 1))))
10. (assert (or (and (= s1 1) (= s4 1))
                (and (= s2 1) (= s3 1))))
11. (assert (< (+ s1 s2 s3 s4 ) min))

12. (check-sat)
13. (get-model)
```

The first 4 lines are to declare four Integer variables with the names $s1$, $s2$, $s3$, and $s4$ by using the Z3 keyword `declare-const`. Lines from 5 to 9 are to define the possible values of these Integer variables. These lines correspond to (Step 1) mentioned above. Note that `assert` is a Z3 keyword to define a logical formula. For example, `(assert (or (= s1 0) (= s1 1)))` is the same as the logical formula $(s1 = 0) \vee (s1 = 1)$. Multiple formulas defined using the keyword `assert` are logically and-conjuncted. For example, the logical formula defined through lines 5 and 6 is same as $((s1 = 0) \vee (s1 = 1)) \wedge ((s2 = 0) \vee (s2 = 1))$.

Lines from 9 to 10 define the conditions that restricts the candidate-points to satisfy the 2-coverage of the problem. Essentially, the conditions are those defined in $SS1$ and $SS2$. These lines correspond to (Step 2) mentioned above.

Line 11 is to define that the sum of the values of the four candidate-points should be smaller or equal to a minimum value denoted by `min`. This line is the same as the logical formula $(s1 + s2 + s3 + s4) <$ `min`. Note that we did not declare the value of `min` in Z3 language before using it. As will be explained below, this value is not known in advance, so we try different values to find the minimum value.

The last line 12 with `(check-sat)` is simply a command to let the Z3 SMT solver to determine the satisfiability of the formulas defined above. Line 13 demands Z3 to output a variable-assignment if the formulas are satisfiable.

We save the above text into a file, input the file into Z3, and wait for Z3 to return a result. As mentioned above, since initially we do not know the exact value of `min`, so we circumvent this by generating multiple files, which have the same textual contents as above except the value of `min`. We then input these files one by one, in an ascending order of `min`'s values, into Z3. If, for a file, a variable-assignment is found which satisfies the 2-coverage problem, we stop there and get the final optimal solution (namely an optimal sensor deployment that satisfies 2-coverage for all the observation-points); if not, we input the next file with the plus-1 value

 ISBN: 978-1-61208-353-7

for `min`, and so on.

For this example, we initially use 1 as the value for `min`, and Z3 returns unsatisfiable for it; we then input the file with `min=2` and input the file, again Z3 returns unsatisfiable. At last, if we set $min = 4$, Z3 returns that these formulas are satisfiable with an assignment $s1 = 1$, $s2 = 1$, $s3 = 0$, and $s4 = 1$. This shows an optimal result with the minimum sensor values of 3.

V. Experiments and Evaluation

We have conducted a series of experiments to evaluate the effectiveness and efficiency of the simplification techniques and the SMT-based algorithm. Our experiment environment is as follows: Windows PC running on a Intel-Core i7-2620M (@2.70GHz, 2.70GHz) CPU with 8.0GB Ram.

In our experiments, we consider 3-coverage problem of four target (rectangle) areas of different sizes, and of 50×50, 90×90, 140×140, and 190×190 units; we consider two kinds of distribution of candidate-points, one is *even* distribution and another is *random* distribution.

For candidate-points, in the case of even distribution, the candidate-points are located in a target area evenly with 14-unit intervals by considering the target area as a grid; in the case of random distribution, the candidate-points are located randomly but with the premise of satisfaction of 3-coverage (In other words, we only consider random distribution that satisfies 3-coverage, and do not taken other cases into account. This is a reasonable premise since our purpose is to examine the effectiveness of simplification). The number of candidate-points increases as with the size increase of target areas. In our experiments, the correspondence between the number of candidate-points and target areas are shown in Table I.

Table I
Correspondence between the Number of Candidate-Points and the Target Areas

Case No.	Target Area	Candidate-Point Number
Case 1	50×50	36
Case 2	90×90	100
Case 3	140×140	225
Case 4	190×190	400

For observation-points, no matter whether the distribution of candidate-points is even or random, observation-points are located evenly in a target area with 1-unit intervals by considering the target area as a grid. For the above target areas, since an observation-point will be deployed at the border (e.g., 0 position) of the area, the numbers of the observation-points are $51 \times 51 = 2601$, $91 \times 91 = 8281$, $141 \times 141 = 19881$, and $191 \times 191 = 36481$.

A. Evaluation of the Simplification Techniques

Through simplification, we can decrease the numbers of observation-points (or candidate-points), and consequentially, simplify the complexity of the target problem. We analyze the effectiveness of the simplification techniques partially mentioned in section III. Note that we do not experiment and analyze the effectiveness of each simplification technique separately since: the set of simplification techniques that we proposed can be iteratively applied; one simplification technique, which could not be applied at a time, may become applicable later after some other simplification techniques have been applied. Therefore, in the following experiments and analysis, the set of simplification techniques will be evaluated as a whole.

First, we evaluate, when the candidate-points are distributed evenly, how the observation-points can be decreased by applying our proposed simplification techniques. The comparison table, before and after applying simplification, is shown in Table II:

Table II
Decrease of Observation-Points with Even Distribution of Candidate-points

Case No.	Before Simplification	After Simplification
Case 1 (50×50)	2601	0
Case 2 (90×90)	8281	64
Case 3 (140×140)	19881	288
Case 4 (190×190)	36481	924

Note that the cases are characterized by the size of target areas (shown in Table I). We can observe that the number of observation-points is decrease extremely. Note also that in Case 1, after simplification the number of observation-points is 0, which seems to be strange. 0 here simply means that an optimal solution to the 3-coverage problem has already been found after simplification.

Next, we evaluate, when the candidate-points are distributed randomly, how the observation-points can be decreased by applying our proposed simplification techniques. The comparison table, before and after applying simplification, is shown as follows (Table III):

Table III
Decrease of Observation-Point with Random Distribution of Candidate-points

Case No.	Before Simplification	After Simplification
Case 1 (50×50)	2601	10
Case 2 (90×90)	8281	26
Case 3 (140×140)	19881	276
Case 4 (190×190)	36481	448

Similarly, we can observe that our proposed simplification techniques are quite effective for decreasing the number of observation-points, which can consequentially reduce the complexity of the coverage problem.

B. Evaluation of the SMT-based Algorithms

We report our experiment results on evaluation of the performance of (1) Algorithm 1 (tuned by following an ascending order of the number of sensors) vs. SMT-based algorithm

 ISBN: 978-1-61208-353-7

both with simplification, and (3) SMT-based algorithm with and without simplification. In the experiments, to enlarge the scope of experiment, we consider 9 cases, in which the sizes of the target areas are from $20 \times 20, \ldots, 190 \times 190$.

1) Algorithm 1 without and with Simplification: The experiment data for the nine cases are listed in Table IV. Note that we removed the cases of 20×20, 30×30, and 40×40 since computation time for them is simply 0. We set 9000 seconds as time out.

Table IV
ALGORITHM 1 AND SMT-BASED ALGORITHM BOTH WITH SIMPLIFICATION (TIME IN SECONDS)

Case No.	Algorithm 1	SMT-based Algorithm
Case 4 (50×50)	0.182	0
Case 5 (90×90)	1.045	0
Case 6 (100×100)	14.353	0.96
Case 7 (110×110)	1095.346	1.58
Case 8 (140×140)	9000	14.82
Case 8 (190×190)	9000	186.77

From the results, we can find that, by using the SMT-based algorithm (particularly Z3 in our experiments), the coverage problem for large target areas such as those more than 140×140 can be computed as well.

2) SMT-based Algorithm with and without Simplification: The next question is that: how about if we directly use SMT solving techniques for the problem without simplification? Many optimization techniques have already been implemented in SMT solvers (such as Z3), and it may be possible that those optimization techniques are sufficient for processing the coverage problems, and thus, our proposed simplification techniques may not be actually necessary. To check this possibility, we conducted experiments by using the SMT-based algorithm without and with our proposed simplification techniques. The results are shown in Table V. Note again that we removed the cases of 20×20, 30×30, 40×40, and 50×50 since computation time for them is simply 0.

Table V
SMT-BASED ALGORITHM WITH AND WITHOUT SIMPLIFICATION (TIME IN SECONDS)

Case No.	With Simplification	Without Simplification
Case 5 (90×90)	0	0
Case 6 (100×100)	0.96	117.45
Case 7 (110×110)	1.58	169.83
Case 8 (140×140)	14.82	736.11
Case 8 (190×190)	186.77	5452

From the above comparison, we can observe that our proposed simplification techniques do provide help to decrease the complexity of the coverage problems, which are necessary even if we use the efficient SMT-based algorithms.

VI. CONCLUSIONS AND FUTURE WORK

We have described our proposed modeling, simplification, and SMT-based computation algorithm for the K-coverage problem in the context of wireless sensor network. Experimental results have shown the efficiency of both the simplification and the SMT-based algorithm. There are much to be done as future work. In addition to conducting more experiments on real problems to further investigate the usability of our proposed approaches, our concrete work undergoing at this moment is to further improve the generality of our modeling and computing methods/algorithms, e.g., to consider user-specified restriction condition in our model.

ACKNOWLEDGMENT

This research is partially supported by "Project for Fostering Value-Creation Advanced ICT Frontier Human Resources by Fused Industry-University Cooperation" supported by MEXT, Japan.

REFERENCES

[1] K. Sohraby, D. Minoli, and T. Znati. "Wireless Sensor Networks: Technology, Protocols, and Applications". Wiley-Interscience, 2007.

[2] Z. Zhou, S. Das, and H. Gupta. "Connected K-Coverage Problem in Sensor Networks". Proceedings of the 13rd International Conference on Computer Communications and Networks (ICCCN2004), IEEE press, October 2004, pp.373–378.

[3] M. Hefeeda and M. Bagheri. "Randomized K-coverage Algorithms for Dense Sensor Networks". Proceedings of IEEE INFO-COM 2007 Minisymposium, IEEE press, May 2007, pp.2376–2380.

[4] H. M. Ammari. "On the Connected K-Coverage Problem in Heterogeneous Sensor Nets: The curse of randomness and heterogeneity". Proceedings of the 29th IEEE International Conference on Distributed Computing Systems (ICDCS2009), June 2009, pp.265–272, IEEE press.

[5] X.-Y. Li, P.-J. Wan, and O. Frieder. "Coverage in Wireless Ad-hoc Sensor Networks". IEEE Transactions on Computers, Vol. 52, No. 6, June 2003, pp.753–763.

[6] C. Barrett, R. Sebastiani, S. A. Seshia, and C. Tinelli. "Handbook of Satisfiability". IOS Press, 2009, Vol. 185, Chapters 25, 26, pp. 781–885.

[7] L. de Moura and N. Bjorner. "Z3: An Efficient SMT Solver". Proceedings of the 14th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS2008), March 2008, pp. 337–340, LNCS 4963.

[8] B. Dutertre and L. de Moura. "A Fast Linear-Arithmetic Solver for DPLL(T)". Proceedings of the 18th International Conference on Computer Aided Verification (CAV2006), August 2006, pp. 81–94, LNCS 4144.

[9] C. Barrett and C. Tinelli. "CVC3". Proceedings of the 19th International Conference on Computer Aided Verification (CAV2007), LNCS 4590, July 2007, pp. 298–302.

[10] C. Barrett, A. Stump, and C. Tinelli. "The SMT-LIB Standard: Version 2.0". Latest official release of Version 2.0 of the SMT-LIB standard. Online: http://www.smtlib.org/ [retrieved: June 2014]

 ISBN: 978-1-61208-353-7

Uncertainty Aware Hybrid Clock Synchronisation in Wireless Sensor Networks

Christoph Steup, Sebastian Zug, Jörg Kaiser
Department of Distributed Systems
Faculty of Computer Science
Otto-von-Guericke-University
Magdeburg, Germany
Email: {steup,zug,kaiser}@ivs.cs.ovgu.de

Andy Breuhan
Rohde & Schwarz DVS GmbH
Email: andybreuhan@andybreuhan.de

***Abstract*—Wireless Sensor Networks, aiming to monitor the real world's phenomena reliably, need to combine and post-process the detected individual events. This is not possible without reliable information of the context of the individual event. One such information of very high importance is time. It enables ordering of events as well as deduction of further data like rates and durations. An unreliable time base influences not only the ordering of events, but also the deduced values, which in consequence are unreliable as well. Therefore the synchronization of the clocks of the individual nodes is of high importance to the reliability of the system. On the other hand tight and reliable synchronization typically induces a large message overhead, which is often not tolerable in WSN scenarios. This paper proposes a new hybrid synchronization mechanism enabling tight synchronization in single hop environments and looser synchronization in multi hop environments. The lack of a guaranteed synchronization precision is mitigated by an explicit synchronization uncertainty, which is passed to the application. This enables the application to react to changes in the current synchronization precision.**

Keywords*—Wireless Sensor Networks, Time Synchronization, Uncertainty

I. Introduction

Wireless Sensor Networks (WSN) gain increasing attention by researchers as well as industry and governments. They provide the ability to monitor large areas for events efficiently and with small effort. Two examples are the SafeCast project [1], which aims to provide people with the ability to cheaply monitor radiation in their vicinity and share this data with others, and the project aiming to detect forest fires, endangering nature and people, as described by Yu et al. [2].

Even though WSN are often reduced to disseminating data, evaluation and decision making are equally important for these systems. Catastrophe warning systems for example, need robust decision making mechanisms to be accepted by people. Therefore, reliable post processing and context detection are crucial to provide a robust output. One of the most important context attributes needed for WSN is time, since an unreliable time base influence the ordering as well as the deduction of events, which in consequence become unreliable as well. Consequently, a reliable, precise global time base is a must-have for each WSN detecting safety relevant events. However, the granularity of the time base depends on dynamics of the system as well as the requirements of the application.

Typical time synchronization approaches like the Network Time Protocol (NTP) [3] or the Precision Time Protocol (PTP) [4] require a lot of messages and consider the underlying network to be quite robust, which are properties not available in typical WSN. This led to the creation of specially adopted time synchronization protocols. Typically these try to provide a trade-off between message overhead and synchronization precisions. Additionally they try to tolerate message losses and changes in the topology of the network. However, most of the existing protocols either try to provide a tight synchronization in a single hop environment and degrade heavily in multi hop environments or provide a generally looser synchronization in both. Unfortunately, none of the existing protocols provide the application with information on the current status of the synchronization, which might be degraded by errors in communication or heavy changes in topology.

This paper introduces a new hybrid synchronization mechanism for WSN. It provides tight synchronization in single hop environments and looser synchronization with a decreasing precision based on the topological distance between nodes. It enables applications to adapt to the currently achieved synchronization precision by providing an estimated synchronization uncertainty together with every time stamp.

The description of our approach starts with a discussion of related work in Section II, followed by the description of our concept in Section III. In Section IV we describe our implementation within the Omnet++ network simulator, the tests carried out and their results. The paper closes with a conclusion of our results and some ideas on future work in Section V.

II. State of the Art

In order to assess the current state of clock synchronisation for WSN, we describe 6 approaches representing basic concepts in the following section.

A. Reference Broadcast Synchronization (RBS)

Reference Broadcast Synchronization as described by Elson et al. [5] is a synchronization mechanism exploiting a physical broadcast in a shared medium. The synchronization starts with one node transmitting a *NOW*-message to all other nodes. This message serves as an indication for all nodes to take a local time stamp. Afterwards the timestamps are

 ISBN: 978-1-61208-353-7

exchanged between all nodes. Finally, all nodes compute individually their offset towards the mean of all exchanged timestamps.

This mechanism reduces the critical path to the transmission time of the *NOW*-message and the local processing time on each node until the local time stamp is taken. This provides very tight synchronization in single hop scenarios as long as the computation time is bounded. However, in worst case for n nodes $\mathcal{O}(n^2)$ messages are needed for a single synchronization round.

Additionally, RBS reacts very sensitive to the mobility of nodes. This is caused by the used averaging mechanism of the protocol. The contribution of all nodes to the averaged new time can create large shifts in the local clock of each node whenever one node's clock is far off. This is especially problematic if this node is only a temporary member of the broadcast group.

B. Delay Measurement Time Synchronization Protocol (DMTS)

The Delay Measurement Time Synchronization Protocol described by Ping [6] extends RBS by exploiting low-level hardware access. It extends the *NOW*-message with a time stamp taken and inserted just before sending. Therefore the exchange of the individual local time stamps can be omitted and the message count can be heavily reduced. To reach similar synchronization precision as RBS, the author estimates the delay of the transmission analytically and modifies the inserted time stamp accordingly.

This approach solves the large amount of message necessary for a single synchronization of RBS. However, in-depth knowledge of the needed hardware and communication mechanisms as well as low-level hardware access is needed to use it. Additionally, the problem of faulty or mobile nodes of RBS is enforced since only a single time stamp is communicated, which hinders fault recovery mechanism to be established.

C. Continuous Clock Synchronization in Wireless Real-time Applications (CCS)

The Continuous Clock Synchronization for Wireless Real-time Applications by Mock et al. [7] is a Master-Slave synchronization method extending the basic clock synchronization mechanism of the 802.11 standard [8]. In contrast to the basic mechanism, the clocks of the slaves are not simply set to the time stamp of the master, but are gradually adopted by adjusting their rate. Additionally, the precision of the synchronization is enhanced by dividing the time beacon in a *NOW*-message and an additional message containing the time stamp of the *NOW*-message. This division enables a more exact estimation of the master's time finishing the transmission of the *NOW*-message. The additional message needed can be saved if the master's time stamp is incorporated in the next *NOW*-message.

This approach enables continuous clock synchronization without gaps in the time base. Additionally, it provides better precision than the basic 802.11 synchronization mechanism without additional message overhead.

D. Probabilistic Clock Synchronization Service (PCS)

The Probabilistic Clock Synchronization Service by PalChaudhuri et al. [9] is an extension of RBS enabling a dynamic trade-off between synchronization precision and message overhead. This approach transmits n *NOW*-messages in one synchronization round, which are used to derive the skew of the sender's and the receiver's clock though linear regression. The results are combined and broadcasted back to the receivers in range. By comparing their own data with the data received from the sender, they are able to adopt their own clocks. To derive the number of needed *NOW*-messages the authors assumed the synchronization error to be normal distributed with zero-mean and a standard deviation of σ. Based on this distribution, the authors analytically derive the probability $P(|\epsilon| < \epsilon_{max})$ of the synchronization error to be less than a specified value ϵ_{max}. For a specified probability of the synchronization to be more precise than ϵ_{max}, the authors derive the number n of message needed. This number heavily depends on the standard deviation σ of the normal distribution.

The approach provides a dynamic trade-off between message overhead and synchronization precision. However the trade-off depends on the standard deviation of the synchronization error, but the acquisition of this value was not covered by the authors. Additionally, only a mathematically prove without any simulation was conducted to evaluate the idea. Consequently, the authors never discussed the effects of non-normal distributed synchronization errors.

E. Time Synchronization in Ad-Hoc Networks (TSAN)

Römer's Time Synchronization in Ad-Hoc Networks [10] is based on Christian's Algorithm [11]. It estimates the round trip time of a message between sender and receiver. Whereas Christian's Algorithm proposed a dedicated server for clients to communicate to, Römer attaches time stamps to events communicated in the network. Therefore Römer's algorithm ideally induces a zero message overhead. However, not all events are acknowledged by the receiver, which might create large durations between events flowing in both directions between two nodes. This is mitigated by the insertion of additional dummy events in case the duration grows too large.

TSAN applies a very loose multi hop synchronization with an ideal message overhead of 0. Unfortunately, the real message overhead is heavily dependent on the actual communication in the network and is therefore very hard to estimate for a real system.

F. Event Composition in Time-Dependant Systems (ECTS)

Liebig et al. [12] described a way to combine multiple events even though their time stamps might not be exact. To achieve this they extended a time stamp to a time interval and derived an ordering relation $<$ for time intervals. Together with a known uncertainty of the event's time stamp, ordering might be possible even in loosely synchronized systems. However, this transition changed the order relation to a half order relation. Consequently, there might be situations in which two events cannot be ordered. This is the case if the intervals of the events' time stamps overlap.

 ISBN: 978-1-61208-353-7

This approach, even though it is not directly handling clock synchronization, enables uncertainty aware clock synchronization algorithms to provide time intervals, which provide awareness of synchronization precision to higher layer applications.

G. Summary

The individual problems and features of the protocols are summarized in Table I. As visible none of the described approaches fully solve the problem of multi-hop uncertainty-aware clock synchronization in wireless sensor networks. However, each approach contains individual features, which might enhance the performance of our approach.

Our approach incorporates the beneficial properties of the different approaches in a single clock synchronization mechanism, which is uncertainty and topology aware and produces time intervals usable by an application. The time interval algebra exploit Liebig et al.'s ordering relation, see II-F, to enable an seamless integration in existing technology. The next section describes our mechanism in detail.

TABLE I. COMPARISION OF THE DISCUSSED TIME SYNCHRONIZATION PROTOCOLS.

Protocol	Synchronization precision	Multihop capability	Message Overhead	Robustness
RBS	high	none	$\mathcal{O}(n^2)$	fragile
DMTS	high	medium	$\mathcal{O}(n)$	fragile
CCS	high	medium	$\mathcal{O}(1)$	robust
PCS	medium	good	dynamic	medium
TSAN	low	good	0 - $\mathcal{O}(1)$	medium

III. UNCERTAINTY AWARE CLOCK SYNCHRONISATION (UACS)

For an efficient synchronization of clocks in WSN multiple parameters are important. On one hand, the synchronization needs to be scalable, while on the other hand the overhead may not exceed a certain threshold to safe battery and prevent an overload of the network. Most of the approaches discussed in Section II favour one over the other. However if we limit our self to certain base topologies better solutions might be found. One interesting topology is the cluster tree structure of IEEE 802.15.4 networks [13] in beacon-enabled mode. This mode divides the nodes in groups called Personal Area Networks (PANs), which have an individual coordinating instance managing the internal communication. The individual PANs communicate only through their respective coordinators, as visible in Figure 1. In the remaining section of the paper we consider an 802.15.4 network, with an already established cluster tree structure. The formation and the handling of dynamic changes in this structure are not considered in this paper.

Based on the initial assumption, that clock synchronization may have a decreasing precision based on topological distance between nodes in the network, we propose a hybrid clock-synchronization, consisting of a tight synchronization mechanism for each individual PAN called Intra Cluster Synchronization and a loose synchronization mechanism between the individual PAN Coordinators, called Inter Cluster Synchronization.

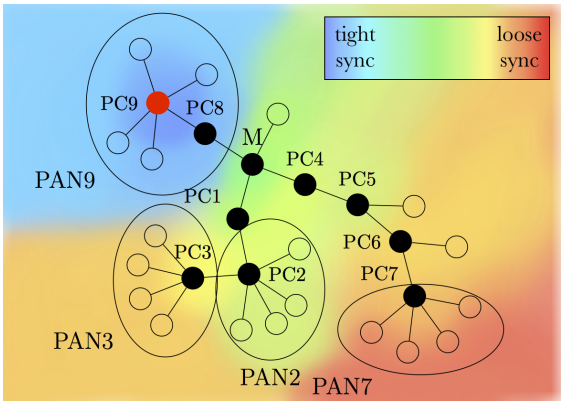


Fig. 1. Example cluster tree structure of an IEEE 802.15.4 Network. The colours indicate the topological distance between each node and $PC9$.

A. Intra Cluster Synchronization

The Intra Cluster Synchronization is based on the CCS approach, see Section II-C. Therefore, each PAN slave $P_{s_j} \in Slaves$ has a virtual synchronized clock $VC_{s_j}(t)$. This clock uses the time stamps created by the node's hardware clock $C_{s_j}(t)$ and modifies it based on the current rate $\rho_{s_j,i}$:

$$VC_{s_j}(t) = \rho_{s_j,i}\left(C(t)_{s_j} - C(t_i)_{s_j}\right) + VC_{s_j}(t_i) \qquad (1)$$

The task of the Intra Cluster Synchronization is the estimation of the parameter $\rho_{s_j,i}$ for each slave at each synchronization round i. The 802.15.4 standard allows a PAN Coordinator to attach additional information to the beacon frame. We use this to attach a 64bit time stamp $t_{c,i}$ to each beacon b_{i+1} transmitted by the coordinator. The attached time stamp represents the coordinators time of successful transmission of the last beacon. This time stamp together with the local reception time of the last beacon $t_{s_{j,i}}$ is then evaluated by each slave P_{s_j} to compute a new rate $\rho_{s_j,i}$.

As described by DMTS, see Section II-B, hardware knowledge may be used to provide the needed local time stamps. The 802.15.4 standard provides the `PD-Data.confirm` primitive as a local event indicating completion of a transmission. The time of this event is used as the source of the time stamp $t_{c,i}$. On reception of the beacon each PAN slave P_{s_j} takes a local time stamp $t_{s_{j,i+1}}$. The networks tightness τ together with the internal computation time t_{comp} of the nodes limits the accuracy of the local time stamps. This computation time can also be mitigated by the PD-DATA.indication primitive of the 802.15.4 standard. Therefore we omit t_{comp} in our approach and the time difference of creation of the local time stamps is bounded by τ.

After acquiring the time stamp for the actual synchronization round the PAN slaves compute the offset $o_{j,i} = t_{c,i} - t_{s_j,i}$ between their previous local time stamp $t_{s_j,i}$ and the time stamp transmitted through the beacon $t_{c,i}$. This is used to compute a new rate $\rho_{s_j,i-1} = k_\rho o_{j,i}$ for the node's virtual clock to compensate the offset, with k_ρ being a proportional factor controlling the rate of adoption.

The Intra Cluster Synchronization provides continuous clock synchronization between the PAN Coordinator and its slaves. The overhead is minimal since no additional message

is necessary and the beacons are only slightly enlarged. The robustness of the synchronization mainly depends on the used algorithm to detect a crash and reselect a PAN Coordinator. The synchronization accuracy depends on the clock skew between the PAN Coordinator and its slaves as well as the tightness of the network and the beacon interval of the PAN. Mobility of individual slaves is handled by the adoption rate of the virtual clock and has no influence on other slaves of the same PAN.

B. Inter Cluster Synchronization

Diverging from the Intra Cluster Synchronization, see Section III-A, each PAN Coordinator does not modify its own virtual clock. Instead every event received by a PAN Coordinator P_{c_r}, which is transmitted by another adjacent PAN Coordinator P_{c_s} is transformed in the time domain of the PAN Coordinators virtual clock $VC_r(t)$, as proposed by TSAN, see Section II-E. To achieve this, the PAN Coordinator P_{c_r} needs to calculate a virtual clock $VC_{r,s}(t)$ for each adjacent PAN Coordinator P_{c_s}.

The virtual clocks are handled similarly to the Intra Cluster Synchronization, since all beacons of all adjacent PAN Coordinators P_{c_s} are received by PAN Coordinator P_{c_r}. On reception of beacons, containing a time stamps $t_{s,i}$, P_{c_r} acquires a local time stamp $t_{r,s,i+1}$. This enables the computation of the offset $o_{r,s,i} = t_{s,i} - t_{r,s,i}$ between P_{c_s} and P_{c_r}. Afterwards P_{c_r} updates the rate $\rho_{r,s,i} = k_\rho o_{r,s,i}$ for the virtual clock $VC_{r,s}(t)$ towards P_{c_s}.

On reception of an event e_n from P_{c_s} containing a time stamps $e_n.ts_s$, P_{c_r} is able to transform the time stamp of the event to its own virtual clock $VC_r(t)$. The transformation is done by adding the offset between the Virtual Clock of the sender $VC_s(t)$ and the receiver $VC_r(t)$ to the event's time stamp:

$$e_n.ts_r = e_n.ts_s + VC_r(t) - VC_s(t) \tag{2}$$

This approach handles mobility well, since the mobility of a node in the local neighbourhood of P_{c_1} does not change P_{c_1}'s virtual clocks of the other nodes. Therefore the transformation of the events is independent of each other. However, the disadvantage is the accumulation of synchronization errors over multiple hops. Therefore we explicitly specify the synchronization error in a time interval $ti = [ts \pm \alpha], ts \in R^+, \alpha \in R^+$ replacing the time stamp $ts \in R^+$. Consequently, each hop additionally modifies the interval bounds by the currently estimated uncertainty of the synchronization $\alpha_{r,s}$:

$$e_n.\alpha_r = e_n.\alpha_s + \alpha_{r,s} \tag{3}$$

C. Estimating the Uncertainty

The estimation of the current uncertainty of the synchronization of the virtual clocks is difficult. Multiple factors influence the actual uncertainty in the synchronization, like beacon losses and the current drift of the individual clocks. In our approach the synchronization error $\epsilon_{r,s,i}$ of synchronization round i between two adjacent PAN Coordinators P_{c_s} and P_{c_r} is characterized by their offset $o_{r,s,i+1}$ at beginning of synchronization round $i+1$.

Following PCS, see Section II-D, we model the synchronisation error to be a zero-mean Gaussian distribution $N(0,\delta)_{r,s}$. To estimate the standard deviation we use the synchronization errors of the previous n synchronization rounds as sample set $E_{r,s} = \{\epsilon_{i-n}, \epsilon_{i-n+1} \dots \epsilon_i\}$. We estimate the standard deviation $\delta_{r,s}$ of our zero-mean Gaussian based on the sample set. Based on this we compute the confidence interval $\left[\bar{x} \pm z_{\left(\frac{1+\gamma}{2}\right)} \frac{\delta}{\sqrt{n}}\right]$ of the synchronization with typical probability γ. The value $z_{\left(\frac{1+\gamma}{2}\right)}$ represents the $\frac{1+\gamma}{2}$-quantile of the standardised normal distribution. The resulting size of the confidence interval $\alpha_{r,s} = z_{\left(\frac{1+\gamma}{2}\right)} \frac{\delta}{\sqrt{n}}$ represents our current uncertainty estimation, which is added to current uncertainty of the event's time interval.

The complexity of this computation is only dependant on n, which represents a trade-off between estimation accuracy and memory and computation overhead. The quantile of the standardised normal distribution is a pre-defined constant, characterising the accuracy of the estimation.

D. Compatibility between Time Intervals and Time Stamps

Since time intervals and time stamps are not directly compatible we introduced a compatibility algebra $([ts \pm \alpha], \{+, -, \cdot, <\})$ based on interval arithmetic and the proposed half order relation of ECTS, see Section II-F. To additionally handle the deduction of new events, we added the operation to multiply the time interval with a constant $k \cdot [ts \pm \alpha] = [k \cdot ts \pm k \cdot \alpha]$. This is useful for applications to scale the difference between two time stamps as needed e.g. in the computation of speeds.

The resulting algebra establishes an ordered vector space, which is easy to compute even for deeply embedded systems. Transformations back to time stamps are easily possible by omitting the uncertainty part of the time intervals.

IV. Evaluation

We evaluated our uncertainty aware hybrid clock synchronization system with a simulation in the Omnet++ network simulator [14] version 4.2. Additionally we used the INETMANET network model [15] as well as the MiXiM model [16]. The implementation is distributed over two layers of the ISO/OSI stack. One part is located at layer 5 of the ISO/OSI stack and handles the transformation of time stamps for the Inter Cluster synchronization. The other is situated at layer 2 to gather high precision time stamps. Both layers are connected through a cross layer communication.

Our evaluation focuses on the Inter Cluster Synchronization, since single-hop synchronization is very well researched and our approach resembles CCS, see Section II-C, and DMTS, see Section II-B. Therefore the performance should be equivalent. For the Inter Cluster Synchronisation we evaluate two main topics. The first considers the influence of the beacon period on the precision of the synchronization. This test will provide information on the trade-off between message overhead and synchronization quality. The second test investigates the influence of the communication topology on the reachable multi-hop precision. It will evaluate the usability of the provided time stamps for smaller and longer routes. All tests used the internal 64 bit simtime of Omnet++ as reference

 ISBN: 978-1-61208-353-7

for the synchronized clocks to evaluate the synchronization error. The simtime was modified by a randomly initialized drift($< 10^{-5}$), to provide a realistic clock for each node. The test considered 1000 randomly created routes between nodes in the network, which were created by an optimal routing algorithm.

Our simulation environment considers beacon losses, created by the collision of transmitted beacons of adjacent coordinators, and the resulting lack of information for the time synchronization. However, we did not transmit data events in the simulation. This decouples our simulations from the used MAC Algorithm and its parameters. Consequently, the current simulations consider an optimal MAC-Algorithm preventing all collision between beacons and events in the network.

A. Beacon Interval Analysis

The beacon interval analysis considered a rectangular grid of 50 PAN Coordinators. The area in which the nodes were distributed was $5000m$ times $5000m$. We used the 2.4GHz specification of the 802.15.4 standard at channel 11 with a maximum transmission power of $1mW$. The thermal noise was fixed at $110dBm$ and the receiver's sensitivity was set to $-85dBm$. Our simulation sweep started with a BO parameter of 8 up till the maximum allowed value of 14. The resulting beacon interval can be computed by $BI = \frac{16 \cdot 60S \cdot 2^{BO}}{SymbolRate}$. The $SymbolRate$ of the 2.4GHz band of $65.2 \cdot 10^3 \frac{S}{s}$ results in beacon intervals between $3.8s$ and $241.2s$ length.

Figure 2 shows a Box-Whisker plot of the simulation's results. Results for BO values from 8 to 10 are omitted because of similarity. The boxes represent the bounds, where 50% of all values are included. The lines represent the interval containing 75% of all values and remaining data points are included as points. As visible with linear increasing BO values the mean synchronization error increases exponentially. This is to be expected because th beacon interval also increases exponentially. Additionally one observes a large standard deviation independent of the hop-count. This is caused by the unsynchronized beacons of the individual PAN Coordinators, which might collide and therefore increase the real beacon interval. Furthermore the data base is better for smaller hop-counts, since in the given scenario short routes are much more probable then longer routes.

This test proved the expected direct correlation between the beacon interval and the synchronization precision. Therefore this value is to be considered critical for the performance of the system.

B. Topology Analysis

Our second evaluation considers the performance of the system in different topologies. This is interesting, because topologies might have an influence on the length of the routes as well as the collision probability of the beacon frames. Therefore we considered four basic topologies with 200 nodes each. We choose a randomly generated, a grid, a linear and a circular topology. For this scenario we used the same parameters as for the Beacon Interval Analysis, see Section IV-A. This time the BO parameter was statically set to 8.

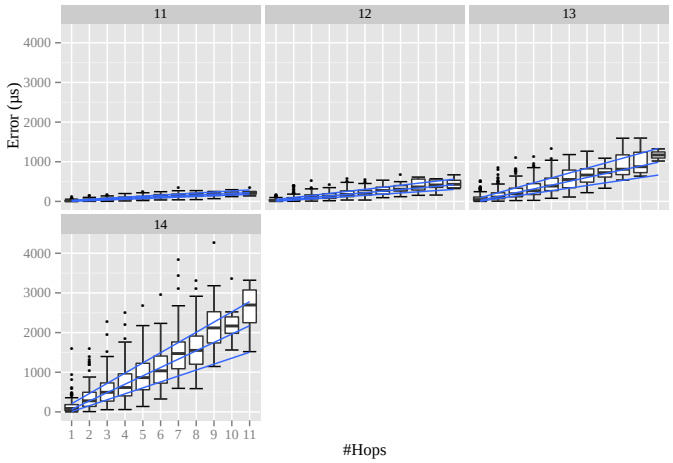


Fig. 2. Box-Whisker plot of the precision of a 50 node grid network with varying BO values.

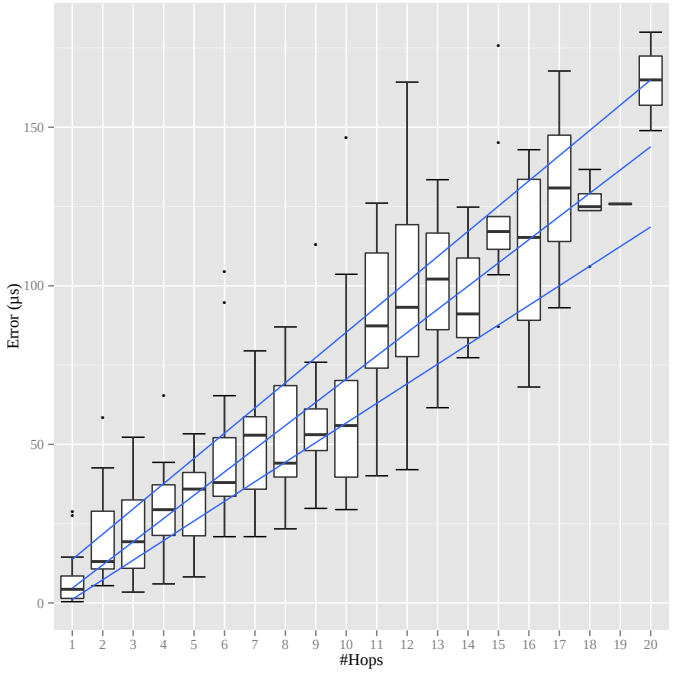


Fig. 3. Box-Wisker plot of the synchronization precision of a 200 Node random topology.

In Figure 3, the performance of the random topology is visible. The mean error of the simulation increases linear with the hop count. This is to be expected, since the synchronization error in the vicinity of each PAN is statistically the same. The summation of the uncertainties matches very well with the increasing error in the simulation. However the deviation of individual results is quite large in this case, which is caused by individual collisions of beacons. This problem is very dependent on the local setup of nodes around a PAN. Therefore it will increase the standard deviation of the random topologies synchronization error.

Table II shows an overview of the results of our evaluation of the different topologies. This table shows quite similar

 ISBN: 978-1-61208-353-7

TABLE II. SYNCHRONIZATION ERROR OF THE SIMULATION OF DIFFERENT TOPOLOGIES IN μs.

#Hop Count	Topology	Mean	Standard Deviation
1	Random	7.069008	0.008495
1	Grid	5.067219	0.005454
1	Linear	12.181786	0.009492
1	Circle	3.327730	0.003993
6	Random	47.401797	0.023945
6	Grid	26.574773	0.013489
6	Linear	55.861373	0.031209
6	Circle	21.579193	0.008327
11	Random	89.191069	0.027941
11	Grid	57.463772	0.019879
11	Linear	91.408097	0.023836
11	Circle	45.693382	0.015802
16	Random	110.255113	0.029990
16	Grid	80.882388	0.020756
16	Linear	131.582874	0.031220
16	Circle	73.628941	0.021878

results for the standard deviation of the tests for equally long routes in the different topologies. However, the mean of the synchronization error is quite different. As expected in all simulations the linear topologies have the largest mean error, which is cause by the highest collision probability of the beacons. The random topology performed the second worst, which is caused by local hot spots in the topology with a lot of nodes increasing the probability of beacon losses. In consequence all topologies show a linearly increasing error. Therefore our original assumption of an additive uncertainty fits very well to the simulated experiments.

C. Comparison with related protocols

Since the environments of the different described protocols differ, we compare our approach to protocols, with available multi-hop synchronization data. DMTS provided a mean synchronization error of $32\mu s$ for one hop and $46\mu s$ for two hop communication. Our approach performed better, but DMTS was evaluated on real hardware with a limited oscillator speed and computing power. Therefore, the results are not directly comparable. TSAN showed a mean synchronization error of $200\mu s$ for one hop and $1113\mu s$ for six hop communication. However, Römer et al. considered an unstructured network, whereas we exploited the structure of the network to increase the synchronization precision without message overhead.

V. CONCLUSION

This paper presents a novel hybrid clock synchronization approach that provides tight synchronization for local clusters of nodes as well as looser synchronization in multi-hop scenarios. The message overhead is minimal since existing periodic beacon messages of the 802.15.4 beacon-enabled mode were used to transmit the synchronization data. To handle the different synchronization precisions, uncertainty awareness was added to enable applications to decide in the case of ambiguous situations. The evaluation was done using the well-established network simulator Omnet++ in multiple scenarios with different configurations and supports the theoretical concepts of the described approach.

In future work, we want to evaluate the clock synchronization in a real scenario with real sensor nodes to evaluate the influence of unforeseen interference, as well as the limited processing power of the nodes. Additionally, we want to investigate the effect of the used MAC-Algorithm on the synchronization quality.

REFERENCES

[1] Y. Abe, "Safecast or the production of collective intelligence on radiation risks after 3.11," in *The Asia-Pacific Journal*, vol. 12, Issue 7, No. 5, Feb. 2014, pp. 1–6.

[2] L. Yu, N. Wang, and X. Meng, "Real-time forest fire detection with wireless sensor networks," in *Proceedings of the International Conference on Wireless Communications, Networking and Mobile Computing*, vol. 2, Sep. 2005, pp. 1214–1217.

[3] J. Burbank, D. Mills, and W. Kasch, "Network time protocol version 4: Protocol and algorithms specification," Internet Engineering Task Force (IETF), Tech. Rep. RFC 5905, Jun. 2010.

[4] "IEEE standard for a precision clock synchronization protocol for networked measurement and control systems," IEEE, Tech. Rep. Standard 1588-2002, 2002.

[5] J. Elson, L. Girod, and D. Estrin, "Fine-grained network time synchronization using reference broadcasts," in *SIGOPS*, vol. 36, no. SI. New York, NY, USA: ACM, Dec. 2002, pp. 147–163.

[6] S. Ping, "Delay measurement time synchronization for wireless sensor networks," Intel Research Berkeley Lab, Tech. Rep. IRB-TR-03-013, 2003.

[7] M. Mock, R. Frings, E. Nett, and S. Trikaliotis, "Continuous clock synchronization in wireless real-time applications," in *Proceedings of the 19th IEEE Symposium on Reliable Distributed Systems*, Oct. 2000, pp. 125–132.

[8] "IEEE Standard for Information Technology- Telecommunications and Information Exchange Between Systems-Local and Metropolitan Area Networks-Specific Requirements-Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications," IEEE, Tech. Rep. IEEE Standard 802.11, 1997.

[9] S. PalChaudhuri, A. Saha, and D. B. Johnson, "Probabilistic clock synchronization service in sensor networks," in *IEEE Transactions on Networking*, vol. 2, no. 2, 2003, pp. 177–189.

[10] K. Römer, "Time synchronization in ad hoc networks," in *Proceedings of the 2nd ACM international symposium on Mobile ad hoc networking & computing*, ser. MobiHoc '01. New York, NY, USA: ACM, 2001, pp. 173–182.

[11] F. Cristian, "Probabilistic clock synchronization," in *Distributed Computing*, vol. 3, no. 3. Springer, Sep. 1989, pp. 146–158.

[12] C. Liebig, M. Cilia, and A. Buchmann, "Event composition in time-dependent distributed systems," in *Proceedings of the Fourth IECIS International Conference on Cooperative Information Systems*, ser. COOPIS '99. Washington, DC, USA: IEEE Computer Society, Sep. 1999, pp. 70–78.

[13] "IEEE Standard for Information Technology- Telecommunications and Information Exchange Between Systems- Local and Metropolitan Area Networks- Specific Requirements Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (WPANs)," Tech. Rep. IEEE Standard 802.15.4, 2006.

[14] G. Pongor, "OMNeT: objective modular network testbed," in *Proceedings of the International Workshop on Modeling, Analysis, and Simulation On Computer and Telecommunication Systems*. San Diego, CA, USA: Society for Computer Simulation International, Oct. 1993, p. 323–326.

[15] A. Ariza and A. Triviño, *Simulation of Multihop Wireless Networks in OMNeT++*. IGI Global, 2012, pp. 140–158.

[16] A. Köpke, M. Swigulski, K. Wessel, D. Willkomm, P. T. K. Haneveld, T. E. V. Parker, O. W. Visser, H. S. Lichte, and S. Valentin, "Simulating wireless and mobile networks in OMNeT++ the MiXiM vision," in *Proceedings of the 1st International Conference on Simulation Tools and Techniques for Communications, Networks and Systems & Workshops*. Brussels, Belgium: Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Mar. 2008, p. 71:1–71:8.

 ISBN: 978-1-61208-353-7

Bringing Context to Apache Hadoop

Guilherme W. Cassales and Andrea S. Charão
Laboratório de Sistemas de Computação
Universidade Federal de Santa Maria
Santa Maria, RS, Brazil
{cassales,andrea}@inf.ufsm.br

Manuele Kirsch-Pinheiro and Carine Souveyet
Centre de Recherche en Informatique
Université de Paris 1 - Panthéon Sorbonne
Paris, France
{manuele.kirsch-pinheiro, carine.souveyet}@univ-paris1.fr

Luiz Angelo Steffenel
Laboratoire CReSTIC - Équipe SysCom
Université de Reims Champagne-Ardenne
Reims, France
luiz-angelo.steffenel@univ-reims.fr

***Abstract*—One of the first challenges when deploying MapReduce over pervasive grids is that Apache Hadoop, the most known MapReduce distribution, requires a highly structured environment such as a dedicated cluster or a cloud infrastructure. In pervasive environments, context-awareness becomes essential to coordinate the resources (task scheduling, data placement, etc.) and to adapt them to the environment variable behavior. In this paper, we present our first efforts to improve Hadoop by introducing context-awareness on its scheduling algorithms. The experiments demonstrate that context-awareness allows Hadoop to better scale based on actual resource availability, therefore improving the task allocation pattern and rationalizing resource usage in a heterogeneous dynamic network.**

Keywords–Context-awareness; MapReduce; Apache Hadoop; job scheduling.

I. INTRODUCTION

Given today's high volume of available data, new methods of processing this huge volume are being researched. Recently one of these new methods, MapReduce [1] and its most known implementation Apache Hadoop [2], is gaining space among both users and developers. MapReduce [1] is a programming model for parallel data processing, while Hadoop is a software platform implementing MapReduce. Thanks to Hadoop, it is possible to easily process large data sets in a computer cluster.

Designed to work with homogeneous cluster environments, Apache Hadoop is currently used not only on dedicated clusters, but also over cloud computing infrastructures. Despite its design, Hadoop has some liabilities that negatively affect its performance. One of those drawbacks is the assumption that every node has the same resource capacity, and that this capacity is set in a default XML file. As the cluster size scales, this task becomes very time consuming and error-prone and ill-configured nodes will harm the overall performance.

Besides, this assumption of a homogeneous environment limits the deployment of Hadoop over desktop and pervasive grids. Pervasive grids [3][4] are characterized by their heterogeneity, integrating nodes with quite different capabilities. Such heterogeneous environments represent an interesting alternative to cloud computing infrastructures. Indeed, as underlined by Schadt et al. [5], cloud computing solutions present important drawbacks when considering data transfer (transferring gigabytes of data across the network can be costly) and data security/privacy (putting sensitive data on the cloud may represent an important issue for some application).

In order to extract the best performance from Apache Hadoop on heterogeneous environments, it is necessary to reconsider how tasks are scheduled on the cluster. Indeed, MapReduce performance in Hadoop is tightly tied to the scheduler [6] and to its capability of observing the environment characteristics. Currently, Hadoop scheduler considers only information from the XML configuration file, ignoring the actual state of the nodes. For instance, when running Hadoop in homogeneous environments, the configuration files represent indeed an easy way to configure a cluster, since one needs only to discover the capacity of a node and to replicate the files to every other node in the environment. However, when using a heterogeneous environment, one will have to discover and edit XML files for each node in a cluster. This behavior often limits the Hadoop scalability in heterogeneous clusters, since the configuration files will not follow real nodes characteristics, and consequently, nodes will be limited to default values.

In order to overcome this drawback, we propose to improve Hadoop scheduling through a context-aware approach. Context can be defined as any information that can be used to characterize the situation of an entity (a person, place or object) that is considered relevant to the interaction between a user and an application [7]. Context information has been used for adapting application behavior during execution time [8][9], adapting content [10] or components deployment [11][12], for instance. We advocate that being aware of context in which a job is executed may contribute to a better use of resources in heterogeneous environments. In this paper, we propose to open Hadoop scheduler to the job execution context, observing real node conditions instead of a static (potentially mismatching) configuration file. By collecting the job execution context, we allow a better utilization of cluster's resources and also a better adaptation to heterogeneous environments.

Our proposal of context-awareness focuses therefore on discovering the real node capacity and providing a scheduler based on true observed information. In this paper, we conducted experiments with a basic set of context information (CPU, memory), which in our tests proved to be significantly different from default values. The experiments demonstrate that the applications performance can be widely improved through the use of context information, which encourages us to develop further the context-aware scheduling with additional parameters such as data locality, CPU speed and even task re-splitting to better explore idle resources and speculative execution.

 ISBN: 978-1-61208-353-7

The rest of the paper is organized as follows: Section II introduces MapReduce model and the basis of Apache Hadoop framework. Section III discusses related works, focusing on context-awareness and on other improved Hadoop schedulers. Section IV analyzes and evaluates Hadoop scheduling mechanism. Section V presents our proposal of context-aware scheduling, while Section VI presents experiments and first results. We conclude in Section VII.

II. About Hadoop

The Apache Hadoop is a framework that has the purpose of facilitating distributed processing through the MapReduce model. MapReduce [1] divides computation into two phases: map and reduce. During map, input data is split into smaller slices of data, whose analysis is distributed over the participating nodes. Each participant computes one (or more) slice of data, generating intermediary key/values results. During reduce phase, intermediary values concerning a given key are put together and analyzed, generating final key/values results. Hadoop framework is in charge of distributing data and map/reduce tasks over the available nodes. As a result, programmers need only to focus on map and reduce functions, since data and task distribution becomes transparent.

Apache Hadoop has two main components (see Fig. 1), which are Hadoop Distributed FileSystem (HDFS) and Yet Another Resource Negotiator (YARN). These components are respectively responsible for the data management on a distributed file system and for MapReduce tasks and job processing. YARN manages tasks and jobs distribution over the available nodes and it is in charge of scheduling jobs according to nodes capacity. Each node information is controlled by an individual NodeManager, while overall cluster information is centralized by the ResourceManager.

YARN integrates a scheduler that is responsible for distributing tasks over the available nodes. Current YARN structure (Fig. 1) aims at acquiring and using each NodeManager resource information to improve the performance of its default scheduler, the CapacityScheduler. Basically, each NodeManager consults configuration files, in order to discover node declared capacity, and inform ResourceManager about its existence. This information is transferred to the scheduler that uses it for deciding an appropriate job scheduling. The CapacityScheduler has the role of centralizing the information about NodeManager's capacities on the "master" node, keeping track of them in a global pool of resources and distributing them according to the ApplicationMasters requests.

In order to deploy Hadoop on a cluster, every node must have some XML configuration files available in their local Hadoop installation. In fact, given Hadoop huge dependence on XML files, even the nodes resources are set by these files. This peculiarity makes a more adaptive environment something hard to achieve with the default Hadoop distribution.

III. Related Work

Because Hadoop performance is tightly dependent on the computing environment, but also on the application characteristics, several researchers focused on bringing context-awareness to Hadoop. Their works can be roughly classified

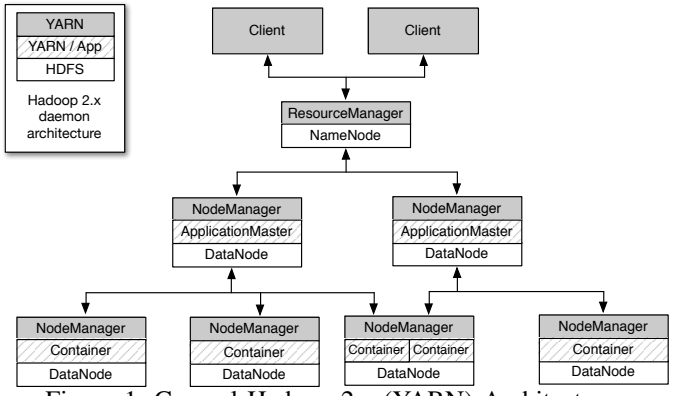


Figure 1: General Hadoop 2.x (YARN) Architecture.

as: (i) job or task schedulers, whose purpose is changing the Hadoop scheduling, and (ii) resource placement facilitators.

In the first case, we find works like Kumar et al. [6], Tian et al. [13] or Rasooli [14]. Those assume that most jobs are periodic and demand similar CPU, network and disk usage characteristics. As a consequence, these works propose classification mechanisms that first analyze both jobs and nodes with respect to its CPU or I/O potential, allowing an optimized matching of applications and resources when a job is submitted. For instance, Kumar et al. [6] and Tian et al. [13] classified both jobs and nodes in a scale of I/O and CPU potential, while Rasooli et al. [14] go beyond I/O and CPU potential and propose a full classification of jobs in order to match these jobs with nodes belonging to the same classification. Similarly, Isard et al. [15] proposes a capacity-demand graph that helps in the calculation of the optimum scheduling from a global cost function.

While the previous works focus on the improvement of the overall cluster performance through an offline knowledge about the applications and the resources, other works focus on individual tasks in order to ensure a smooth operation. For instance, works like Zaharia et al. [16] and Chen et al. [17] focus on improving tasks deployment inside a job, as a way to reduce the response time in large clusters, executing many jobs of short duration. These works rely on heuristics to infer the job estimated progress and decide whether to launch a speculative task on another possibly faster machine. Similarly, Chen et al. [17] propose using historical execution traces to improve its predictions. They propose a re-balancing of data across the nodes, leaving more data to faster nodes and less data on slower nodes.

Finally, works like Xie et al. [18] aim at providing better performance on jobs through better data placement, using mainly the data locality as decision making information. The performance gain is achieved by the data re-balancing in nodes, feeding faster nodes with more data. This lowers the cost of speculative tasks and also of data transfers through the network.

We may observe that most of these works rely on the categorization of jobs and nodes, which is hard in a dynamic environment like pervasive grids. Even when runtime parameters such as elapsed time or data placement are considered, they assume a controlled and well-known environment. Because

 ISBN: 978-1-61208-353-7

of these assumptions, these works fail on responding to the requirements of pervasive grids. Indeed, previous works focus on the reduction of response time or improvement of overall performance, which is a goal slightly different from ours, which is to adapt Hadoop to heterogeneous environments.

IV. HADOOP SCHEDULING

In order to improve Hadoop, it is important to understand current mechanisms that would be influenced and whose behavior should be altered. Thus, before presenting how we extend current Hadoop behavior with context information, we shall introduce the Hadoop resource allocation pattern.

A. Understanding Hadoop Allocation Pattern

Apache Hadoop operates in a master/slave hierarchy on both components (YARN and HDFS), each component being subdivided in numerous sub-components. On the top of YARN daemons, as shown in Fig. 1, we found the ResourceManager (RM), which is in charge of managing the resources from the entire cluster and of assigning applications to the underlying computing resources. These resources belong to the NodeManagers (NM), and each NM will inform the RM the amount of available resources upon start.

When a new job is submitted to the cluster, the ResourceManager registers the start of the job and then delegates the supervision to an ApplicationMaster (AM). The ApplicationMaster is the manager of the application, which asks for resources for the CapacityScheduler (a component of ResourceManager). The CapacityScheduler tracks the free/used resources on the cluster and grants them to the AM based on the global (cluster) and local (application) limits. The granted resources are presented as Containers, a processing instance in which all the processing takes place.

Resource allocation is based on a set of parameters defined in the XML configuration files. These parameters concern both memory and number of cores for applications and containers, and are composed by the minimum and maximum limit. If no value is given for a precise node or application, default values from the XML files are assumed, which can substantially differ from real node characteristics.

With an experiment running on nodes with the same configuration, it would be easy to discover the true capacity of a node, change the values on a XML file and replicate it to all other nodes inside the environment. The problem becomes evident once the environment is not homogeneous, as it would require the discovery of the true capacity of each node and the creation of separated XML files for each node. Indeed, quite often a cluster configuration does not follow real nodes characteristics, being limited to default values.

B. Experimenting Resource Allocation

To better understand the impact of configuration parameters on the resource allocation mechanism, we present in this section an experiment where we compare different memory requests against the minimum and maximum memory parameters. We considered four different scenarios: (*i*) default allocation, (*ii*) request higher than maximum allowed, (*iii*) request smaller than minimum allowed and (*iv*) request inside the range.

Table I: RESULTS FOR RM MEMORY ALLOCATION EXPERIMENT.

	Default	Higher	Smaller	In Range
Minimum Memory (MB)	1024	512	2048	512
Maximum Memory (MB)	8192	768	8192	8192
Map Memory Request (MB)	1024	1024	1024	3456
Reduce Memory Request (MB)	1024	1024	1024	3712
Allocated Map Memory (MB)	1024	ERROR	2048	3584
Allocated Reduce Memory (MB)	1024	ERROR	2048	4096

The results from these scenarios can be seen in Table I. The columns represent each scenario, while the first two lines (Minimum/Maximum Memory) refer to configuration parameters. The third and fourth lines (Map/Reduce Memory Request) refer to the application request parameters. Finally, the last two rows (Allocated Map/Reduce Memory) present the resources effectively allocated to the job based on the other parameters.

From these scenarios, we observe that a request with a value higher than the maximum will cause an error that aborts the job. For a request of a value smaller than the minimum, the cluster grants the minimum allowed. The fourth scenario shows a request inside the valid range. Although the requests were similar, the resources granted were different. Indeed, when the request is in the minimum-maximum range, Hadoop performs a small set of calculations to determine how much memory will be granted. Whenever the minimum allocation does not satisfy the request, the granted value is incremented by the minimum allocation until it matches one of the following cases: *(a)* the value is equal to the request; *(b)* the value is higher than the request and lower than the maximum allocation; or *(c)* the value exceeds maximum allocation.

This experiment demonstrates that the default scheduling is closely related to the resource availability. Having a wrong information could ruin the performance of the algorithm. Since there is no mechanism to automatically detect and modify resource parameters, dealing with a heterogeneous environment, such as a pervasive grid, quickly becomes a challenging task. It appears then clear that, in order to support heterogeneous environments, Hadoop must be aware of its real (and not supposed) execution environment.

V. CONTEXT-AWARE SCHEDULING

In order to detect the node real capacity, we chose to integrate a context collector into Hadoop. This collector is charged of observing the execution environment, allowing an automatic detection of each node capacity. Thanks to this context collector, context information representing real memory and CPU conditions of each node can be observed, allowing the proposal of improved scheduling mechanisms.

A. Collecting Context Information

Context information corresponds to a large concept, often related to the observation of a user, a device or the execution environment. Commonly, it is defined as any information that may characterize the situation of an entity. This entity can be a user, a device or the environment itself [7]. Quite often, context information is used for adaptation purposes [9]. Context

 ISBN: 978-1-61208-353-7

awareness can then be seen as the capability a system has of observing and reacting to the environment in order to adapt its own behavior to context changes [8][9]. Different context information can be observed for adaptation purposes, varying from user's profile, location and activities till characteristics of the used device and execution environment [9][10][19]. Indeed, several works in the literature [11][12][20][21] propose observing device execution conditions (including available memory, CPU, network connection, battery consumption, etc.) in order to adapt application execution to them, by selecting or deploying appropriate components. In our case, we are interested on the execution context of a job, which is composed by the nodes executing it. We believe that Hadoop must be aware of this execution context in order to schedule appropriately submitted jobs. Among such information, we can include CPU and memory capacities, node's current charge, but also network speed and data locality (related to HDFS replication) to improve task allocation.

Nonetheless, in order to be useful, context information should be acquired and modeled appropriately, with a minimal impact to the overall performance of the running applications. This is particularly true for Hadoop, whose goal is precisely to improve performance of MapReduce applications [22]. A lightweight mechanism, in the opposite to traditional context management systems [8][9], is then needed.

Thus, to include context information on Hadoop, we integrated a lightweight collector module, using the Java Monitoring API (Application Programming Interface) [23], which allows to easily access the real characteristics of a node, with no additional libraries required. The collector module, illustrated by Fig. 2, allows observing different context information, such as the number of processors (cores) and the system memory, using a set of interface and abstract classes that generalize the collecting process. Due to its design, it is easy to integrate new collectors and improve available context information for the scheduling process, providing data about the CPU load or disk usage, for example.

Context information is described by using a predefined name and a description. Such name corresponds to a concept identified in a context ontology. This model, inspired from Kirsch et al. [19], considers each context information as a context element, for which multiple values can be observed. Context ontology allows then to semantically describe each element, while the description gives a human readable definition for it.

This collector module was integrated to the NodeManager, since it is in charge of processing tasks and managing node definition. In this first prototype, we collect node capacity (available memory and number of cores), and this information is then sent to the ResourceManager. As a consequence, the information from the context collector module allowed us to improve the Hadoop scheduler operation without having to modify its implementation. This is especially interesting as further works will be able to compare other schedulers from the literature without having to modify their implementation.

B. Integrating Context Information

Context information detected using the context collector is transferred to Hadoop scheduler, which can scale the allocation limits to the real cluster resource availability. This scaling affects the containers allocation as a function of the available

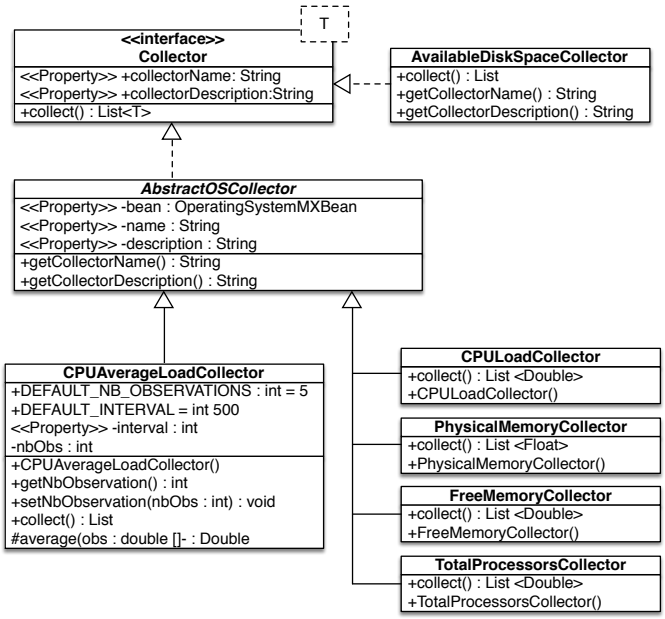


Figure 2: Elements of the context collector for Hadoop.

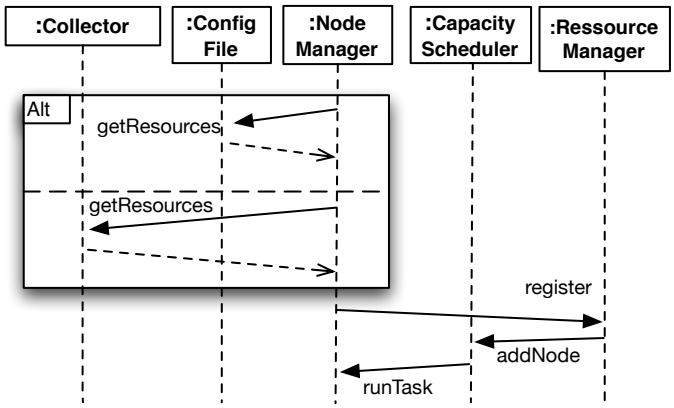


Figure 3: Simplified sequence diagram for resource registering and granting.

memory and computing cores, impacting therefore on the choice of tasks placement and how speculative task are started. As a result, we could obtain a better usage of the resources, minimizing the need for speculative tasks too. By adapting the capacity to the cluster real resource, no resource would be wasted or left inactive while the scheduler is making tasks wait due to wrong information being received.

In order to integrate the collector, we identified the NodeManager (NM) as the best entry point, since this is the service responsible for processing tasks. Collected context information about available memory and number of cores is sent to the ResourceManager (RM). When each NM registers to the RM, it tries to obtain this information from the context collector, which supplies NM with observed values. This information is sent and provided to the CapacityScheduler that uses it to dispatch tasks. Fig. 3 illustrates this process. It is worth noting that, if the collector is unavailable, NM will keep using traditional configuration files.

 ISBN: 978-1-61208-353-7

Information from the context collector module allows us to improve the behavior of the Hadoop scheduler without having to modify its implementation. This is especially interesting as further works will be able to compare other schedulers from the literature without having to modify their implementation.

VI. Experiments and Results

This section provides information about the experiments we conducted to improve the Hadoop scheduling behavior, as well as the results achieved. These experiments consisted in deploying Hadoop services in a cluster with original CapacityScheduler and comparing it against a context-aware CapacityScheduler. Theses experiments focus on two basic parameters, available memory and number of cores, whose injection in the CapacityScheduler is straightforward and require no modifications to the CapacityScheduler algorithm. Additional context information can be included through the collector module, as described in the Section V-A.

The experiments were performed in a cluster from the Grid'5000 [24] computing environment. We used five nodes (a master and 4 slaves), each having the following configuration: 2 AMD CPUs 1.7GHz, 12 cores/CPU and 48 GB of RAM. All nodes run Ubuntu-x64-12.04, with JDK 1.7 installed, and the Hadoop distribution was the 2.2.0 YARN version. As The TeraSort benchmark was used as application subject.

A. Results and interpretation

With an experiment running on nodes with the same configuration, it would be easy to discover the true capacity of a node, change the values on a XML file and replicate it to all other nodes inside the environment. The problem becomes evident once the environment is not homogeneous, as it would require the discovery of the true capacity of each node and the creation of separated XML files for each node. Indeed, quite often a cluster configuration does not follow real nodes characteristics, being limited to default values.

Our first experiment compares the allocated node memory when using the default implementation or our context collector. Thanks to the context collector, we could be able to detect real node characteristics, which are significantly different from the default values, as one can see in Table II. This discrepancy in capacity collected/used is due the utilization of default XML parameters by the default scheduler. As stated before, the Hadoop configuration is heavily dependent on XML files, making it hard to extract the full potential of the cluster without a delicate and time-consuming configuration. The default XML files have the node memory set to 8 GB and the node number of cores set to 8, that are reasonable numbers when using a cluster of personal computers, but when deployed in a larger cluster these values will, more often than not, waste potential.

Table II: Resources available on original and context-aware CapacityScheduler.

	Original CapacityScheduler	Context-aware CapacityScheduler
Node Memory	8 GB	48 GB
Node Vcores	8	24

The second experiment compares the behavior of the original CapacityScheduler with our own context-aware CapacityScheduler. This experiment used the same configuration from the previous experiment. We launched a TeraSort job with 5 GB data to sort, therefore requesting enough containers and providing enough data to stress the cluster. The original CapacityScheduler uses the default configuration, with a minimum allocation of 1 GB and 1 core, maximum allocation of 8 GB and 32 cores per node for a total resource summing up 32 GB and 32 cores for the cluster. For the context-aware CapacityScheduler, the collector detects all 192 GB and 96 cores on the cluster (48 GB and 24 cores from each node), with a minimum allocation of 4 GB and 2 cores and maximum allocation of 24 GB and 12 cores per node.

Figs. 4 and 5 present a chart with tasks execution (actually the containers) and the nodes they are tied to. The different segments indicate the tasks that have been allocated to a given NodeManager, and the numbers inside the segment indicates which containers are running at that moment. When a segment ends, it means that at least one task has finished or a new task has been started on that NodeManager. If a number was on a segment and disappears on the next, that task has finished. If a number wasn't on the first segment and suddenly appears on the next, that task has started. Because the default configuration uses one single Reduce task, we simplify the diagram by representing only the Map tasks.

Fig. 4 portraits the execution of the TeraSort algorithm with the original CapacityScheduler. One can notice that some containers had to wait for the completion of others in order to start processing their tasks. Indeed, Hadoop splits the work in 38 Map tasks (numbered 2-39), which are distributed to the nodes according to the known resource capabilities. When the first tasks are completed, new tasks are provided to the nodes, if any available (as illustrated in Fig. 4, where tasks 32-39 represent the second execution wave).

Fig. 5 portraits the execution of the TeraSort algorithm with context-aware CapacityScheduler. In this case, the overall completion time was reduced due to the fact that all containers could be started right after the arrival of the request, thanks to the higher resource availability.

After an analysis and comparison of both charts, it is possible to notice that the default chart has containers 41-43 started on node stremi-5 and container 44 started on node stremi-42, while the context-aware chart has only the standard containers, which are numbered 2-39. These extra tasks are what Hadoop calls speculative tasks, when one or more tasks are advancing slower than the rest, Hadoop creates new copy tasks in order to prevent bigger losses in throughput. This means that if the original tasks were indeed experiencing issues, in the event they fail to complete, another task is already processing the faulty task data, otherwise, if the task eventually finishes its processing before the speculative, the copies will be disregarded.

B. Heterogeneity Simulation

A third experiment was performed to simulate a heterogeneous environment and test how well the context-aware would adapt. Once again, the experiment consisted in executing the TeraSort algorithm in the cluster with the simulated heterogeneous environment using context-aware CapacityScheduler.

 ISBN: 978-1-61208-353-7

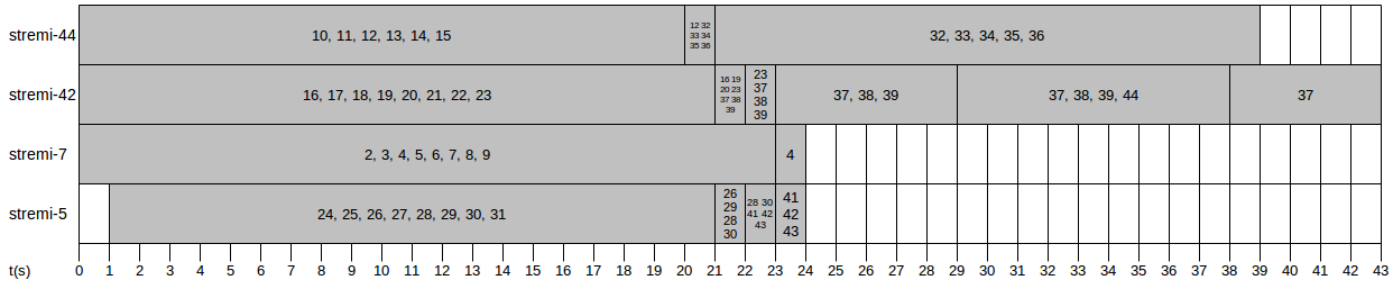


Figure 4: Container assignment with the original CapacityScheduler.

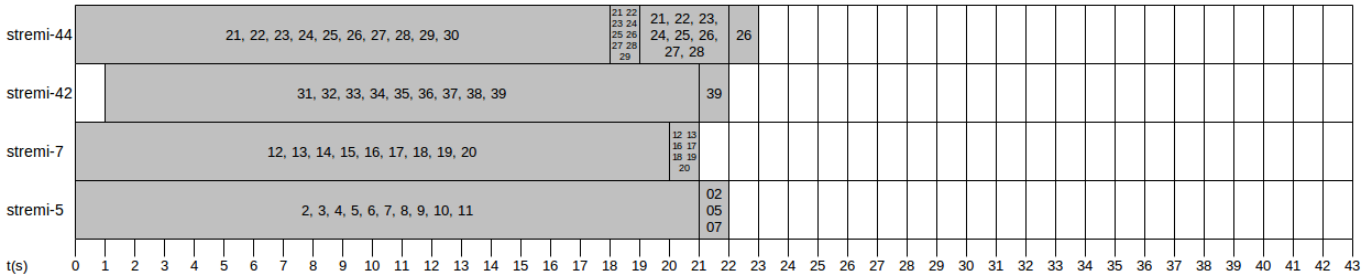


Figure 5: Container assignment with the context-aware CapacityScheduler.

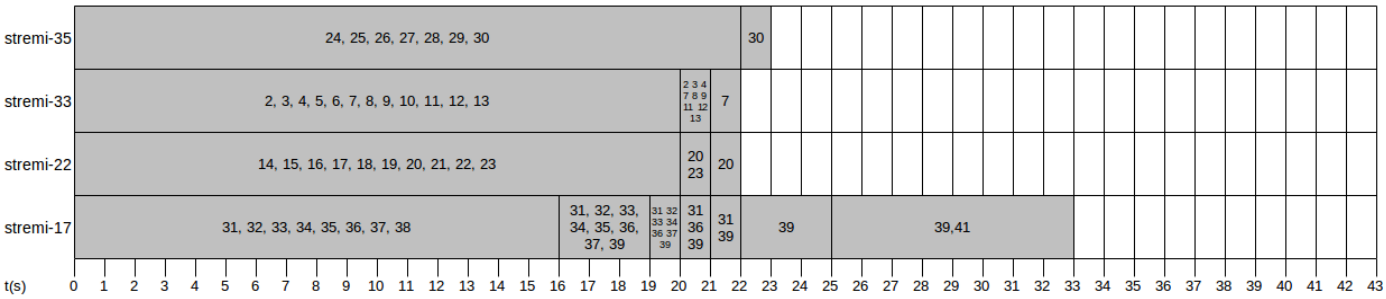


Figure 6: Container assignment in the simulated heterogeneous environment.

This experiment used the same configuration from the previous experiment. The only difference is that the nodes are purposely given false capacities when being added to the RM, simulating the following heterogeneous cluster:

- stremi-17: 28 GB of memory and 14 cores.
- stremi-22: 32 GB of memory and 18 cores.
- stremi-33: 48 GB of memory and 24 cores.
- stremi-35: 24 GB of memory and 12 cores.

Fig. 6 portraits the execution of TeraSort within the simulated heterogeneous environment, also using context-aware CapacityScheduler. Compared to the default case, the heterogeneous execution shows an improvement, but due to lower cluster capacity, it is slightly worse than the context-aware scheduler on homogeneous environment.

On this experiment a speculative task was launched, the container 41. It is also noteworthy that the scheduler did not change nodes to launch the speculative task, because the node had spare capacity when the request for the speculative arrived.

This experiment shows that it is possible to use this context-aware scheduler in a heterogeneous environment. Indeed, the allocations were adapted to a slightly smaller cluster if compared to the real environment. As a future work, it is possible to set the allocation limits in function not only of total cluster resources but also of each individual node resource capacity.

VII. Conclusion and Future Work

In this paper, we have proposed to improve Apache Hadoop behavior with context information, thanks to a new context-aware scheduler. These changes allowed Hadoop to be aware of its execution context, and particularly of the real capacity of the nodes composing the cluster. The context-aware CapacityScheduler we have proposed here is capable of receiving the real capacity from each NodeManager, thanks to a lightweight context collector plugged on NodeManager. This provides the cluster a better scaling potential while also using every node's full capacity. Experimental results demonstrate that the context-aware CapacityScheduler could better scale up improving containers management, and consequently the overall Hadoop scheduling behavior.

This context-aware scheduling represents a first step of a further vision, proposed by the PER-MARE project [25][26]. Indeed, we intend to go further in this direction, considering not only nodes capabilities, but also current state (current available memory, CPU load or network bandwidth, for instance). We strongly believe that such a context-aware behavior is essential for supporting MapReduce application over pervasive grids.

 ISBN: 978-1-61208-353-7

ACKNOWLEDGMENT

The authors would like to thank their partners in the PER-MARE project [26] and acknowledge the financial support given to this research by the CAPES/MAEE/ANII STIC-AmSud collaboration program (project number 13STIC07). Experiments presented in this paper were carried out on Grid'5000 [24] experimental testbed.

REFERENCES

[1] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," Commun. ACM, vol. 51, no. 1, 2008, pp. 107–113.

[2] The Apache Software Foundation, "Apache hadoop," 2014, [retrieved: Mar. 2014]. [Online]. Available: http://hadoop.apache.org/

[3] M. Parashar and J.-M. Pierson, "Pervasive grids: Challenges and opportunities," in Handbook of Research on Scalable Computing Technologies, K.-C. Li, C.-H. Hsu, L. T. Yang, J. Dongarra, and H. Zima, Eds. IGI Global, 2010, pp. 14–30.

[4] V. Hingne, A. Joshi, T. Finin, H. Kargupta, and E. Houstis, "Towards a pervasive grid," in Proceedings of the International Parallel and Distributed Processing Symposium, ser. IPDPS'03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 207.2–, [retrieved: Mar. 2014]. [Online]. Available: http://ebiquity.umbc.edu/_file_directory_/papers/623.pdf

[5] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis," Nature Reviews Genetics, vol. 11, no. 9, Sept 2010, pp. 647–657.

[6] K. A. Kumar, V. K. Konishetty, K. Voruganti, and G. V. P. Rao, "Cash: context aware scheduler for hadoop," in International Conference on Advances in Computing, Communications and Informatics, ser. ICACCI '12. New York, NY, USA: ACM, 2012, pp. 52–61.

[7] A. Dey, "Understanding and using context," Personal and Ubiquitous Computing, vol. 5, no. 1, 2001, pp. 4–7.

[8] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems," International Journal of Ad Hoc and Ubiquitous Computing, vol. 2, no. 4, 2007, pp. 263–277.

[9] D. Preuveneers, K. Victor, Y. Vanrompay, P. Rigole, M. Kirsch-Pinheiro, and Y. Berbers, Context-Aware Mobile and Ubiquitous Computing for Enhanced Usability: Adaptive Technologies and Applications. IGI Global, 2009, ch. Context-Aware Adaptation in an Ecology of Applications, pp. 1–25.

[10] M. Kirsch-Pinheiro, M. Villanova-Oliver, J. Gensel, Y. Berbers, and H. Martin, "Personalizing web-based information systems through context-aware user profiles," International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2008), 2008, pp. 231–238.

[11] D. Preuveneers and Y. Berbers, "Context-driven migration and diffusion of pervasive services on the osgi framework," International Journal of Autonomous and Adaptive Communications Systems (IJAACS), vol. 3, no. 1, Dec. 2010, pp. 3–22.

[12] C. Louberry, P. Roose, and M. Dalmau, "Kalimucho: Contextual deployment for qos management," in Distributed Applications and Interoperable Systems, ser. Lecture Notes in Computer Science, P. Felber and R. Rouvoy, Eds. Springer, 2011, vol. 6723, pp. 43–56.

[13] C. Tian, H. Zhou, Y. He, and L. Zha, "A dynamic mapreduce scheduler for heterogeneous workloads," in 8th International Conference on Grid and Cooperative Computing, ser. GCC '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 218–224.

[14] A. Rasooli and D. G. Down, "Coshh: A classification and optimization based scheduler for heterogeneous hadoop systems," in Proceedings of the 2012 SC Companion: High Performance Computing, Networking Storage and Analysis, ser. SCC '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 1284–1291.

[15] M. Isard, V. Prabhakaran, J. Currey, U. Wieder, K. Talwar, and A. Goldberg, "Quincy: fair scheduling for distributed computing clusters," in ACM SIGOPS 22nd Symposium on Operating Systems Principles, ser. SOSP '09. ACM, 2009, pp. 261–276.

[16] M. Zaharia, A. Konwinski, A. D. Joseph, R. Katz, and I. Stoica, "Improving mapreduce performance in heterogeneous environments," in Proceedings of the 8th USENIX Conference on Operating Systems Design and Implementation, ser. OSDI'08. Berkeley, CA, USA: USENIX Association, 2008, pp. 29–42.

[17] Q. Chen, D. Zhang, M. Guo, Q. Deng, and S. Guo, "Samr: A self-adaptive mapreduce scheduling algorithm in heterogeneous environment," in Proceedings of the 2010 10th IEEE International Conference on Computer and Information Technology, ser. CIT '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 2736–2743.

[18] J. Xie, S. Yin, X. Ruan, Z. Ding, Y. Tian, J. Majors, A. Manzanares, and X. Qin, "Improving mapreduce performance through data placement in heterogeneous hadoop clusters," in Parallel and Distributed Processing, Workshops and Phd Forum (IPDPSW), 2010, pp. 1–9.

[19] M. Kirsch-Pinheiro, J. Gensel, and H. Martin, "Representing context for an adaptative awareness mechanism," in Groupware: Design, Implementation, and Use, ser. LNCS, G.-J. Vreede, L. Guerrero, and G. Marín Raventós, Eds., vol. 3198. Springer, 2004, pp. 339–348.

[20] M. Baldauf and P. Musialski, "A device-aware spatial 3d visualization platform for mobile urban exploration," Fourth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2010), 2010, pp. 47–52. [Online]. Available: http://www.thinkmind.org/index.php?view=article&articleid=ubicomm_2010_3_20_10127

[21] J. Floch, C. Frà, R. Fricke, K. Geihs, M. Wagner, J. L. Gallardo, E. S. Cantero, S. Mehlhase, N. Paspallis, H. Rahnama, P. A. Ruiz, and U. Scholz, "Playing music - building context-aware and self-adaptive mobile applications," Software: Practice and Experience, vol. 43, no. 3, 2013, pp. 359–388.

[22] M. Kirsch-Pinheiro, "Requirements for context-aware mapreduce on pervasive grids," PER-MARE Deliverable D3.1, Deliverable D3.1, 2013, [retrieved: Mar. 2014]. [Online]. Available: http://hal.archives-ouvertes.fr/hal-00858310

[23] Oracle, "Monitoring and management for the java platform," 2014, [retrieved: Mar. 2014]. [Online]. Available: http://docs.oracle.com/javase/7/docs/technotes/guides/management/

[24] Grid'5000, "Grid'5000," 2014, [retrieved: Mar. 2014]. [Online]. Available: https://www.grid5000.fr

[25] L. Steffenel, O. Flauzac, A. S. Charao, P. P. Barcelos, B. Stein, S. Nesmachnow, M. K. Pinheiro, and D. Diaz, "PER-MARE: Adaptive deployment of mapreduce over pervasive grids," in 8th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), Oct 2013, pp. 17–24.

[26] STIC-AmSud, "PER-MARE project," 2014, [retrieved: Mar. 2014]. [Online]. Available: http://cosy.univ-reims.fr/PER-MARE

 ISBN: 978-1-61208-353-7

U-Lab Cloud: A Ubiquitous Virtual Laboratory Based on Cloud Computing

Rafaela Ribeiro Jardim, Eduardo Lemos, Fabricio Herpich, Ricardo Bianchim, Roseclea Medina
PPGI - Post graduation Program in Informatics
Federal University of Santa Maria (UFSM)
Santa Maria, Brazil
(rafa.rjardim, elemos04, fabricio.herpich, ricardo, roseclea.medina) @gmail.com

Felipe Becker Nunes
Computer Education Post Graduate Program (PPGIE)
Federal University of Porto Alegre (UFRGS)
Porto Alegre, Brazil
nunesfb@gmail.com

***Abstract*— This paper describes a proposal of U-Lab Cloud, a ubiquitous virtual laboratory based on Cloud Computing. This environment will provide hardware and software resources dynamically, making it possible to increase the resources as needed. The environment will be adapted to the user's context and will consider the individual characteristics of each student, such as connection speed, device type and cognitive style of the student.**

Keywords- u-learning; laboratory virtual; cloud computing.

I. INTRODUCTION

The use of virtual laboratories is essential for acquiring practical skills [1]. These laboratories allow students to perform simulations and practice theories without the need of a physical laboratory, because the construction of such laboratory requires high investment costs and adequate infrastructure. Moreover, others difficulties found are: the need of exclusive use of these resources and its quick obsolesce. [2]. In this scenario, it is necessary to provide an accessible environment that meets the individual characteristics of each student. According to Piovesan [3], U-learning environments provide access to educational resources with full mobility and an adaptation system to the computational context of students. Facing the need of providing a personalized educational environment to the context of the students, this work aims to provide the U-Lab Cloud, a ubiquitous virtual laboratory. Its objective is to identify variables of the user context, such as the student's cognitive profile and connection speed. With the identification of these characteristics, the U-Lab Cloud will indicate what type of content presentation is the most appropriate for that student, then U-Lab Cloud will dynamically provide the software and hardware resources available from a Cloud Computing environment.

Some analysts predict that by 2020, most of the world's digital data will be handled, monitored and/or stored in clouds - if not throughout their whole life cycle, at least in part of it [4]. Thus, Cloud Computing emerges as a technology model that allows access to the network in demand, in favor of configurable shared computing resources (e.g., networks, servers, storage, applications and services) [5]. Most existing solutions for U-learning, based on Cloud Computing, implement some models of cloud services without integration with a learning service [2][6][7].

Therefore, this work aims to integrate the latest technology trends, ubiquitous computing and Cloud Computing, in order to present the same potential in educational settings, given the diversity of available platforms, operating systems and patterns.

II. CLOUD COMPUTING

Cloud Computing is based on surveys of virtualization, shared computing, "grid computing" and more recently, networks, the web and software services [7]. Cloud Computing is defined as "a computer model with the ability to allow, in a ubiquitous and convenient way, the access to shared and configurable computing resources" [8]. One of the advantages of using cloud services is the reduced concern of the loss of data or the intrusion of a virus [9]. The use of virtualization and cloud technologies provide versatile management of computing systems, because it can be easily deployed, scaled, replicated and updated in any of the levels of the Cloud Computing [1].

Cloud Computing is an emerging computing model in which users can have ubiquitous access to their applications [10]. According to Dey [11], it promotes the idea of ubiquitous exchange of information anytime and anywhere by the use of transparent, intelligent and integrated computer technologies. This way, Cloud Computing can be used to support ubiquitous environments. However, the implementation of a ubiquitous, scalable and reusable educational Cloud Computing architecture still faces big challenges in the areas of technology advancement and better practices [9]. Although several studies address Cloud Computing in Education [1][12], the lack of a systematic description during the development of the Cloud Computing platform is noted, because most papers discuss only techniques, types of services and others.

Among many characteristics of Cloud Computing, according to [1] using this technology allows the management of versatile computing systems, because it can be deployed, scaled, replicated and updated easily in any of the levels of the computing cloud. To Liang and Yang [9], one of the biggest advantages of using cloud services is the reduced concern about data loss or the intrusion of viruses. Cloud computing is an emerging computing model that the users can have ubiquitous access to their applications from any connected device [10]. When analyzing the computational cloud model, one realizes that one of its biggest benefits is the

 ISBN: 978-1-61208-353-7

flexibility that it provides to the user. Therefore, the user needs to have in his computer device only a browser installed to access the desired application, without the need for installing software or even performing updates.

Another potential of Cloud Computing is the possibility of sharing data and conducting collaborative work practices that facilitate working in educational settings. For example, to centrally store all data in one place with the same format, it excludes the need for conversions and adaptations.

In addition, large financial investment to use Cloud Computing in institutions is not necessary, since this infrastructure requires little computational resources, i.e., it can be used by computers with simple hardware (personal computers). Spending on software license is not necessary because by joining this alternative you have the possibility of using open source operating systems, which can be installed for free with no restrictions on the number of machines.

So, it was estimated that the use of this new paradigm of computing can bring many operational and financial benefits. However, there are some significant obstacles when using Cloud Computing. A major disadvantage of using this model points to data security and asks whether it is possible to maintain total security in data traffic. The unavailability of service can be another obstacle to the success of this computational model, since it is necessary to connect to the Internet to access certain data.

III. U-LEARNING

The term Ubiquitous Computing has emerged in the 90s, initially proposed by Mark Weiser. This concept promotes the idea of the exchange of information anytime and anywhere through the use of a transparent, intelligent and integrated way of computer technology [13]. Yahya et al. [14] describes the technological developments, especially the expansion of computing and communication abilities of small electronic devices as responsible for the progress of electronic-learning (e-learning) to mobile learning (m-Learning) and m-learning to the ubiquitous learning (u-Learning). So, the u-learning automatically adapts resources and content according to the preferences and needs of students at any time and place, observing the characteristics, context and available resources [3][15].

In [7], Ubiquitous computing environments are defined as "an area that incorporates a set of embedded systems (computers, sensors, user interfaces, and service infrastructure) that is reinforced by technological computation and communication." The ubiquitous computing environments allow learning the right thing, at the right place and time, and in the right path [3]. Piovesan [3] complements that the u-learning environment allows "access to educational resources with full mobility and adapting the system to the computational context of the students".

A feature of Ubiquitous Computing is the sensitivity to the context. In [11], a definition of context is any information that can be used to characterize the situation of entities that are considered relevant to the interaction between a user and an application.

To Pernas et al. [15], sensitivity to context refers to everything that occurs around the user, and that influences how it interacts with the environment and with other people. According to Dey [11], context-sensitive computing should be aware of the state of the user and its surroundings, and must modify their behavior based on this information.

IV. RELATED WORKS

Cloud Computing proposes the integration of various technological models for the provision of hardware infrastructure, platforms of development and applications as a service [6]. The virtualization is one of the prerequisites for the realization of Cloud Computing [5]. Therefore, it allows the execution of multiple virtual machines (VMs) on a single physical machine. Among the many advantages of Cloud Computing applications with focus on teaching, one can mention scalability, centralized storage, flexibility, accessibility and low costs [16]. Cloud Computing is an infrastructure that has been explored in educational environments [1][5][15][16][18][19]. It is important to highlight that for the proper choice of the Cloud Computing platform, some aspects need to be considered: the Information Technology infrastructure and the most appropriate service model type to be implemented in the educational institution. An application can use one or more types of service, including: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [17].

For example, Wenhong et al. [16] proposes a framework for managing PaaS in a virtual laboratory based on Cloud Computing that implements user management, resources and accesses, but it approaches the subject in general and does not focus on the needs of an educational system.

In [18], an infrastructure solution is presented in a private cloud that addressed a combination of three service models (IaaS, SaaS, PaaS); it was developed in the University of Hochschule Furtwangen (HFU). This solution allows the creation of Virtual Machines from choosing an image and software packages by the user. Similarly, Vouk et al. [19] describes a Virtual Laboratory with a Cloud Computing solution; it was developed by the North Carolina State University (USA). It enables students to reserve VMs according to their needs, choosing from basic images or images with specific applications. This platform is being used by a large number of users and currently uses a service (SaaS) model, providing a variety of specific applications for different areas of study.

In [1], a system based on Cloud Computing and virtualization directed to the engineering education was presented. This virtual laboratory provides hardware and software resources, where practical work from VMs can be developed.

The problems mentioned during the implementation were related to network communication and the creation of VMs on OpenNebula. Similarly, Wang et al. [12] provides a laboratory for network administration, implemented with the integration of VMware and Lab Manager, having as main characteristic the flexibility of resources.

 ISBN: 978-1-61208-353-7

Comparing with the mentioned works, the U-Lab Cloud differs from them because it will provide an environment adapted to the peculiarities of the students, for example, cognitive profile, the device they will be using and their connection speed. Another difference is that the U-Lab Cloud will provide students a ubiquitous environment and may be accessed from any device with an Internet connection, but without the need to make downloads or install applications.

V. PROPOSAL

The proposal of this paper is to supply from a platform of Cloud Computing the U-Lab Cloud, a ubiquitous virtual laboratory.

Among several factors that motivated this work, one of them is the problems faced by students when they fail to download certain materials, because they are using a low speed connection. This barrier discourages students, making them give up from seeking access to the materials. So, the proposed environment aims to get information from the user's context, such as: the connection speed to the cognitive style can create a user's profile. Then, it will provide the presentation of the content, adapted to each user's particularity.

Another particularity of this environment is that it will supply dynamic hardware and software resources, making it possible to increase or decrease these resources as it is needed for use.

The Cloud Computing is divided in three layers: IaaS, PaaS and SaaS [17]. So the proposed architecture for the U-Lab Cloud will approach the model in layers of services, as it can be visualized in Figure 1.

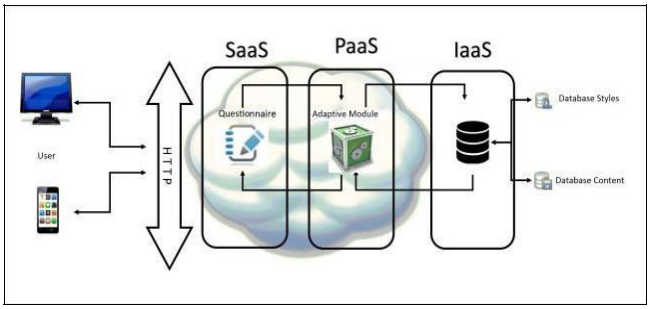


Figure 1. Architecture of U-Lab Cloud.

The SaaS layer is the most superficial one of this environment; it will provide the interface for the student. It will be available over Internet and it will be able to be accessed from any device with a Web browser. In this layer, students will have access to the provided applications, the adapted content and the proposed activities.

For the intermediate PaaS layer, it is suggested an Adaptive Module that will do the treatment of the student's information context. This is how it is going to work: first, it will be collected preferences from the student through the answering of a survey. Then the connection speed of the device that the student is using will be collected. After that, the Adaptive Module will determine the context style of the user, based on the connection speed and the materials that were chosen.

Finally, the Adaptive Module will communicate with the Database (DB) stored on the IaaS layer, and it will adapt the presentation of the content in the U-Lab Cloud, according to each student's context.

U-Lab Cloud will allow students to use the most recent software versions, independently from the hardware they have, because it will provide these resources in a dynamic way. So, if the number of users get higher, it will be possible to increase these resources as needed, without the concern of the physical infrastructure. To provide this environment, after performing a performance analysis which will be done with very similar hardware machines, a Cloud Computing platform will be selected.

Students will be able to access the U-Lab Cloud by any technology device that has an Internet connection that has a very minimal requisite: only a Web browser. It will be used in the desired time, from anywhere, from any device, making this approach ubiquitous.

VI. METHODOLOGY

To develop he U-Lab Cloud, five distinct steps were scheduled. The first is characterized by the construction of the theoretical framework, where we will seek to study the main concepts that will be used in its development. The second step also involves the implementation of a private cloud on a platform of Cloud Computing; the Eucalyptus version 3.4.2 with CentOS was used. Eucalyptus was chosen for providing open source versions and presenting the initial requirements for the context of this work.

The installation of this platform will be held on the network server located in the Network Group and Applied in the IT Laboratory that is located at the Federal University of Santa Maria in the city of Santa Maria, Brazil. This server has an Intel Core2 Quad Q9300 2.66 GHz processor with 4 physical cores of 8 GB of RAM and one hard drive with the capacity of storing 750GB with CentOS operating system. For the third stage of this work, it is planned to build the U-Lab Cloud, as well as the insertion of objects and Computer Networks related to teaching activities. The development of this stage involves the installation of Moodle [20], OpenSim [21], Sloodle [22] and WampServer [23] platforms.

Moodle was chosen because it is open source and is the virtual environment used at this university. OpenSim is the tool that makes the creation of the graphical interface of the virtual laboratory and it was chosen because it is open source, stable and allows the integration with the Moodle version. Sloodle will be used to integrate Moodle with OpenSim, and it allows the inclusion of activities on the topic of Computer Networks which is available to students through the U-Lab Cloud. Finally, the WampServer was used to create a local server.

In the fourth stage of this work, U-Lab Cloud will be validated in the discipline of Computer Networking, with students from the Computer Science Course of this institution. The fifth stage includes a statistical analysis that will be performed with data collected during the validation process.

 ISBN: 978-1-61208-353-7

VII. CONCLUSION

The lack of physical laboratories at the institutions made the virtual laboratories arise as a complementary alternative to practice theories. Several works using virtual laboratories and Cloud Computing were evidenced but most of them had static characteristics.

The difference of this proposal is that it will be considered the variables from the student's context, because these particularities influence directly in the teaching and learning process.

With this work, it is intended to provide a ubiquitous environment, based on the principles of Cloud Computing and virtualization. These technologies will be selected after the realization of performance tests using almost homogenous machines. This work has presented the proposed architecture for the development of U-Lab Cloud, describing the service layers that it will approach.

With the fulfillment of U-Lab Cloud, it is intended to supply scalability, availability and ubiquity to students and teachers. With the completion of this study, it is intended to demonstrate the viability of the Cloud Computing infrastructure platform in the educational environment, as well as the advantages in using it.

REFERENCES

[1] A. C. Caminero, R. Hernandez, S. Ros, L. Tobarra, A. Robles-Gomez, E. San Cristobal, M. Tawfik and M. Castro, "Obtaining university practical competences in engineering by means of virtualization and Cloud Computing technologies," in Teaching Assessment and Learning for Engineering (TALE), IEEE International Conference, pp. 301-306, August 2013.

[2] T. R. S. Rauen, "An alternative approach to teaching computer networks". Master's Dissertation - Federal University of Santa Catarina, pp.85, 2003.

[3] S. Piovesan, "U-SEA: A Ubiquitous Learning Environment Using Cloud Computing". Master's Dissertation, Federal University of Santa Maria, pp.84, 2011.

[4] H. Costa, "Large Masses of Data in the Cloud: Challenges and Techniques for Innovation," in XXX Brazilian Symposium on Computer Networks and Distributed Systems, 2012.

[5] Nist. *The Nist Definition of Cloud Computing*., [Online]. Available from: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf /. 2014.05.25.

[6] T. Sá, J. Smith, and D. Gomes "Cloudreports: A graphical tool for simulation of Cloud Computing environments based on cloudsim framework," in IX Workshop on Clouds and Applications – WCGA, 2011.

[7] M. D. Zrakić, K. Simić, A. Labus, A. Milić and B. Jovanić, "Scaffolding Environment for Adaptive E-learning through Cloud Computing," in Educational Technology & Society, pp. 301–314, October 2012.

[8] P. Mell and T. Grance, "The Nist Definition of Cloud Computing," in National Institute of Standards and Technology, pp.7, 2011.

[9] P. Liang and J. Yang, "Virtual Personalized Learning Environment (VPLE) on the cloud" in Lecture Notes in Computer Science, v. 6988, pp.403-411, 2011.

[10] I.B.M Corporation. *Seeding the Clouds: Key Infrastructure Elements for Cloud Computing.* [Online].Available from: http://www.software.ibm.com/common/ssi/sa/wh/n/oiw03022usen/OIW03022USEN.pdf/ 2014.05.22.

[11] A. Dey, "Understanding and using context", in Personal and Ubiquitous Computing, vol. 5, pp. 4–7, Feb. 2001.

[12] X. Wang, C. Hembroff and R. Yedica, "Using the VMware vCenter Lab Manager in undergraduate courses in system administration and network security." in Proc. da 10ª Conference on Information technology education (SIGITE), pp. 43-52, 2010.

[13] M. Weiser, "The Computer for the 21st Century" in Scientific American, [S.l.], v.265, pp. 94–104, 1991.

[14] S. Yahya, A. Ahmad and A. Jalil, "The definition and characteristics of ubiquitous learning: a discussion" in International Journal of Education and Development using ICT, [S.l.], v.6, n.1, p.117–127, February 2010.

[15] A. Pernas, I. Gasparini, J. Oliveira and M. Pimenta "An Adaptive ODL Environment Considering User Context," Federal University of Pelotas, p.6, 2009.

[16] T. Wenhong, S. Sheng and L. Guoming, "A framework for implementing and managing platform as a service in a virtual Cloud Computing lab," in The Second International Workshop on Education Technology and Computer Science, pp. 273-276, March 2010.

[17] Webgranth. *A Complete Reference to Cloud Computing*. [Online] Available from: http://www.webgranth.com/a-complete-reference-to-cloud-computing/ 2014.05.22.

[18] F. Doelitzscher, A. Sulistio, C. Reich, H. Kuijs and D. Wolf, "Private cloud for collaboration and e-Learning services: From IaaS to SaaS," in Journal Computing, v.91, pp. 23-42, January 2011.

[19] M. Vouk, S. Averitt, M. Bugaev, A. Kurth, A. Peeler, H. Shaffer, E. Sills, S. Stein and J. Thompson "Powered by VCL -Using Virtual Computing Laboratory (VCL) Technology to Power Cloud Computing," in International Conference on Virtual Computing, pp. 1-10, May 2008.

[20] Moodle. [Online]. Available from: http://docs.moodle.org/. 2014.05.19.

[21] OpenSim. [Online]. Available from: http://opensimulator.org/. 2014.05.22.

[22] Sloodle. [Online]. Available from: https://www.sloodle.org/. 2014.05.21.

[23] WampServer. [Online]. Available from: http://www.wampserver.com/en/.2014.05.18.

 ISBN: 978-1-61208-353-7

Ergodic Capacity Analysis for Ubiquitous Cooperative Networks Employing Amplify-and-Forward Relaying

Peng Liu, Saeed Gazor, and Il-Min Kim

Department of Electrical and Computer Engineering
Queen's University, Kingston, ON, Canada
Emails: {peng.liu, gazor, ilmin.kim}@queensu.ca

***Abstract*—This paper studies the ergodic capacity for ubiquitous cooperative networks employing amplify-and-forward relaying in Rayleigh fading. A general *asymmetric* channel model is considered, in which the average signal-to-noise ratios associated with different wireless channels are generally unequal. We derive an *exact* expression of the ergodic capacity in a single-integral-form, which serves as a benchmark for ubiquitous cooperative networks. To the best of our knowledge, this exact expression has not been reported in the literature. To evaluate this integral more efficiently, we develop a hybrid Gaussian quadrature expression in *closed-form*, which has a high relative accuracy. Finally, it is demonstrated that the obtained analytical results overlap the simulation curves, while the existing bounds are loose in various scenarios.**

Keywords–Amplify-and-forward; ergodic capacity; Rayleigh fading; relaying.

I. INTRODUCTION

Cooperative networks for ubiquitous communications have gained considerable attention in the last decade, due to their great potential to combat fading impairments [1]–[4]. The main idea of cooperative communications is that several geographically distributed wireless terminals, including the source and relay(s), collaborate with one another to form a virtual multi-antenna array, which enables *distributed* spatial diversity. Amplify-and-Forward (AF) is one of the most popular relaying protocols, in which the relay simply amplifies its received signal and forwards it to the destination [4]–[6].

A. Ergodic Capacity

The fading environment manifests itself in *ergodic* fading when the channel coherence time is (much) shorter than the codeword length, due to the usage of sufficiently long codewords and/or high mobility of wireless terminals [7]. In an ergodic fading scenario, each codeword typically spans *many* coherence time intervals, giving rise to rapid fading fluctuations *within* each codeword. The codeword is long enough to average out the randomness of fading, rendering Shannon capacity a deterministic constant independent of the instantaneous fading state. Mathematically, Shannon capacity of ergodic channels, a.k.a., ergodic capacity, is equal to the maximum achievable rate averaged over the fading distribution, which depends only on the fading statistics [7], [8]. Ergodic capacity is a fundamental information-theoretic performance measure, which captures the maximum rate of reliable communications under ergodic fading [7].

B. Related Work

It is of paramount importance to characterize the ergodic capacity, which serves as a benchmark for practical communication systems. However, due to the nonlinear expression of the end-to-end Signal-to-Noise Ratio (SNR) in AF relaying, only few works have studied the *exact* ergodic capacity for AF networks [9][10]. Specifically, the exact analysis of the ergodic capacity was limited *only* to a *symmetric* single-relay network *without* the the direct source-destination link [9, eq. (11)]. The major limitation of the *symmetric* channel setting is that the average received SNRs of the first- and the second-hop channels must be identical in Rayleigh fading. However, this does not necessarily hold in practice due to large-scale path-loss and shadowing effects. Furthermore, Fan *et al.* derived an *exact* expression of the ergodic capacity for a *multi*-relay network without the direct link where only a *single*-relay with the maximum second-hop SNR is allowed to assist the communication. Since the selection of the relay in [10][11] neglects the first-hop channel conditions, the achievable diversity order is *always* equal to one in Rayleigh fading, irrespective of the number of relays. That is, from a diversity point of view, the relaying scheme in [10][11] achieves the same performance as in a single-relay network with no direct link, which makes the exact analysis by Fan *et al.* [10] valid only for a single-relay network without the direct link.

In addition to the exact analysis, other works have been devoted to finding bounds or approximations of the ergodic capacity. Specifically, various upper bounds were obtained using the geometric-mean [12, eq. (8)], harmonic-mean [13, eq. (18)], and Jensen's inequality [12, eq. (5)], [14, eq. (10)]. In addition, series expansions of the ergodic capacity were developed in [9, eq. (4)], [12, eq. (17)], and [15, eq. (9)]. These bounds/approximations, unfortunately, suffer from low accuracy and/or high computational complexity. To the best of our knowledge, an *exact* expression of the ergodic capacity has not been reported for a general *asymmetric* AF cooperative network *with* the direct link, which motivates our work.

C. Contribution

To the best of our knowledge, the *exact* analytical expression of the ergodic capacity for a single-relay network *with* the direct link has not been reported in the literature, which motivates this work. In this paper, we carry out *exact* analysis under the general *asymmetric* channel setting where different channels have generally unequal average SNRs. The exact

 ISBN: 978-1-61208-353-7

expression of the ergodic capacity is derived in a single integral form for the single-relay network with the direct link. To numerically evaluate this expression, we further develop a hybrid Gaussian quadrature rule in *closed-form*, which achieves a high *relative* accuracy of 10^{-6}, and thus, is very suitable for real-time evaluations. These obtained expressions serve as a useful tool to analytically evaluate the performance limits of piratical AF relaying systems.

The remainder of the paper is organized as follows. Section II introduces the system model. Section III conducts the ergodic capacity analysis. Section IV validates the capacity analysis by simulations. Finally, Section V concludes the paper.

Notation: We use $x \triangleq y$ to denote that x, by definition, equals y. $\mathbb{E}(\cdot)$, $\ln(\cdot)$, and $\log_2(\cdot)$ denote the expectation, natural logarithm, and base-2 logarithm, respectively. For a random variable X, $f_X(\cdot)$ and $F_X(\cdot)$ are the probability density function (PDF) and cumulative distribution function (CDF), respectively. Finally, $X \sim \mathcal{CN}(\mu, \sigma^2)$ means that X is a circularly symmetric complex Gaussian random variable with mean μ and variance σ^2.

II. System Model

Consider a cooperative network composed of three single-antenna terminals: a source S, a relay R, and a destination D, where the direct S-D link exists. Let h_{sd}, h_{sr}, and h_{rd} denote flat fading gains for the S-D, S-R, and R-D links, respectively. The fading gains are assumed independent and modeled as $h_{ij} \sim \mathcal{CN}(0, \Omega_{ij})$, $ij \in \{sd, sr, rd\}$. The additive white Gaussian noises (AWGNs) associated with h_{ij}'s follow the distribution of $\mathcal{CN}(0, \sigma_{ij}^2)$, $ij \in \{sd, sr, rd\}$. Let E_s and E_r denote the transmit powers at S and R, respectively. The *instantaneous* SNRs of the S-D, S-R, and R-D links are respectively denoted $\gamma_{sd} \triangleq \frac{E_s|h_{sd}|^2}{\sigma_{sd}^2}$, $\gamma_{sr} \triangleq \frac{E_s|h_{sr}|^2}{\sigma_{sr}^2}$, and $\gamma_{rd} \triangleq \frac{E_r|h_{rd}|^2}{\sigma_{rd}^2}$, and the corresponding *average* SNRs are $\bar{\gamma}_{sd} \triangleq \frac{E_s\Omega_{sd}}{\sigma_{sd}^2}$, $\bar{\gamma}_{sr} \triangleq \frac{E_s\Omega_{sr}}{\sigma_{sr}^2}$, and $\bar{\gamma}_{rd} \triangleq \frac{E_r\Omega_{rd}}{\sigma_{rd}^2}$. We focus on a general *asymmetric* channel setting where $\bar{\gamma}_{sd}$, $\bar{\gamma}_{sr}$, and $\bar{\gamma}_{rd}$ are *unequal* in general, which makes our analysis practical.

III. Exact Analysis of Ergodic Capacity

We consider the CSI-assisted orthogonal AF relaying with a maximum ratio combiner at the destination, yielding the end-to-end received SNR at the destination as follows [1], [2], [4]

$$\gamma_{\text{AF}} \triangleq \gamma_{sd} + \frac{\gamma_{sr}\gamma_{rd}}{\gamma_{sr} + \gamma_{rd} + 1}. \tag{1}$$

The ergodic capacity is thus given by

$$C_{\text{AF}} \triangleq \frac{1}{2}\mathbb{E}\big\{\log_2(1 + \gamma_{\text{AF}})\big\}, \tag{2}$$

where the pre-log factor $1/2$ accounts for the orthogonal relaying.

Theorem 1: The ergodic capacity for a single-relay AF network with the direct link is given by

$$C_{\text{AF}} = \frac{1}{2\ln(2)}\left\{e^{\frac{1}{\bar{\gamma}_{sd}}}E_1\left(\frac{1}{\bar{\gamma}_{sd}}\right) + \frac{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}{2\bar{\gamma}_{sd}}I\right\}, \tag{3}$$

where

$$I \triangleq \int_0^\infty \phi(x)dx. \tag{4}$$

The integrand of (4), $\phi(x)$, is defined as follows:

$$\begin{aligned}\phi(x) \triangleq{}& \exp\left(\frac{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}{2\bar{\gamma}_{sd}}x + \frac{1}{\bar{\gamma}_{sd}}\right)E_1\left(\frac{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}{2\bar{\gamma}_{sd}}x + \frac{1}{\bar{\gamma}_{sd}}\right)\\ &\times\sqrt{x\left(x + \frac{2}{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\right)}K_1\left(\sqrt{x\left(x + \frac{2}{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\right)}\right)\\ &\times\exp\left(-\frac{\bar{\gamma}_{sr} + \bar{\gamma}_{rd}}{2\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}x\right),\end{aligned} \tag{5}$$

where $E_1(\cdot)$ and $K_1(\cdot)$ denote the exponential integral function [16, eq. (15.1.1)] and the first-order modified Bessel function of the second kind [16, eq. (9.6.11)], respectively.

Proof: See Appendix A. ∎

At the first glance, the exact expression of ergodic capacity in (3) involving the single-integral of (4) is complicated. In fact, the integrand of (5), $\phi(x)$, is a well-behaved function for $x > 0$. First of all, $\exp\big(\frac{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}{2\bar{\gamma}_{sd}}x + \frac{1}{\bar{\gamma}_{sd}}\big)E_1\big(\frac{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}{2\bar{\gamma}_{sd}}x + \frac{1}{\bar{\gamma}_{sd}}\big)$ and $\sqrt{x(x + 2/\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}})}K_1\big(\sqrt{x(x + 2/\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}})}\big)$ in $\phi(x)$ are both monotonically decreasing for $x > 0$. This is easily justified as follows. For $x > 0$, we have $\big[e^x E_1(x)\big]' = \frac{xe^x E_1(x) - 1}{x} < \frac{1}{x}\big(\frac{x+1}{x+2} - 1\big) < 0$ [17, eq. (6.8.2)] and $[xK_1(x)]' = -xK_0(x) < 0$ [18, eq. (8.486.14)]. Thus, $e^x E_1(x)$ and $xK_1(x)$ are monotonically decreasing functions for $x > 0$. Furthermore, $\exp\big(-\frac{\bar{\gamma}_{sr}+\bar{\gamma}_{rd}}{2\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}x\big)$ in $\phi(x)$ decays exponentially fast for $x > 0$. Therefore, the multiplication of these terms, namely $\phi(x)$ in (5), is a monotonically decreasing and exponentially decaying function of x for $x > 0$. Furthermore, $\phi(x)$ is a smooth function of x, which has derivatives of any order. These properties enables efficient numerical calculation of (3), detailed as follows.

It is well-known that the Gaussian-Laguerre quadrature is extremely accurate for large $x > 0$ with integrand of the form $e^{-x}g(x)$ for some function $g(\cdot)$, and the composite Gaussian-Legendre quadrature is very effective for finite integrals [19, Ch. 3]. To exploit the benefits of both quadrature rules, we rewrite the integral of (3) into two sub-integrals

$$I = \int_0^\tau \phi(x)dx + \int_\tau^\infty \phi(x)dx, \tag{6}$$

where the first (finite) integral over $[0, \tau]$, $\tau > 0$, is readily computed using the composite Gaussian-Legendre quadrature [19, eq. (3.3.1)], and the second (semi-infinite) integral over $[\tau, \infty)$ can be accurately evaluated by the Gaussian-Laguerre quadrature [16, eq. (25.4.45)]. As a result, we propose a *hybrid Gaussian quadrature* to compute the integral (3) as follows:

$$\begin{aligned}I = \frac{\tau}{2M}\sum_{j=1}^{M}\sum_{i=1}^{N_1} w_{1,i}\phi\Big[\frac{\tau}{2M}(t_{1,i} + 2j - 1)\Big]\\ + \sum_{i=1}^{N_2} w_{2,i}e^{t_{2,i}}\phi(t_{2,i} + \tau) + \mathcal{R},\end{aligned} \tag{7}$$

where $\tau > 0$ is the integral limit chosen for Gaussian-Legendre quadrature, M represents the number of subintervals considered in the composite Gaussian-Legendre quadrature [19],

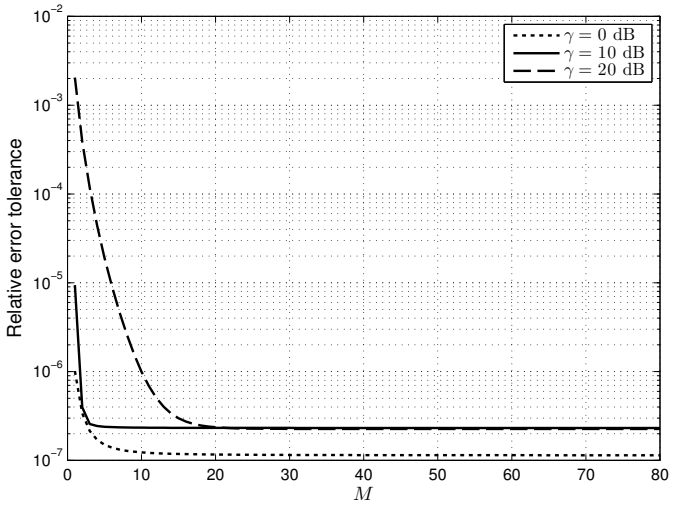


Figure 1. Relative error tolerance, $|\mathcal{R}/I|$, versus M for $d_r = 0.5$ and $\gamma = 0, 10, 20$ dB, where $\tau = 1$, $N_1 = N_2 = 15$.

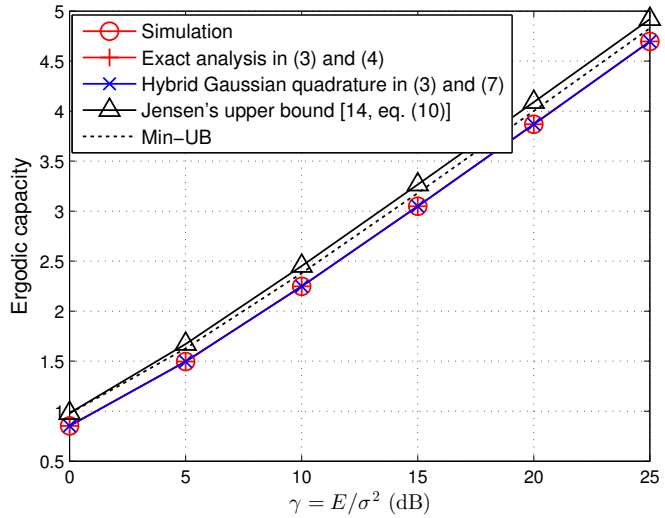


Figure 2. Ergodic capacity of single-relay AF network with the direct link under asymmetric channel setting where $d_r = 0.6$.

and $\mathcal{R}$ is the remainder. Also, $w_{1,i}$ and $t_{1,i}$, $i = 1, \cdots, N_1$, are, respectively, the weights and zeros of the Legendre polynomial of order N_1 [16, Table 25.4], and $w_{2,i}$ and $t_{2,i}$, $i = 1, \cdots, N_2$, are, respectively, the weights and zeros of the Laguerre polynomial of order N_2 [16, Table 25.9]. It is obvious that the accuracy of (7) is dependent on τ, M, N_1, and N_2. In general, any specific accuracy can be achieved by *simultaneously* increasing M, N_1, and N_2 [19], for any τ. In practice, however, using small M, N_1, and N_2 to achieve the target accuracy for a specific τ is desirable. The choice of parameters are discussed in detail in the next section.

IV. Numerical Results

Consider a line S-R-D model in which R is located in the straight line between S and D. Let d_r denote the distance between S and R. The path loss model for typical urban environments is adopted, i.e., $\Omega_{sd} = 1$, $\Omega_{sr} = d_r^{-4}$, and

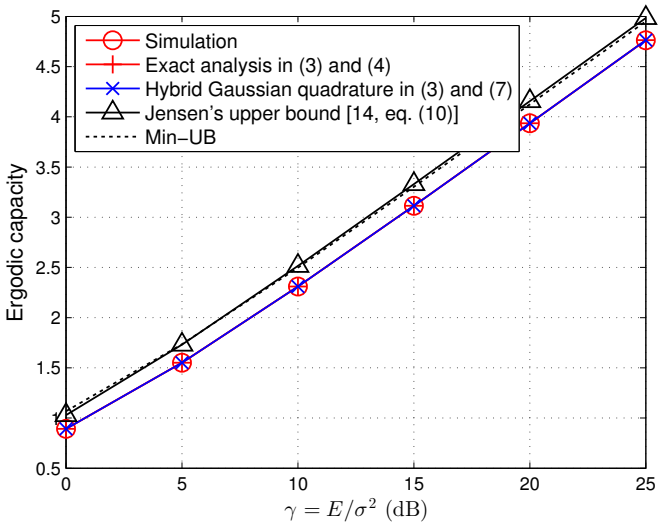


Figure 3. Ergodic capacity of single-relay AF network with the direct link under symmetric channel setting where $d_r = 0.5$.

$\Omega_{rd} = (1 - d_r)^{-4}$ [8]. We set $E_s = E_r = \frac{E}{2}$, where E is the total power consumed in the whole network. The variances of AWGNs are set to be identical, i.e., $\sigma_{ij}^2 = \sigma^2$ for $ij \in \{sd, sk, kd\}_{k=1}^K$. By varying the source-relay distance d_r, the link SNRs $\bar{\gamma}_{sd}$, $\bar{\gamma}_{sr}$, and $\bar{\gamma}_{rd}$ are unequal in general, constituting a general *asymmetric* network. The ratio of the total transmission power E to the AWGN variance, $\gamma \triangleq E/\sigma^2$, is referred to as the network SNR.

A. Choice of Parameters in (7)

The parameters τ, M, N_1, and N_2 in (7) need to be chosen appropriately to achieve a desirable accuracy. In practice, τ is chosen such that $\phi(x) < 0.1$ for $x \geq \tau$, and thus, the approximation error of the (second) integral over $[\tau, \infty)$ in (6) can be made negligible by using a reasonably small value of N_2, e.g., $N_2 = 15$ [19]. Since $\phi(x) < 0.0249$ for all $x \geq 1$, we choose $\tau = 1$ and $N_2 = 15$. Then, the overall accuracy of (7) mainly depends on the first (finite) integral over $[0, \tau]$ in (6). It is possible to increase the accuracy of this integral by simply increasing M for any fixed value of N_1. Thus, by setting $\tau = 1$ and $N_1 = N_2 = 15$, the overall accuracy of (7) is solely dependent on the parameter M. The *relative* error tolerance, i.e., $|\mathcal{R}/I|$, versus M is illustrated in Fig. 1 for $\gamma = 0, 10, 20$ dB, where $d_r = 0.5$. As a benchmark, we compute the integral I of (4) using MATLAB *quadgk* function at a relative error tolerance of 10^{-10}, which is considered as the "exact" value. The difference between this exact value and that computed using (7) is the *absolute* error tolerance $\mathcal{R}$. It is clearly seen that for $\tau = 1$, $N_1 = N_2 = 15$, and $M = 20$, the *relative* error tolerance, $|\mathcal{R}/I|$, is smaller than 10^{-6}. Therefore, we suggest that $\tau = 1$, $N_1 = N_2 = 15$, and $M = 20$ constitutes a suitable choice in (7) to yield an accurate and efficient approximation for (4).

B. Ergodic Capacity Evaluation

For a single-relay AF network *with* the direct link, the following results are compared: i) proposed "exact" result

$$A = \frac{2}{\bar{\gamma}_{sd}} e^{\frac{1}{\bar{\gamma}_{sd}}} \Bigg\{ \int_0^\infty E_1\Big(\frac{1+z}{\bar{\gamma}_{sd}}\Big) e^{(\frac{1}{\bar{\gamma}_{sd}} - \frac{1}{\bar{\gamma}_{sr}} - \frac{1}{\bar{\gamma}_{rd}})z} \sqrt{\frac{z(z+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}} K_1\Bigg(2\sqrt{\frac{z(z+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\Bigg) dz$$
$$- \Bigg[\underbrace{E_1\Big(\frac{1+z}{\bar{\gamma}_{sd}}\Big) \int_0^z e^{(\frac{1}{\bar{\gamma}_{sd}} - \frac{1}{\bar{\gamma}_{sr}} - \frac{1}{\bar{\gamma}_{rd}})x} \sqrt{\frac{x(x+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}} K_1\Bigg(2\sqrt{\frac{x(x+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\Bigg) dx}_{\triangleq B(z)} \Bigg]_{z=0}^{\infty} \Bigg\} \quad \text{(A.4)}$$

$$= \frac{2}{\bar{\gamma}_{sd}} e^{\frac{1}{\bar{\gamma}_{sd}}} \int_0^\infty E_1\Big(\frac{1+z}{\bar{\gamma}_{sd}}\Big) e^{(\frac{1}{\bar{\gamma}_{sd}} - \frac{1}{\bar{\gamma}_{sr}} - \frac{1}{\bar{\gamma}_{rd}})z} \sqrt{\frac{z(z+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}} K_1\Bigg(2\sqrt{\frac{z(z+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\Bigg) dz. \quad \text{(A.5)}$$

by (3) and (4), where (4) is computed using *quadgk* at a relative error tolerance of 10^{-10}; ii) proposed hybrid Gaussian quadrature using (7) in (3), where $\tau = 1$, $N_1 = N_2 = 15$, and $M = 20$; iii) Jensen's upper bound [14, eq. (10)]; and iv) "Min-UB" which refers to the widely-used upper bound using the minimum of the two-hop SNRs, i.e., $\frac{\gamma_{sr}\gamma_{rd}}{\gamma_{sr}+\gamma_{rd}+1} \leq \min(\gamma_{sr}, \gamma_{rd})$. The comparison is performed for an *asymmetric* network with $d_r = 0.6$ in Fig. 2 and a symmetric network with $d_r = 0.5$ in Fig. 3. We observe that the ergodic capacity computed using the hybrid Gaussian quadrature in (7) is in excellent agreement with the exact value of (3). This validates once again the accuracy and efficiency of the proposed hybrid Gaussian quadrature. Furthermore, it is clearly seen that Jensen's upper bound and the Min-UB are loose upper bounds, throughout the whole SNR range. Indeed, the loose bounds/approximations highlight the usefulness of the obtained exact expression.

V. Conclusion

We analyzed the ergodic capacity of AF relaying for ubiquitous cooperative networks in Rayleigh fading where the average SNRs of different wireless channels are unequal in general. For the single-relay case with the direct link, we derived the ergodic capacity in an exact single-integral-form, which serves as a benchmark for AF relaying systems. To facilitate evaluation of this exact expression, we derived a hybrid Gaussian quadrature rule in closed-form, which is extremely accurate and easy to compute.

An interesting extension of this work is to study the *exact* ergodic capacity for general *multi*-relay AF networks, which will be addressed in our further work.

Appendix A
Proof of Theorem 1

Let $\gamma_r \triangleq \frac{\gamma_{sr}\gamma_{rd}}{\gamma_{sr}+\gamma_{rd}+1}$, and thus, $\gamma_{\text{AF}} = \gamma_{sd} + \gamma_r$. It follows from [20, eq. (11)] that the CDF of γ_r is

$$F_{\gamma_r}(z) = 1 - 2e^{-\left(\frac{1}{\bar{\gamma}_{sr}} + \frac{1}{\bar{\gamma}_{rd}}\right)z} \sqrt{\frac{z(z+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}} K_1\Bigg(2\sqrt{\frac{z(z+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\Bigg),$$

for $z > 0$. For independent Rayleigh fading channels, γ_{sd}, γ_{sr}, and γ_{rd} are *independent* exponential random variables with means $\bar{\gamma}_{sd}$, $\bar{\gamma}_{sr}$, and $\bar{\gamma}_{rd}$, respectively, implying that γ_{sd} and γ_r are also independent. Thus, the CDF of γ_{AF}, $F_{\gamma_{\text{AF}}}(z) = \int_0^z F_{\gamma_r}(x) f_{\gamma_{sd}}(z-x)dx$, is evaluated to

$$F_{\gamma_{\text{AF}}}(z) = 1 - e^{-\frac{z}{\bar{\gamma}_{sd}}} - \frac{2}{\bar{\gamma}_{sd}} e^{-\frac{z}{\bar{\gamma}_{sd}}} \int_0^z e^{\left(\frac{1}{\bar{\gamma}_{sd}} - \frac{1}{\bar{\gamma}_{sr}} - \frac{1}{\bar{\gamma}_{rd}}\right)x} \times \sqrt{\frac{x(x+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}} K_1\Bigg(2\sqrt{\frac{x(x+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\Bigg) dx. \quad \text{(A.1)}$$

Since $I_{\text{AF}} = \frac{1}{2}\log_2(1+\gamma_{\text{AF}}) > 0$, we have

$$\mathbb{E}\{I_{\text{AF}}\} = \int_0^\infty \big[1 - F_{I_{\text{AF}}}(z)\big] dz = \frac{1}{2\ln(2)} \int_0^\infty \frac{1 - F_{\gamma_{\text{AF}}}(z)}{1+z} dz, \quad \text{(A.2)}$$

where (A.2) follows from $F_{I_{\text{AF}}}(z) = F_{\gamma_{\text{AF}}}(2^{2z}-1)$ and the change of variable $2^{2z}-1 \to z$. Substituting (A.1) into (A.2) and using [18, eq. (3.352.4)], we obtain

$$\mathbb{E}\{I_{\text{AF}}\} = \frac{1}{2\ln(2)} \Big\{ A + e^{\frac{1}{\bar{\gamma}_{sd}}} E_1\Big(\frac{1}{\bar{\gamma}_{sd}}\Big) \Big\}, \quad \text{(A.3)}$$

where $A \triangleq \frac{2}{\bar{\gamma}_{sd}} \int_0^\infty \frac{e^{-\frac{z}{\bar{\gamma}_{sd}}}}{1+z} \int_0^z e^{\left(\frac{1}{\bar{\gamma}_{sd}} - \frac{1}{\bar{\gamma}_{sr}} - \frac{1}{\bar{\gamma}_{rd}}\right)x} K_1\big(2\sqrt{\frac{x(x+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\big) \sqrt{\frac{x(x+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}} dx dz$. Using [16, eq. (5.1.24)] and [16, eq. (5.1.26)], we have $\big[E_1\big(\frac{1+z}{\bar{\gamma}_{sd}}\big)\big]' = -\frac{1}{1+z}\exp\big(-\frac{1+z}{\bar{\gamma}_{sd}}\big)$. Thus, A is evaluated in (A.4) and (A.5) at the top of this page, where (A.5) follows by the following property

$$\lim_{z \to z_0} B(z) = 0, \text{ for } z_0 = 0, \infty. \quad \text{(A.6)}$$

This equality holds for $z_0 = 0$ because $\lim_{z\to 0} E_1\big(\frac{1+z}{\bar{\gamma}_{sd}}\big) = E_1\big(\frac{1}{\bar{\gamma}_{sd}}\big) < \infty$. To prove the equality of (A.6) for $z_0 = \infty$, we first obtain a trivial lower bound $B_l(z) = 0$ and an upper bound $B_u(z)$ for $B(z)$ as follows:

$$B(z) \leq e^{\frac{z}{\bar{\gamma}_{sd}}} E_1\Big(\frac{1+z}{\bar{\gamma}_{sd}}\Big) \int_0^z e^{-(\frac{1}{\bar{\gamma}_{sr}} + \frac{1}{\bar{\gamma}_{rd}})x} \times \sqrt{\frac{x(x+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}} K_1\Bigg(2\sqrt{\frac{x(x+1)}{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\Bigg) dx$$
$$= \frac{1}{2} e^{-\frac{1}{\bar{\gamma}_{sd}}} \mathbb{E}\{\gamma_r\} \lim_{z\to\infty} \Big\{ e^{\frac{1+z}{\bar{\gamma}_{sd}}} E_1\Big(\frac{1+z}{\bar{\gamma}_{sd}}\Big) \Big\} \quad \text{(A.7)}$$
$$\triangleq B_u(z),$$

where (A.7) follows by the fact that $\mathbb{E}\{\gamma_r\} = \int_0^\infty [1 - F_{\gamma_r}(x)]dx$. Since $\frac{1}{2}\ln(1+\frac{2}{x}) < e^x E_1(x) < \ln(1+\frac{1}{x})$ [16, eq. (5.1.20)], $\lim_{x\to\infty}\{\frac{1}{2}\ln(1+\frac{2}{x})\} = 0$, and

$\lim_{x\to\infty}\{\ln(1+\frac{1}{x})\} = 0$, using Squeeze Theorem [21], we have $\lim_{x\to\infty}\{e^x E_1(x)\} = 0$. This implies that $\lim_{z\to\infty}\left\{e^{\frac{1+z}{\bar{\gamma}_{sd}}} E_1\left(\frac{1+z}{\bar{\gamma}_{sd}}\right)\right\} = 0$. Since $0 < \gamma_r < \min(\gamma_{sr}, \gamma_{rd})$, we have $0 < \mathbb{E}\{\gamma_r\} < \mathbb{E}\{\min(\gamma_{sr}, \gamma_{rd})\} < \min(\bar{\gamma}_{sr}, \bar{\gamma}_{rd}) < \infty$. By definition of (A.7), we have $\lim_{z\to\infty} B_u(z) = 0$. Thus, for $z \geq 0$ we have $B_l(z) \leq B(z) \leq B_u(z)$, where $\lim_{z\to\infty} B_l(z) = \lim_{z\to\infty} B_u(z) = 0$. Using the Squeeze Theorem [21], we have $\lim_{z\to\infty} B(z) = 0$, which consequently proves (A.5). Applying change of variable $\frac{2z}{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}} \to x$ in (A.5), we have

$$
\begin{aligned}
A = & \frac{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}{2\bar{\gamma}_{sd}} \int_0^\infty \exp\left(\frac{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}{2\bar{\gamma}_{sd}}x + \frac{1}{\bar{\gamma}_{sd}}\right) \\
& \times E_1\left(\frac{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}{2\bar{\gamma}_{sd}}x + \frac{1}{\bar{\gamma}_{sd}}\right) \exp\left(-\frac{\bar{\gamma}_{sr}+\bar{\gamma}_{rd}}{2\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}x\right) \\
& \times \sqrt{x\left(x + \frac{2}{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\right)} K_1\left(\sqrt{x\left(x + \frac{2}{\sqrt{\bar{\gamma}_{sr}\bar{\gamma}_{rd}}}\right)}\right) dx.
\end{aligned}
$$

Finally, substituting the above A into (A.3) yields (3).

REFERENCES

[1] A. Sendonaris, E. Erkip, and B. Aazhang, "User cooperation diversity, part I: system description," IEEE Trans. Commun., vol. 51, Nov. 2003, pp. 1927–1938.

[2] ——, "User cooperation diversity, part II: implementation aspects and performance analysis," IEEE Trans. Commun., vol. 51, Nov. 2003, pp. 1939–1948.

[3] F. Sun, T. M. Kim, A. Paulraj, E. de Carvalho, and P. Popovski, "Cell-edge multi-user relaying with overhearing," IEEE Commun. Lett., vol. 17, June 2013, pp. 1160–1163.

[4] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: efficient protocols and outage behavior," IEEE Trans. Inf. Theory, vol. 50, Dec. 2004, pp. 3062–3080.

[5] P. Liu and I.-M. Kim, "Optimum/sub-optimum detectors for multi-branch dual-hop amplify-and-forward cooperative diversity networks with limited CSI," IEEE Trans. Wireless Commun., vol. 9, Jan. 2010, pp. 78–85.

[6] F. Sun, E. De Carvalho, P. Popovski, and C. D. T. Thai, "Coordinated direct and relay transmission with linear non-regenerative relay beamforming," IEEE Signal Process. Lett., vol. 19, Oct. 2012, pp. 680–683.

[7] D. N. C. Tse and P. Viswanath, Fundamentals of Wireless Communications. Cambridge, MA: Cambridge Univ. Press, 2005.

[8] A. Goldsmith, Wireless Communications. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[9] O. Waqar, D. C. McLernon, and M. Ghogho, "Exact evaluation of ergodic capacity for multihop variable-gain relay networks: a unified framework for generalized fading channels," IEEE Trans. Veh. Technol., vol. 59, Oct. 2010, pp. 4181–4187.

[10] L. Fan, X. Lei, and W. Li, "Exact closed-form expression for ergodic capacity of amplify-and-forward relaying in channel-noise-assisted cooperative networks with relay selection," IEEE Commun. Lett., vol. 15, Mar. 2011, pp. 332–333.

[11] D. B. da Costa and S. Aissa, "Amplify-and-forward relaying in channel-noise-assisted cooperative networks with relay selection," IEEE Commun. Lett., vol. 14, July 2010, pp. 608–610.

[12] G. Farhadi and N. C. Beaulieu, "On the ergodic capacity of multi-hop wireless relaying systems," IEEE Trans. Wireless Commun., vol. 8, May 2009, pp. 2286–2291.

[13] F. Yilmaz, O. Kucur, and M.-S. Alouini, "Exact capacity analysis of multihop transmission over amplify-and-forward relay fading channels," in Proc. IEEE Int. Symp. Personal, Indoor Mobile Radio Commun. (PIMRC), Sept. 2010, pp. 2293–2298.

[14] G. Farhadi and N. Beaulieu, "On the ergodic capacity of wireless relaying systems over Rayleigh fading channels," IEEE Trans. Wireless Commun., vol. 7, Nov. 2008, pp. 4462–4467.

[15] S. Chen, W. Wang, and X. Zhang, "Ergodic and outage capacity analysis of cooperative diversity systems under Rayleigh fading channels," in Proc. IEEE Int. Conf. Commun. (ICC), June 2009, pp. 1–5.

[16] M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. New York: Dover publications, 1964.

[17] F. W. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark, NIST Handbook of Mathematical Functions. New York: Cambridge Univ. Press, 2010.

[18] I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series, and Products, 6th ed. San Diego, CA: Academic Press, 2000.

[19] J. Pachner, Handbook of Numerical Analysis Applications with Programs for Engineers and Scientists. New York: McGraw-Hill, 1984.

[20] R. H. Y. Louie, Y. Li, and B. Vucetic, "Performance analysis of beamforming in two hop amplify and forward relay networks," in Proc. IEEE Int. Conf. Commun. (ICC), May 2008, pp. 4311–4315.

[21] S. M. Nikolsky, A Course of Mathematical Analysis 1, 5th ed. Moscow: MIR Publisher, 1977.

 ISBN: 978-1-61208-353-7

CGSIL: A Viable Training-Free Wi-Fi Localization

Han N. Dinh, Thong M. Doan
University of Science, Vietnam National University
Ho Chi Minh, Vietnam
jennynguyen.jn293@gmail.com; dmthong@apcs.vn

Nam T. Nguyen
University of Information Technology
Ho Chi Minh, Vietnam
namnguyen@uit.edu.vn

***Abstract* –Localization for indoor environment normally does not use GPS signals since it cannot penetrate through walls and buildings. Instead, many works have focused on using Wi-Fi signals as the mean to locate the position of the mobile devices. However, most of these approaches require a training step to build a Wi-Fi's map for each location. This requirement practically prevents these approaches from being realistic, since the training step is extremely time-consuming (hundreds of labor hours). Recently, ISIL has been proposed as the first Wi-Fi-based technique that is training-free, in which the localization can be done instantly at any location without the need of training and building Wi-Fi map. ISIL collects from the web the related information of all observable access points and infers the current position based on that. As the first search-based Wi-Fi localization, ISIL removes the unacceptable time-consuming training step. However, it still does not provide adequate accuracy due to the lack of exploiting regional correlation of information returned by the search engine. In this paper, we proposed CGSIL, another kind of search-based Wi-Fi localization that provides the accuracy level of nearly twice as much as ISIL by collaborative filtering and clustering geographic information collected from the search engines. Through experiment results, CGSIL proves to be a feasible replacement for future indoor localization due to its high accuracy and reasonable cost.**

Keywords—search engine; clustering; regional relationship.

I. INTRODUCTION

Localization is becoming an essential technique to enable any useful service, such as Google Maps, Facebook and other services [1]. Several localization techniques have been proposed recently using Global Positioning System (GPS) [2], cellular [3]-[5], and Wi-Fi [6]-[14] technologies. GPS-based localization can achieve the accuracy of up to a few meters [2]. However, in GPS, the signals are transferred from the satellites to a device, and thus the signals can be weakened by obstacles. This explains why GPS can only be used for outdoor environment. Approaches using cellular technology [3]-[5] can work for both outdoor and indoor locations (covered by cell towers) but offer low accuracy (several hundred meters). They also require the knowledge of cell towers' map. Recently, many approaches using Wi-Fi (802.11) signals [6]-[14] have been proposed for indoor locations thanks to their high accuracy rate and the increasingly popularity of the 802.11 Access Points (APs).

According to Le et al. [15], Wi-Fi based localization algorithms can be divided into five main categories: range-based, range-free (centroid [6][7]), aggregate and singular, scene matching (fingerprint [8]) and SIL (search-based) [15]. In the first four categories, one common step these algorithms all require is the costly training phase. In this step, some known positions in the network are recorded with their coordinates and associated information. This information map is used to estimate the location in the runtime phase. The biggest challenge of this training step is that it requires a lot of time and physical-labor. Additionally, this step needs to be repeated regularly to adapt to environment changes.

To avoid the costly training phase, Search-based Indoor Localization (SIL), the 5th category, is proposed. The first algorithm in this category is ISIL [15]. ISIL eliminates the need of the costly training step by exploiting nearby observable access points' names at the runtime phase. The algorithm utilizes what the APs' names represent (usually the business) and aggregates the information to predict the device's current position.

However, ISIL does not exploit the geographical relationship between nearby APs; thus it leads to low accuracy when presenting the predicted address to users. Additionally, to increase the accuracy, ISIL presents a list of 16 possible addresses for users to choose from manually. This approach is not user-friendly and prevents automatic localization since it requires explicit user feedback. Another problem is the lack of a ranking strategy for multiple collected addresses on the same street; therefore, ISIL can only return predicted address with up to street name (no street number). In other words, it cannot provide fine-grained result up to street number.

In this paper, we present CGSIL, a Collaborative Geo-clustering Search-based Localization that provides an accuracy level that is two times better than ISIL. In Section II, we will review and categorize the existing Wi-Fi localization algorithms. In Section III, we describe our new approach, CGSIL, and its advantages. In Sections IV and V, we discuss the experiment setup and analyze the experiment results of CGSIL. To prove the practical aspect of CGSIL, we also provide a cost analysis in terms of storage and bandwidth usage in Section V.

II. RELATED WORKS

According to Le et al. [15], Wi-Fi localization techniques can be classified into four categories: range-based, range-free, aggregate and singular, scene matching. All of them require a costly training phase, in which some known positions in the network are recorded with their coordinates and associated information. This information map is then used to estimate the location when the algorithms are in the runtime phase. Recently, ISIL has been introduced as the first training-free localization algorithm [15]. Due to the basic nature of the training-free solution, we classified ISIL to belong to the new category, called SIL. In the next sections, we will first summarize the first four

 ISBN: 978-1-61208-353-7

categories and then have a brief review about SIL, the new category.

A. The first Four Categories (Training-Required Group)

Most of the Wi-Fi localization techniques in the first four categories have two main phases: a training (offline) phase and a deployment (online) phase [15]. The main task of the training phase is to build a map containing known location indicators. These indicators are then used in the deployment phase to estimate the location by retrieving the most appropriately similar location indicators from the pre-built map.

Technically, the training phase could vary depending on the unique property in each category. However, in most cases, this phase requires an extensive amount of time and human labor to accomplish, as the location indicators must be collected at every location. Additionally, this costly training step must be repeated regularly due to the changes of the environment (weather, human, building). Finally, if devices used in the deployment phase are different from the sample devices used in the training phase, the accuracy can degrade remarkably [8]. Re-training for new devices will improve the accuracy but it is time-consuming and impractical for wide-scale deployment due to the variety of mobile devices [15][21][22].

B. SIL (Training-Free Group)

SIL is a Wi-Fi based localization approach that aims to remove the need of the costly training step (ISIL [15] is an example in this group). By analyzing the SSIDs of observable APs collected at a location, SIL will aggregate the information related to the SSID to predict the device's current position. The information related to the SSID can be extracted instantly by querying any search engines [15] or other means. SIL is a simple alternative for indoor localization where GPS signals are not available and when it is nearly impractical to require the training step.

Fig. 1 illustrates the general framework of SIL (used in ISIL [15]). It is composed from three components: Scanning, Geo-information Retrieving and Address Processing. In the Scanning component, the mobile device will scan for information extracted from nearby APs. Next, the Location Geo-information Retrieving component will gather relating information from the Internet and extract a list of potential addresses. Finally, the Address Processing component will rank the addresses and return the correct ones to the users. In this component, we can apply different algorithms with different strategies to process and evaluate the list of potential addresses. One example of such algorithm is ISIL [15].

At first, ISIL is a novelty algorithm due to its training-free properties. ISIL works independently on the type of wireless card of mobile devices, and is not affected by environmental changes. It can work on any Wi-Fi based mobile device that has access to a search engine [15].

Even though ISIL does not require training step, its accuracy is up to street name only, which causes considerable distance error, since some streets can be several kilometers in length. Moreover, ISIL returns result as a list of predicted addresses and requires user to select one manually. In other words, if the size of the returned address list is small (like 1 or 2), the accuracy rate is low (50% to 55%) [15]. If the size is larger, the number of returned addresses could easily confuse the users.

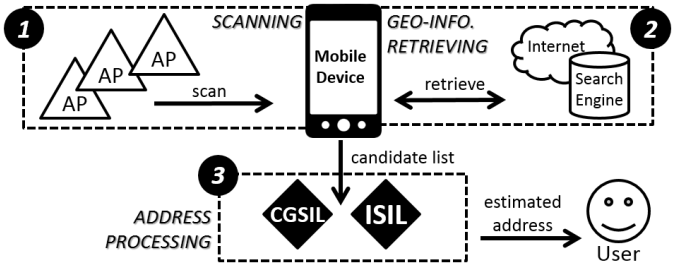


Fig. 1. The General Framework of SIL

To address those constraints of ISIL, we propose CGSIL, a more accurate and finer-grained result than ISIL. Specifically, CGSIL has incorporated Search Engine Optimization property, geographic information and region-based relationship of APs into a comprehensive strategy to predict addresses. Thus, CGSIL only needs to return the result as a single address with the accuracy that is 2 times better than ISIL.

III. OUR APPROACH

In this section, we will give an overview of SIL, ISIL and its weakness. Finally, we propose CGSIL, our new approach.

A. Overview of SIL

SIL relies on the observation that the names of the APs located at a location often contain information relating to that location. For instance, if an AP with the name TokyoDeli is detected, it is a good indicator telling us that our current position is nearby one of the TokyoDeli restaurants. Thus, if SIL can analyze all the SSIDs of observable APs, it can extract the information linking to the user's current position. By aggregating all information returned by the names of all APs, SIL can predict the location of the device. Continuing with the previous example, if we can detect another AP with the name McDonald, it means that the current location must be around McDonald and TokyoDeli restaurants. Thus, if we could find a location that is geographically close to both restaurants, we can use it as the current predicted address.

To do that, SIL needs a database containing the APs' names and their corresponding location. A valuable and always-on database SIL can use is a search engine. As most mobile devices have access to the Internet, querying search engine is totally feasible. The system can feed the AP's name into the search query. The webpages returned from the search engine are parsed to extract all the addresses presented on these webpages. These collected addresses are aggregated and examined to predict the location [15]. The main idea of SIL can be summarized into three phases, corresponding to the three components in Fig. 1.

1) Scanning

In this phase, the deploying device scans nearby APs for their SSIDs. These SSIDs are then pre-processed and split into keyword for querying search engine.

2) Geo-information Retrieving – GR

The keywords extracted from the scanning phase are sent to a search engine. Relevant URLs returned by the search engine are parsed to collect possible location information (addresses).

 ISBN: 978-1-61208-353-7

Since the search engine may return many results (pages), it is impossible to parse them all. Therefore, the top web page results returned by the search engine are selected to parse for location information. The number of selected pages directly affects the breadth of the search space and thus is defined as *breadth*.

The set of webpages returned directly by the search engine is called at *depth* 0. In many cases, it is not sufficient to parse only the webpages at *depth* 0 because the street address of the location may be a few links away. Thus, SIL needs to follow the links appearing on webpages at *depth* 0 to get to subsequent webpages. The successive pages that are one link away from the pages at *depth* 0 are called pages at *depth* 1. We defined *depth* as the number of links away from the pages returned directly by the search engine.

The deeper we crawl for the URLs, the longer it takes for the system to process. The same is for the *breadth*. Therefore, *depth* and *breadth* are the two vital factors we need to analyze to find the optimal values. SIL has shown that *depth* 1 is good enough for the system to return acceptable accuracy [15].

The outcome of this GR phase is a list of potential addresses with high probability to be near to the actual location. This list is defined as the *candidate list*. To point out the correct address from this l*ist*, SIL utilizes the Address Processing component, which is discussed next.

3) Address Processing

From the previous phase, we now have a list of candidate addresses where one of them could be in a close proximity with the actual address. Therefore, the task of this component is to find that address and return it to the users. The performance of SIL greatly depends on the algorithm chosen for this Address Processing component. ISIL is the first algorithm proposed [15]. In the next section, we will describe ISIL in detail.

B. ISIL and its Drawbacks

1) ISIL

Let us define:

- $A = \{ap_1, ap_2, \dots, ap_n\}$: as the set of all access points at one location.
- $extract(x), (x \in A)$: as the function to return all addresses extracted from an access point *x*.

Let *D* be the set of all addresses collected at one location. From the set *A* & function *extract*, we have:

$$D = \bigcup_{i=1}^{n} extract(ap_i), ap_i \in A \quad (1)$$

Finally, we have *S(y)*, the set of all access points belonging to an address (*y*) is constructed by the following function:

$$S(y) = \{x | \, x \in A \wedge y \in extract(x)\}, y \in D \quad (2)$$

According to ISIL algorithm, the authors used two metrics to measure the relevancy of each collected address:

- |*S(y)*|
- The *depth* of the web page where the address appears.

In other words, the ranking of ISIL is based on the following observation:

- If an address is extracted from the search result of more APs, it is more likely to be related to the current location;
- If an address appears in a web page that is further away from *depth* 0, it is less likely to be related to that location.

2) Drawbacks of ISIL

The accuracy of ISIL only works well at street level. The biggest drawback of this is some streets can be very long (10 – 20 km), which negatively affects the accuracy. The second drawback is that two different streets can be in a close geo-proximity, but in ISIL, they will be treated to be unrelated when doing the ranking. Third, ISIL returns the predicted result as a list of possible addresses that requires the user to choose from. This may confuse the user if the list is long. From our experiments, ISIL may return a list of 16 addresses in order to achieve the accuracy level of 80% or more. It is not user-friendly and troublesome to return multiple options for the user to select.

ISIL does not fully exploit the geographic relationship of the APs. In fact, the way the APs in close proximity support each other could be a hint to improve ranking strategy. If an address geographically belongs to the intersected region of more nearby APs, the address is likely to be nearby the current position.

In this paper, we propose CGSIL to address these limitations of ISIL.

C. CGSIL

To address the drawbacks of ISIL, we propose CGSIL, which returns finer-grained and more accurate localization result. This achievement utilized popular technique such as search engine optimization (SEO) [16][19], geographic mutual-relationship, collaborative filtering and cluster analysis [18].

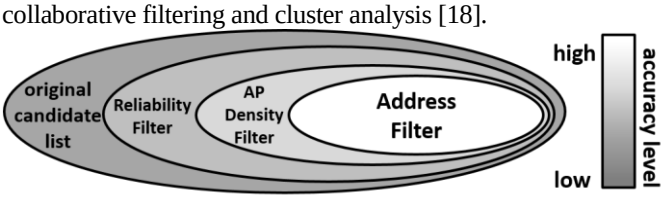


Fig. 2. Venn Diagram of the Filters in CGSIL

Fig. 2 illustrates how we apply different filters in CGSIL to predict the address. The candidate list is narrowed down after each filtering and the accuracy level of the remaining addresses eventually increase.

The detailed process of how CGSIL works can be described in 3 steps: 1) un-related addresses will be filtered out by the Reliability Filter; 2) The AP Density Filter, based on the visibility of surrounding access points, will try to detect a set of addresses having high likelihood to be close to the current location. 3) The Address Filter will rank the addresses and return the top one as the result.

1) Reliability Filter

After the Scanning and GR phases, (Section III.A.2), we now have the *candidate list* composed from addresses extracted from the SSID' search results. However, many addresses from this list are unrelated as they come from irrelevant webpages, such as advertising sites or personal blog-sites. Therefore, CGSIL will use this Reliability Filter to eliminate unwanted addresses.

This filter works by utilizing SEO presentation, embedded inside each web page. SEO is the process of affecting the

 ISBN: 978-1-61208-353-7

visibility of a website or a page in a search engine's results returned to users [17]. The SEO presentation of a web includes: the header text, the footer text, the contents, the codes and the URL itself. Among these SEO attributes, CGSIL will focus on the URLs (the anchor texts) [19] because they often provide more accurate descriptions of Web pages [16]. This observation is utilized in CGSIL to select the pages most relevant to the source SSIDs. Moreover, choosing the URL over other SEO attributes improves performance since processing one line of text is more light-weighted than processing the whole page's content.

If the hyperlink text of one URL does not contain its SSID, the URL and its extracted addresses will be removed from the *candidate list*. After that, the remaining addresses in the list will be sorted in the way that the URL containing more characters from its SSID has higher order than the one containing few characters from SSID; the top addresses in this list are defined as the *F1-candidates*.

However, as displayed in Fig. 3, these F1-candidates ("diamond markers") could be scattered on the geographic map. Hence, our next task is to find a region covering most potentially correct addresses, which is discussed in the AP Density Filter.

2) AP Density Filter

This filter works based on the observation that the area covering addresses from most APs is likely to contain correct estimation for the current location. The idea is illustrated in Fig. 3.

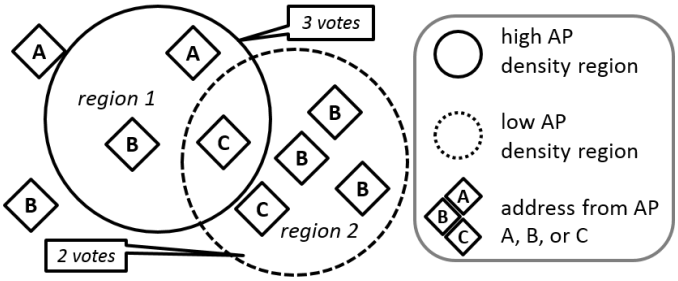


Fig. 3. The Highest Vote Region

Fig. 3 presents a map with 10 scattered addresses and 2 suspected regions that may provide the correct estimations for the current location. For each region, we count the number of votes from the APs. We define a vote from an AP for a region as: the region must contain at least one address extracted from that AP's search results. For example, in Fig. 3, region 1 gets three votes because it contains addresses extracted from three APs ("A", "B" and "C"). Likewise, region 2 gets two votes. This collaborative process chooses the region with the highest votes from all the observable APs. Therefore, the region 1 could be the most likely correct region of the current position.

To find the region of the highest votes, we must have the geographic data associated with each address so that we can perform calculation with the addresses. Such information could be retrieved from any online address database, for example the Google Map, the one we use in our experiment. In addition, we use Google Map API to provide the latitude and longitude coordinates for a given string address. Note that our technique does not depend on any specific map API; for instance, the country's local map API can be selected as an alternative.

Based on the information provided by the Google Map API, we find the region with the highest votes from the APs. If there is one region with the highest vote, we simply return the center of the region as the localization result. Nevertheless, in many cases, there may be multiple regions with the same highest number of votes. These regions could be overlapped or scattered geographically. Thus, we need the Address Filter to estimate the best result to return to the users.

3) Address Filter

The idea of this filter is to find a high-density geographical cluster from multiple same rank regions, discovered from previous step. The center of the result cluster is returned to the users as the localization result.

A cluster is a group of addresses locating at relatively close distance to each other. This problem is classified into typical clustering problem:

- Give a constant d as the maximum Euclid distance between any 2 addresses
- Let A be the set of all *F2-candidate* (all addresses in the *Highest Vote Regions*), C_i be the set of all addresses in one cluster, we have:
$$A = \bigcup_{i=1}^{n} C_i \quad (3)$$
- Define *ed(a, b)* as a function to calculate Euclid distance between 2 addresses a & b, we have:
$$ed(a,b) = \sqrt{(a.x - b.x)^2 + (a.y - b.y)^2} \quad (4)$$
- Finally, we have the condition for any address, called a, to belong to a cluster, called C_i:
$$a \in C_i \leftrightarrow \forall b \in C_i, ed(a,b) \le d(a \in A) \quad (5)$$

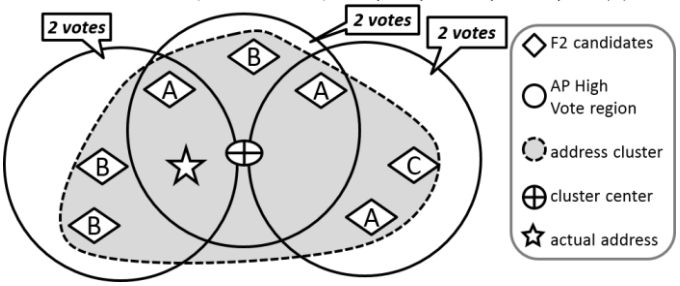


Fig. 4. Clustering and Generating Final Answer for CGSIL

We use condition (5) to distribute all *F2-candidates* into separate clusters. After clustering, the cluster containing addresses from most APs is chosen and its center coordinate is returned as the localization result.

Fig. 4 demonstrates how to cluster the *F2-candidates* to calculate the center of the cluster. Since the 3 circles both cover 2 APs, all 3 circles are considered F2-regions and their addresses are clustered. Satisfying the condition (5), all the addresses are grouped into one cluster (the dark region in Fig. 4). The center of the dark region is returned since the actual address is highly likely to be inside the region.

IV. EXPERIMENT SETUP

In our experiment, we collected data from 4 districts in HCM City which is the same set of districts used in [15]. To increase

 ISBN: 978-1-61208-353-7

the confidence level of the dataset, we collected more than 6,700 locations, which is approximately two times the number of samples collected in [15]. To collect the whole dataset, it took 600 hours of labor.

A. Data Collection

Our Wi-Fi data collection includes around 60 streets in HCM city. On each street, we recorded data at different locations. The collected data includes the AP's name. The exact street number addresses were also recorded for the purpose of evaluating the accuracy of our approach.

The collected data covers District 1 (the city center), 3, 5 and 10. The total street length of our collected data is about 67,500 meters. On each road, we recorded data at different locations, which are 10-15 meters apart from each other. The reason we chose 10-15 meters is that there is not much difference (in terms of observable APs' name) within that distances. The number of locations on a road varies from 40 to 120 depending on its length and availability. At each point, a mobile device continuously scans the Wi-Fi signals for 60 to 90 seconds. On average, there are about 25 APs detected at one location. A group of 150 volunteer students, divided into 60 groups equipped with laptops, participated in the experiment. Each group was responsible for one street. The data set consists of approximately 6,700 locations which take approximately 600 hours of human labor.

B. Accuracy Measurement

To evaluate the accuracy of CGSIL, we recorded actual address at each location (test dataset) to compare with the predicted addresses returned by our algorithm. We defined some terminologies used in presenting results in Section V:

- ***Distance error:*** the Euclid distance between the actual address and predicted address.
- ***Acceptable error range:*** the error range that is acceptable by the users. For example, if the acceptable error range is 500m, that means the users accept the predicted address to be correct if it is within 500m from the actual location.

The accuracy level of CGSIL is calculated as:

$$accuracy = \frac{the\ number\ of\ locations\ yielding\ correct\ address}{the\ number\ of\ all\ locations} \quad (6)$$

V. PERFORMANCE RESULTS

In this section, we will first present the accuracy of CGSIL in comparison with that of ISIL [15]. Next, we analyze how the change in breadth affects the overall accuracy of CGSIL. After that, we will describe Incremental Geo-information Retrieving, used in the Geo-information Retrieving Component of SIL (Section III.A.2), to acquire the information more efficiently. Finally, we study the cost of CGSIL to ensure the feasibility of CGSIL.

A. Accuracy Comparison between CGSIL vs. ISIL

1) Overall Accuracy

Fig. 5 shows the mean localization accuracy of CGSIL and ISIL at District 1 with a variety of acceptable error range.

CGSIL is nearly two times more accurate than ISIL when acceptable range is from 500m or more. This resulted from the collaborative filtering and geographic information clustering implemented in CGSIL. Note that when the acceptable range is 200m, there is not much difference between the two. This is because the Wi-Fi signal normally can cover up to 500m.

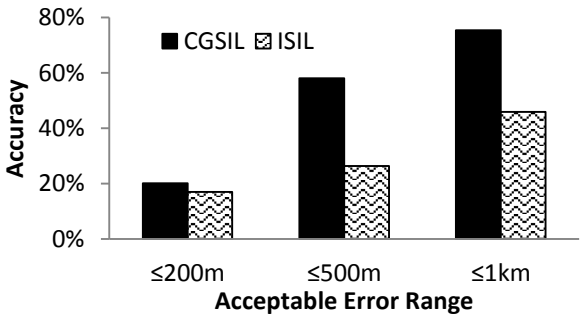


Fig. 5. Accuracy of CGSIL vs. ISIL at District 1 for Variety of Error Range

When the acceptable error range increases from 500m to 1km, the accuracy of CGSIL rises from 58 percent to about 75 percent. For ISIL, to reach this accuracy, it has to return at least 3 candidates for users to choose from.

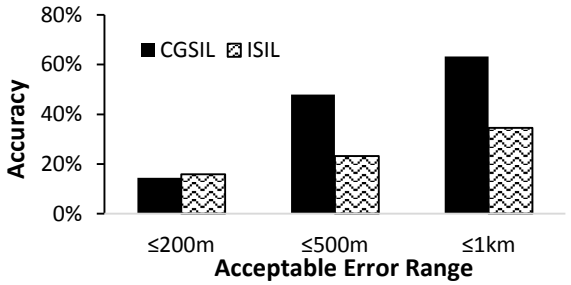


Fig. 6. Accuracy of CGSIL and ISIL at all Districts.

Fig. 6 shows the mean localization accuracy of CGSIL and ISIL for all districts. It has the same pattern as in District 1, but with lower accuracy because it includes non – business districts. This will be discussed more in the next section, V.A.2.

2) Accuracy with respect to Districts

Fig. 7 shows the mean localization accuracy at 4 different districts (acceptable error range is 1 km). The highest accuracy is seen in District 1, which is about 75 percent. The accuracy is lower for District 3, 10 and 5. The accuracy level of these districts is correlated to the business density of the corresponding districts [15]. Crowded business districts tend to yield higher accuracy due to the availability of more APs from nearby business. This is consistent with the finding in [15].

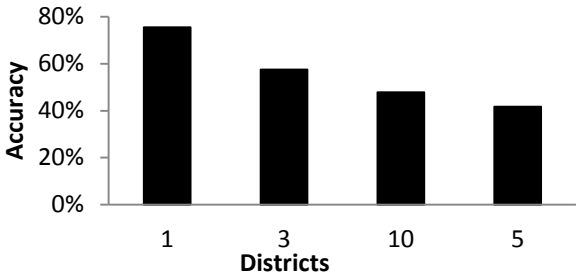


Fig. 7. CGSIL Accuracy in Different Districts (Error Range 1 km)

 ISBN: 978-1-61208-353-7

B. Incremental Geo-information Retrieving – IGR

In this section, we discuss an optimization: IGR. As discussed in III.A.2, whenever mobile device moves to a new location, it must scan the names of all nearby APs and uses those to retrieve the geo-information for localization. This retrieving step requires fetching HTML pages. From our experiment, there are about 25 APs detected at each location on average. Thus, this process may create overhead on bandwidth usage if many APs are detected at each location.

However, adjacent locations are usually covered by many common APs due to the overlap coverage. In other words, when moving from a location to a new one, the mobile device may observe many APs but most of which were previously seen at the old location. From our experiment data, the number of newly detected APs at the new location is only about 2 Aps (out of a total of 25 APs).

Therefore, once moving to a new location, CGSIL only needs to retrieve geo-information for the newly detected APs. This mechanism is called Incremental Geo-information Retrieving (IGR). By doing this, we diminish the bandwidth usage of the device tremendously.

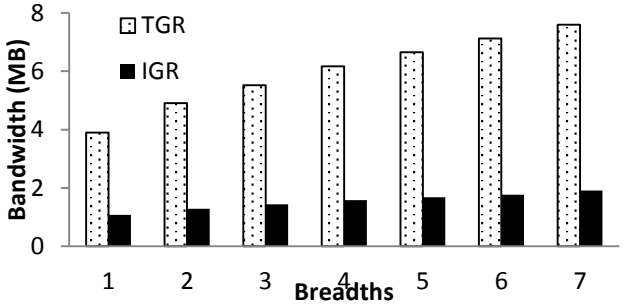


Fig. 8. Bandwidth Usage for Different Level of Breadths.

Fig. 8 shows the bandwidth usage for different breadth levels when using IGR vs. traditional geo-info retrieving (TGR). From the figure, IGR decreased the bandwidth usage by four times comparing to TGR. At breadth 1, IGR used up about 1MB, whereas in TGR, it is about 4MB. Furthermore, when the breadth level increases, the bandwidth usage of IGR rises up slowly from 1MB to 2MB, whereas in TGR, it increases hastily from 4MB to 7.5MB. Note that, the bandwidth usage can even further reduce by using local or cloud storage, which will be discussed in the next section.

C. Cost Analysis of CGSIL

In this section, we will analyze the cost of deploying CGSIL in term of bandwidth cost and storage.

Fig. 9 [20] illustrates the cost mobile users pay per megabyte over the years. The y-axis is in log scale. The x-axis represents the years. In this figure, we see that the cost per megabyte decreases exponentially in prices. With the introduction of 4G, we expect the price will go down in the same trend for 2015 and later.

Thus, if each location requires 2MB of bandwidth to localize (Section V.B), the cost is 0.01 USD/location for 2014. If a user uses CGSIL to localize 100 times/day, the cost that user has to pay for CGSIL is 1 USD/day. However, if the future cost of bandwidth keeps decreasing at the same rate as in the last 4 years, the expected cost of CGSIL can go down to 0.04 USD/day in 2018 and 0.008 USD/day in 2020, which is a negligible quantity. It means that in the next three years, the expected cost for CGSIL is small and affordable for everyone.

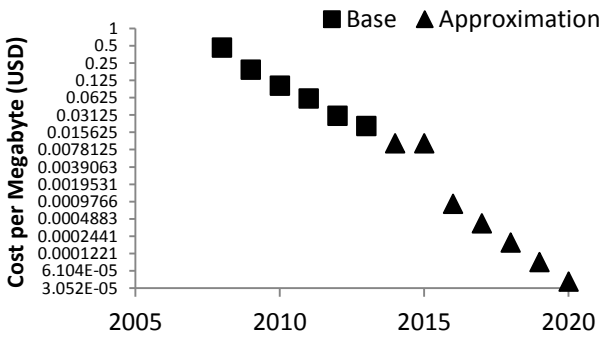


Fig. 9. Cost per Megabyte of Mobile Data worldwide.

s

Moreover, the above estimated cost assumes that the user always moves to the new locations and never go back to any previously visited locations. But, in fact, users are in the habit of moving to the same set of places most of the time: home, office, etc. In that case, if geo-information of visited APs are saved on cache, CGSIL does not need to use bandwidth anymore when users go back to the place they visited before. In other words, after using CGSIL for a few weeks, the users may not need to pay for bandwidth usage or very little.

Note that even though we need to fetch 2MB of HTML files to extract the geo-information, the actual geo-information collected afterward is about 2.5KB in size. Thus, if this geo-information are pushed to the cloud and shared between users, it can be fetched by other user at a rate of 2.5KB/location instead of 2MB/location, which will reduce the bandwidth usage almost 1,000 times (0.1cent/day for 2014).

The storage requirement to implement the geo-information cache at local device is also small. On average, one AP takes about 2.5 KB of storage to save the geo-info on cache. With 100MB cache, the total locations can be cached is about 40,000 locations. Additionally, as the phone storage keep increasing every year, the cache capacity can grow accordingly to hold even more locations if needed.

Fig. 10 illustrates the storage capacity of a common brand phone over time. We see that the capacity jumps double every 2 years. Therefore, a 100 MB of cache on a 64-GB phone takes about 0.015% of its memory, a negligible quantity. With the increment of storage size trend, in the next three years, the expected cache containing geo-info of billion APs is feasible.

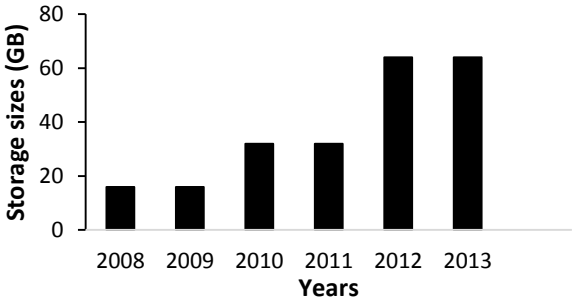


Fig. 10. Phone's Storage Capacity over time

 ISBN: 978-1-61208-353-7

Therefore, we believe that CGSIL is a feasible solution in term of monetary, bandwidth and storage cost.

VI. CONCLUSION

We have proposed CGSIL, a feasible and training-free Wi-Fi localization that is capable of returning higher accurate and finer-grained results. The training-free characteristic of CGSIL makes it more practicable comparing with other Wi-Fi based localization since it can save a lot of money, human-labor and especially time. This is crucial when the localization needs to be implemented in wide-scale with many locations such as city level. Additionally, CGSIL shows a clear advantage over ISIL, the first training-free approach, by offering an accuracy level that is two times better than that of ISIL. This achievement is based on the new ranking strategy, which utilizes the collaborative filtering, SEO properties, and the geographically clustering of location information from observable APs. The cost analysis also showed the feasibility of CGSIL in near future. CGSIL is a good choice when users desire a localization accuracy level of up to 70% with a training-free experience. When the accuracy level of 80% or more is required, other Wi-Fi based approach should be used, yet, with the cost of the expensive training step.

REFERENCES

[1] Nam Nguyen, Leonard Kleinrock, and Peter Reiher, "Debugging Ubiquitous Computing Applications with the Interaction Analyzer" in the International Journal on Advances in Software, IARIA, 2012, vol. 5, no 3-4, pp. 345-357.

[2] Nirupama Bulusu, John Heidemann, and Deborah Estrin, "GPS-less Low Cost Outdoor Localization for Very Small Devices", IEEE Personal Communications, 2000, vol. 7, issue 5, pp. 28-34.

[3] Mike Y. Chen et al., "Practical Metropolitan-Scale Positioning for GSM Phones", in Proc. Int. Conf. on Ubiquitous Computing (Ubicomp), 2006, LNCS 4206, pp. 225-242.

[4] D. Gundlegard and J. M. Karlsson, "Handover location accuracy for travel time estimation in GSM and UMTS", IET Intelligent Transport Systems, 2009, pp. 87-94.

[5] Ian Smith et al., "Place Lab: Device Positioning Using Radio Beacons in the Wild", in Proc. IEEE Conference on Pervasive Computing (Percom), 2005, pp. 116-133.

[6] Yu-Chung Cheng, Yatin Chawathe, Anthony LaMarca, and John Krumm, "Accuracy Characterization for Metropolitan-Scale Wi-Fi Localization", in Proc. ACM Int. Conf. on Mobile Systems, Applications, and Services, 2005, pp. 233-245.

[7] P. Bahl and V. N. Padmanabhan, "Radar: An in-building RF-Based User Location and Tracking System", in Proc. IEEE Conf. of Computer and Communications Societies (Infocom), Apr. 2000, pp. 775-784.

[8] Alex Varshavsky, Denis Pankratov, John Krumm, and Eyal de Lara, "Calibree: Calibration-Free Localization Using Relative Distance Estimations", in Proc. ACM Int. Conf. Pervasive, 2008, pp. 146-161.

[9] Doherty L. Pister and El Ghaoui, "Convex position estimation in wireless sensor net-works", in Proc. IEEE Int. Conf. on Computer Communications (Infocom), 2001, pp. 1655-1663.

[10] Shang Y. Ruml, W. Zhang, and Y. Fromherz, "Localization from mere connectivity", in Proc. ACM Int. Symposium on Mobile Ad-Hoc Networking and Computing (Mobihoc), 2003, pp. 201-212.

[11] Moustafa Youssef, Ashok Agrawala, and A. Udaya Shankar, "WLAN Location Determination via Clustering and Probability Distributions", in Proc. IEEE Int. Conf. on Pervasive Computing and Communications (Percom), 2003, pp. 143-150.

[12] Truc D. Le, Hung M. Le, Nhu T. Q. Nguyen, Dinh Tran, and Nam T. Nguyen, "Convert Wi-Fi Signals for Fingerprint Localization Algorithm", in Proc. IEEE Int. Conf. on Wireless Communication, Networking and Mobile Computing (WiCOM11), Wuhan, China, session 12, 2011, pp. 1–5.

[13] Truc D. Le, Nam T. Nguyen, "A Scalable Wi-Fi Based Localization Approach", in the REV Journal on Electronics and Communications (REV-JEC), 2011, vol. 1, no. 3, pp. 167-174.

[14] Andreas Haeberlen, Eliot Flannery, Andrew M. Ladd, Algis Rudys, Dan S. Wallach, and Lydia E. Kavraki, "Practical Robust Localization over Large-Scale 802.11 Wireless Networks", in Proc. ACM Int. Conf. on Mobile computing and networking (MobiCom), 2004, pp. 70-84.

[15] Truc Le, Thong Doan, Han Dinh, and Nam Nguyen, "Instant Search-based Indoor Localization" in the Proceedings of the 10th Annual IEEE Consumer Communications and Networking Conference (CCNC), Las Vegas, Nevada, USA, 2013, pp. 143-148

[16] Sergey Brin and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Science Department, Stanford University, Stanford, CA 94305.

[17] Wikipedia Information about Search Engine Optimization, http://en.wikipedia.org/wiki/Search_engine_optimization, 6, 2014.

[18] Cluster Analysis, http://en.wikipedia.org/wiki/Cluster_analysis, 6, 2014

[19] Anchor Text, http://en.wikipedia.org/wiki/Anchor_text , 6, 2014

[20] Mobile data usage trends 20112015., http://www.slideshare.net/KarlPortio , 6, 2014

[21] Liu, Kaikai and Liu, Xinxin and Li, and Xiaolin, "Guoguo: Enabling Fine-grained Indoor Localization via Smartphone", in book "Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services" (MobiSys'13), 2013, Taipei, Taiwan, pp. 235-248.

[22] Priyantha, Nissanka B. and Chakraborty, Anit and Balakrishnan, and Hari, "The Cricket Location-support System", in "Proceedings of the 6th Annual International Conference on Mobile Computing and Networking" (MobiCom'00), 2000, Boston, Massachusetts, USA, pp. 32-43.

 ISBN: 978-1-61208-353-7

Automatic Modulation Classification of Digital Modulation Signals Based on Gaussian Mixture Model

W.H. Ahn, J.W. Choi, C.S. Park, B.S. Seo
Dept. of Electronics Engineering
Chungbuk National University
Cheongju, Korea
{glingi, cjw442, chansp, boseok}@cbnu.ac.kr

M.J. Lee
Agency for Defense Development of Korea
Daejeon, Korea
drdlalswns@dreamwiz.com

***Abstract*—In this paper, we propose an automatic modulation classification scheme for digitally modulated signals, such as MSK, GMSK, BPSK, QPSK, 8-PSK, 16-QAM, 32-QAM, and 64-QAM. As features which characterize the modulation type, higher order cyclic cumulants up to eighth order of the signal are used. For feature classification, a Gaussian mixture model based algorithm is used. Simulation results are demonstrated to evaluate the performance of the proposed scheme under AWGN channels.**

Keywords- automatic modulation classification; Gaussian mixture model; cyclostationary; higher order cyclic cumulants.

I. INTRODUCTION

Automatic modulation classification (AMC) is a technique to identify the modulation type of the detected signal as well as to estimate the signal parameters such as carrier frequency and symbol rate, etc [1]. It is widely tried to apply in the field of military and civilian for electronic warfare, spectrum monitoring, surveillance, and cognitive and software defined radios.

There have been many studies on AMC for last two decades. AMC schemes are normally classified into two major categories which are likelihood-based approach [2] and feature-based approach [3]-[10]. The likelihood-based method shows optimal performance in the sense that it maximizes the probability of correct classification. However, it has higher computational complexity for likelihood computation. In addition, it is highly sensitive to modeling mismatch such as timing, phase and frequency offsets, and noise variance.

The feature-based approach attempts to extract a set of features from the received signal. Because the features represent a distinct pattern in a feature space, various pattern recognition algorithms can be applied for classification. Although this approach may not show optimal performance, it is easy to implement and shows nearly optimal performance when appropriate features and a classifier are combined.

In [3], higher order statistics up to sixth order are used to discriminate binary phase-shift keying (BPSK), quadrature phase-shift keying (QPSK), 16-ary quadrature-amplitude modulation (16-QAM), and 64-QAM. As a classifier, a genetic programming combined with the *K*-nearest neighbor (KNN) algorithm is used. It shows good performance even in the presence of noise and frequency offset. In [4]-[6], cyclostationary statistics such as spectral correlation density [4] and cyclic cumulants (CCs) [5, 6] of the signal are used as features. These methods use a simple decision tree [4], the minimum distance metric [5], and the Mahalanobis distance metric [6] for classification. The classification performance is insensitive to model mismatch and independent of any a priori knowledge of signal parameters. In [7], a Gaussian mixture model (GMM) is used for classification of instantaneous amplitudes and phases of binary amplitude-shift keying (BASK), binary frequency-shift keying (BFSK), QPSK, 16-QAM, and 64-QAM signals and it shows better performance compared to [5, 6].

In this paper, we propose a feature-based AMC scheme which uses CCs as features and the GMM for feature classification. We deal with both linear and nonlinear modulations, such as minimum-shift keying (MSK), Gaussian MSK (GMSK), BPSK, QPSK, 8-PSK, 16-QAM, 32-QAM, and 64-QAM.

This paper is organized as follows. In section II, we explain the signal model and entire framework of the proposed scheme. Brief review about CCs used for features is given in section III. The proposed AMC adopting GMM is explained in section IV. Simulation results are demonstrated in section V to evaluate the performance. Finally, the paper concludes in Section VI.

II. SYSTEM MODEL

The proposed modulation classification scheme consists of two parts of a feature extraction block and a GMM classifier as shown in Fig. 1. Two signals, $y(n)$ and $r(n)$ are entered to the first block. The received baseband signal $y(n)$ at time n is represented as

$$y(n) = \sum_{l=-\infty}^{\infty} a(l)\, p(nT_s - lT) e^{j(2\pi \Delta f n T_s + \theta_c)} + w(nT_s) \qquad (1)$$

where $a(l)$ is the l-th transmitted symbol and assumed to be independent and identically distributed (IID), which generally holds in digital communications, with unit variance. T_s and T denote the sampling period and symbol

 ISBN: 978-1-61208-353-7

duration, respectively. And Δf is the residual carrier frequency offset and θ_c the residual carrier phase. The oversampling ratio is defined by $\rho = T_s / T$. The pulse $p(t)$ reflects the channel effects and $w(t)$ is zero-mean additive white Gaussian noise (AWGN).

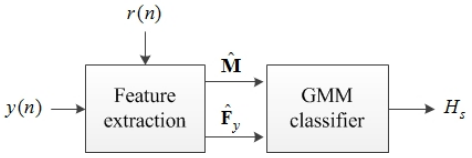


Figure 1. The structure of the proposed modulation classification scheme.

The input $r(n)$ is the signal to generate the reference data for GMM, which has the same modulation type as $y(n)$ with AWGN. The signal is required only for the learning process.

At feature extraction stage, we extract the feature vector $\hat{\mathbf{F}}_y$ and the reference data matrix $\hat{\mathbf{M}}$ from $y(n)$ and $r(n)$, respectively. Then, the GMM classifier is used to predict the modulation type H_s corresponding to the received signal.

III. CYCLIC CUMULANTS (CCS)

Most communication signals made by human represent cyclostationary characteristic, i.e., statistical properties of the signal are varying periodically with respect to time. Up to now, the CCs are widely used for AMC among many cyclostationary statistics of the signal.

The *k*-th order with *q*-conjugate (*k*, *q*) CCs of $y(n)$ are defined as Fourier coefficients of time-varying *k*-th order cumulants [8] as follows

$$C_y(\beta, \boldsymbol{\tau})_{k,q} \triangleq \left\langle C_y(n, \boldsymbol{\tau})_{k,q} e^{-j2\pi\beta n} \right\rangle \tag{2}$$

where $\langle \cdot \rangle$ means sample average, β is the cycle frequency, and $\boldsymbol{\tau} = [\tau_1, \cdots, \tau_n]^T$ is the vector of time lags.

Let $P = \{v_j\}_{j=1}^{p}$ be a set of partitions v_j for the index set $\{1, 2, ..., n\}$, and p is the number of elements in the set. For example, there are five different sets of P for $k = 3$, i.e., {(1, 2, 3)}, {(1), (2, 3)}, {(2), (1, 3)}, {(3), (1, 3)}, and {(1), (2), (3)}, and p is 1, 2, 2, 2, and 3, in order. Then, (2) can be further expressed in terms of cyclic moments according to the relation between moments and cumulants [8] as

$$C_y(\beta, \boldsymbol{\tau})_{k,q} = \sum_{P} (-1)^{p-1} (p-1)! \left[\sum_{\boldsymbol{\alpha}^T \mathbf{I} = \beta} \prod_{j=1}^{p} M_y(\alpha_j, \boldsymbol{\tau}_{v_j})_{k_j, q_j} \right] \tag{3}$$

where $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_m]^T$ is the vector composed of cycle frequencies, $\mathbf{I}$ is the *m*-dimensional vector whose all *m* elements are ones, and $M_y(\alpha_j, \boldsymbol{\tau}_{v_j})_{k_j, q_j}$ is the (k_j, q_j) cyclic moment of $y(n)$ at cycle frequency α_j and delay vector $\boldsymbol{\tau}_{v_j} = [\tau_1, \cdots, \tau_{k_j}]^T$. Then (k_j, q_j) cyclic moment is denoted by

$$M_y(\alpha_j, \boldsymbol{\tau}_{v_j})_{k_j, q_j} = \left\langle \prod_{\eta=1}^{q_j} y^*(n + \tau_\eta) \prod_{\xi=q_j+1}^{k_j} y(n + \tau_\xi)\, e^{-j2\pi\alpha_j n} \right\rangle \tag{4}$$

where * means complex conjugate.

Since the CCs are just Fourier series coefficients of the cumulants as shown in (2), we can notice that the properties of the CCs follow those of the cumulants. Therefore, it is worth examining the properties of the cumulants. Some important properties of the cumulants are as follows [9]:

- If $y(n)$ is Gaussian random process, its cumulants higher than second order are zero.
- The cumulants of the sum of the independent random processes are equal to the sums of their cumulants.
- All odd-order cumulants are equal to zero when the distribution is symmetric.

Cumulants higher than second order are called higher order cumulants. They characterize statistical amplitude features such as the shape of a signal regardless of Gaussian noise. Digitally modulated signal has its own distinctive pulse shape and amplitude distribution, therefore it shows different higher order CCs. On the other hand, the odd-order CCs are zero due its symmetric amplitude distribution. In this paper, we use higher even-order CCs such as fourth, sixth, and eighth order CCs as features to identify the modulation type. This is why the computational complexity greatly increases as the order of CC becomes higher.

Fig. 2 represents the magnitudes of the various (k, q) CCs with the eight modulation signals. These values are calculated from (3) with 10 trials per each modulation type. In this figure, BPSK signal is omitted because its CC values are much greater than others. We notice that each modulation type has distinct values with different kinds of (k, q) CCs. In the case of high level modulation signal (e.g., QPSK and 8-PSK), the overlapping of fourth and sixth order CCs become to split in eighth order CCs. And, except (8, 0) and (8, 4) CCs, all CCs are nearly constant over SNR variation.

 ISBN: 978-1-61208-353-7

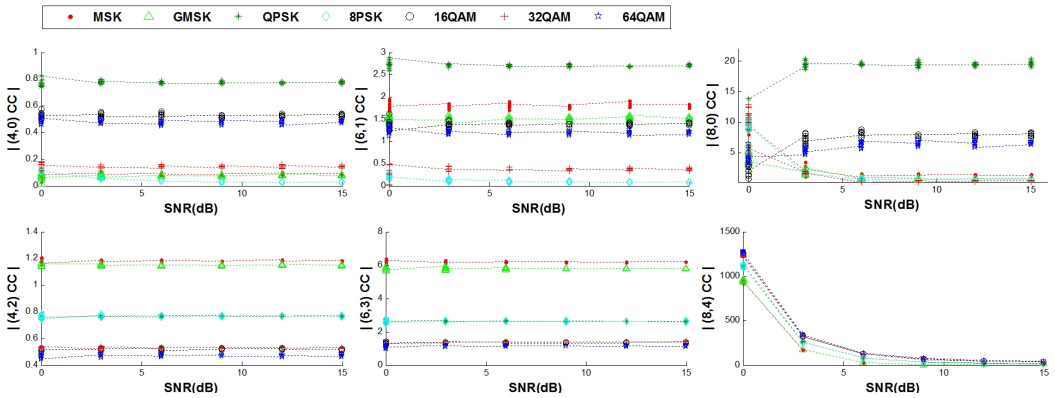


Figure 2. The magnitudes of the k-th order with q-conjugate (k, q) cyclic cumulants for eight modulation signals with various SNR.

IV. PROPOSED MODULATION CLASSIFICATION SCHEME

A. Feature extraction

From the results of Figure 2, we choose the three magnitudes of (4, 2), (6, 3), (8, 0) CCs as the features. With the number of elements $N_F = 3$, the feature vector $\mathbf{F} \in \mathbb{C}^{N_F}$ of CC magnitudes at zero-time lag is given by

$$\mathbf{F} = [\ |C_y(\beta,\mathbf{0})_{4,2}|,\ |C_y(\beta,\mathbf{0})_{6,3}|,\ |C_y(\beta,\mathbf{0})_{8,0}|\ ]^T \quad (5)$$

We estimate the CCs from (3) using the magnitude of the maximum value of cyclic moments which are obtained by FFT operation of (4). We select the number of FFT points as follows

$$N_{FFT} = 2^{\lceil \log_2(N\rho) \rceil} \quad (6)$$

where $\lceil x \rceil$ means the smallest integer not less than x and N is the number of symbols.

In order to setup reference data, we generate baseband modulation signal $r(n)$ by increasing SNR with N_{snr} steps. Then, estimate the feature vector N_{trial} times for each modulation type. As a result, the total number of training samples is

$$N_{total} = N_{snr} N_{mod} N_{trial} \quad (7)$$

where N_{mod} is the number of candidate modulation signals to be identified.

The feature matrix for the reference data obtained from $r(n)$ is constructed as follows

$$\hat{\mathbf{M}} = \left[\hat{\mathbf{m}}_1, \ \cdots \ , \hat{\mathbf{m}}_{N_{\text{mod}}}\right] \in \mathbb{C}^{N_F \times N_{total}} \quad (8)$$

where $\hat{\mathbf{m}}_i \in \mathbb{C}^{N_F \times (N_{trial} \times N_{snr})}$ corresponds to the i-th modulation type, and is given by

$$\hat{\mathbf{m}}_i = [\hat{\mathbf{F}}_1^{\gamma}\ \hat{\mathbf{F}}_2^{\gamma}\ \cdots\ \hat{\mathbf{F}}_{N_{iter}}^{\gamma}, \ \cdots, \ \hat{\mathbf{F}}_1^{\gamma+N_{snr}-1}\ \hat{\mathbf{F}}_2^{\gamma+N_{snr}-1}\ \cdots\ \hat{\mathbf{F}}_{N_{iter}}^{\gamma+N_{snr}-1}] \quad (9)$$

where the superscript of the feature vector means SNR of γ to $\gamma + N_{snr} - 1$.

Fig. 3 represents the feature distribution of the reference data in a 3-dimensional feature space spanned by (4, 2), (6, 3), (8, 0) CCs. We observe that there are N_{mod} different feature clusters corresponding to different modulation types and they are not overlapped each other. Since we utilize noisy signals for the reference data, each feature cluster reveals a probability distribution with mean as its center.

Therefore, by using the probability distribution of the reference data, we can improve the performance of modulation classification.

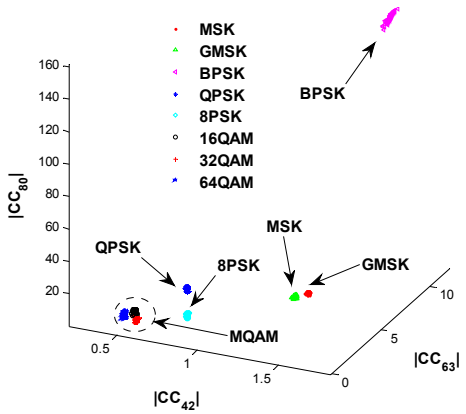


Figure 3. Reference data in a 3-dimentional feature space for the eight modulation signals with SNR of 3 to 10 dB.

B. GMM-based classifier

The modulation classification can be regarded as to find the cluster to which the feature vector $\hat{\mathbf{F}}_y$ of the received signal belongs. For classification, we use a GMM-based

method which considers probability distribution of the reference data.

The method is based on the fact that, for the reference data $\hat{\mathbf{M}} = \{\hat{\mathbf{F}}_i\}_{i=1}^{N_{total}}$, it's unknown probability distribution can be represented by a weighted linear combination of multivariate Gaussian density functions, given by [10]

$$p(\hat{\mathbf{M}} \mid \boldsymbol{\Theta}) = \sum_{s \in S} p(\hat{\mathbf{M}} \mid C_s, \boldsymbol{\theta}_s) P_s \tag{10}$$

where S = {MSK, GMSK, BPSK, QPSK, 8-PSK, 16-QAM, 32-QAM, 64-QAM} is the set of eight modulation types, and $p(\hat{\mathbf{M}} \mid C_s, \boldsymbol{\theta}_s)$ is the N_F-variate Gaussian density function of the cluster C_s with unknown parameter $\boldsymbol{\theta}_s$, which is determined by a mean vector $\boldsymbol{\mu}_s \in \mathbb{C}^{N_F}$ and a covariance matrix $\boldsymbol{\Sigma}_s \in \mathbb{C}^{N_F \times N_F}$. P_s is the prior probability of the Gaussian density function for modulation *s*. The unknown parameter vectors can be collectively represented by

$$\boldsymbol{\Theta} = \{\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s, P_s\} \tag{11}$$

The N_F-variate Gaussian density function is as follows

$$p(\hat{\mathbf{M}} \mid C_s, \boldsymbol{\theta}_s) = \frac{1}{(2\pi)^{N_F/2} \mid \boldsymbol{\Sigma}_s \mid^{1/2}} \cdot \exp\left(-\frac{1}{2}(\hat{\mathbf{M}} - \boldsymbol{\mu}_s)^T \boldsymbol{\Sigma}_s^{-1} (\hat{\mathbf{M}} - \boldsymbol{\mu}_s)\right) \tag{12}$$

The unknown parameter vectors can be obtained by the expectation-maximization (EM) algorithm which maximizes the expectation of the loglikelihood function of the GMM.

To apply the EM algorithm, the initial estimate $\boldsymbol{\Theta}(0)$ and a termination threshold ε are required for iteration. We estimate the mean vector and covariance matrix from the reference data for the initial estimate $\boldsymbol{\Theta}(0)$ and assume equal value of prior probabilities for clusters.

The EM algorithm is summarized as follows:

(1) Expectation step: At iteration λ, where $\boldsymbol{\theta}(\lambda)$ is available, compute the expected value of the followings:

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}(\lambda)) = \sum_{i=1}^{N_{total}} \sum_{s=1}^{S} P(C_s \mid \hat{\boldsymbol{F}}_i; \boldsymbol{\Theta}(\lambda)) \cdot \ln(p(\hat{\boldsymbol{F}}_i \mid C_s; \boldsymbol{\theta}_s) P_s \tag{13}$$

(2) Maximization step: Compute the next (λ+1)-th estimate of $\boldsymbol{\theta}$ by maximizing $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}(\lambda))$, that is,

$$\boldsymbol{\Theta}(\lambda+1) = \boldsymbol{\Theta}(\lambda) \text{ such that } \frac{\partial Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}(\lambda))}{\partial \boldsymbol{\Theta}} = 0 \tag{14}$$

After some manipulations, the following general forms of parameters are derived as follows

Means:

$$\mu_j(\lambda+1) = \frac{\sum_{i=1}^{N_{total}} P(C_s \mid \hat{\boldsymbol{F}}_i; \boldsymbol{\Theta}(\lambda)) \hat{\boldsymbol{F}}_i}{\sum_{i=1}^{N_{total}} P(C_s \mid \hat{\boldsymbol{F}}_i; \boldsymbol{\Theta}(\lambda))} \tag{15}$$

Covariance Matrices:

$$\Sigma_s(\lambda+1) = \frac{\sum_{i=1}^{N_{total}} P(C_s \mid \hat{\boldsymbol{F}}_i; \boldsymbol{\Theta}(\lambda))(\hat{\boldsymbol{F}}_i - \boldsymbol{\mu}_s(\lambda))(\hat{\boldsymbol{F}}_i - \boldsymbol{\mu}_s(\lambda))^T}{\sum_{i=1}^{N_{total}} P(C_s \mid \hat{\boldsymbol{F}}_i; \boldsymbol{\Theta}(\lambda))} \tag{16}$$

Prior Probabilities:

$$P_s(\lambda+1) = \frac{1}{N_{total}} \sum_{i=1}^{N_{total}} P(C_s \mid \hat{\boldsymbol{F}}_i; \boldsymbol{\Theta}(\lambda)) \tag{17}$$

For the *i*-th reference data and cluster C_s, the a posteriori probability for Gaussian density function of C_s is given by

$$p(C_s \mid \hat{\boldsymbol{F}}_i; \boldsymbol{\Theta}(\lambda)) = \frac{p(\hat{\boldsymbol{F}}_i \mid C_s; \boldsymbol{\theta}_s(\lambda)) P_s(\lambda)}{\sum_{s=1}^{S} P(\hat{\boldsymbol{F}}_i \mid C_s, \boldsymbol{\theta}_s(\lambda)) P_s(\lambda)} \tag{18}$$

If the following termination condition is not satisfied, the iteration (1) and (2) continues.

$$\| \boldsymbol{\Theta}(\lambda+1) - \boldsymbol{\Theta}(\lambda) \| < \varepsilon \tag{19}$$

Based on the estimated GMM of the reference data, the modulation type of the received signal can be determined. After computing the a posteriori probability between GMM of each cluster and the feature vector $\hat{\mathbf{F}}_y$ of the received signal, select the cluster representing the maximum value of a posteriori probability, and decide the corresponding modulation type as that of the received signal, as follows

$$H_s = \arg\max_{s \in S} P(C_s \mid \hat{\mathbf{F}}_y; \boldsymbol{\theta}_s) \tag{20}$$

V. SIMULATION RESULTS

We use eight baseband modulation signals of MSK, GMSK, BPSK, QPSK, 8-PSK, 16-QAM, 32-QAM, and 64-QAM to be classified. That is, the number of candidate

 ISBN: 978-1-61208-353-7

modulation signals to be identified is $N_{mod} = 8$. The symbol rate and oversampling ratio are $T = 1$ and $\rho = 11$, respectively. A raised-cosine filter with roll-off factor 0.35 is used for pulse shaping in generation of *M*-PSK and *M*-QAM signals. For GMSK, the bandwidth-time product for Gaussian filter is 0.5. We assume that the relative carrier frequency offset Δf is 0 and phase offset $\Delta\theta$ is uniformly distributed over $[-\pi, \pi)$. For the reference data, the modulation signal is generated with 3 to 10 dB of SNR, or $\gamma = 3$ and $N_{snr} = 7$. The initial estimate of the GMM parameter, $\boldsymbol{\Theta}(0)$, and the termination condition ε are set to be as in Table I and 1e-6, respectively.

TABLE I. PARAMETER ESTIMATE OF REFERENCE DATA

Mod.	Mean			Covariance ($\times 10^{-4}$)			Prior prob.
MSK	1.25	7.0	2.07	0	0	4	0.125
				0	5	-16	
				4	-16	956	
GMSK	1.20	6.46	1.68	0	0	2	
				0	5	-5	
				2	-5	792	
BPSK	1.54	10.63	157.35	0	3	56	
				3	30	599	
				56	599	1.19	
QPSK	0.77	2.63	19.6	0	0	-2	
				0	2	-9	
				-2	-9	809	
8-PSK	0.76	2.63	0.7	0	0	-1	
				0	2	-3	
				-1	-3	287	
16-QAM	0.52	1.34	7.92	0	1	-2	
				1	3	-9	
				-2	-9	753	
32-QAM	0.52	1.35	0.32	0	1	1	
				1	3	1	
				1	1	475	
64-QAM	0.47	1.15	6.46	0	1	1	
				1	2	-2	
				-1	-2	769	

As a performance measure for AMC, we use the probability of correct classification P_c. To obtain the probability, 300 trials are performed in AWGN channels.

Fig. 4 shows the performance with respect to the number of symbols of the received signal at SNR = 10 dB. The performance improves as the number of symbols increases. At low SNR, the bad classification performance of *M*-QAM deteriorates the whole performance of the scheme. Clusters of *M*-QAM are closely located in the feature space as shown in Fig. 3. Therefore, when the number of symbols is small, the variance of the feature vectors is large and results in bad classification of *M*-QAM signal. We observe that the probability of correct classification approximates one for the number of symbols larger than 3000. Therefore, we determine the number of received symbols as 4000 for the next experiments.

Fig. 5 compares the performance of the proposed scheme

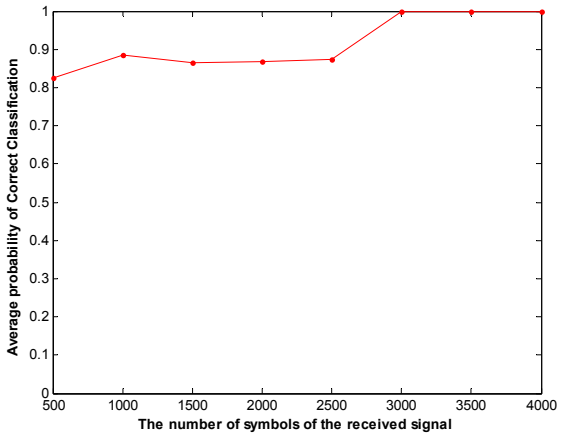


Figure 4. Performance of the proposed scheme with the number of symbols of the received signal at SNR = 10 dB.

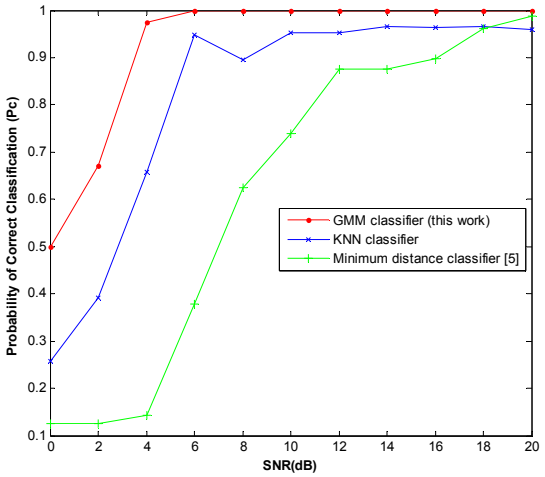


Figure 5. Comparison of the performance of the proposed scheme, KNN classifier, and the minimum distance classifier [5].

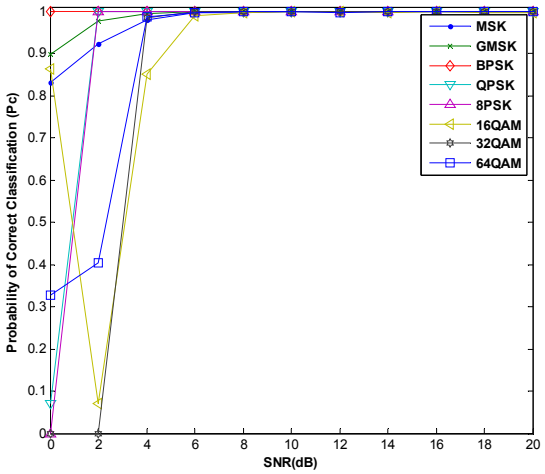


Figure 6. Performance of the proposed scheme for each modulation signal.

 ISBN: 978-1-61208-353-7

with the *K*-nearest neighbor (KNN) classifier and the minimum distance classifier of [5]. In [5], the eighth order cumulant with even number of conjugates is used as a feature. We observed that the performance of the proposed scheme is better than others over the entire range of SNR. When SNR is higher than 8 dB, the performance of the proposed scheme is nearly perfect.

Fig. 6 shows the performance for the individual modulation signal in the same conditions as those of Fig. 5. As shown in Fig. 5, the probabilities of correct classification for all modulation types become 1 with SNR higher than 6 dB. At lower SNR than 4 dB, the performance of the *M*-QAM degrades abruptly because of the reason mentioned above.

VI. CONCLUSION

In this paper, we proposed an feature-based automatic modulation classification scheme for eight digital modulation signals of MSK, GMSK, BPSK, QPSK, 8-PSK, 16-QAM, 32-QAM, and 64-QAM. We use the magnitude of fourth, sixth, and eighth order cyclic cumulants as features. To approximate the probability distribution of the features, the Gaussian mixture model is used. Based on the probability distribution, we identify the modulation type of the received signal according to the probability of its features. The simulation results show that the classification performance is much better as compared to the conventional K-nearest neighbor classifier and the minimum distance classifier.

ACKNOWLEDGMENT

This work has been supported by the National GNSS Research Center Program of DAPA and ADD.

REFERENCES

[1] O. A. Dobre and R. Inkol, "Blind signal identification: Achievements, trends, and challenges," 9th International Conference on Communications (COMM), 2012, pp. 349-352)

[2] J. L. Xu, S. Wei, and Z. MengChu, "Likelihood-Ratio Approaches to Automatic Modulation Classification," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 41, 2011, pp. 455-469.

[3] M. W. Aslam, Z. Zhu, and A. K. Nandi, "Automatic Modulation Classification Using Combination of Genetic Programming and KNN," IEEE Transactions on Wireless Communications, vol. 11, 2012, pp. 2742-2750.

[4] E. Like, V. Chakravarthy, P. Ratazzi, and Z. Wu, "Signal Classification in Fading Channels Using Cyclic Spectral Analysis," EURASIP Journal on Wireless Communications and Networking, 2009, pp. 879-812.

[5] O. A. Dobre, A. Abdi, Y. Bar-Ness, and W. Su, "Cyclostationarity-Based Modulation Classification of Linear Digital Modulations in Flat Fading Channels," Wireless Personal Communications, vol. 54, 2009, pp. 699-717.

[6] O. A. Dobre, M. Oner, S. Rajan, and R. Inkol, "Cyclostationarity-Based Robust Algorithms for QAM Signal Identification," IEEE Communications Letters, vol. 16, 2012, pp. 12-15.

[7] J. G. Liu, W. Xianbin, J. Nadeau, and L. Hai, "Modulation classification based on gaussian mixture models under multipath fading channel," IEEE Global Communications Conference (GLOBECOM), Dec. 2012, pp. 3970-3974.

[8] W. A. Gardner and C. M. Spooner, "The cumulant theory of cyclostationary time-series. I. Foundation," IEEE Transactions on Signal Processing, vol. 42, 1994, pp. 3387-3408.

[9] C. L. Nikias and J. M. Mendel, "Signal processing with higher-order spectra," IEEE Signal Processing Magazine, vol. 10, 1993, pp. 10-37.

[10] S. Theodoridis and K. Koutroumbas, Pattern Recognition. Elsevier Science, 2008.

 ISBN: 978-1-61208-353-7

Indoor Location Estimation Using Smart Antenna System with Virtual Fingerprint Construction Scheme

Shiann-Tsong Sheu, Yen-Ming Hsu, Hsueh-Yi Chen
Department of Communication Engineering
National Central University
Taoyuan, Taiwan
Email: stsheu@ce.ncu.edu.tw; huge1124a@gmail.com; god455108@gmail.com

***Abstract*—Regarding indoor location estimation, many smart positioning techniques have been proposed and they could be classified into several categories. Fingerprinting is one of the categories and its features are to build the indoor radio map during offline phase and to utilize the radio map to estimate the location during online phase. Creating indoor radio map is the most critical step, where the accuracy of location estimation depends on the distribution and number of reference points (RPs). Usually, the time required to collect received signal strengths (RSSs) is proportional to the number of 'real' RPs. To reduce the complexity of offline process, this paper proposes an efficient scheme, which is based on the well-known signal propagation model, to construct 'virtual' RPs in indoor radio map. The RSSs on virtual RPs are calculated and used to substitute for the training RSSs on 'real' RPs collected during offline phase. The proposed scheme can not only shorten the collecting time but also achieve high accuracy for indoor positioning. The indoor location estimation using smart antenna system (SAS) also requires longer time to collect all RSS information because of multiple antennas. By combining this scheme with SAS, we can easily obtain enough and valid RSSs information to build the indoor radio map in a more efficient way. Experimental results showed that applying virtual fingerprint construction scheme on SAS can decrease 33% 'real' RPs in indoor radio map without sacrificing the positioning accuracy.**

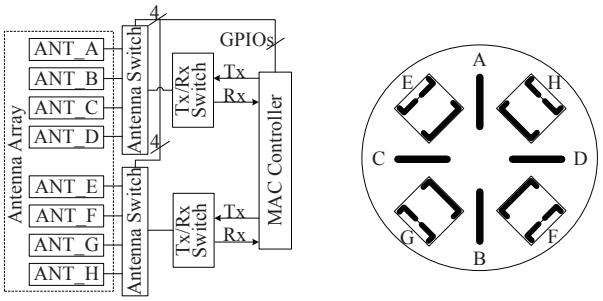
Keywords - Fingerprint; Indoor Positioning; Location Estimation; Receive Signal Strength; Smart Antenna System; WLAN

I. INTRODUCTION

Nowadays, indoor location estimation is very important for many contemporary location-based services, such as health care monitoring, indoor navigation, personal tracking, inventory control, and so on. Without Global Positioning System (GPS) in indoor environment, there are many alternative techniques have been proposed [1][2] and they could be classify into several categories including Time of Arrival (ToA), Angle of Arrival (AoA) and the RSS-based location [3][4][5]. There are two reason why ToA and AoA are unsuitable for indoor environments: 1) they require the line-of-sight (LOS) between a pair of transmitter and receiver and 2) the special hardware design to support ToA and AoA algorithms is expensive. RADAR [6] is the first explored indoor positioning system, which computes the user location based on the RSS [7][8] from wireless local area network (WLAN). It is an attractive and suitable solution for indoor positioning since it reuses the existing and pervasive WLAN infrastructure. However, RSS-based location estimation is challenging because the radio signal is easily affected by reflection, refraction, shadowing and scattering. To resolve the problem of unstable RSSs in indoor environment, fingerprinting is considered as a feasible solution. Using the fingerprint technique with smart antenna system (SAS), which uses multi-antenna to form several logical APs with different radio coverage areas, three major processes for location estimation are required: 1) building the indoor radio map, 2) logical AP selection and 3) location estimation. The fingerprint technique consists of the offline phase and the online phase. During the offline phase, the target indoor area is logically partitioned into a number of equal subareas. For simplicity, all the corners of subareas could be arranged as the reference points (RPs) in indoor radio map. After then, a smart phone is placed on every RP to transmit a number of dummy packets to every logical AP co-located within the centric SAS. For each dummy packet, the SAS measures the RSS and stores it with the coordinate of corresponding RP and the index of logical AP into the database in designated server. During the online phase, SAS measures the RSSs of received packets sent from one mobile device at unknown position and then forwards them to designated server for further processing. Based on these received RSSs, the server calculates the possible position of mobile device by means of online location algorithm. The estimated location may either fully match or just closely match an RP in database.

Without loss of generality, the online location algorithm often relies on well-known pattern-matching algorithms including the k-nearest neighbor [1], neural network, the probabilistic approach [9], and so on. In other words, the pattern-matching algorithm tries to figure out the possible location according to the known relation between the RSS and the position of mobile device. From our observations, the weighted-based kernel function, which has been proposed in [10], to improve the indoor positioning accuracy is very suitable for SAS-based online location estimation.

 ISBN: 978-1-61208-353-7



(a) System block diagram (b) The vertical view of antenna array layout

Fig. 1. The architecture of smart antenna system

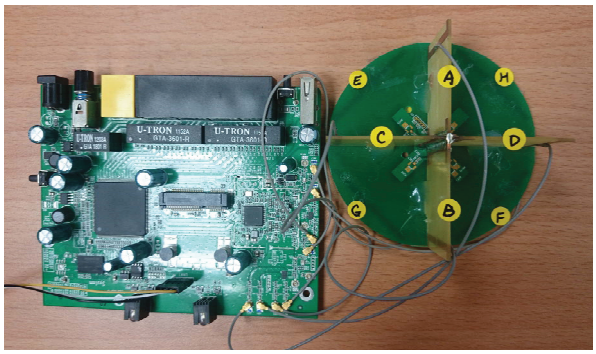


Fig. 2. Smart antenna system

TABLE I
ANTENNA SET CONFIGURATION

Antenna set No.	Antennas	Antenna set No.	Antennas
0	A:E	8	C:E
1	A:F	9	C:F
2	A:G	10	C:G
3	A:H	11	C:H
4	B:E	12	D:E
5	B:F	13	D:F
6	B:G	14	D:G
7	B:H	15	D:H

Recall that constructing radio map is the most critical step, where the accuracy of location estimation depends on the number of RPs. Obviously, the process of collecting RSSs often takes much time as there are many real RPs in radio map. To reduce the complexity of offline process, this paper proposes an efficient scheme, which utilizes the signal propagation model, to construct virtual RPs in radio map. Those calculated RSSs on virtual RPs are treated as the training data on real RPs. The proposed scheme not only shortens the RSS collecting time but also achieves high accuracy for indoor positioning. Similarly, using SAS to estimate indoor location also requires longer time to collect all RSS information. Combining this scheme with SAS has the advantage of easily obtaining enough RSSs information to build the radio map in a more efficient way.

The organization of this paper is described as follows. In Section II, we introduce the developed SAS for indoor location estimation. Section III is devoted to the proposed scheme of efficiently constructing the radio map and the applied indoor estimation algorithm. In Section IV, we describe the experimental environment and the obtained results. Finally, Section V concludes this paper and also gives some remarks for future works.

II. Introduction of Smart Antenna System

Smart antenna system is an emerging technique to promote the communication efficiency in wireless networks. It works by taking the advantage of the diversity effect at the transceiver in wireless systems. The diversity effect is used to decrease the error rate during data communication and to increase data transmission rate between transmitter and receiver.

The SAS designed for indoor location estimation is composed of eight directional antennas which are divided into horizontal plane (with horizontal polarization) and vertical plane (with vertical polarization). Each plane has four antennas to cover four different directions, one antenna for one direction. The SAS generates 16 (4×4) distinct antenna sets by combining any pairs of antennas from two planes, where one antenna from horizontal plane and the other antenna from vertical plane. As a result, those antenna sets have different signal coverage and trajectory to provide a better characteristic for indoor location estimation. For brevity, in this paper the antenna set is called as 'logical AP' ,which represents for the traditional AP from the aspect of WLAN client. The block diagram of SAS is shown in Fig. 1(a)[13]. The embedded platform includes a programmable 2×2 IEEE 802.11n Media Access Control (MAC) Controller and it uses 8 general purposes I/Os (GPIOs) to connect to two switches for dynamically selecting 2 antennas from 8 antennas (one from horizontal plane and one from vertical plane) for transmissions and receptions. The state of the switch is determined by the DC voltage. The RF output connects with the designed antenna array on the circle plate as shown in Fig. 1(b)[13]. In the antenna array, four antennas, denoted as ANT_A, ANT_B, ANT_C, and ANT_D, respectively, are formed an angle 90 degree with respect to the other four antennas, denoted as ANT_E, ANT_F, ANT_G, and ANT_H, respectively. Fig. 2[13] is the photograph of smart antenna system and Table I[13] shows the configuration between the index of antenna set and corresponding antennas controlled by MAC controller. For readability, terms 'logical AP' and 'antenna set' are interchangeable.

III. Smart Indoor Location Estimation

If the indoor signal propagation model is able to derive the precise path loss, the RSSs on any RP in indoor environment could be obtained without any actual measurement. However, from our observations, it is very difficult to use a signal propagation model with single pass loss parameter set to cover the whole indoor environment. That is why the convenient way to achieve more precise indoor positioning is to arrange as more RPs as possible regardless of the signal propagation model. However, because the SAS aims to generate distinct antenna patterns in different directions, the path loss parameters for different antenna sets would be naturally varied even in the same indoor environment. In other words, if an RP in the indoor radio map can have different path

 ISBN: 978-1-61208-353-7

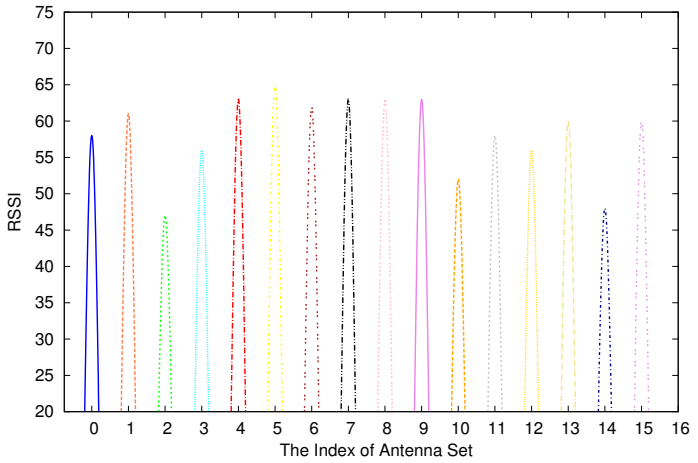


Fig. 3. The real RSS values derived from measurements

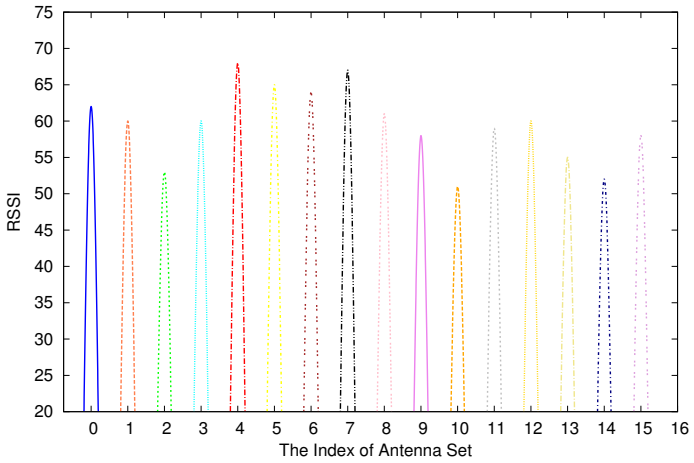


Fig. 4. The virtual RSS values generated from signal propagation model

loss parameters to reflect different antenna patterns, the indoor signal propagation model could be applicable for fingerprint-based indoor location estimation. It also implies that the number of real RPs for building indoor radio map could be reduced significantly. In a word, the way of increasing the RPs in this paper is to utilize the real measurements from real RPs to generate virtual measurements of other new RPs, called virtual RPs, in database for shortening the offline process meanwhile maintaining high positioning accuracy.

A. Location Estimation Algorithm

At the beginning, we simply assume RSS values are random variables and they could be modeled as a Gaussian distribution. The whole location process is composed of two phases. The offline phase is to build the radio map which is the key part for the fingerprint technique. As mentioned above, with the signal propagation model, we are going to further generate more virtual RPs in radio map for online computation.

We divide offline phase into two steps: 1) constructing the radio map which stores RSS values from 16 antenna sets (i.e., 16 logical APs) at every real RP and 2) performing the logical AP selection to find a proper candidate set of logical APs for online location estimation. This set could be regarded as the basis for subsequent online computation. Online phase is also composed of two steps: 1) collecting RSS values from the logical APs in candidate set at unknown location, and 2) estimating the user location based on the RSS values from logical APs in candidate set and the radio map prepared during offline phase.

B. Probabilistic-based Location Estimation

1) Scheme of Virtual Fingerprint Construction: Generally, the relation between signal strength and distance can be expressed by the following equation:

$$P(d)[dBm] = P(d_0)[dBm] + 10\gamma \log_{10}(\frac{d}{d_0}) + X_\sigma \quad (1)$$

where $P(d_0)$ represents the transmitting power of a wireless device at the reference distance d_0, d is the distance between the wireless device and the SAS, γ is the path loss exponent and X_σ is the shadow fading which follows zero mean Gaussian distribution with σ standard deviation. To figure out d_0 and d, we should define $d_{AP}(R) = \sqrt{(X_R - X_{AP})^2 + (Y_R - Y_{AP})^2}$ as the distance between smart antenna AP and the real RP, where (X_{AP}, Y_{AP}) and (X_R, Y_R) denote the coordinates of smart antenna AP and a real RP respectively.

Owing to the diverse radio field pattern and trajectory in SAS, each antenna set should have its own distinct value of path loss parameter. For the case that the SAS is located in the center within the target area, we can easily figure out the path loss exponent (γ) for each antenna set by two real RPs and then use it to generate new RPs in the radio map. Let m denote the index of antenna set and γ_m denote the path loss exponent of the m-th antenna set. More specifically, we first carefully choose two real RPs (along one radiation direction) from the radio map and then calculate the path loss exponent of every antenna set. After deriving all the path loss exponents (i.e., the set of $\{ \gamma_1, \gamma_2, ..., \gamma_{16} \}$), the RSSs of a new virtual RP with specified distance to SAS are derivable. Afterward, the virtual RP and corresponding RSSs are added into the radio map. Fig. 3 shows the real RSS values derived from measurements at a certain RP and Fig. 4 shows the virtual RSS values generated by the proposed virtual fingerprint construction scheme. From these two figures, it reveals that the virtual RP indeed has similar radio characteristic compared with the actual measurements.

In this paper, we let N denote as the number of total RPs for online computation and N_R and N_V respectively represent the numbers of real RPs and RPs in the database. In this paper, we let $N_R = 24$ and $N = 36$, which indicates that there are 12 virtual RPs (N_V=12) in the database.

2) The Algorithm of Antenna Set Selection: For some locations, the measured RSSs from one logical AP might be similar to the other logical AP. Such similarity will lead to biased location estimation and redundant computations. Therefore, it is important of using antenna set selective technique to determine a proper candidate set of logical APs for location estimation.

The selection methodology is to choose a set of logical APs

 ISBN: 978-1-61208-353-7

with the highest correlation of RSSs. Therefore, the highest correlation of RSS could provide the highest probability of coverage over time [11]. It is worthwhile to notice that the logical APs with the strongest signal with respect to WLAN client may not provide the best positioning accuracy [12].

Here, we briefly introduce the weighted-based kernel function [10] and explain how to calculate the weighted values for the SAS. First, the information of each logical AP is quantified by calculating the signal discrimination between different RPs. Let $\bar{o}_m$ denote the mean value of RSSs obtained from the m-th antenna set. We have

$$\bar{o}_m = \frac{1}{N \cdot N_{rssi}} \sum_{i=1}^{N} \sum_{j=1}^{N_{rssi}} o_{i,m(j)}, \tag{2}$$

where N, N_{rssi}, and $o_{i,m(j)}$ represent the total number of RPs, the number of collected RSSs at every RP and the j-th ($1 \leq j \leq N_{rssi}$) collected RSS at the i-th ($1 \leq i \leq N$) RP obtained from the m-th ($1 \leq m \leq M$) logical AP, respectively. Therefore, the information of the m-th logical AP denoted as η_m is given by

$$\eta_m = \frac{1}{N \cdot N_{rssi}} \sum_{i=1}^{N} \sum_{j=1}^{N_{rssi}} (o_{i,m(j)} - \bar{o}_m). \tag{3}$$

It is noted that the RSS does not change significantly at different RPs if η_m is small. On the contrary, the higher value of η_m reveals that the RSSs varies obviously at different RPs.

To obtain a quantitative metric by calculating the spatial likelihood of the measured RSSs, a quasi entropy function denoted as $H(\cdot)$ is used to determine the weight ϖ_m. We have

$$\varpi_m = 1 + \frac{H(1-\eta_m^*)}{\max[H(1-\eta_m^*)]}, \quad \text{if } 0 < m \leq M, \tag{4}$$

where $\eta_m^* = \frac{\eta_m}{\sum_{m=1}^{M} \eta_m}$ and M indicate the normalized value and the number of logical APs respectively. The value of weight ranges from 1 to 2. The weighted value ϖ_m is related with the value η_m. [13] has shown that the order of η_m does not affect the positioning accuracy in SAS. Therefore, the weighted value ϖ_m could be ignored in SAS.

3) Online Location Estimation: This paper adopts the theorem of Bayesian Network [14] to estimate the conditional probability of each location according to the observed samples during the offline and online phases.

Let l_i and O_i denote the location of the i-th RP and the mean RSS observation set of the i-th RP respectively. The vector of RSS values for the i-th logical AP could be denoted as $O_i = [O_{i,1}, O_{i,2}, \ldots, O_{i,M}]^T$. According to the inference form [15], given the observation O, the posterior distribution indicates the likelihood of location l_i. We have

$$p(l_i|O) = \frac{p(O|l_i)p(l_i)}{p(O)} = C \cdot p(O|l_i), \tag{5}$$

where C indicates a constant value when the value of l_i follows a uniform distribution and and O is the online measured RSS. To estimate $p(O|l_i)$, the kernel-based method [9][16]

Algorithm 1 Summaries of the offline and online phases

Offline phase:

1: For the i-th RP, measure RSS from M logical APs to form observation set $O_i = [O_{i,1}, O_{i,2}, \ldots, O_{i,M}]^T$.
2: Figure out the distance from an RP, say R, to the central AP
$d_{AP}(R) = \sqrt{(X_R - X_{AP})^2 + (Y_R - Y_{AP})^2}$
3: Using d,d_0 and the RSSs of two specified real RPs to derive the path loss exponents $\gamma_m (1 \leq m \leq M)$ through the propagation model:
$P(d)[dBm] = P(d_0)[dBm] + 10\gamma \log_{10}(\frac{d}{d_0}) + X_\sigma$
4: Use the propagation model with derived pass loss exponent to compute the RSSs of a virtual RP.
5: Repeat step 2 to step 5 to generate all required virtual RPs.
The number of total RPs: N
The number of real RPs: N_R
The number of virtual RPs: N_V=N-N_R
6: Estimate the mean value ($\bar{o}_m$) of Gaussian distribution in the space for the m-th logical AP.

Online phase:

7: **for** $i = 1$ to N **do**
8: **for** $j = 1$ to N_{rssi} **do**
9: **for** $m = 1$ to M **do**
10: $K(O, O_i(j)) = \exp\left\{\frac{-1}{2\sigma^2} \sum_{m=1}^{M} [o_m - o_{i,m}(j)]^2\right\}$;
11: **end for**
12: $p(O|l_i) = \frac{1}{N_{rssi}} \sum_{j=1}^{N_{rssi}} K(O, O_i(j))$;
13: **end for**
14: $\hat{L} = \sum_{i=1}^{N} l_i p(O|l_i)$;
15: **end for**

could exploit the probabilistic weighting function by means of the kernel density estimator and training data. We have

$$p(O|l_i) = \frac{1}{N_{rssi}} \sum_{j=1}^{N_{rssi}} K(O, O_i(j)), \tag{6}$$

where $O_i(j)$ represents the j-th measured fingerprint at the i-th location and the Gaussian kernel function, $K(O, O_i(j))$, is given by

$$K(O, O_i(j)) = \exp\left\{\frac{-1}{2\sigma^2} \sum_{m=1}^{M} [o_m - o_{i,m}(j)]^2\right\} \tag{7}$$

where σ is an adjustable parameter which controls the accuracy of the location estimation. In this paper, we set σ as 0.5 because the experimental shows the accuracy can reach 90% when σ = 0.5. Therefore, the possible location can be derived

 ISBN: 978-1-61208-353-7

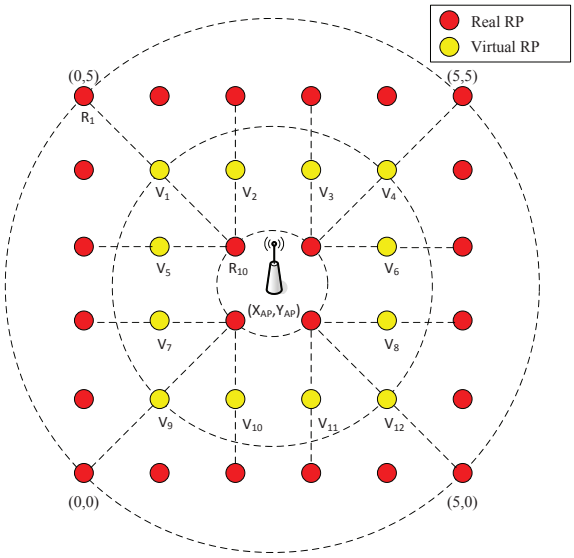


Fig. 5. The Layout of Radio Map

by

$$\hat{L} = \sum_{i=1}^{N} l_i p(O|l_i). \quad (8)$$

To wrap up, we give the summary of algorithm to clearly describe the offline and online phases.

IV. Experimental Environment and Results

A. Experimental Environment

Fig. 5 illustrates the layout of the indoor radio map, where the red nodes represent the real RPs and the yellow nodes represent the virtual RPs which are not really measured. The dash circles mean the radio wave contours generated from the SAS in the center and the dash line delegates how to use real RPs to derive the virtual RPs according to the signal propagation model. More specifically, a dash line always crosses two red nodes (i.e., real RPs) and one yellow node (i.e., virtual RP), and the yellow node just locates between two red nodes. It means that the new data of the yellow node is constructed within the range of known data of two red nodes by means of interpolation method. For applying the signal propagation equation, we need to have two distances (d and d_0) with respect to centric AP and two RSS values ($P(d)$, and $P(d_0)$) per logical AP to derive the real path loss exponent γ_m of the m logical AP. To do that, we need two real RPs for constructing every virtual RP. For the sake of explanation, we use the virtual RP, say V_1, as an example to explain how the scheme uses two real RPs, say R_1 and R_{10}, to generate virtual RP, V_1. Because the distances between AP and real RPs (R_1 and R_{10}) and the sixteen mean RSSs of sixteen logical APs measured at real RPs (R_1 and R_{10}) are known, the path loss exponent γ_m of the m-th logical AP along the path from R_{10} to R_1 is known also. By simply replacing the real RP R_1 as virtual RP V_1, the sixteen mean RSSs of sixteen logical APs probably measured at virtual RP V_1 can be derived if the sixteen path loss exponents are known already.

TABLE II
Experimental Parameters

Number of real RPs (N_R)	24
Number of virtual RPs (N_V)	12
Number of total RPs in radio map (N)	36
Number of access points	1
Number of antennas in the AP	8
Number of logical APs (M)	16
Distance between adjacent location in x-axis	0.5 m
Distance between adjacent location in y-axis	0.5 m
Number of measured RSS data per real RP (N_{rssi})	500

The set of experimental parameters is shown in Table II. The number of real RPs (N_R) is 24 and the number of virtual RPs (N_V) in radio map is 12. These virtual RPs play the role of real RPs. Consequently, the number of total RPs N is 36.

B. Experimental Results

Fig. 6 shows the path loss exponent γ as a function of virtual RP and logical AP. It can be observed that the path loss exponents are somehow different in the considered indoor space because of the diversity effect of SAS. As a consequence, the constructed RSSs of a virtual RP are depending on the location and they are different from that of the others. Fig. 7 indicates the numerical result that depicts the accuracy (in percentage) and standard deviation of error (in meter) respectively derived from the methods with and without applying the virtual fingerprint construction scheme. For fair comparisons, the numbers of real RPs considered in the method without the virtual fingerprint construction scheme are 36 (denoted as 36 R_{RP}) and 24 (denoted as 24 R_{RP}), and the numbers of real RPs and the number of virtual RPs considered in the method with the virtual fingerprint construction scheme is 24 and 12 (denoted as 24 R_{RP} + 12 V_{RP}). It is not difficult to find that using more RPs in location estimation will result in a higher accuracy and lower standard deviation of error in Fig. 7. That is, the method using only 24 real RPs gets the lowest performance in terms of accuracy and standard deviation of error. Moreover, the performance of the method using 24 real RPs and 12 virtual RPs is very close to the method using 36 real RPs. It implies that the proposed virtual fingerprint construction scheme can perform as good as conventional method but consuming less overhead during offline phase.

Fig. 8 illustrates the positioning accuracy derived from the method with the virtual fingerprint construction scheme (24 R_{RP} + 12 V_{RP}) as a function of the number of logical APs in candidate set. The more bright areas represent more precise indoor positioning. From Fig. 8(d), we can find that, when sixteen logical APs of the SAS are included in the candidate set, the proposed scheme can obtain the best positioning accuracy level. It implies that the antenna diversity effect of developed SAS is very obvious, which is very useful for indoor location estimation.

 ISBN: 978-1-61208-353-7

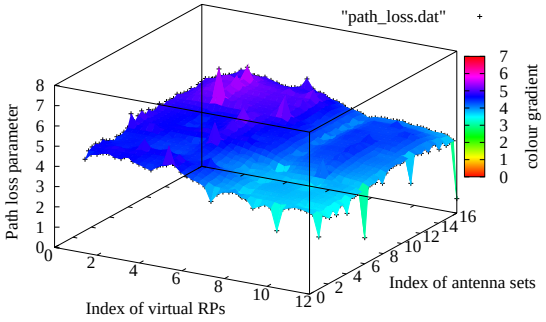


Fig. 6. Path loss exponents as a function of RP and logical AP

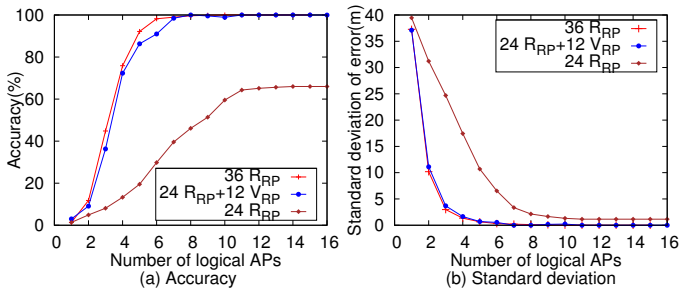


Fig. 7. Comparisons of location accuracy and standard deviation of error among original method (36 R_{RP} or 24 R_{RP}) and proposed method (24R_{RP}+12V_{RP})

V. Conclusions

To realize the fingerprint-based indoor location estimation, the offline radio map construction is definitely essential for the performance. Generally, this process requires a lot of time to gather training data from a considerable amount of reference points. In this paper, we proposed a virtual fingerprint construction scheme to shorten the offline process by means of reducing the number of real reference points in radio map. More precisely, the vanished reference points are automatically recovered by applying the signal propagation model. Experimental results show that, the smart antenna system with the proposed scheme can easily generate the valid virtual RPs which have very similar characteristic of radio signals as the measured ones. Experimental results also reveal that the proposed scheme is workable in realistic indoor environment and the virtual fingerprint construction scheme can reduce 33% $(12/36)$ real RPs during the offline phase.

The concept behind the virtual fingerprint construction scheme is the interpolation method. Our future work is to use the extrapolation method on the virtual fingerprint construction scheme to build unlimited indoor radio map with very small amount of real reference points.

Acknowledgments

This research was supported by the Realtek Semiconductor Corporation in Taiwan. We also appreciate the National Science Council of Taiwan for funding this work.

References

[1] K. Pahlavan, X. Li, and J. Makela, "Indoor Geolocation Science and Technology," *IEEE Communications Magazine*, vol. 40, no. 2, pp. 112–118, 2002.

[2] Y. Zhao, "Mobile Phone Location Determination and its Impact on Intelligent Transportation Systems," *IEEE Transaction on Intelligent Transportation Systems*, vol. 1, no. 1, pp. 55–64, 2000.

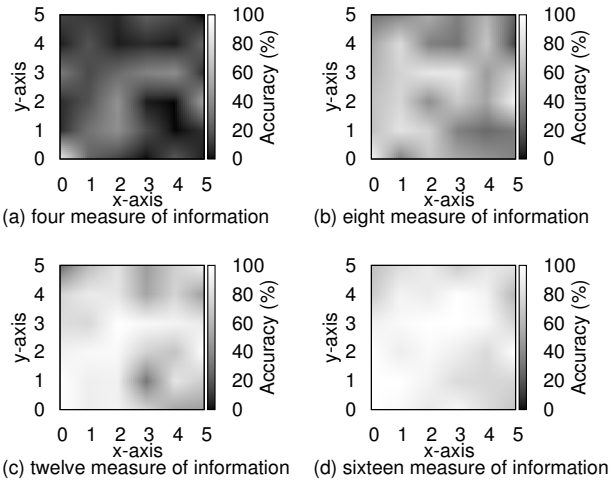


Fig. 8. The accuracy estimation under different number of antenna sets.

[3] G. Sun, J. Chen, W. Guo, and K. Liu., "Signal Processing Techniques in Network-aided Positioning: a Survey of State-of-the-art Positioning Designs," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 12–23, 2005.

[4] A. Sayed, A. Tarighat, and N. Khajehnouri, "Network-based Wireless Location: Challenges Faced in Developing Techniques for Accurate Wireless Location Information," *IEEE Signal Processing Magazine*, vol. 22, no. 4, pp. 24–40, 2005.

[5] T.-N. Lin and P.-C. Lin, "Performance Comparison of Indoor Positioning Techniques Based on Location Fingerprinting in Wireless Networks," *Proceedings of International Conference on Wireless Networks, Communications and Mobile Computing*, pp. 1569–1574, Jun. 2005.

[6] P. Bahl and V. Padmanabhan, "RADAR: An In-building RF-based User Location and Tracking System," *Proceedings of the 19th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp. 775–784, Apr. 2000.

[7] P. Prasithsangaree, P. Krishnamurthy, and P. K. Chrysanthis, "On Indoor Position Location with Wireless LANs," *Proceedings of the 13th IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications*, pp. 720–724, Sept. 2002.

[8] M. Brunato and R. Battiti, "Statistical Learning Theory for Location Fingerprinting in Wireless LANs," *The International Journal of Computer and Telecommunications Networking*, vol. 47, no. 6, pp. 825–845, Apr. 2005.

[9] T. Roos, P. Myllymki, H. Tirri, P. Misikangas, and J. Sievanen, "A Probabilistic Approach to WLAN User Location Estimation," *International Journal of Wireless Information Networks*, vol. 9, no. 3, pp. 155–164, 2002.

[10] S.-H. Fang and T.-N. Lin, "Accurate Indoor Location Estimation by Incorporating the Importance of Access Points in Wireless Local Area Networks," *Proceedings of 2010 IEEE Global Telecommunications Conference (GLOBECOM 2010)*, pp. 1–5, Dec. 2010.

[11] M. Youssef, A. A., and A. Udaya Shankar, "WLAN Location Determination via Clustering and Probability Distributions," *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications*, pp. 143–150, Mar. 2003.

[12] Y. Chen, Q. Yang, J. Yin, and X. Chai, "Power-efficient Access-point Selection for Indoor Location Estimation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 7, pp. 877–888, 2006.

[13] Y.-M. H. Shiann-Tsong Sheu, Ming-Tse Kao and Y.-C. Cheng, "Indoor Location Estimation Using Smart Antenna System," *Proceedings of the IEEE Vehicular Technology Conference (VTC Fall 2013)*, pp. 1–5, Sep. 2013.

[14] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*. second ed. Chapman and Hall, 2004.

[15] D. Madigan, E. Elnahrawy, R. P. Martin, W.-H. Ju, P. Krishnan, and A. Krishnakumar, "Bayesian Indoor Positioning Systems," *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, pp. 1217–1227, Mar. 2005.

[16] A. Kushki, K. N. Plataniotis, and A. N. Venetsanopoulos, "Kernel-Based Positioning in Wireless Local Area Networks," *IEEE Transactions on Mobile Computing*, vol. 6, no. 6, pp. 689–705, 2007.

 ISBN: 978-1-61208-353-7

Performance Evaluation of ZigBee Transmissions on the Grass Environment

Teles Bezerra, Saulo Silva,
Erika Silva and Marcelo Sousa

Laboratory for Cognitive Systems
and Personal Networks (LABee)
Federal Institute of Paraíba (IFPB)
Campina Grande, Brasil
Email: {teles, saulo.eleuterio,
erika.delmiro, marcelo.portela}@ieee.org

Matheus Cavalcante

Department of Electrical Engineering
Institute for Advanced Studies in Communications (Iecom)
Federal University of Campina Grande (UFCG)
Campina Grande, Brasil
Email: matheus.cavalcante@ee.ufcg.edu.br

***Abstract*—This article proposes a prototype to measure the Received Signal Strength Indicator (RSSI) value in devices of ZigBee-based networks. The developed device was employed in two outdoor experiments, to verify how far transmitter and receiver can be separated to still maintain a connection. The proposed system utilizes a XBee module Series 1, an Arduino Uno R3 microcontroller board, a XBee Arduino Shield and a LCD. The measurements were accomplished on and over a grass environment. The RSSI meter demonstrated efficiency for good quality connections, but at certain environmental conditions, the connection was lost at a short distance of $9\,\mathrm{m}$.**

Keywords–Arduino; RSSI; Prototype; ZigBee.

I. INTRODUCTION

Wireless Sensor Networks (WSN) became a trend in the last years due to advances in wireless communications, such as new information technologies and electronic attributes developed for these technologies [1]. WSNs are one of the most promising technologies from this generation, as they have great usability due to their implementation on industrial control systems. Moreover, the low cost multi-functional sensors that accomplish surveillance control reinforce the strong importance of these networks.

With a great versatility for being used in several application fields, WSNs have joined an increasingly interest in the last years [2]. Particularly, extensive researches have induced to the definition of a new wireless systems generation, capable of extending even more the WSN application fields. It is relevant to characterize how the radio signal range varies over indoor and outdoor environments, because some ambient conditions can cause impairments to any transmissions [3].

The ZigBee technology promotes communication between devices and manages big WSN size. This standard provides a license-free and low-power, two-way wireless communications with high reliability and more extensive reliable range at an affordable cost. It is deployed in wireless control and monitoring applications with low data rate, low power consumption, allows longer life with smaller batteries [4].

In wireless networks, an important project aspect is to consider the fading effect, as shown in several articles, such as [5]–[7]. There are various electromagnetic wave's fading processes, for example fading caused by signal reflections on objects, and they all quite affect the transmission between nodes. The waves travel through different ways, that not necessarily have the same length, and interactions between them and the objects and barriers during their travel are responsible for great part of the fading phenomenon on transmission and reception processes. Fading during electromagnetic waves propagation is also caused by reflection, diffraction and scattering [7].

RSSI is a measurement of the power in a received radio signal. Several works have been proposed to investigate radio signals' propagation effects at ZigBee devices. Ben Hamida and Chelius [8] have studied the effects of human movement on the RSSI, in an indoor environment. The sensors were deployed in different floors of a building, and the researchers analyzed their data by observing the impact from human presence, and showed that it causes a degrading effect on the system's performance. In [9], the impacts from antenna orientation in WSNs were empirically determined, by tilting TelosB modules in several directions. When these modules are in contrary orientations, there are great variations on the RSSI. Ben Graham et al. [10], they monitor the effects by installing the sensors on the ceiling, where the antenna stays inverted and pointing to the ground. All these works focused in different factors that have influence on the RSSI, such as antenna orientation and human presence.

In this paper, we propose the development of a RSSI meter, which was utilized to collect data on sport fields, on the grass environment. The data collected during tests were used as a case study, and the relevant factors on several RSSI measurements were identified. Environment effects related to the considered physical scenario were taken into account, as such as wind, temperature and humidity, which were determining factors on the RSSI variation.

The remaining of the paper is organized as follows: in Section II, the ZigBee technology and the XBee modules characteristics are presented. Section III describes the hardware used during the RSSI meter assembly, and the concepts about the RSSI meter prototype. Section IV brings the methodology used for the experiments and Section V presents the measurement results. At last, the conclusions are devoted to Section VI.

 ISBN: 978-1-61208-353-7

II. ZigBee Technology

ZigBee technology is an option to fill a gap in WSN's network architecture, being an appropriate communication protocol for this application. The differential of this technology is its advantage in face of other communications protocols, such as Wi-Fi [11] and Bluetooth [12]. ZigBee technology has a protocol that supports mesh, star and tree networks, creating more than one path possible between a transmitter and a receiver.

This technology has been gaining great notoriety due to simplified code and protocol, and also its reduced development cost. The modularization allowed by ZigBee during development also attracts attention to this technology. The IEEE (Institute of Electrical and Electronics Engineers) has regularized its functioning in the IEEE 802.15.4 protocol [13]. However, Zigbee is not part of IEEE 802.15.4 standard, being Zigbee and IEEE 802.15.4 related but different things. IEEE 802.15.4 is a IEEE communication standard that specifies the medium access control (MAC) sublayer and physical (PHY) layer for low rate and low power wireless communication devices.

XBee is the brand name from Digi International for a family of ZigBee-compatible radio modules, that take part during hardware implementation necessary on a ZigBee network's assembly. The device controls radio wave's propagation by the transmission/reception antenna. This antenna, usually, works at a maximum power of $60\,\mathrm{mW}$, and its frequencies are usually between $2.40\,\mathrm{GHz}$ and $2.48\,\mathrm{GHz}$.

III. Hardware Description

A. Arduino

Arduino is a small microcontroller board, connected to a PC through an USB connection, allowing so a connection between the board and the PC. Moreover, an Arduino board contains several other terminals that allow connection with external devices such as motors, relays, light sensors, LEDs, speakers and so on. This board's project is open-source, that means that any user could construct Arduino-compatible boards. A board's cost reduction was accomplished by the opening of the Arduino's source code.

The Arduino's development platform was the base-device during the assembly of the prototypes. The model used was an Arduino Uno R3, chosen due to its great application field and its easy integration with the XBee module's platform, and shown in Figure 1.

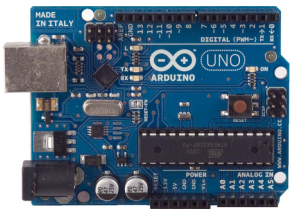


Figure 1. *Arduino Uno R3* microcontroller board

B. XBee Shield for Arduino and XBee Module

The basic Arduino boards can be complemented some boards, called *shields*, that can be coupled over the basic Arduino board. These shields are circuit boards containing other devices (GPS receivers, LCD displays, Ethernet modules, etc.) that are connected to the Arduino in order to obtain additional functionalities. Thus, a shield is a Printed Circuit Board (PCB) coupled over an Arduino board allowing communication between these boards, through an connection fed by connector-pins.

In order to correctly couple the XBee modules to the Arduino board, it was also used a shield whose function is to convert the socket's format from the Arduino board to the one from the *XBee S1 MaxStream* module. This shield is shown in Figure 2.

Figure 2. *XBee MaxStream* Module

All this connection procedure is made aiming the surveillance at the received signal's state during transmissions between the two modules. Their communication is very simply configured, as the signal forwarded to the transmitter device is directly sent to the receiver device. Such a configuration is possible because the XBee modules can directly communicate, without any need for addressing, while at AT (transparent) mode.

C. RSSI

RSSI is a measure from the received radio signal's power. Such a metric used to estimate the transmission quality between two nodes as the distance between them is varied. It works by using the distance between transmitter and receiver to designate the quality from the received signal, even considering variations on signal strength by comparing the received signal level with probability distributions and localization measures based on statistic analysis [14].

IV. Methodology

The methodology adopted during the assembly of the prototype and subsequent measurements with it were made under the following schedule.

- 1st: The used device allowed the RSSI measurement on a specific pin. On this pin, there was a PWM (Pulse Width Modulation) modulated signal, where the RSSI value is codified in how long the pin's output stands on a certain digital level. This output was treated as analogic, but actually is a digital output that generates an alternating signal (*low* and *high* digital levels).
- 2nd: Compatibility tests were made between the used Arduino Uno R3 platform and the XBee modules, for the purpose of test basic trigger circuits with the modules, and verify whether there was communication establishment between the devices. On this stage, the prototypes were assembled in assembly boards. Other electronic components

 ISBN: 978-1-61208-353-7

were also added to project, such as a 16×2 LCD display for showing RSSI values, and components responsible for maintain and feed this display.

3rd: The programming language used to create the source code executed by the prototype is called *Wiring*, that stands as the standard development language for Arduino projects. The RSSI value was obtained from the $A1$ pin, through `pulseIn()` function. This function measures the length of the PWM pulse. The code executed by the microcontroller responsible for the reading function follows.

```
int dur = pulseIn(A1,LOW,200);
int rssi=(dur+50)*(-1);
```

4th: The measurements were made on an outdoor sports field. The transmitter was fixed and the receiver was taken to increasingly distances to the transmitter. Two tests were made: on the first, both devices were on ground level. A photo was taken showing how the receiving device was put during this test, and this photo is shown in Figure 3. At first, distancing them by one meter, the RSSI was measured by the receiving device. The distance was increased up to one in which there weren't connection. This test was made on a sunny day, in the afternoon, low wind, temperatures between $28\,^{\circ}\mathrm{C}$ and $31\,^{\circ}\mathrm{C}$, and air humidity at 65%. The second test was made similarly to the first, but the devices were $45\,\mathrm{cm}$ over the ground.

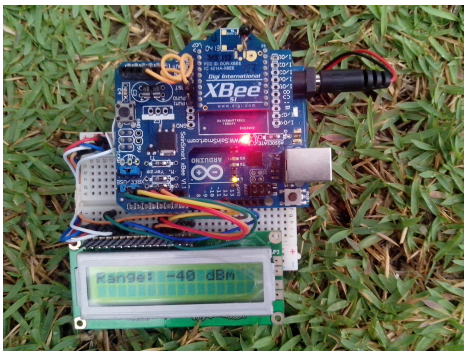


Figure 3. Prototype standing on the ground, during first measurement.

5th: On the tests, every measure took 45 RSSI samples. As the tests ended, the samples were processed. For every measure point in every test, it was calculated the average, the standard deviation of the mean (through formula shown in (1)), and the highest and lowest received signal strength. Equation (1) shows how the standard deviation of the mean was calculated.

$$\sigma_{\text{mean}} = \sqrt{\frac{1}{N \cdot (N-1)} \sum_{i=1}^{N} (x_i - \overline{x})^2}, \qquad (1)$$

in which N is the number of samples, x_i is the value of the i-th sample, and $\overline{x}$ is the arithmetical mean.

V. Performance Evaluation

Figure 4 shows the measurement's data for the experiment in which the devices were on ground level.

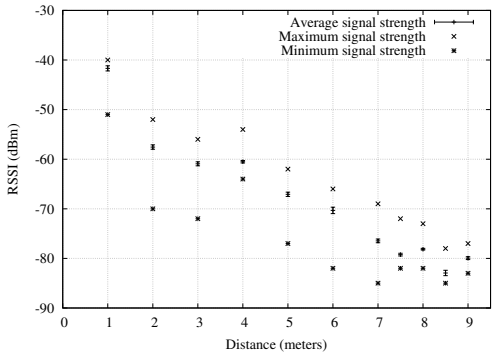


Figure 4. Received signal's power versus distance between transmitter and receiver at first measurement.

As shown in Figure 4, as the receiver device was taken farther, lower RSSI values were measured, reaching minimum levels of -85 dBm. When the receiver was at a distance of $9\,\mathrm{m}$ far, there was not anymore connection between the devices. Before that, the last average signal strength (measured at $8.5\,\mathrm{m}$) was -75.95 dBm. The RSSI values did not fluctuated too much around the mean, as shown by the small standard deviations.

A similar measurement was made, but this time the devices were $45\,\mathrm{cm}$ over the ground. The results can be seen in Figure 5.

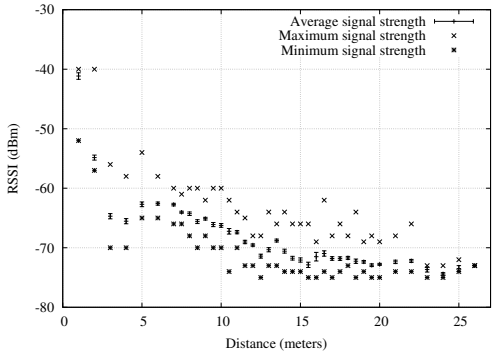


Figure 5. Received signal's strength versus distance between transmitter and receiver at second measurement.

As the environmental parameters on both measurements were almost the same, the difference between how far the transmission went can be explained by the device's height. In the second experiment, the connection was broken when the devices were distanced by $26\,\mathrm{m}$. Overall, the RSSI during the second experiment was higher than the one of the first, in which the lowest strength was about $-76\,\mathrm{dBm}$. On the second experiment, just before connection loss, the last average signal strength was $-73.00\,\mathrm{dBm}$.

Apparently, in such environmental conditions there is no useful connection between the devices if the RSSI is under $-70\,\mathrm{dBm}$. Based on this statement, it was analyzed how long

 ISBN: 978-1-61208-353-7

the signal strength stood under this level, from now on called of percentage of idle time. It is shown in Figure 6 the analysis' result for the first experiment. From $8\,\mathrm{m}$ and beyond it's difficult to hold an connection between the two devices, as the signal has a very low RSSI on such a distance.

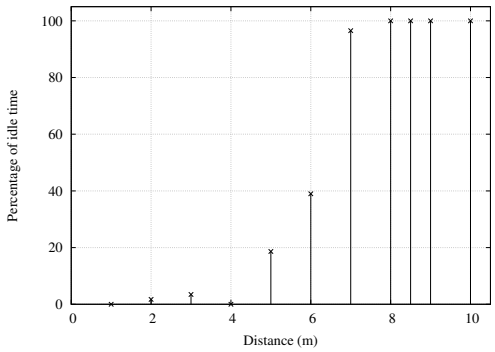


Figure 6. Percentage of idle time versus distance in the first measurement.

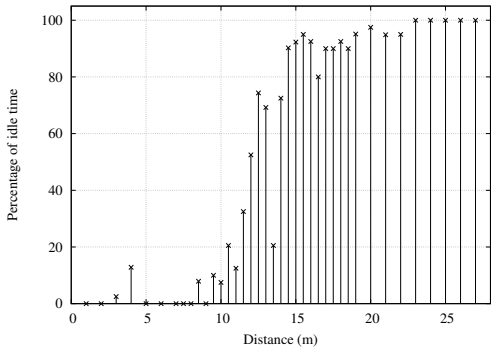


Figure 7. Percentage of idle time versus distance in the second measurement.

The results for the second measurement can be found in Figure 7. Despite the fact the connections is maintained up to a distance between transmitter and receiver of $26\,\mathrm{m}$, distances greater than $15\,\mathrm{m}$ would have a bad quality of service, as the connection would have to be constantly reestablished.

VI. Conclusion and Future Work

We successfully built a functional prototype and analyzed some data extracted from the experiments. Reading the RSSI value seems to be a positive indicator for good quality connections, but at certain environmental conditions (devices at ground level) the connection was lost at a small distance of $9\,\mathrm{m}$.

One important result achieved through this article was the determination of how far can two sensors be and still maintain connection. At ground level, this distance is about $7\,\mathrm{m}$, and to a $45\,\mathrm{cm}$ height, this distance goes up to $15\,\mathrm{m}$, showing how great is the influence of the environmental conditions.

We aim at improving the measurements quality, by changing other environmental parameters. Experiments in cloudy or rainy days, with or without human presence, with more samples, and at different times shall be done in order to allow a more complete efficiency analysis of the experiments data.

Acknowledgments

The authors would like to thank Federal Institute of Paraíba (IFPB), Laboratory for Cognitive Systems and Personal Networks (LABee), Federal University of Campina Grande (UFCG), Institute for Advanced Studies in Communications (IECOM) and IFPB's IEEE Student Branch, they all in Campina Grande, and also the National Council for Scientific and Technological Development (CNPq).

References

[1] F. Yahaya, Y. Yussoff, R. Rahman, and N. Abidin, "Performance analysis of wireless sensor network," in 5th International Colloquium on Signal Processing Its Applications., March 2009, pp. 400–405.

[2] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," IEEE Communications Magazine, vol. 40, no. 8, Aug 2002, pp. 102–114.

[3] R. Pellegrini, S. Persia, D. Volponi, and G. Marcone, "RF propagation analysis for ZigBee sensor network using RSSI measurements," in 2nd International Conference on Wireless Communication, Vehicular Technology, Information Theory and Aerospace Electronic Systems Technology (Wireless VITAE), Feb 2011, pp. 1–5.

[4] J. de S. B. Ramos, Instrumentação eletrônica sem fio: transmitindo dados com módulos ZigBee e PIC16F877A, 2012.

[5] N. Lo, D. Falconer, and A. U. H. Sheikh, "Adaptive equalization for a multipath fading environment with interference and noise," in IEEE 44th Vehicular Technology Conference, Jun 1994, pp. 252–256.

[6] J.-L. Chu and J.-F. Kiang, "Multipath effects on beacon performances," in IEEE International Conference on Networking, Sensing and Control, vol. 1, March 2004, pp. 635–638.

[7] R.-H. Wu, Y.-H. Lee, H.-W. Tseng, Y.-G. Jan, and M.-H. Chuang, "Study of characteristics of RSSI signal," in IEEE International Conference on Industrial Technology, April 2008, pp. 1–3.

[8] E. Ben Hamida and G. Chelius, "Investigating the impact of human activity on the performance of wireless networks 2014; an experimental approach," in IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WoWMoM), June 2010, pp. 1–8.

[9] M. Wadhwa, M. Song, V. Rali, and S. Shetty, "The impact of antenna orientation on wireless sensor network performance," in IEEE 2nd International Conference on Computer Science and Information Technology (ICCSIT), Aug 2009, pp. 143–147.

[10] B. Graham et al., "Analysis of the effect of human presence on a wireless sensor network," in International Journal of Ambient Computing and Intelligence (IJACI), vol. 3, 2011, pp. 1–13.

[11] "Unapproved draft standard for information technology– telecommunications and information exchange between systems– local and metropolitan area networks– specific requirements part 15.1reva: Wireless medium access control (MAC) and physical layer (PHY) specifications for wireless personal area networks (WPANs) (revision of IEEE 802.15-2002)," IEEE Std P802.15.1REVa/D5, 2004.

[12] "IEEE standard for information technology - telecommunications and information exchange between systems - local and metropolitan area networks specific requirements part 15.4: Wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs)," IEEE Std 802.15.4-2003, 2003, pp. 0_1–670.

[13] M. Saleiro and E. Ey, Zigbee: Uma Abordagem Prática, 2009. [Online]. Available: http://lusorobotica.com/ficheiros/Introducao_ao_Zigbee_-_por_msaleiro.pdf

[14] C. Park, D. Park, J. Park, Y. Lee, and Y. An, "Localization algorithm design and implementation to utilization RSSI and AOA of Zigbee," in 5th International Conference on Future Information Technology (FutureTech), May 2010, pp. 1–4.

 ISBN: 978-1-61208-353-7