UBICOMM 2015

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

ISBN: 978-1-61208-418-3

July 19 - 24, 2015

Nice, France

UBICOMM 2015 Editors

Carlos Becker Westphall, Federal University of Santa Catarina, Brazil

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain

Przemyslaw Pochec, University of New Brunswick, Canada

UBICOMM 2015

Forward

The Ninth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2015), held between July 19-24, 2015 in Nice, France, 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 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 2015 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.

The conference had the following tracks:

- Information ubiquity
- Ubiquitous multimedia systems and processing
- Ubiquitous mobile services and protocols
- Ubiquitous software and security
- Users, applications, and business models
- Ubiquitous networks
- Fundamentals

Similar to previous editions, this event attracted excellent contributions and active participation from all over the world. We were very pleased to receive top quality contributions.

We take here the opportunity to warmly thank all the members of the UBICOMM 2015 technical program committee, as well as the numerous reviewers. The creation of such a 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 effort to contribute to UBICOMM 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the UBICOMM 2015 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that UBICOMM 2015 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of Mobile Ubiquitous Computing, Systems, Services and Technologies. We also hope that Nice, France, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

UBICOMM 2015 Chairs

UBICOMM Advisory Chairs

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, Technical University of 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
Inas Khayal, Masdar Institute of Science and Technology - Abu Dhabi, United Arab Emirates
Cornel Klein, Siemens AG/Corporate Research and Technologies - Münich, Germany
Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA
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

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

UBICOMM 2015

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, Technical University of 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
Inas Khayal, Masdar Institute of Science and Technology - Abu Dhabi, United Arab Emirates
Cornel Klein, Siemens AG/Corporate Research and Technologies - Münich, Germany
Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA
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

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

UBICOMM 2015 Technical Program 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, Vienna University of Technology, Austria
Michel Banâtre, IRISA - Rennes, France
Felipe Becker Nunes, Federal University of Rio Grande do Sul (UFRGS), Brazil
Simon Bergweiler, German Research Center for Artificial Intelligence (DFKI), Germany
Aurelio Bermúdez Marin, Universidad de Castilla-La Mancha, Spain
Carlo Alberto Boano, Graz University of Technology, Austria
Bruno Bogaz Zarpelão, State University of Londrina (UEL), Brazil
Jihen Bokri, ENSI (National School of Computer Science), Tunisia
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
Davut Cavdar, Middle East Technical University, Turkey
José Cecílio, University of Coimbra, Portugal
Bongsug (Kevin) Chae, Kansas State University, USA
Konstantinos Chatzikokolakis, National and Kapodistrian University of Athens, Greece
Harsha Chenji, Ohio University, USA
Sung-Bae Cho, Yonsei University - Seoul, Korea
Mhammed Chraibi, Al Akhawayn University - Ifrane, Morocco
MyoungBeom Chung, Sungkyul University, Korea
Michael Collins, Dublin Institute of Technology, Dublin, Ireland
Andre Constantino da Silva, IFSP, Brazil
Kyller Costa Gorgônio, Universidade Federal de Campina Grande, Brazil
Stefano Cresci, IIT-CNR, Italy
Pablo Curiel, DeustoTech - Deusto Institute of Technology, Spain

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
Vana Kalogeraki, Athens University of Economics and Business, Greece
Faouzi Kamoun, Zayed University, UAE
Fazal Wahab Karam, Gandhara Institute of Science and Technology, Pakistan
Nobuo Kawaguchi, Nagoya University, Japan
Subayal Khan, VTT, Finland
Inas Khayal, Masdar Institute of Science and Technology - Abu Dhabi, United Arab Emirates
Brian (Byung-Gyu) Kim, SunMoon University, South Korea
Soo-Kyun Kim, Samsung Electronics, South Korea
Sung-Ki Kim, Sun Moon University, South Korea
Manuele Kirsch Pinheiro, Université Paris 1 Panthéon Sorbonne, France
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
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, Conversant Media, 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
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
Gustavo López Herrera, Research Center on Information and Communication Technologies (CITIC) - Universidad de Costa Rica, Costa Rica
Jeferson Luis Rodrigues Souza, University of Lisbon, Portugal

Paul Lukowicz, German Research Center for Artificial Intelligence (DFKI), Germany
Lau Sian Lun, Sunway University, Malaysia
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
Carlo Mastroianni, ICAR-CNR - Rende, Italy
Roseclea Duarte Medina, Universidade Federal De Santa Maria (UFSM), Brazil
Natarajan Meghanathan, Jackson State University, U.S.A.
Nemanja Memarovic, University of Zurich, Switzerland
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
Dugki Min, Konkuk University, South Korea
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
Tamer Nadeem, Old Dominion University, USA
Tatsuo Nakajima, Waseda University, Japan
Wolfgang Narzt, Johannes Kepler University - Linz, Austria
Rui Neves Madeira, New University of Lisbon, Portugal
David T. Nguyen, Facebook / College of William and Mary, USA
Giang Nguyen, TU Dresden, Germany
Quang Nhat Nguyen, Hanoi University of Science and Technology, Vietnam
Ryo Nishide, Ritsumeikan University, Japan
Gregory O'Hare, University College Dublin (UCD), Ireland
Kouzou Ohara, Aoyama Gakuin University, Japan
Akihiko Ohsuga, The University of Electro-Communications (UEC) - Tokyo, Japan
Satoru Ohta, Toyama Prefectural University, Japan
George Oikonomou, University of Bristol, UK
Carlos Enrique Palau Salvador, University Polytechnic of Valencia, Spain
Agis Papantoniou, National Technical University of Athens (NTUA), Greece
Kwangjin Park, Wonkwang University, South Korea
Ignazio Passero, Università degli Studi di Salerno - Fisciano, Italy
Serena Pastore, INAF- Astronomical Observatory of Padova, Italy
Misha Pavel, Northeastern University, USA
Jyrki T.J. Penttinen, Finesstel Ltd, Finland
Jorge Pereira, European Comission, Belgium
Nuno Pereira, CISTER/INESC TEC - ISEP, Portugal

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

Sensing by Proxy: Occupancy Detection Based on Indoor CO_2 Concentration

Ming Jin, Nikolaos Bekiaris-Liberis, Kevin Weekly, Costas Spanos, Alexandre Bayen

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, California 94720, USA
Emails: {jinming,bekiaris-liberis,kweekly,spanos,bayen}@berkeley.edu

***Abstract*—Sensing by proxy, as described in this study, is a sensing paradigm which infers latent factors by "proxy" measurements based on constitutive models that exploit the spatial and physical features in the system. In this study, we demonstrate the efficiency of sensing by proxy for occupancy detection based on indoor CO_2 concentration. We propose a link model that relates the proxy measurements with unknown human emission rates based on a data-driven model which consists of a coupled Partial Differential Equation (PDE) – Ordinary Differential Equation (ODE) system. We report on several experimental results using both a CO_2 pump that emulates human breathing, as well as measurements of actual occupancy by performing controlled field experiments, in order to validate our model. Parameters of the model are data-driven, which exhibit long-term stability and robustness across all the occupants experiments. The inference of the number of occupants in the room based on CO_2 measurements at the air return and air supply vents by sensing by proxy outperforms a range of machine learning algorithms, and achieves an overall mean squared error of 0.6569 (fractional person), while the best alternative by Bayes net is 1.2061 (fractional person). Building indoor occupancy is essential to facilitate heating, ventilation, and air conditioning (HVAC) control, lighting adjustment, and occupancy-aware services to achieve occupancy comfort and energy efficiency. The significance of this study is the proposal of a paradigm of sensing that results in a parsimonious and accurate occupancy inference model, which holds considerable potential for energy saving and improvement of HVAC operations. The proposed framework can be also applied to other tasks, such as indoor pollutants source identification, while requiring minimal infrastructure expenses.**

Keywords–Occupancy detection; Building energy efficiency

I. INTRODUCTION

The thorough understanding of the interaction of occupants and indoor environment has been the key component towards occupancy comforts and energy efficiency of buildings, which account for 40% of total energy usage in the U.S. [1]. Intelligent buildings are conscious about both its occupancy and environment, in order to take controls over its physical systems, such as HVAC and lighting, to optimize user comforts and energy consumption. The knowledge of zone-based occupancy coupled with adaptive building services offers considerable potential for energy reduction [2]–[5].

Many existing methods resort to machine learning algorithms through dense sensor deployment. Various sensors have been employed, including passive infrared (PIR) sensors, read switches, camera [3], [5]–[7], as well as environmental sensors such as acoustics, carbonmonoxide (CO), total volatile organic compounds, small particulates (PM2.5), CO_2, illumination, temperature, and humidity [6], [8], [9]. Indoor CO_2 concentration is indicative of the occupancy, as humans are the main source of CO_2 production, although existing approaches suffer from the delay of detection as a result of the relatively long time (10–15minutes) it takes for CO_2 to build up to the corresponding level of concentration [6].

Unlike traditional sensing problems, sensing by proxy has explicit dependency on sensors relative locations as it appears in the computational physics. There are three essential components inherent in the problem, the Location, Link function, and Latent factors, which we call the **L3** factors. In all sensing by proxy systems, the crucial step is to identify these three components and the relationship between them. Location refers to the dependence between sensors location and latent factors. Latent factors, as its name suggests, includes factors that are not directly observable by the sensors and have nonnegligible impact on the system. Link model, a term analogous to the link function from the generalized linear model (GLM) in statistics, refers to the transformation of sensor readings to the quantity of interests. As the design of sensing by proxy systems relies on the identification of the above **L3** factors and their interdependency, the most critical part is the link model, which links location and latent factors and determines the effectiveness of the system.

The key contributions of our work are as follow:

- We develop a link model based on constitutive partial differential equation (PDE) coupled with ordinary differential equation (ODE) that captures the spatial and temporal features of the system and links unobserved human emission to "proxy" measurements of CO_2 concentrations (Section II).
- Our most significant contribution is the design, implementation, and evaluation of occupancy detection algorithm (**Algorithm** 1, Section II) based on the sensing by proxy methodology in controlled and field experiments (Section III). Our method achieves a root mean-squared error (in fractional person) of 0.6311, as compared to 1.2061 by the best alternative strategy (Section IV).

The rest of the paper is organized as follow. The link model is detailed in Section II, which includes the proxy design, modeling, as well as inference. Section III describes the design of CO_2 pump and occupants experiments, whose results are reported in Section IV. Related work is summarized in Section V. Section VI draws conclusion and discusses future works.

 ISBN: 978-1-61208-418-3

II. SENSING BY PROXY: LINK MODEL

The focus of this section is to introduce the link model in our sensing by proxy framework, which relates the proxy to latent factors and enables the estimation of the latent factors (Section IV).

We start with a description of the PDE-ODE model.

A. Proxy Design and Modeling

We model the dynamics of the CO_2 concentration in the room using a convection PDE with a source term which models the effect of the CO_2 that is generated by humans. The source term, $X(t)$, measured in ppm (part per million), is the output of a linear, time-invariant, scalar, stable ODE system whose input, $V(t)$, in ppm/s, represents the unknown humans' emission rate of CO_2 inside the room (within the vicinity of humans), given by

$$\dot{X}(t) = -aX(t) + V(t) \tag{1}$$

where we assume that the unmeasured CO_2 emission rate, $V(t)$, from the humans has the form of a piece-wise constant signal,

$$\dot{V}(t) = 0 \tag{2}$$

which is based on our experimental observation that the response of the CO_2 concentration in the room due to changes of the human's CO_2 input has some similarities with the step response of a low-pass filter. The measure of how fast changes to the CO_2 emission rate by the humans affect the CO_2 concentration in the room is specified by the time constant, $\frac{1}{a}$, in units of 100s.

The ODE is coupled with a PDE that models the evolution of the CO_2 concentration in the room given by

$$u_t(x,t) = -bu_x(x,t) + b_X X(t) \tag{3}$$

$$u(0,t) = U(t) \tag{4}$$

where $u(x,t)$ denotes the concentration of CO_2 in the room in ppm at a time $t \geq 0$ and for $0 \leq x \leq 1$, the steady state input CO_2 concentration of the fresh incoming air in ppm is U_e, and the measured concentration of the fresh incoming air at the air supply vent is the input U in ppm. Positive parameter, b, in $\frac{1}{100s}$, represents the speed of air convection in the room. The rate of dispersion of CO_2 from the local vicinity of the human to the room is measured by b_X, in $\frac{1}{100s}$, which is a positive number. We scale and center the dimension along the supply-return path so that the air supply is located at $x = 0$ and the air return is at $x = 1$; therefore, the spatial variable x is unitless and represents a normalized distance along the path. The CO_2 concentration inside the room at the location of the air supply is represented by $u(0,t)$, and the CO_2 concentration inside the room at the location of the air return is given by $u(1,t)$.

The evolution of the CO_2 concentration in the room is thus modeled as a linear system, one of whose inputs is the CO_2 concentration of the fresh incoming air measured at the location of the air supply, and the other input is the human emission, if any. The output of the system can be viewed as the CO_2 concentration of the air at the return vent, which is mixed with CO_2 that convects from the air supply towards the air return and the CO_2 that is produced from humans. The concentration of CO_2 at the ceiling in a (non-ratiometric) normalized distance along an axis from the supply to the return vent is indicated by the value of the PDE on the corresponding interior point of its spatial domain.

The physical representation of our model is illustrated in Figure 1. The convection of air from air supply to the air return vent near the ceiling is represented by the PDE part. The diffusive term is intentionally omitted since it plays a relatively minor role in dispersing indoor pollutants as suggested in [10]. Another design consideration involved is the modeling of the CO_2 concentration near the ceiling since this is where we see most effect from human-generated CO_2. One explanation is that the warm breath from a human occupant acts as a "bubble" of gas that rises to the ceiling, since it is more buoyant than the ambient, cooler air. Thus, the air coming from lower in the room is modeled as a source term on the PDE across its entire path. The fact that this bubble of air does not immediately rise to the ceiling but only gradually (as observed in the response of the CO_2 concentration in the room due to changes of the human's CO_2 input shown in Figures 5 and 8, during the occupant experiments) is captured by the ODE part of the model, which behaves as a filter between the unknown CO_2 emission rate of humans and the CO_2 concentration in the room.

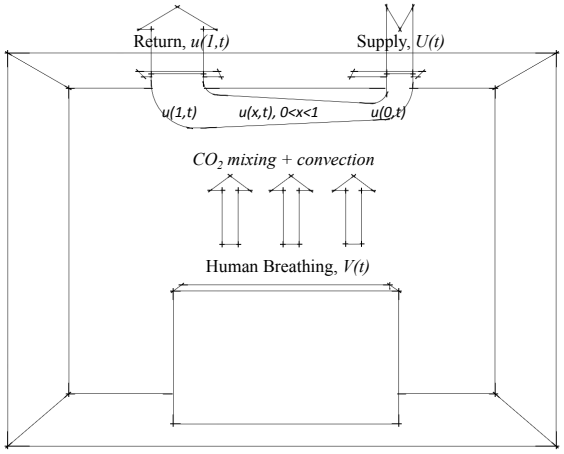


Figure 1. The physical representation of the model. Fresh air with CO_2 concentration $U(t)$ enters the room from the supply vent, and exits the room after convection and mixing with human breath, $V(t)$, which rises to the ceilings, and the measured CO_2 concentration at the return vent is $u(1,t)$.

B. Proxy Inference

The latent factors that are not directly observable are sensed by proxy based on the link model which describes the evolution of proxy under the effects of latent factors. The temporal and spatial dynamics captured by the PDE-ODE link model effectively regularizes the inference output. The approach is clearly different from discriminative models, which assume samples are independent and identically distributed (i.i.d.). It shares some similarities with dynamic Bayesian models such as particle filters (PF) and Conditional Random Field (CRF), which accounts for time evolution of the underlying phenomenon. Nevertheless, proxy inference is directly derived from physical dependency among sensors and is thus more accurate and reliable with provable behaviors as we show next.

 ISBN: 978-1-61208-418-3

The central task in this chapter is to derive an estimation strategy for latent factors, namely the human emission rate $V(t)$, based on proxy measurements at the supply vent, $U(t)$, and return vent, $u(1,t)$. The notation in the derivation follows from the previous description, with a hat to indicate estimation. We consider the following observer, which is a copy of the plant (1)-(4) plus output injection

$$\hat{u}_t(x,t) = -b\hat{u}_x(x,t) + b_X\hat{X}(t) + r(x)(u(1,t) - \hat{u}(1,t)) \quad (5)$$

$$\hat{u}(0,t) = U(t) \quad (6)$$

$$\dot{\hat{X}}(t) = -a\hat{X}(t) + \hat{V}(t) + L_1(u(1,t) - \hat{u}(1,t)) \quad (7)$$

$$\dot{\hat{V}}(t) = -L_2(u(1,t) - \hat{u}(1,t)) \quad (8)$$

The corresponding occupancy detection algorithm is shown in Algorithm 1. The observer design for our PDE-ODE model is based on the design in [11], specifically Theorem 2, which has its origins on the backstepping observer design for some classes of PDEs presented in [12]. We refer the interested reader to [11] for the proof of the following corollary.

Corollary 1: Consider the system (1)-(4) and the proxy observer (5)-(8) with

$$r(x) = L_1\pi_1(x) + L_2\pi_2(x) \quad (9)$$

$$\pi_1(x) = \frac{b_X}{a}(e^{\frac{a}{b}x} - 1) \quad (10)$$

$$\pi_2(x) = \frac{b_X}{ba}x + \frac{b_X}{a^2}(1 - e^{\frac{a}{b}x}) \quad (11)$$

Let $b_X \neq 0$ and choose L_1, L_2 such that the matrix $A - \begin{pmatrix} L_1 \\ L_2 \end{pmatrix} C$ is Hurwitz, where

$$A = \begin{pmatrix} -a & 1 \\ 0 & 0 \end{pmatrix} \quad (12)$$

$$C = \begin{pmatrix} \pi_1(1) & \pi_2(1) \end{pmatrix} \quad (13)$$

Then for any $u_0(x)$, $\hat{u}_0(x) \in L_2(0,1)$, $X(0)$, $\hat{X}(0)$, $V(0)$, $\hat{V}(0) \in \mathbb{R}$, there exists positive constant λ and κ such that the following holds for all $t \geq 0$

$$\Omega(t) \leq \kappa\Omega(0)e^{-\lambda t} \quad (14)$$

$$\Omega(t) = \int_0^1 (u(x,t) - \hat{u}(x,t))^2 dx + (X(t) - \hat{X}(t))^2 + (V(t) - \hat{V}(t))^2 \quad (15)$$

Algorithm 1 Sensing by proxy for Occupancy Detection

function SENSINGBYPROXY(X^R,X^S, $Param$)
  Inputs: X^R: measurements at air return of size $1 \times T$
  X^S: measurements at air supply of size $1 \times T$
  $Param$: hyperparameters
    1) Model specification as in Table III: convection coefficient b, source term coefficient b_X, time constant of human effect a, human emission rate V^H, equilibrium concentration in air U_e
    2) Control parameters: L_1, L_2 as in (7) and (8).
    3) Spatial resolution d_s, temporal resolution d_t
    4) Smoothing window for median filter: w
  Initialization:
  $\hat{u} \leftarrow U_e\mathbf{1}(d_s, Td_t)$ ▷ matrix of size $d_s \times Td_t$
  $\hat{X} \leftarrow \mathbf{0}(1, Td_t)$ ▷ vector $1 \times Td_t$ for emission effect
  $\hat{V} \leftarrow \mathbf{0}(1, Td_t)$ ▷ vector $1 \times Td_t$ for emission rate
  $x^R \leftarrow kron(X^R, \mathbf{1}(1, d_t))$ ▷ Discretize return/supply
  $x^S \leftarrow kron(X^S, \mathbf{1}(1, d_t))$ ▷ by kronecker product
  $\tau \leftarrow \frac{1}{d_t}$ ▷ time discretization step
  $r(n) \leftarrow L_1\frac{b_X}{a}(e^{a/b} - 1) + L_2(\frac{b_X}{ba} + \frac{b_X}{a^2}(1 - e^{a/b}))$
  Main program:
  for $t \in \{1, ..., Td_t\}$ **do**
    $\hat{u}(0,t) \leftarrow x^S(t)$ ▷ Equ.(6)
    for $n \in \{1, ..., d_s\}$ **do** ▷ PDE updates
      $\hat{u}_x(n,t) \leftarrow (\hat{u}(n,t) - \hat{u}(n-1,t))d_s$ ▷ spatial
      $\hat{u}(n,t+1) \leftarrow \hat{u}(n,t) + \tau\Big(- b\hat{u}_x(n,t) + b_X\hat{X}(t) + r(n)(x^R(t) - \hat{u}(d_s,t))\Big)$ ▷ Equ. (5) updates
    end for
    $\hat{X}(t+1) \leftarrow \hat{X}(t) + \tau\Big(-a\hat{X}(t) + \hat{V}(t) + L_1(x^R(t) - \hat{u}(d_s,t))\Big)$ ▷ updates by (7)
    $\hat{V}(t+1) \leftarrow \hat{V}(t) + \tau L_2(x^R(t) - \hat{u}(d_s,t))$
  end for
  Outputs: $y^{\text{Occupants}} \leftarrow \lfloor \frac{\text{median}(\hat{V}, w)}{V^H} + \frac{1}{2} \rfloor$ ▷ round of signal after median filter with window size w
end function

III. EXPERIMENTAL DESIGN

A. Hardwares

As our approach is not particularly demanding of the accuracy of the proxy measurements, we employ the low-cost K30 CO_2 sensor [13], shown in Figure 2, as the main module in our sensor platform. We implemented a local data storage solution with SD card, and plan to integrate a wireless transmission module in the long run to directly deposit data in our database. The sensor is capable of measuring CO_2 concentrations from 0 to 5000 ppm at a frequency of 1Hz with an accuracy of ± 30 ppm, or $\pm 3\%$ of measured value, which is considered sufficient for the purpose of occupancy detection.

Sensor calibration is performed by baseline method. We leave the sensors in a well ventilated room with outdoor supply air for few hours. The systematic offset, ξ_i, is given by

$$\xi_i = \frac{1}{T_{\text{cal}}}\sum_{t=1}^{T_{\text{cal}}} y_t - x_{\text{outdoor}} \quad (16)$$

where T_{cal} is the length of the calibration period, y_t is the sensor reading at time t, and x_{outdoor} is the outdoor CO_2 concentration, usually at 400ppm. The offset ξ_i is subtracted from sensor i under *the well mixed assumption*, which states as "at steady state, the air in the room is well mixed, with the CO_2 concentration the same as the fresh air from the air supply vent".

 ISBN: 978-1-61208-418-3

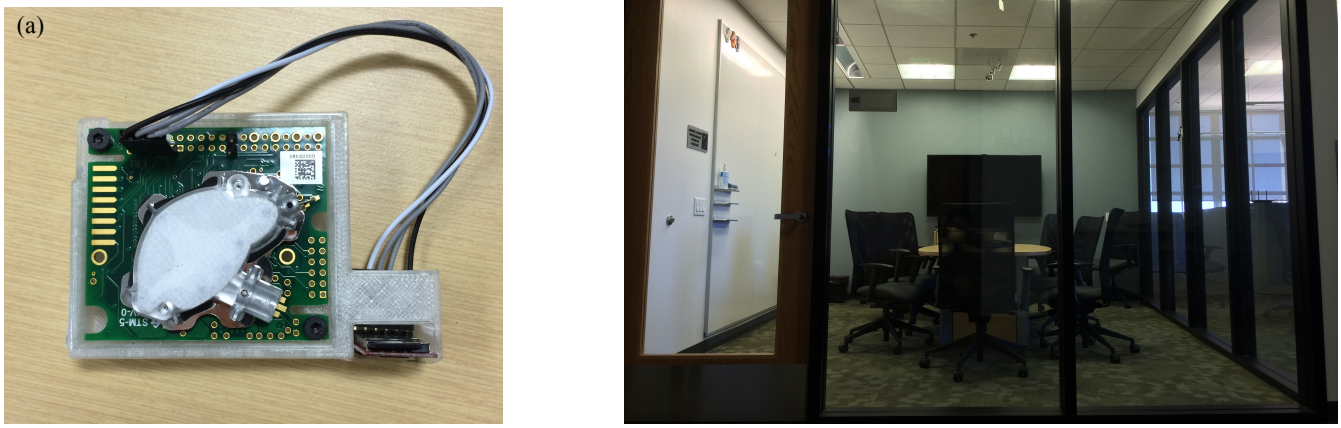


Figure 2. (a) CO_2 sensor up close. The platform is integrated with the main module to sense CO_2 concentration, and a local data storage solution with SD card. (b) The testbed is a conference room located at the Center for Research in Energy Systems Trasnformation (CREST) in Cory Hall on the UC Berkeley campus. The space is of size $14 \times 10 \times 9$ ft^3, equipped with a full ventilation system including an air return vent and air supply vent, as illustrated in Figure 1.

B. Testbed Deployment

We implemented the experiments in a typical conference room, shown in Figure 2, located at the Cory Hall on the UC Berkeley campus, whose occupancy is demand-based and not regular. The room bears close resemblance to other typical indoor spaces, with a ventilation system including air supply and air return vents on the ceiling. The sensors are placed on both vents, in addition to the blackboard on the sidewall.

C. Experiments

Two types of experiments are performed, namely CO_2 pump and occupants experiments, with different focuses.

For the CO_2 pump experiments, an outlet placed $\sim$ 20cm above the desk injects beverage-grade CO_2 through a 200W personal heater to emulate warm human breaths. The experiment is designed with two purposes. First, we want to examine the spatial dependence of CO_2 concentration in the room. Second, we can collect data to identify the parameters of the model whose output matches the measured data, under different frequency of excitation. Hence, we conducted experiments with the pump alternating between ON and OFF states, with the length of a full period of 30min (A), 1hour (B), 3hours (C), and 10hours(D), whose results are detailed in Section IV.

For the occupants experiments, the purpose is to validate sensing by proxy in a real setting. Hence we performed both controlled experiments (E) and field measurements (F,G). Our excitation procedure for the controlled experiments consists of adding or removing one of two participants of the experiment, and noting the time that the occupancy changes. The subjects are graduate students with similar physique. The door is closed during the experiment, while the participants are engaged in normal activity such as working on their computers and talking to each other. The field measurements require much less commitment from the occupants, who are using the conference room for meetings or group study. The occupancy schedules for E, F, G are demonstrated in Figures 5 and 8.

IV. Results and Discussion

In this section, we report results from experiments and simulation, and the performance of sensing by proxy in occupants experiments.

A. Experimental Results and Data Analysis

As described in the section of experimental design, we performed two groups of experiments, namely, one with CO_2 pump and the other with varying number of occupants. Based on the measurements, we make qualitative and quantitative analysis as a preparation.

1) CO_2 pump experiments:

Hypothesis: when the CO_2 is injected for a long time with constant emission rate, the system reaches steady state.

The steady-state characterization experiment is conducted, when the pump is turned on for 5 consecutive hours. Figure 3 illustrates the measurements from the supply vent, return vent, and blackboard.

The rate of CO_2 concentration starts to decrease after few hours, and reaches a plateau in the last hour. The steady state concentration settles at around 1200 ppm as a result of mixing of fresh incoming air and CO_2 release.

Hypotheses: when the CO_2 is released periodically, the measurement exhibits periodic patterns according to the PDE-ODE system. Further, besides transient behavior due to changes of ventilation rate, the CO_2 concentrations from different points in the room react the same, albeit with different magnitudes.

Both the short period and long period excitation experiments are performed, with the periods of 30 minutes (15min ON, 15 min OFF, same for the following), 1 hour, and 3 hours, as shown in Figures 4 and 7.

As can be seen the CO_2 concentrations at all the sensed locations are responsive to the periodic injection, though the measurement at the air supply vent has a smaller magnitude compared with blackboard and the air return vent. While the CO_2 accumulates from the start of the injection, the first

 ISBN: 978-1-61208-418-3

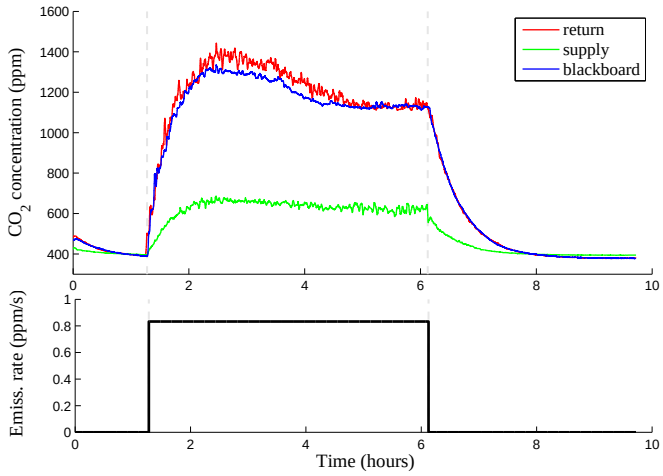


Figure 3. CO_2 pump experiment D. The measured CO_2 concentration from different locations for a 5-hour CO_2 release are shown.

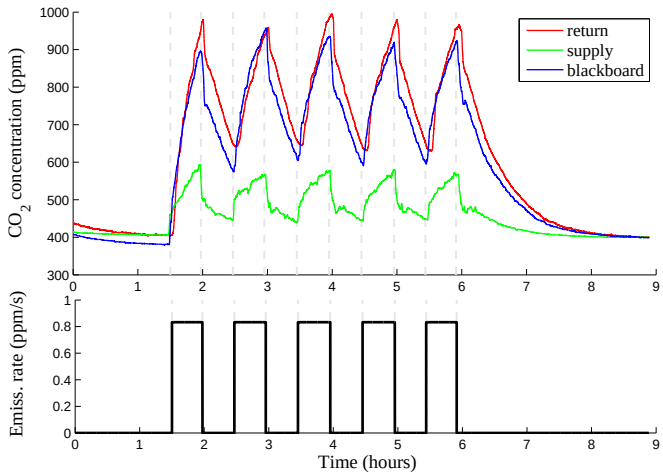


Figure 4. CO_2 pump experiment B. Short term excitation with period of 1 hour. Measurements at return (red), supply (green) vents, and blackboard (blue) are shown.

order derivative decreases as the room reaches higher CO_2 concentration.

To quantitatively evaluate the spatial dependencies of sensors in the room, we now derive the cross-correlation between measurements from three different locations for the CO_2 pump experiments. The definition of the cross-correlation $r_{y_1y_2}$ between two signals y_1, y_2, that is employed here is given by

$$r_{y_1y_2} = \frac{\sum_{k=1}^{T}(y_1(k)-\bar{y}_1)(y_2(k)-\bar{y}_2)}{\sqrt{\sum_{k=1}^{T}(y_1(k)-\bar{y}_1)^2(y_2(k)-\bar{y}_2)^2}} \qquad (17)$$

where $\bar{y}_1$ and $\bar{y}_2$ are the sample mean of y_1 and y_2 respectively. The cross-correlation is a measure of the degree of linear dependency between two signals, and hence, it is a meaningful measure for comparing the measurements from different locations inside the room. The values of the cross-correlations are shown in Table I. One can observe that the cross-correlation between return and blackboard measurements is high, whereas

TABLE I. CROSS-CORRELATION VALUE OF CO_2 MEASUREMENTS AT DIFFERENT LOCATIONS FOR EXPERIMENTS A, B, C.

Location	Cross-correlation
Return-Supply	0.9592
Return-Blackboard	0.9882
Supply-Blackboard	0.9635

the cross-correlations that involve supply measurements are lower. This implies that the signals have a high degree of linear dependency (note that when $y_1(k) = c_1y_2(k) + c_2$, for all k, the cross-correlation is one) on each other, although the correlation with the supply measurements is lower due to the ventilation operation. Note that the cross-correlation between any two locations is derived as the average cross-correlation obtained from the measurements of the experiments.

2) Occupants experiments:

As sensing by proxy aims at accurately infer occupancy through proxy measurements, in addition to CO_2 pump experiments, occupants experiments are necessary to validate our methodology.

As described in the experimental design, we perform strictly controlled and field experiments. The former implements a designed schedule of occupancy, and requires the occupants to sit in designated chairs and remain in the room during the experiment, while allowing them to be engaged in normal activities, such as using computers and chatting. The latter is taken during daily events and requires much less commitment from the occupants.

The following shows results from several such experiments, which substantially cover the usage of the conference room, and can be easily extended to other areas in the building. The field measurements are shown in Figure 5 and 8 (right), and the strictly controlled experiment is illustrated in Figure 8 (left). Note that to avoid significant overlap between graphs of this section and those of the simulation section, we arbitrarily decide which graphs show the blackboard measurement and the others show the simulated return, as long as the evidence is sufficient to for the argument.

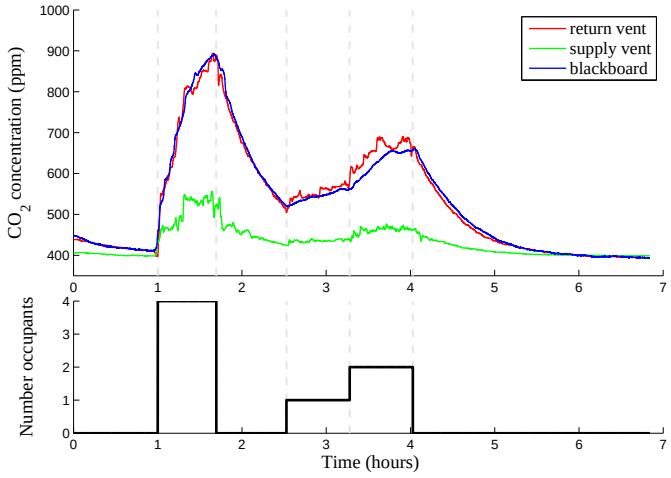


Figure 5. Occupants experiment F. Field measurements during project discussion. Top: Proxy measurements. Bottom: Corresponding occupancy.

Similar to the pump experiment, CO_2 concentration in-

creases almost immediately at the start of occupancy, and the concentration level and rate have a clear correspondence to the number of occupants in the room. The possibility of relating proxy measurements, namely CO_2 concentration, to latent factors, namely occupancy, lays the foundation for sensing by proxy.

Though the system is responsive to the change of occupancy, the time it takes to accumulate or deplete CO_2 to the corresponding stationary value is fairly long. From vacancy to a high level occupancy, the measurement slowly sweeps across several intermediate levels. The difficulty of most distribution-based classification methods is illustrated in Figure 6, where the significant overlapping of regions and misplacement of modes corresponding to different levels of occupancy will lead to confusion for standard machine learning algorithms. By modeling the temporal and spatial dynamics of the system, as we demonstrate in the subsequent sections, we can develop an inference method that is both robust to noise and responsive to change of occupancy.

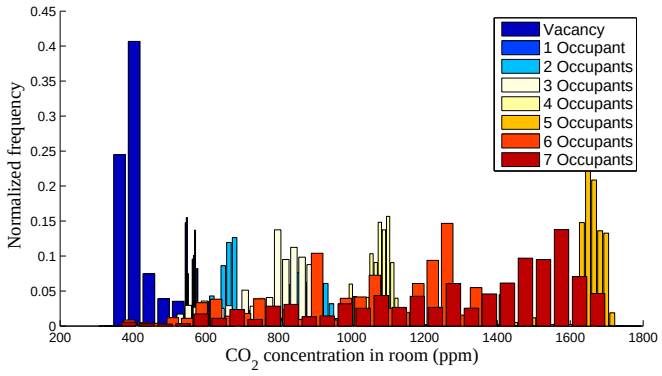


Figure 6. Empirical distribution of CO_2 concentration for all occupants experiments corresponding to different occupancy (color coded).

B. Simulation with Proxy Link Model

This section applies the model as described by (1)-(4), which links the location-specific proxy measurements to latent CO_2 emission factors to the CO_2 pump experiments and occupants measurements. In particular, we are concerned with the reproduction of the return vent measurements $u(1,t)$, i.e., the output of the system, given the supply vent measurements $U(t)$ and emission rate $V(t)$.

The results are illustrated in Figure 7, where two experiments from CO_2 pump measurements are arbitrarily shown since the results are very similar. The set of parameters for the group of CO_2 pump experiments is determined by visual evaluation of the matching of simulation to the air return measurements, which is listed in Table II. The process of parameter evaluation is actually very simple, given the derived equation for stationary distribution

$$u^{\text{stationary}} = U_{\text{stationary}} + \frac{b_X V}{ab} \tag{18}$$

according to the link model (1)-(4), where V is the fixed emission rate.

The stability of the CO_2 system can be seen in the good matching of all the air return vent measurements. There are,

TABLE II. PHYSICAL PARAMETERS OF PROXY LINK MODEL USED IN ALL THE CO_2 PUMP EXPERIMENTS (A, B, C, D)

Physical parameter	Symbol	Value
Convection coefficient ($\frac{1}{100s}$)	b	2.5
Source coefficient ($\frac{1}{100s}$)	b_X	1.00
Time constant of human effect ($100s$)	$\frac{1}{a}$	16.67
Pump emission rate (ppm/sec)	ρ_{pump}	0.833
Equilibrium concentration in air (ppm)	U_e	400

TABLE III. PHYSICAL PARAMETERS OF PROXY LINK MODEL USED IN ALL THE OCCUPANTS EXPERIMENTS (E, F, G)

Physical parameter	Symbol	Value
Convection coefficient ($\frac{1}{100s}$)	b	2.5
Source coefficient ($\frac{1}{100s}$)	b_X	1.50
Time constant of human effect ($100s$)	$\frac{1}{a}$	16.67
Human emission rate (ppm/sec)	V^H	0.183
Equilibrium concentration in air (ppm)	U_e	400

nevertheless, occasionally over- and under- matching, especially around the peak and valleys, which might be caused by the fluctuation of ventilation rates. The mismatch, even though not frequent, might introduce bias in our emission rate and occupancy estimations as we show in the next section. It is, therefore, recommended to examine the cause of the mismatch in actual building operations and periodically calibrate the model in order for sensing by proxy to make the most reliable inference. It is also possible to design an automatic calibrator for each distributed sensor system.

Based on our experiences in the CO_2 pump experiments, we designed occupants controlled and field experiments to collect occupancy ground truth and validate our link model in practice, as shown in Figure 8.

In actual building usage, especially conference rooms and common areas, the occupancy is often irregular, as exemplified by the experimental profiles. The simulation of proxy measurements, therefore, is direct estimation of the effects of the irregular change of latent factors. The closeness of simulation matching to actual proxy measurements, as can be seen, is a clear indication of the accuracy of the link model, and also ensures reliable inference of latent factors. The spatial and temporal simulation is illustrated in Figure 9.

As a general remark, our proxy link model is extremely simple and parsimonious with parameters. The set of parameters, including the convection coefficient b, the source coefficient b_X, time constant of human effects $\frac{1}{a}$, in addition to the human emission rate V and CO_2 concentration of fresh air U_e, which are standard fixed parameters, are shared among all the experiments in the same group of CO_2 pump and occupants experiments, with relatively small difference between different groups due to the extent of emulation by the pump to human breathing. This makes our model extremely easy to train and employ in practice. The additional advantage of parsimonious model relies on its stability and robustness by avoiding the potential overfitting problem. As we demonstrate next, the sensing by proxy approach substantially outperforms other popular methods and yet remains physically meaningful.

C. Proxy Inference of Occupancy

The observer model as described by (5)-(8) and Algorithm 1 are applied in this section to infer the CO_2 emission rate and occupancy based on proxy at return and supply vents.

 ISBN: 978-1-61208-418-3

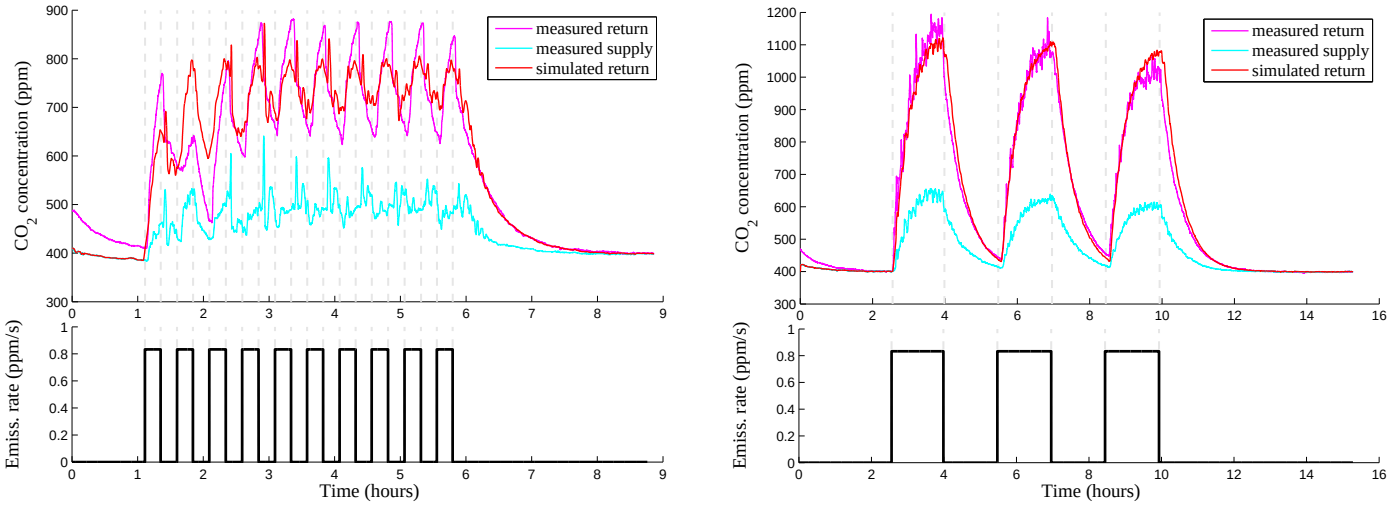


Figure 7. Proxy model simulation with CO_2 pump experiment A with 30 minutes (left) and experiment B with 1 hour (right) periodic excitation. Measurements at supply (green), return (blue), and simulated return (red) vents are shown.

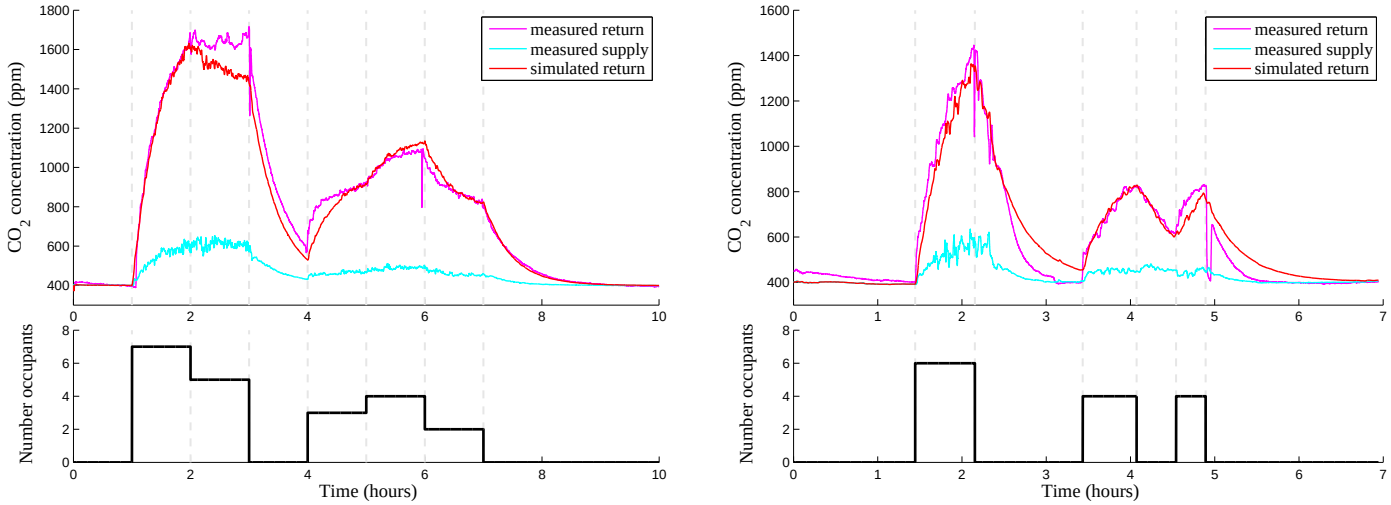


Figure 8. Proxy link model simulation with occupants experiment E (left) and G (right). The proxy measurements at return (blue), supply (green), and simulated return (red) vents are demonstrated.

Sensing by proxy distinguishes from other machine learning methods that assumes independence of samples by implicitly considering time-autocorrelation of the latent emission rate. The advantage as a result is to have smooth state trajectory after simple signal processing, where we employed median filter directly on the estimated emission rate, $\hat{V}$, with a window of 8min for Experiment A, 20min for B and C, and 25min for all the occupant experiments. The median filter is a useful denoising method in signal processing, which is often preferred to mean filter to preserve relevant details and sharp transitions in the trace, as we will demonstrate next. Figure 10 is plotted for the CO_2 pump experiment with periods of 30 minutes and 3 hours, respectively.

Contrary to the common belief that CO_2-based methods are slow in response, sensing by proxy exhibits fast response to the change of occupancy. The previous argument is based on the fact that it takes time to accumulate CO_2 to a level that can be detected, and this accumulation time is fairly long as we observed in the experiments. During the accumulation, the concentration value sweeps across the stationary values for lower occupancy when several people enter the room, or those for higher occupancy when people leave, which account for the significant overlap in the histograms of CO_2 concentration in Figure 6.

Sensing by proxy, however, tackles this issue by modeling the dynamics of the measurements based on our link model, which implicitly considers the increasing rate and stationary values to infer the actual occupancy. As a result, sensing by proxy is immediately responsive to changes of occupancy even when the transition is fairly frequent in the case of Figure 10 (left), which is not possible with other methods since the concentration remains at a relatively high level even when the pump is turned off. The parameters chosen for the estimation are $L_1 = 2$, $L_2 = 0.02$, and the other physical parameters of

 ISBN: 978-1-61208-418-3

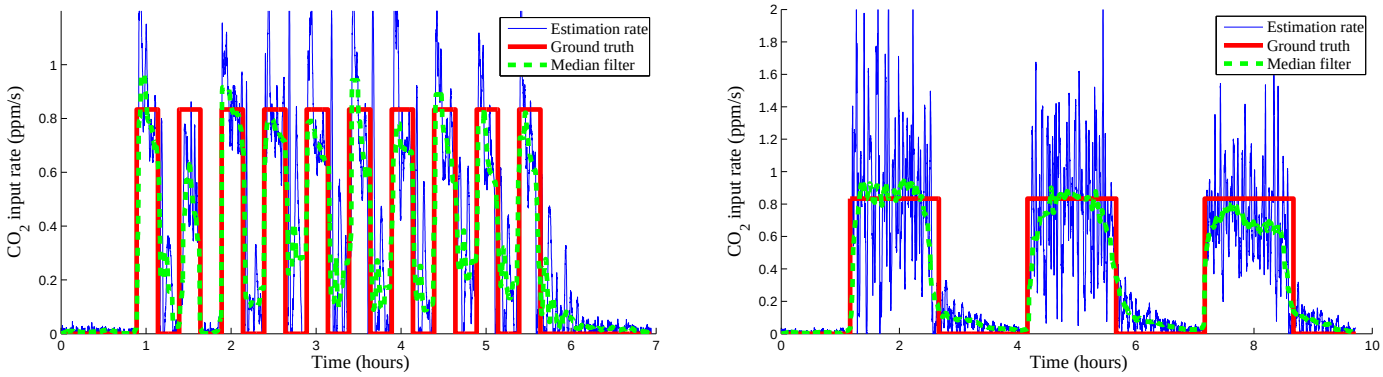


Figure 10. Sensing the latent CO_2 emission rate by proxy for CO_2 pump experiment A (left) and experiment C (right). The estimated emission rate (blue), median filtered rate (green), and ground truth (red) are given.

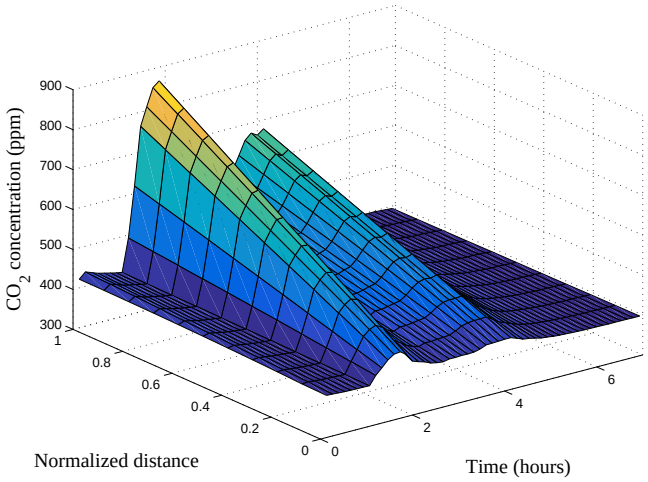


Figure 9. Spatial and temporal dynamics of CO_2 concentration as represented as the states in the proxy link model.

the model are shown in Table II.

In the case of occupants estimation, the task is more difficult due to the following reasons. First, humans are not uniform in physique, so the emission rate must vary for different occupants. Second, the positions of the people sitting in the room are arbitrary, which might question our assumption that the human emission has uniform effect on measurements on the ceiling regardless of positions of sources. Also, the ventilation rate, opening and closing of doors, and different activities might all introduce additional noise to the measurements. Nevertheless, regardless of these factors, Figure 11 shows that sensing by proxy is reasonably robust to these influences, where we plot the estimated number of occupants together with the ground truth.

The fast transition behavior exhibited in the CO_2 pump experiments is also observed for the occupants experiments, even without any sensors to explicitly sense the exits or entry of people as in other methods such as particle filters or Markov models [8]. The occupancy inference is accurate without explicitly specifying the transition rates of the occupancy model. For all these inference, the parameters chosen are the same, namely $L_1 = 2$, $L_2 = 0.02$, and physical parameters from Table III.

To compare with other models, we employ the root mean-squared-error (RMSE) with units of fractional people, given by

$$RMSE = \sqrt{\frac{1}{T}\sum_{k=1}^{T}\left(\phi(k) - \hat{\phi}(k)\right)^2} \tag{19}$$

where $\phi(k)$ is the ground truth occupancy at time k, $\hat{\phi}(k)$ is the estimated occupancy at time k given by

$$\hat{\phi}(k) = \left\lfloor \frac{\tilde{\hat{V}}(k)}{\rho_{\text{human}}} + \frac{1}{2} \right\rfloor \tag{20}$$

where $\tilde{\hat{V}}(k)$ is the median-filtered estimated emission rate at time k, $\rho_{\text{human}} = 0.183$ppm/sec is the average sedentary person emission rate, and $\lfloor x \rfloor$ is the floor operation to obtain the largest integer smaller than x.

The comparison of sensing by proxy with other methods are shown in Table IV. Since all the other models require substantial training phase, the data is split to training and testing sets and the RMSE is computed by 10-fold cross-validation. The algorithms take the measurements from the air supply and air return vents as features, where the corresponding labels are the number of occupants. The outputs for each testing point are the number of occupants obtained by classification, which are compared against the ground truth. For standardization purpose, we employ the Weka Machine Learning Toolkit [14] for the implementation of these algorithms. No time dynamic models, such as particle filters are learned for comparison, as it requires additional sensors to measure transitions and extra knowledge of transitional probabilities, which require substantial learning data and might not be reliable for the case of non-stationary activities in practice.

Even though the parameters of our link model are shared across all the experiments, the training for other models might be significantly different for each experiments, which differ by scale and time. Therefore we decide to separate the RMSE for each experiment, as shown in the first three columns of Table IV, which might make it obvious that which model is consistently better even with different dataset. In the last column, we combine all the occupants experiments data and test each model. Note that as it is possible for other models

 ISBN: 978-1-61208-418-3

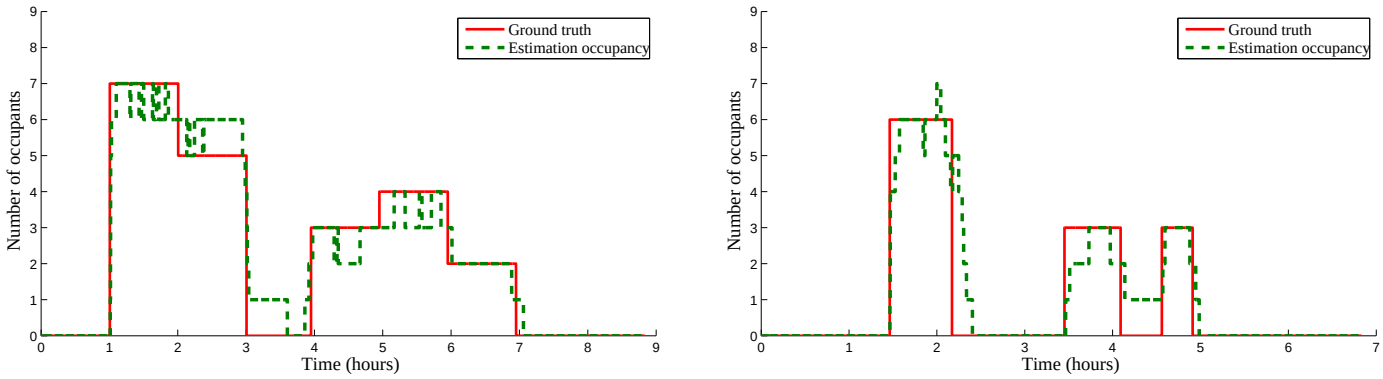


Figure 11. Occupancy detection by "sensing by proxy" (Algorithm 1) for experiment E (left) and G (right). The response times from vacancy to occupancy and vice versa are about 10 seconds and 5 to 10 minutes respectively, since the method detects the dynamics of the system rather than static concentration. The estimations (green) are within 1 occupant of the ground truth (red).

TABLE IV. COMPARISON OF ROOT MEAN-SQUARED ERROR OF ESTIMATION IN OCCUPANTS EXPERIMENTS

	Exp. E	Exp. F	Exp. G	Mixed
Naïve Bayes	1.3080	0.7454	1.7457	1.3555
Bayes Net	1.2345	0.6555	1.5406	1.2061
Logistic regression	1.0796	0.6109	1.8414	1.4736
Multi-Layer Perceptron	0.9686	0.5672	1.6221	1.2321
RBF Network	1.0837	0.6760	1.6496	1.3341
Seq. Min. Opt. (SMO)	1.2326	0.6185	1.8803	1.6118
AdaBoostM1	1.6415	0.7053	2.2257	2.3927
Sensing by Proxy	**0.5922**	**0.3809**	**0.7331**	**0.6311**

to yield different outputs due to different training, sensing by proxy will output the same value given the chosen parameters, which is desirable since it is less susceptible to training noise.

Sensing by proxy, as can be seen, delivers standout performance in all the testings, while the second best (underlined) positions are shared between Multi-Layer Perception (MLP) and Bayes Net, whole error metric almost doubles that of sensing by proxy in the mixed dataset case. By ignoring the dynamics of CO_2 concentration, these algorithms are confused by the overlapping concentration region as shown in Figure 6 especially during CO_2 accumulation and depletion period.

Close examination of the confusion matrix for our model and the second best model, in the mixed dataset case, the Bayes Net, as visualized in Figure 12, reveals an additional advantage of sensing by proxy. In the illustration, the size of the bubble represents the percentage of data classified as $\hat{\phi}$ (y-axis) for ground truth ϕ (x-axis), normalized for the sample size corresponding to ϕ. Bayes Net has a straight diagonal pattern, but it is undermined by the nonnegligible points far off the diagonal, representing misclassification error with large magnitude. On the contrary, sensing by proxy, though not possessing the straight diagonal pattern as in Bayes Net, is fairly clean of points far off the diagonal region. The point mass is also concentrated in the narrow band of sub-diagonals, which indicate that the estimation is within an error of 1 person. This is clearly preferred in practice, as sufficiently accurate estimation of occupancy can save much more energy than exact occupancy estimation but with misinference of crowded space when the room is just vacant.

V. Related Work

Existing approach to indoor occupancy estimation employed machine learning methods with multi-sensor fusion through dense sensor deployment. Passive Infrared (PIR) sensors and magnetic reed switch to detect door open/close events are suitable for binary occupancy detection [3]. Fusion of PIR sensors with cameras in the particle filter framework was proposed for occupancy prediction. Inhomogeneous Markov Chain, such as closest distance Markov chain and Blended Markov Chain, was employed for real-time occupancy based conditioning strategies [7]. Occupancy estimation using real (motion, door closure) and virtual (PC activity detector) sensors was presented in a small office based on the decision tree and artificial neural network models [15]. Individual presence detection based on power consumption using zero-training algorithm is proposed in [16]. A complex sensor network was established [8] comprising ambient-sensing (lighting, temperature, relative humidity, motion detection and acoustics), CO_2 sensing, and air quality sensing systems, which were incorporated into a Hidden Markov Model.

There are two main streams of modeling room air dynamics, namely computational fluidic dynamics (CFD) and zonal models [17]. CFD requires substantial model specification (e.g., locations of all walls, furniture, and occupants) and computation to produce detailed map of air motion. Zonal models, on the contrary, relies on ODE mass balance laws between different zones, though the distributed local nature of airborne contaminant transfer within a single space is not captured. Clearly a trade-off between spatial details and model simplicity is more practical for occupancy sensing.

Techniques for the estimation of the concentration of contaminants emitted from a source in indoor environments exist in the literature [17]. Boundary observers for some classes of PDEs are constructed in [18] via backstepping. In [19], this methodology is applied for the estimation of the state-of-charge of batteries. Observer designs for time-delay systems with unknown inputs are presented in [20]. In light of the current development in control theory, sensing by proxy recasts the occupancy inference as a problem of state and input estimation to allow robust, automatic, real-time inference.

 ISBN: 978-1-61208-418-3

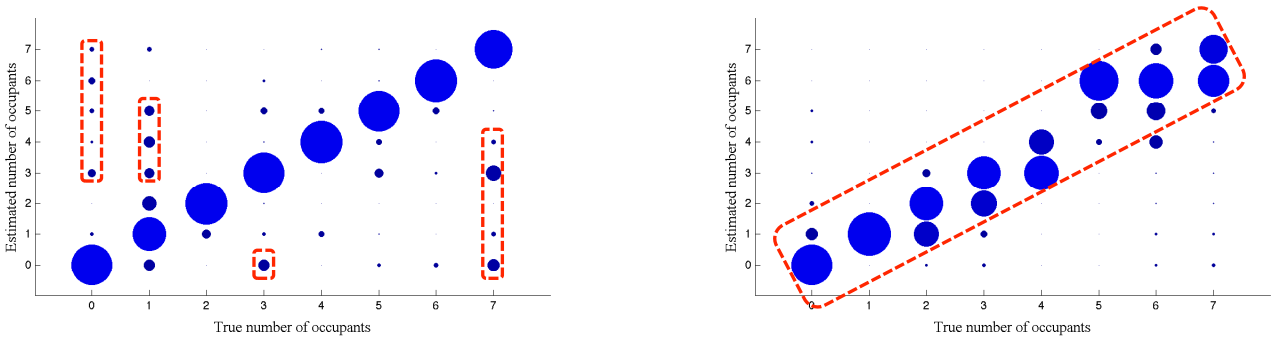


Figure 12. Visualization of confusion matrix for Bayes Net (left) and sensing by proxy (right), where the position of circles represents the true number of occupants (x-axis) and estimated number of occupants (y-axis), and the size indicates the percentage normalized for each column.

VI. Conclusion

This study describes an occupancy detection algorithm using indoor CO_2 concentration based on the sensing by proxy methodology, which explores the spatial and temporal features of the system with constitutive models. Controlled field experiments are conducted in a typical indoor space to show that the proposed link model can reproduce the CO_2 measurements given the latent emission rates. It is demonstrated that sensing by proxy can reliably detect the number of occupants based on "proxy" observations with RMSE of 0.6311 (factional person), as compared to 1.2061 (factional person) of the best alternative machine learning algorithm. Investigation of the confusion matrices reveals that the estimation by sensing by proxy is within 1 occupant of the ground truth with high probability, while the estimation by Bayes Net sometimes has large deviations. By successfully identifying the **L3** factors (location, latent factors, and link model) in the problem, sensing by proxy can be also applied to other tasks, such as indoor pollutants source identification, while requiring minimal capital investments.

Acknowledgment

This research is funded by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore.

References

[1] J. McQuade, "A system approach to high performance buildings," United Technologies Corporation, Tech. Rep, 2009.

[2] V. Garg and N. Bansal, "Smart occupancy sensors to reduce energy consumption," Energy and Buildings, vol. 32, no. 1, 2000, pp. 81–87.

[3] Y. Agarwal, B. Balaji, R. Gupta, J. Lyles, M. Wei, and T. Weng, "Occupancy-driven energy management for smart building automation," in Proceedings of the 2nd ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Building. ACM, 2010, pp. 1–6.

[4] J. Lu et al., "The smart thermostat: using occupancy sensors to save energy in homes," in Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems. ACM, 2010, pp. 211–224.

[5] V. L. Erickson et al., "Energy efficient building environment control strategies using real-time occupancy measurements," in Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings. ACM, 2009, pp. 19–24.

[6] S. Meyn, A. Surana, Y. Lin, S. M. Oggianu, S. Narayanan, and T. A. Frewen, "A sensor-utility-network method for estimation of occupancy in buildings," in Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on. IEEE, 2009, pp. 1494–1500.

[7] V. L. Erickson, S. Achleitner, and A. E. Cerpa, "Poem: Power-efficient occupancy-based energy management system," in Proceedings of the 12th international conference on Information processing in sensor networks. ACM, 2013, pp. 203–216.

[8] B. Dong et al., "An information technology enabled sustainability test-bed (itest) for occupancy detection through an environmental sensing network," Energy and Buildings, vol. 42, no. 7, 2010, pp. 1038–1046.

[9] S. Wang and X. Jin, "Co2-based occupancy detection for on-line outdoor air flow control," Indoor and Built Environment, vol. 7, no. 3, 1998, pp. 165–181.

[10] A. Baughman, A. Gadgil, and W. Nazaroff, "Mixing of a point source pollutant by natural convection flow within a room," Indoor air, vol. 4, no. 2, 1994, pp. 114–122.

[11] N. Bekiaris-Liberis and M. Krstic, "Lyapunov stability of linear predictor feedback for distributed input delays," Automatic Control, IEEE Transactions on, vol. 56, no. 3, 2011, pp. 655–660.

[12] M. Krstic and A. Smyshlyaev, Boundary control of PDEs: A course on backstepping designs. Siam, 2008, vol. 16.

[13] "K-30 10,000 ppm CO2 sensor," http://www.co2meter.com/products/k-30-co2-sensor-module, June 2015.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," ACM SIGKDD explorations newsletter, vol. 11, no. 1, 2009, pp. 10–18.

[15] S. Srirangarajan and D. Pesch, "occupancy estimation using real and virtual sensors," in Proceedings of the 12th international conference on Information processing in sensor networks. ACM, 2013, pp. 347–348.

[16] M. Jin, R. Jia, Z. Kang, I. C. Konstantakopoulos, and C. Spanos, "Presencesense: Zero-training algorithm for individual presence detection based on power monitoring," in BuildSys14, November 5–6, 2014, Memphis, TN, USA, 2014, pp. 1–10.

[17] L. Mora, A. Gadgil, and E. Wurtz, "Comparing zonal and cfd model predictions of isothermal indoor airflows to experimental data," Indoor air, vol. 13, no. 2, 2003, pp. 77–85.

[18] A. Smyshlyaev and M. Krstic, "Backstepping observers for a class of parabolic pdes," Systems & Control Letters, vol. 54, no. 7, 2005, pp. 613–625.

[19] S. Moura, N. Chaturvedi, and M. Krstic, "Pde estimation techniques for advanced battery management systemspart i: Soc estimation," in American Control Conference (ACC), 2012. IEEE, 2012, pp. 559–565.

[20] D. Koenig, N. Bedjaoui, and X. Litrico, "Unknown input observers design for time-delay systems application to an open-channel," in Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on. IEEE, 2005, pp. 5794–5799.

 ISBN: 978-1-61208-418-3

Usability Evaluation Approaches for (Ubiquitous) Mobile Applications: A Systematic Mapping Study

[1]Rodrigo A. Cruz Reis, [2]Ludymilla L. A. Gomes, [3]Arilo Claudio Dias-Neto, [4]Awdren de L. Fontão
[123]Institute of Computing
Federal University of Amazonas
[4]INDT
Manaus-AM, Brazil
Emails: [1]rdgdosanjos@gmail.com, [2][llag] [3][arilo] @icomp.ufam.edu.br, [4]fontao@microsoft.com

***Abstract*—In case of ubiquitous mobile applications, there is an increased need for effective/efficient approaches to evaluate the usability of these applications. The technical literature provides several evaluation approaches found in several sources, with different characteristics and classifications. This paper presents the results of a systematic mapping study that investigated usability evaluation approaches for (ubiquitous) mobile applications. In total, we identified 101 usability evaluation approaches for mobile applications, 28 of which applied to ubiquitous mobile applications. They were classified according to some attributes, such as: type of evaluation technique, type of mobile apps to be evaluated, experiment used to evaluate the approach, usability attribute/factors to be evaluated, and characteristics of ubiquity evaluated by each approach, representing the state-of-art in this research field.**

Keywords-Usability; Mobile apps; Systematic Mapping; Survey.

I. INTRODUCTION

Smartphones have become very popular in our current society. Advances in mobile technologies have allowed the emergence/development of a wide range of software for these mobile devices (called mobile applications or, simply, mobile apps) [1]. This platform introduced several advantages. Perhaps, the most noticeable would be the mobility to its users while using different mobile apps.

This large and growing number of mobile apps has challenged software engineers to develop applications with a high level of quality in order to become more attractive and competitive in this new market [2]. Moreover, this platform introduced some challenges and constraints to be considered during the software development, such as small screen size, limited connectivity, high power consumption rates and limited input modalities [10].

According to Duh et al. [3], usability is a critical factor for the popularity and success of mobile apps. A good usability design improves the device user's operability and, thus, enhances the overall product quality. Users tend to choose applications that are easy to learn, which take less time to complete a particular task and seem to be more "friendly" to the user [4]. Thus, various approaches aimed at supporting the usability definition and evaluation for mobile apps have been proposed in the technical literature.

Usability evaluation of software for desktop and mobile devices platforms is an emerging areas of research [5]. In the past, software usability was subjectively evaluated by informal processes [4]. Researchers just selected some usability attributes that they wanted to assess and then measured what they considered important. In recent years, usability measurement and analysis approaches have been proposed and improved. Laboratory experiments, field studies and hands-on measurements are some of the methods most often applied by researchers [4][6]. Each of these evaluation methods has its advantages and disadvantages. Due to the highly dynamic context of use, offered by mobile apps, laboratory and field usability testing involves different challenges and may find different usability problems [3].

In order to analyze the scenario of evaluation usability approaches for ubiquitous mobile apps and ubiquitous mobile apps, this paper presents the results of a systematic mapping study [7] that identified and characterized 101 different approaches. This study aims to complement previous characterization studies, such as the studies published in [4][6], in two aspects: (1) it updates the list of approaches identified in the technical literature; (2) it presents a different perspective on the identified approaches, analyzing, for example, the categories of mobile apps and the type of proposed evaluation approach (e.g., static or dynamic analysis). Finally, some challenges and trends are presented as a result of this study.

This paper is structured as follows. Section 2 presents some definitions relevant to this paper and related work. Section 3 presents the systematic mapping protocol. Section 4 presents the results obtained. Finally, Section 5 presents a summary of this work and a brief discussion on future work.

II. USABILITY EVALUATION IN MOBILE APPS

Several definitions for usability can be found in the technical literature. For example, ISO-9241 [8] defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specific context of use".

With the emergence and rapid deployment of mobile technologies, the usability of applications developed for this platform has been the focus of several studies. According to Duh et al. [3], a good project, besides meeting the needs of the market and providing the device user satisfaction, can also reduce physical and mental stress, reduce the learning curve, improve the operability of device use and, thus, improve the overall product quality.

Zhang et al. [10] claimed that the mobile usability includes some of the new challenges related to mobility, such as: mobile context, connectivity, small screen size, different display resolutions, limited capacity and power processing, data entry methods, interaction with multi-touch screen,

 ISBN: 978-1-61208-418-3

show different resolutions and dimensions, device orientation changes and gestures such as tap, flick, and pinch. Thus, an approach aiming to evaluate the usability of mobile apps needs to deal with these challenges.

Software usability evaluation approaches have become increasingly popular in technical literature [11],[12]. Usability evaluation approaches aims to obtain a third-party judgment regarding user's characteristics to assess effectively and efficiently whether a user is able to view the content or perform a task on a specified device [10].

Some previous studies presented a usability evaluation analysis for mobile apps. In [4], the authors presented a study that analyzed the methodologies used to empirically evaluate the mobile usability, classifying as laboratory experiments, field studies and measuring practice. The study described advantages and limitations of each method, but did not identify/characterize the publications on usability evaluation in the technical literature.

In [6], the authors presented a longitudinal review of Human-Computer Interaction (HCI) research methods for the mobile platform published until 2012, analyzing more than 140 papers. In this study, publications were classified in terms of their research method (case study, field of study, action research, laboratory experiment, survey, basic and applied research, and normative writings) and purpose (understanding, engineering, re-engineering, evaluation and description). This study revealed that 68% of the material evaluated in research on human-computer interaction in mobile apps until 2009 involved mobile usability evaluations, where 63% of these researches made through laboratory experiments, 29% through field studies, and 7% through surveys.

Duh et al. [3] described a study that investigated the differences between the usability testing on mobile phones conducted in laboratory and real-life situations. Significant differences were found, including the frequency and severity of usability problems found in both scenarios, user behavior and subjective responses to the device and the interaction between users and the devices.

Kjeldskov et al. [5] presented and analyzed six techniques for evaluating the usability of mobile apps in laboratory. The six techniques were analyzed using two usability experiments. The goal was to examine whether the evaluation of mobile systems in a controlled environment is similar to a real user behavior.

Finally, in [10], the authors presented an overview of existing usability studies, focusing on usability testing, and discussed the main issues investigated in the technical literature. Then, they proposed a generic framework and provided detailed guidelines on how to conduct such usability studies.

We can observe that the studies and approaches that address the evaluation of usability in mobile apps are dispersed in different sources in the technical literature, making it difficult to analyze empirical evidence known about this research area. Thus, this paper describes a systematic mapping study conducted to identify/characterize/evaluate usability evaluation approaches for mobile applications proposed at the technical literature. The following sections present the planning and results of this systematic mapping study that investigated different perspectives related to usability evaluation approaches for mobile apps.

III. SYSTEMATIC MAPPING ON USABILITY EVALUATION FOR MOBILE APPS

According to Kitchenham et al. [7], a systematic mapping consists of a type of secondary study where the dimensions to be evaluated in a secondary study (population, intervention, comparison and outcomes) are not fully described. This study explores a less strict research protocol when compared to protocols commonly used in systematic reviews.

A good systematic mapping always considers the following questions [13]: identifying all published materials related to the investigation goal; choosing criteria for the inclusion of materials; evaluating the quality of each material; producing the results of each material impartially; interpreting the results; and, presenting a reasonable and neutral summary of the results.

This research follows a systematic mapping process described by [7], which is composed of three stages: (1) Plan the study; (2) Conduct the review; (3) Report the results. The activities related to the planning and conducting of this literature systematic mapping study will be described in the following subsections. The results from this study are described in the subsequent section.

A. Research Questions

The objective of this study is to identify approaches and types of research in usability evaluation for mobile apps and also point out the areas where the available empirical evidences were insufficient and therefore, more studies are needed. In order to address the objectives of this research, four relevant research questions were prepared:

- Q1. What types of approaches have been proposed for usability evaluation of mobile apps?
- Q2. To which category of mobile app approaches are employed usability evaluation approaches?
- Q3. What usability attributes/factors are evaluated by these approaches?
- Q4. Which characteristics of computational ubiquity are evaluated by this approach?

B. Identifying and Selection of Primary Studies

The sources used for selection of primary studies in this study were two digital libraries: IEEExplorer and Scopus (according to its maintainer, this online indexing service would cover the major computing digital libraries, such as ACM Digital Library or Science Direct. Only IEEExplorer would be partially indexed by Scopus).

The search string used for the search of primary studies was structured according to the rules described in [14], and was composed of the elements Population (P), Intervention (I), Comparison [optional in a systematic mapping study] (C) and Outcomes (O), as follows:

- **Population:** "Mobile Application" OR "Mobile Software" OR "Mobile App" OR "Mobile System".

 ISBN: 978-1-61208-418-3

- **Intervention:** "usability" OR "user experience" OR "HCI" OR "human computer interaction".
- **Comparison:** not applied to systematic mapping study.
- **Outcome:** "evaluation" OR "assessment" OR "measure" OR "experiment" OR "test" OR "inspect" OR "review".

C. Primary Studies Inclusion Criteria

A list of primary studies was obtained through the search string from the selected sources of bibliographic material. Then, the following criteria for inclusion of primary studies that were related to the objective of this study, in order to answer the research questions, were applied: (1) It describes research that explores usability evaluation approaches for mobile apps; (2) It must contain a full research publication; (3) It must be written in English; (4) It must be available for download.

Papers duplicated on different search sources (e.g., papers indexed by IEEExplorer and Scopus) would have only one instance selected in this study.

D. Systematic Mapping Execution

The activities of execution of research string and papers selected in this study were made between January and February 2014.

TABLE I. NUMBER OF SELECTERD PUBLICATIONS PER PHASE

Source	*Returned Papers*	*Filter 1*	*Filter 2*	*Filter 3*
IEEEXplorer	317	10	8	1
Scopus	53	170	93	27
TOTAL	**370**	**180**	**101**	**28**

The preliminary research offered 317 relevant publications in the Scopus library and 53 in the IEEExplorer library. The inclusion analysis of these papers was done in three steps (Table I):

1. Tracking the initial set of papers based on the titles, abstracts and introduction sections. In total, 180 publications were pre-selected to step 2;
2. Complete reading of the paper. A total of 101 publications were selected. In this step, the information to answer questions Q1, Q2 and Q3, previously presented in section III.A, was extracted from these papers.
3. Complete reading of the article from the ubiquitous applications point of view. 28 publications were selected in this phase, which were used to answer Q4.

E. Data Extraction Form

For each selected paper, we extract the main information aiming to characterize the usability evaluation approach:

- **[YEAR]** Publication Year.
- **[CATEGORY]** Type of evaluation technique:
  - Static: methods that do not involve software execution.
  - Dynamic: methods involving software execution with real/simulated data in a real or simulated environment.
- **[TYPE]** Type of mobile apps evaluated by the approach, classified according to [14] as:
  - **Native Apps:** application specifically developed to execute on a specific device platform.
  - **Web Apps:** application that runs over a browser embedded in the device and does not have access to some device's internal resources.
  - **Hybrid Apps (HTML5 and widgets apps):** they get stored in the device's main screen and can take advantage of all devices' internal resources, but they can be based on HTML5 and displayed through a web browser.
- **[EVALUATION]** Type of empirical evaluation applied to the approach, according to [6]: Case Study, Field Study, Action research, Lab experiments, Survey research, Applied research, Basic research, Normative writings.
- **[ATTRIBUTES]** Attributes evaluated by the approach (classification proposed by [15]): Efficiency, Satisfaction, Learnability, Memorability, Errors.
- **[FACTORS]** Usability factors evaluated by the approach (rating also proposed by [1]):
  - **User:** It is important to consider the end user of an application during the development process.
  - **Task:** refers here to the goal the user is trying to accomplish with the mobile application.
  - **Context of Use**: refers here to the environment in which the user will use the application.
- **[UBIQUITOUS]** Ubiquitous characteristics evaluated by the approach (classification proposed by [16]): Pervasiveness services, Invisibility, Context awareness, Adaptive behavior, Experiences Capture, Functionality composition, Spontaneous interoperability, Heterogeneity of devices, and Fault tolerance.

IV. RESULTS ANALYSIS

The usability evaluation approaches for mobile applications were analyzed according to the characteristics defined in the data extraction form (Section III.E). Thus, the overall results for each research question (presented in section III.A) and attributes extracted from the evaluation approaches are discussed in subsequent sections.

A. Analysis by Publication Year

In this study, we identified usability evaluation approaches for mobile apps published from 2004 to 2014 (when the study was run).

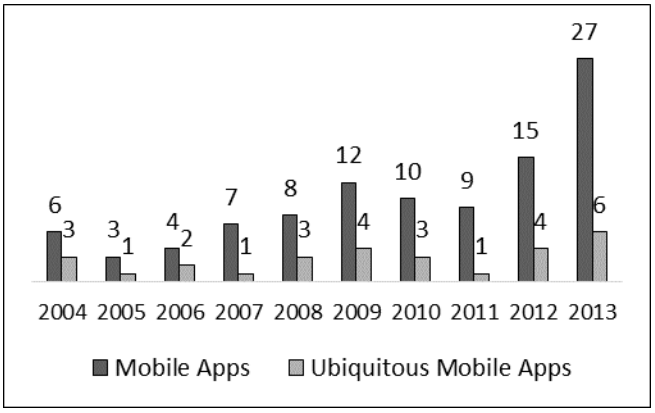


Figure 1. Analysis of papers by publication year.

The distribution of results is displayed graphically in Figure 1, where it is observed that there was a considerable increase in research on usability evaluation for mobile apps in the community recently (mainly in the past 2 years). This indicates the need for research in this area and shows the evolution in the level of importance of the issue. With the advances in mobile technology in bringing the concept of ubiquity, this area tends to become more interesting for future research [15].

We also noticed the number of usability evaluation approaches for ubiquitous mobile apps remained stable from 2009 until 2013. This 5-year interval has 18 of the 28 papers found in this study. This shows that the need for evaluation in ubiquitous mobile apps is really relevant to the academy and its interest in academic research is growing in the last years.

B. Analysis by Type of Evaluation Technique

In order to answer the question Q1 discussed in section III.A, an analysis of evaluation techniques per category was made (Figure 2).

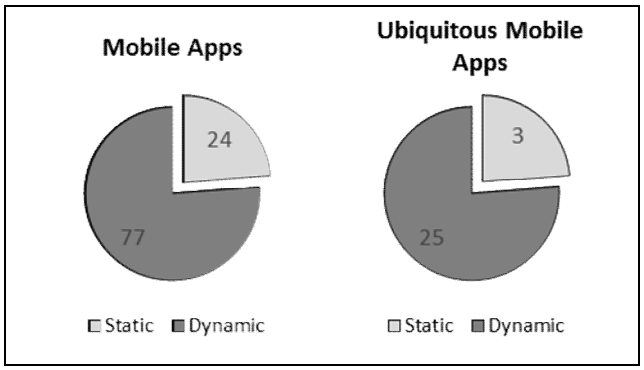


Figure 2. Analysis by evaluation technique category

We observed the category of usability evaluation techniques most frequent are dynamic approaches (77/101 for mobile apps; 25/28 for ubiquitous mobile apps). This result is justified due to the need of assessing the app on a scenario closer to reality, possibly by using dynamic approaches, making the evaluation more efficient.

C. Analysis by Mobile App Category

In order to answer the question Q2 discussed in section III.A, an analysis per mobile app category evaluated by the identified approaches was performed (Figure 3).

Among the categories analyzed in this study, it is remarkable that native apps have been more explored in research with the purpose of evaluating usability attributes in general and ubiquitous mobile apps. The reason could be the requirements, accessibility, and restriction issues imposed by mobile platforms. Then, the second more explored category is web apps, due to the popularity of this type of application for the mobile platform.

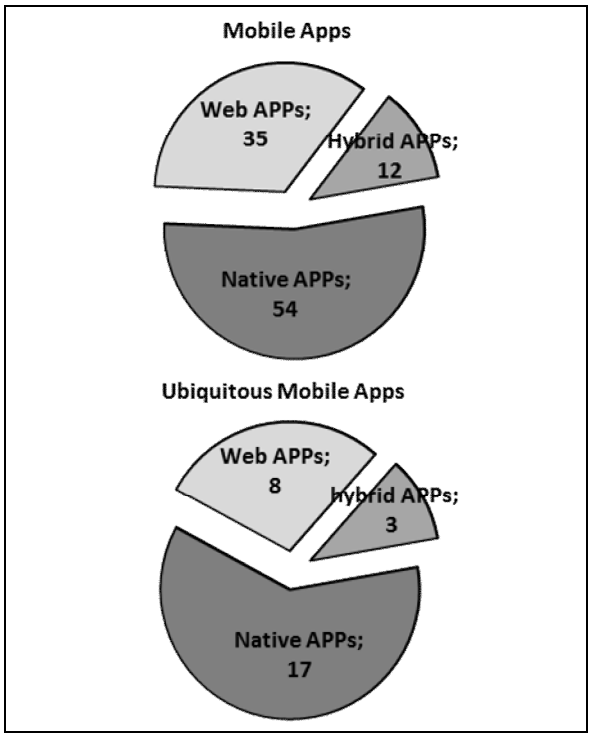


Figure 3. Analysis of mobile application category

Finally, evaluations in hybrid apps are starting to emerge, because it represents a new trend of development of mobile apps, justifying the small number of research in the area. All authors who proposed usability evaluation approaches for hybrid apps highlighted that this category is emerging and needs more research, not only for evaluation of usability, but also for application development.

D. Analysis by Empirical Evaluation Type

We also analyzed the type of empirical evaluation applied to the selected approaches, as shown in Figure 4. The results indicate that several authors have chosen to apply empirical techniques as a strategy for the final assessment of the proposed approaches. The results indicate the predominance of Case Studies (40/101), followed by Field Studies (29/101) and Lab Experiments (25/101). Three type of empirical evaluation were not found in the selected papers: Normative writings, Applied Research, and Basic Research. Analyzing the results for Ubiquitous Mobile Apps, they indicate approximate values between the same three types of evaluation: Field Study, with (10/28, followed by Case Study (9/28) and Lab Experiment (8/28).

 ISBN: 978-1-61208-418-3

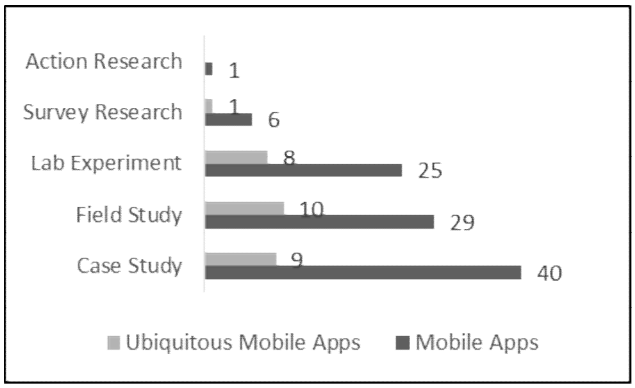


Figure 4. Analysis by type of experiment.

Trying to understand the result, we could observe the types of investigation most applied to evaluate the proposed approaches (case, field, and lab study) were formal investigations, having more credibility in the academy. This scenario can justify the difference obtained when compared to other types of investigations.

E. Analysis by Usability Attributes

In order to answer the question Q3 addressed in Section III.A, an analysis by usability attributes was performed, as shown in Figure 5. We observed that user satisfaction is the most investigated attribute in the identified approaches (67/101 papers) on mobile apps, and it is the second more investigated in ubiquitous mobile apps (15/28 papers). In this analysis, a paper could address one or more attributes, justifying that the sum of the numbers distributed among the attributes is greater than the number of identified papers.

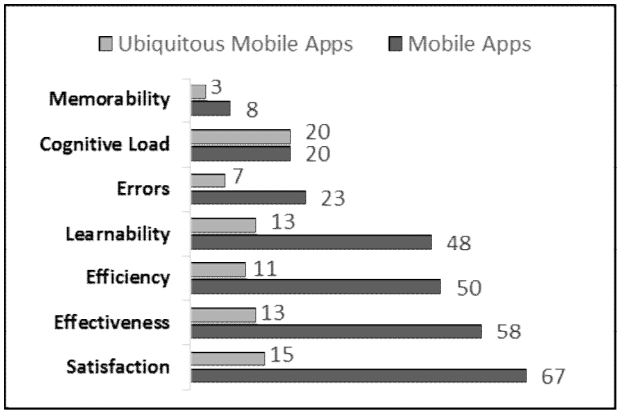


Figure 5. Usability Attributes Analysis

It is not possible to conclude that a usability attribute, for being more addressed than others, would be more or less relevant to be evaluated for mobile apps. The results only describe the state of the art on the use of these attributes in the academy. However, an interesting aspect may be observed: the Cognitive Load attribute was only used in mobile apps that deal with context awareness requirements, one of the features present in ubiquitous mobile apps (question 4, to be discussed below).

F. Analysis by Usability Factors

Yet to answer the question Q3 addressed in section III.A, an analysis by usability factors was done, as shown in Figure 6. Analyzing the results, we observed that the factors user (60/108) and tasks (66/108) are more frequent in the selected papers. In general, most of the papers when dealing with one of these factors also deal with the second one. The evaluation of the factor context of use was observed only in approaches that deal with context awareness requirements. In this analysis, a paper could address one or more factors, justifying that the sum of the numbers distributed among the factors is greater than the number of identified papers.

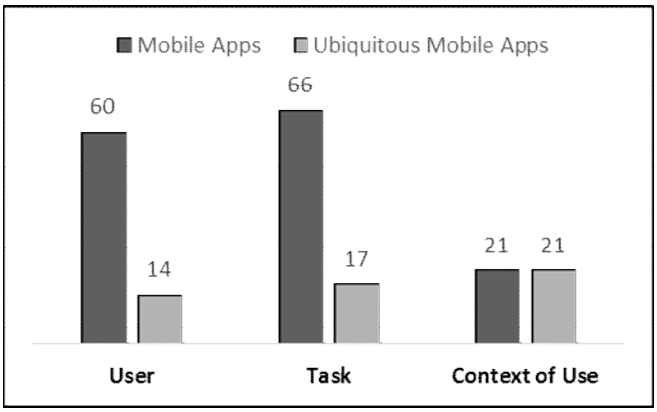


Figure 6. Analysis of Usability Factors

Furthermore, this also shows us that ubiquitous mobile applications can be evaluated recitals three factors as context of use with 21 identified papers, the user featuring 14 papers and 17 papers task with a total of 28.

G. Ubiquitous Features Analysis

In order to answer the question Q4 addressed in Section III.A, an analysis of ubiquitous feature evaluated by the identified approaches was done (Figure 7). In [16], a table with ubiquity characteristics from a functional point of view is presented.

Only 28 papers addressed the usability evaluation for mobile apps with characteristics of computational ubiquity. Among the 10 characteristics defined in [16], only 5 were addressed in papers identified in this study, suggesting that these would be the computational ubiquitous features that could be evaluated by means of usability requirements.

Furthermore, we observed that the definition of pervasiveness services suggests it as a main feature of ubiquity. Thus, in the case of ubiquitous mobile apps, pervasiveness services feature will always be present. This explains why all 28 papers that deal with usability evaluation approaches for ubiquitous mobile apps cite this feature, but did not propose an approach to analyze this feature for ubiquitous mobile apps.

The distribution of the papers among the ubiquitous characteristics is presented in Figure 7.

 ISBN: 978-1-61208-418-3

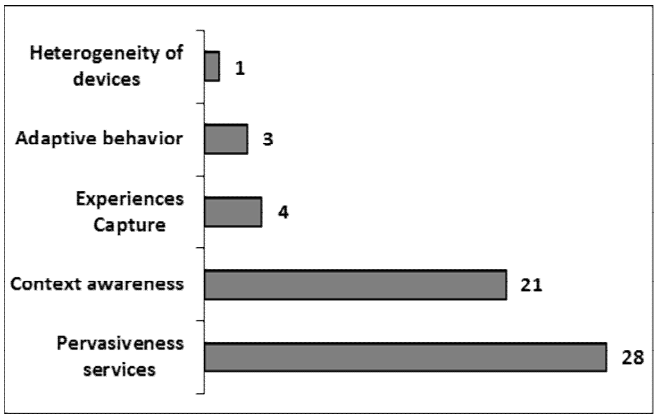


Figure 7. Feature of ubiquity analysis

By analyzing the results, we observed that the characteristics context awareness was addressed in 21 papers. We observed that in other characteristics, few studies addressing usability evaluation were found. One factor that may hinder this analysis is that each ubiquity feature has a vast array of settings and areas that still need to be analyzed from a usability point of view. We could observe that research in the field of usability in mobile apps that deal with each ubiquity feature is still quite scarce. This shows good opportunities for research in this domain area.

V. CONCLUSIONS

The number of mobile apps used in daily life is continuously growing and so is the search on their quality. Despite this evolution, if we compare the demand implementation of ubiquity characteristics, which is a factor present currently in many apps, we can observe that there is still a need for studies on mobile apps usability evaluation.

The results of this systematic mapping study revealed some perspectives about the approaches to support the evaluation of usability in mobile apps in the last 10 years. For example, they indicate that such approaches mainly utilize dynamic techniques (e.g., testing). Many publications brought justification for use of such technique, saying that the context of use was the main reason for choosing the testing technique.

We also observed that native and web apps have been the focus of usability evaluation approaches for mobile apps, which indicates a need for attention to hybrid apps, which are partly native and partly web application.

In order to evaluate the techniques, experimentation is ahead of the other techniques. The justifications of the authors are related to the restrictions that mobile apps need, what could be settled in an empirical evaluation.

We also observed that a small number of studies covers ubiquitous mobile apps. Soon, there will be the necessity for more studies on this topic.

This systematic mapping was done in order to identify which types of research in mobile app usability evaluation are being used in the academy. With the results obtained from the mapping, it is possible identify the areas most addressed by the community in which there are a large number of studies and point out the areas where the available evidence is insufficient and therefore more studies are needed.

The need for studies focused on different ubiquity factors oriented mobile apps is noticeable. Future work can be made, such as: choose the type of mobile application category that can be web applications, native applications or hybrid applications and instantiate ubiquity factors to usability evaluations. There is a clear need for approaches, processes, tools to support the assessment of usability in ubiquitous mobile applications.

ACKNOWLEDGMENT

We would like to thank FAPEAM and INDT for their financial support.

REFERENCES

[1] R. Harrison, D. Flood, and D. Duce, Usability of mobile applications: literature review and rationale for a new usability model, “Journal of Interaction Science 1 (1), 2013, pp. 1-16.

[2] Global mobile statistics 2011, <http://mobithinking.com/mobilemarketing-tools/latest-mobile-stats> 2014/05/13

[3] H. Duh, G. Tan, and V. Chen, “Usability evaluation for mobile device: a comparison of laboratory and field tests,” Proceedings of the 8th conference on Human-computer interaction with mobile devices and services (MobileHCI), 2006, pp. 181-186.

[4] F. Nayebi, J. M. Desharnais and A. Abran, “The State of the Art of Mobile Application Usability Evaluation,” In: 25th IEEE Canadian Conference on Electrical Computer Engineering CCECE, 2012, pp.1-4.

[5] J. Kjeldskov and J. Stage, “New techniques for usability evaluation of mobile systems,” In: International Journal of Human-Computer Studies, 2004, pp. 599-620.

[6] J. Kjeldskov and J. Paay, “A longitudinal review of Mobile HCI research methods,” In: International Conference on Human-computer interaction with mobile devices and services (MobileHCI), 2012, pp. 69-78.

[7] B. Kitchenham and S. Charters, “Guidelines for performing Systematic Literature Reviews” in Software Engineering, EBSE Technical Report EBSE-2007-001, Department of Computer Science, University of Durham, 2007.

[8] ISO 9241-11 “Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)” International Standards Organisation, Geneva, 1997.

[9] ISO/IEC 25000.” Software Engineering -- Software product Quality Requirements and Evaluation (SQuaRE) -- Guide to SQuaRE”, (2003/2014).

[10] D. Zhang and B. Adipat, “Challenges, methodologies, and issues in the usability testing of mobile applications,” International Journal of Human-Computer Interaction, vol. 18, no. 3, 2005, pp. 293-308.

[11] M. Billi, L. Burzagli, T. Catarci, G. Santucci, E. Bertini, F. Gabbanini, and E. Palchetti, “A unified methodology for the evaluation of accessibility and usability of mobile applications,” Universal Access in the Information Society (9:4), 2010, pp 337-356.

[12] K. Hornbæk, “Dogmas in the assessment of usability evaluation methods,” Behaviour & Information Technology (29:1), 2010, pp 97-111.

[13] E. Mendes, “A systematic review of web engineering research”, Proceedings of the International Symposium on

 ISBN: 978-1-61208-418-3

Empirical Software Engineering (ISESE), Australia, 2005, pp. 498-507.

[14] M. Pai, M. McCulloch, J. D. Gorman, N. Pai, W. Enanoria, G. Kennedy and J. M. Colford Jr, "Systematic reviews and meta-analyses: an illustrated,' step-by-step guide. Natl Med J India. Mar-Apr, 2004, pp 86–95.

[15] J. Nielsen, "Usability engineering. Morgan Kaufmann Pub," (1994).

[16] R. O. Spínola, H. T. Guilherme, "Towards a framework to characterize ubiquitous software projects, Information and Software Technology," Volume 54, Issue 7, July 2012, pp 759-785.

 ISBN: 978-1-61208-418-3

Extension of Sikuli Tool to Support Automated Tests to Windows Phone Context-Aware Applications

Elizângela Santos da Costa
Product Validation
Institute of Technology Development
Manaus-AM, Brazil
email: elizangela.costa@indt.org.br

Rodrigo dos Anjos Cruz Reis
Institute of Computing
Federal University of Amazonas
Manaus-AM, Brazil
email: rdgdosanjos@gmail.com

Arilo Claudio Dias-Neto
Institute of Computing
Federal University of Amazonas
Manaus-AM, Brazil
email: arilo@icomp.ufam.edu.br

***Abstract* — The wide diffusion in the use of mobile devices has brought the need to improve the process of verification and validation in mobile applications. The usual way of interaction between these apps and users is through device interface. For context-aware mobile applications, the number of interactions that a user can perform is much larger if compared to common applications. Thus, manual testing execution turns out to be an exhaustive and error-prone activity. The main contribution of this work is to propose the creation of new functions in Sikuli tool to automate tests for context-aware mobile application developed to the Windows Phone platform in order to develop reliable applications using an effective test strategy.**

Keywords-Context-awareness; Automated Testing; Sikuli; Mobile Testing.

I. INTRODUCTION

On mobile platforms, the main form of interaction between users and applications is through Graphical User Interface (GUI). Thus, since mobile applications development is increasing greatly, GUI becomes more complex and more concern is required for its quality.

A method to evaluate the quality of software through its GUI is by performing a testing technique called GUI Test [1]. This type of testing must simulate the sequence of events performed by users. The large number of input possibilities for this sequence makes the GUI test a complex activity and requires a lot of manual effort for the testing process.

An important factor to be considered during the evaluation process of mobile applications is the context-awareness characteristic [2]. This characteristic aims to describe different contexts in which applications are subjects, meaning that they can react differently to changes in their environments. Different contexts in different applications are more likely to generate failures and the criteria of coverage to reach all possibilities should be proposed. An alternative to reduce the effort in these tests is the adoption of automated testing.

The rest of this paper is organized as follows. Section II describes the background. Section III describes the tool extended in this work. Section IV addresses the proof of concept performed with the extended tool. Finally, conclusion and future work are described in Section V.

II. BACKGROUND

Ubiquitous Computing is defined as: "an area of research that studies the integration of technology to human activities in a transparent way, when and where needed" [3]. Context-awareness is a sub domain of Ubiquitous Computing. *Context* is defined as any information that can be used to characterize the situation of an entity [2]. Devices, services and software components should be aware of their contexts and automatically adapt to your changes, characterizing the *context-awareness* [4].

A concept for *mobile application testing* and *GUI test* can be found in [5][6] respectively. *Sikuli* is a tool that uses visual approach to search and automate GUI tests using screenshots. Through this tool, testers can write visual scripts that specify the components to interact and what visual feedback to expect, it has the advantage of being independent of any platform [7].

III. EXTENDING THE SIKULI TOOL

The work was developed in four steps. First, the context elements of the device defined to be automated with respect to screen orientation, phone battery level, internet connection and location.

Changes in context elements were chosen in the second step. Thus, screen orientation can be positioned vertically or horizontally, battery has several levels represented by a percentage, connection can be enabled or disabled for both Wi-Fi and 3G or 4G connections, location can be enabled or disabled. Figure 1 resumes the structure of development. In the third step, the following automation scripts in Sikuli were developed:

- ***UtilWP***: Main functions for test automation;
- ***PC***: Script for proof of concept, it calls the functions created in UtilWP script;
- ***NivelBateria*** *and* ***NivelBateriaH:*** They store a database for existing images of battery levels in portrait and landscape;
- ***Scroll***: Contains functions that perform the sliding movement;
- ***Alfabeto***: It stores a database for alphabet letters helping in function of open app.

 ISBN: 978-1-61208-418-3

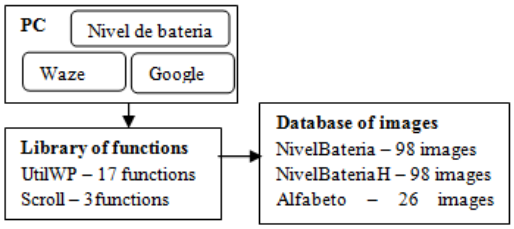


Figure 1. Structure of scripts developed.

It used the Windows Phone emulator called *Project My Screen App* [8] to project the application screen to be possible run the scripts. This is the last step.

IV. PROOF OF CONCEPT

To analyze the functions created, three mobile apps were chosen as shown on Table 1. To validate the reaction of mobile applications to the contexts, first it was analyzed how they should behave in cases of certain changes in device contexts.

TABLE I. CONTEXT VARIATION IN THE MOBILE APPS PER ELEMENTS

App	Screen Orientation	Internet Connection	Location	Battery
Google	X	X		
Waze			X	
Nível de Bateria (Battery Level)				X

After that, the execution of scripts was implemented using the functions created, as described in Section III. An example of automation can be seen in Figure 2 that validates the behavior of Waze app. In short, the script defines the expected behavior through image (line 2 and 3) for both location ON and OFF, modifies the status of the location property of the phone (to ON), opens the app and verifies whether Waze has adapted to the new context or not.

```
def WazeTest():
    locOn= 
    locOff= 
    switchApp("Project My Screen App")
    location=LocationOn()
    if (location=='on'):
        OpenApp('w',  )
        print '--Location connected--'
        wait(locOn,5)
        AssertEqual(locOn)
    else:
        print 'Connect location failed'
```

Figure 2. Test case for Waze.

The final steps were to run the scripts for the remainder of the apps and compare the results after execution.

As demonstrated, a tester is able to test a context-aware application once the right screen is defined to each state of context element. The additional libraries provide an easy way to write automated tests because the changes in device were automated before such as basic functions like open a mobile app. Now that changes in location are automated in the extended library, any mobile app that is context-aware to location can be tested with less effort, not only Waze.

In order to avoid errors for whom uses the tool, some care has been taken, such as treatments for occasional exceptions images that are not found. The device style pattern should also be noticed, since it is possible to change themes. Inserting wait commands with the pictures expected before checking equality between screens helps in fault occurrence prevention. Check if the image inserted in the script will be recognized by matching preview (Sikuli property) provides the preliminary recognition of faults that can be generated.

V. CONCLUSION AND FUTURE WORK

To provide a better process on validation methodology, it was proposed to use an existing tool for automating tests, Sikuli, extending its functions to create new ones that reduce manual efforts for the testing activity. Improvements of the implemented functions can be achieved by creating an image library for common buttons making the scripts cleaner and easier to maintain.

Some limitations were found throughout this work, and the most critically noted ones are instability in the Sikuli tool and operating system restriction, since Windows must be 8 or greater, to support the functionality of the control device by emulator.

Despite the existing limitations, the support offered to reduce the manual testing tasks and proved that it is feasible to automate context-aware mobile applications observing its adaptations to changes in addition to offer reduction of testing execution time and manual effort.

REFERENCES

[1] A. Ruiz and Y. W. Price, “GUI Testing made easy,”. Testing: Academic & Industrial Conference - Practice and Research Techniques. IEEE 2008, pp. 99-103, doi: 10.1109/TAIC.PART.2008.11.

[2] D. Amalfitano, A. Fasolino, P. Tramontana, and N. Amatucci, “Considering context events in event-based testing of mobile applications”, IEEE Sixty International Conference on Software Testing, Verification and Validation Workshops (ICSTW), Luxemburgo, Mar. 2013, pp. 126-133.

[3] M. Weiser, “The Computer for the 21st Century”. Mobile Computing and Communications Review – Special Issue Dedicated to Mark Weiser, vol. 3, no. 3, July 1999, pp. 3-11.

[4] J. L. B. Lopes, “Exehda-on: an approach based on ontologies for context-awareness in pervasive computing”, Dissertation (Master in Computer Science) – School of Informatics, Catholic University of Pelotas, 2008, p. 198.

[5] J. Gao, X. Bai, W. Tsai, and T. Uehara, “Mobile Application Testing: A Tutorial”, In:,” Computer, vol. 47, no. 2, Feb. 2014, pp. 46-55.

[6] A. M. Memon, M. E. Pollac, and M. L. Soffa, “Hierarchical GUI Test Case Generation Using Automated Planning,” IEEE Transactions on Software Engineering, vol. 27, no. 2, Feb 2001, pp. 144-155.

[7] T. H. Chang, T. Yeh, and R. Miller, “GUI Testing using Computer Vision,”. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New York, USA, 2010, pp. 1535-1544.

[8] Project my phone screen to a TV or PC. Available from: http://www.windowsphone.com/en-us/how-to/wp8/connectivity/project-my-phone-screen/2015.05.30

 ISBN: 978-1-61208-418-3

On an EAV Based Approach to Designing of Medical Data Model for Mobile HealthCare Service

Alexander Borodin, Yulia Zavyalova

Department of Computer Science
Petrozavodsk State University
Petrozavodsk, Russia
Email: {aborod, yzavyalo}@cs.petrsu.ru

***Abstract*—As a consequence of the advances in technologies of wearable devices, the amount of data received from different kinds of sensors tend to grow significantly. This is also the case with medical data. In most cases, this highly diverse medical data has multi-dimensional and sparse structure. There are generic XML and RDF formats to describe this data. Moreover, RDF is a key technology for semantic applications. Nevertheless, in real-world high-loaded services, relational databases are used for storing large volumes of attribute-value pairs for the purposes of better performance. This paper describes the approach to database schema design for a service of continuous monitoring of health parameters. The proposed approach is based on Entity-Attribute-Value model widely used in medical databases.**

Keywords–m-Health; Personalized Services; Entity-Attribute-Value.

I. INTRODUCTION

Currently, life expectancy among the elderly tends to improve. It leads to a significant growth in the number of patients with chronic diseases and it increases health care spending. Remote monitoring of chronically ill patients is the promising way of bounding the healthcare costs growth.

In the Park of Innovative Technologies of Petrozavodsk State University, two systems for the analysis of remotely harvested health data processing are being developed. The Research Platform for Building Medical Diagnostic Services is designed for processing medical diagnostic data [1]. This platform relies on Continuity of Care Record (CCR) standard and is intended for research purposes.

The aim of second project, CardiaCare is to provide a system for continuous monitoring of heart function [2]. The service concept developed in this project is more practical oriented and should be implemented in collaboration with an Emergency Hospital and with assistance of Ministry of Health of Karelia.

With the development of technologies of a mobile healthcare and, in particular, of wearable sensors, the amounts of data received from different kinds of sensors grew significantly. This highly diverse diagnostic information is stored in so-called Clinical Study Data Management Systems (CSDMSs) for further analysis and decision making. One of the vital features of CSDMSs is the ability to encompass hundreds of new clinical parameters with no need of database schema updates during the life cycle of the system [3].

Both projects mentioned above use a relational databases as the CSDMS backend. However, in relational databases, a description of the kinds of data that the database stores — the attributes – is recorded implicitly in the schema in the form of the structure of tables and relationships between them. Nevertheless, in rapidly evolving circumstances related to the development of new sensor hardware, frequent schema changes will be required. Redesign of the schema will, in turn, lead to the need of reimplementing almost all components of the system.

Under these circumstances, the so-called Entity-Attribute-Value (EAV) database design is useful. Nevertheless, the EAV design has both advantages and drawbacks.

In this paper, we discuss properties of EAV-modeled database aimed at the remote monitoring of health parameters and present design decisions that have been implemented in CardiaCare service backend.

The rest of the paper is organized as follows. Section II presents an overview of the CardiaCare service and the main requirements of the storage are justified based on the architecture of the service. Section III recalls the properties of EAV model. Section IV introduces our design decisions. Section V summarizes the results of this paper.

II. OVERVIEW OF CARDIACARE SERVICE

CardiaCare is a mobile service aimed at the continuous monitoring of heart function, detection of several kinds of arrhytmias and risk factors assessment based on the joint analysis of electrocardiogram recordings, auxiliary conditions, parameters of the environment (e.g., outside temperature) and concomitant data (e.g., individual notes on stairs tests or indisposition cases).

Individual medical data is measured by the network of personal devices that are equipped with medical sensors and connectivity modules. Recordings of health parameters are sent wirelessly to the smartphone. The patient is also able to input optional notes manually. A CardiaCare mobile app provides a simple analysis of ECG recordings and arrhytmia detection, risk factors assessment and in case of emergency situation sends an SMS to a specified mobile number to inform relatives or doctor. The application provides a simple feedback to the patient, e.g., advises to slow down physical activity or to visit a doctor.

The harvested data are sent to the server for further analysis on resource-intensive algorithms and storing. On the servers side, alarms can also be generated. Workplace for a doctor is also provided.

 ISBN: 978-1-61208-418-3

The high-level architecture of CardiaCare service is presented in Figure 1.

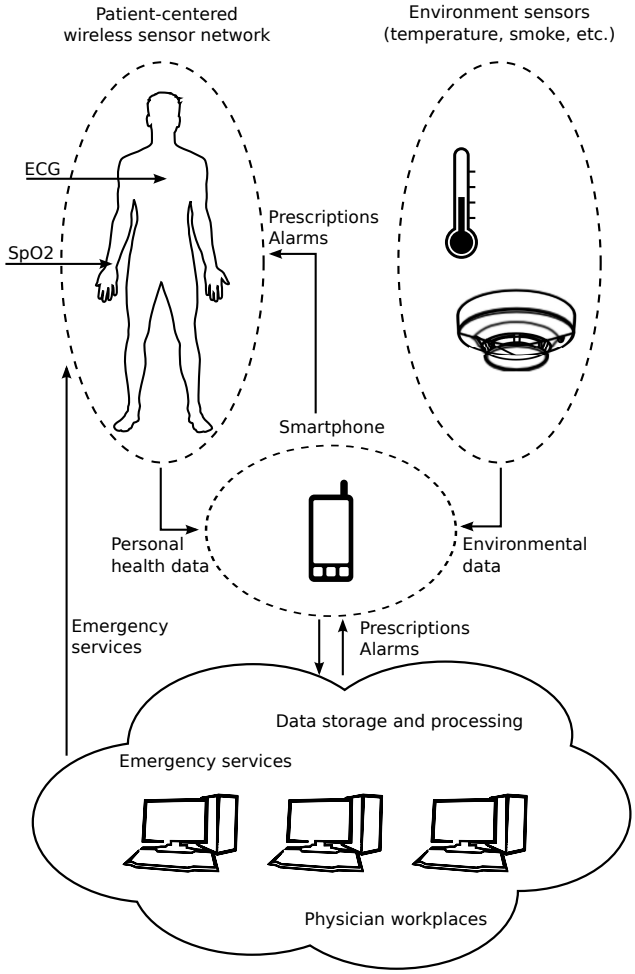


Figure 1. Using attribute dictionaries

For the purpose of improving the speed of emergency help, we propose the emergency service in terms of smart spaces paradigm on top of the CardiaCare service.

This service provides the healthcare personnel (ambulance services, volunteers, etc.) with the information of the emergency cases and location of the patients. The smartphones of participants are equipped with a software agent that has access to the risk factors and alarms generated by CardiaCare. In terms of smart spaces these agents are referred to as knowledge processors (KPs). Collecting and exchanging of the alarms and location of the participants is managed by semantic information broker (SIB). The emergency service provides the possibility of finding the closest care producer and assessment of the amount of time needed to reach the location of care receiver.

The architecture of the emergency service is presented in Figure 2.

From the description of the architecture follows that the number of sensors and other tests could vary during the life cycle of the service. Nevertheless, the design of the medical data storage should provide a possibility of flexible adaptation to these circumstances.

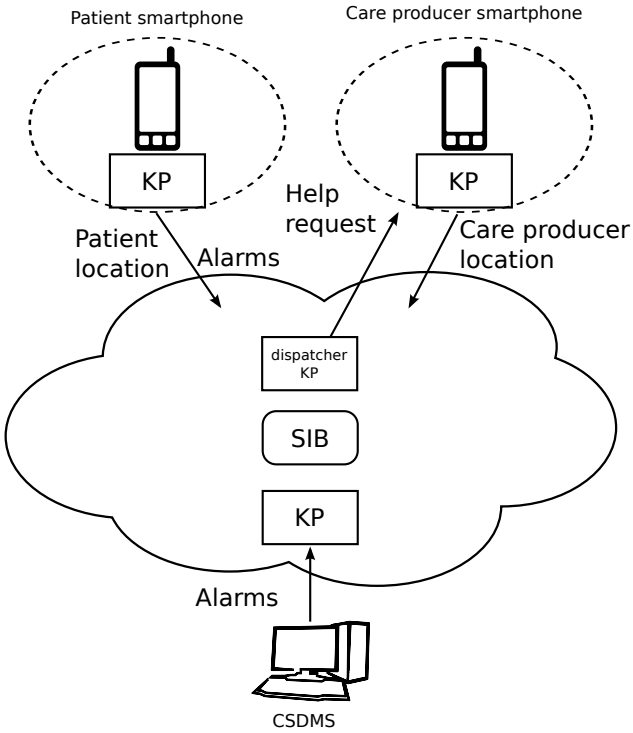


Figure 2. Using attribute dictionaries

III. ENTITY-ATTRIBUTE-VALUE DESIGN

In relational databases attributes usually are represented by table columns, one column per attribute. Nevertheless, this model is suitable when modeled entities have a fixed number of attributes and most or all of them have values for any given instance. If entities have potentially large number of attributes and, for a given instance, only a few of them are non-empty (applicable or known) then representation of attributes with columns leads to very sparse tables.

This situation is distinctive for clinical databases, when for any given patient only few vital parameters are actually recorded from hundreds of available ones.

A standard way of representing arbitrary information about some object is a set of attribute-value pairs, which become triples with the entity. There are generic formats to describe such triples:

- Extensible Markup Language (XML) provides a tool of describing complex hierarchical structures;
- Resource Description Format (RDF) operates with object-attribute-value triplets.

The attribute-value representation of the information can also be appropriate within a relational database.

The basic principle of EAV design is in the way of attributes representation. Unlike column-based attributes in relational model, in EAV attributes are row-modeled, one fact about entity per row [3].

EAV approach trades off the simplicity of physical representation and complexity of mapping to logical one and querying data. Search inefficiency is one of the drawbacks of the EAV approach. To eliminate this inefficiency several optimizations have been proposed [3] [4].

 ISBN: 978-1-61208-418-3

Commonly used optimizations can be described with the following procedure illustrated in (Figure 3).

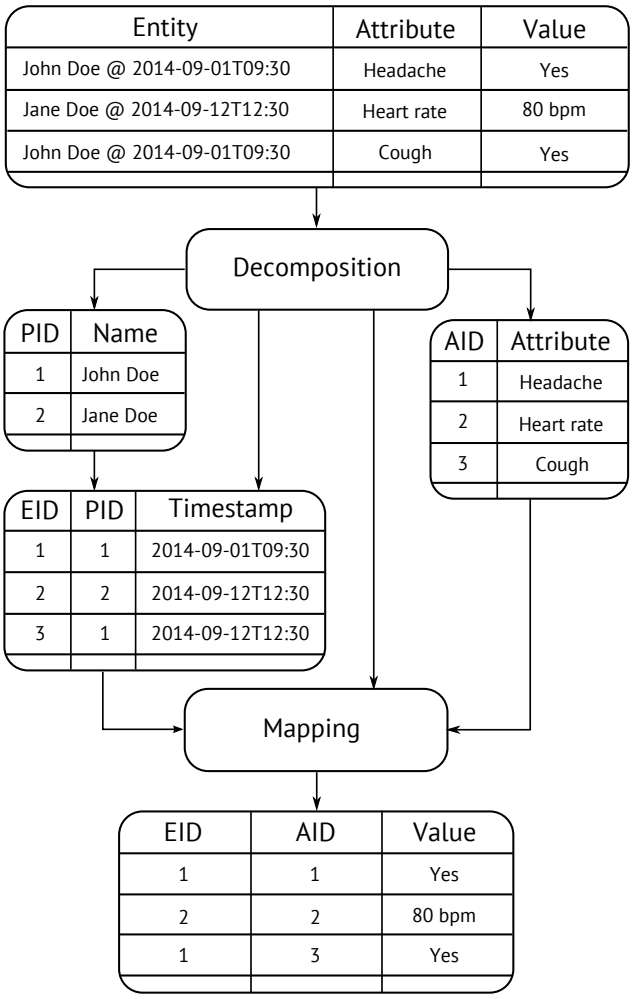


Figure 3. Using attribute dictionaries

In the first step, entities and attribute names are moved to separate dictionaries. Therefore, they are mapped to integer identifiers and the redundancy of duplicating attribute names is eliminated. In a second step, an auxiliary table is constructed to map values to entities and attributes.

Other optimizations propose additional indexes constructing, using, for example, binary representations of attribute and value, which are concatenated. In this solution, an index is constructed automatically sorting rows by attribute. It makes the model more optimized for ”read” operations than original EAV and it fits the requirements of most of CSDMSs, since, in clinical trials, we write to the database once, but the data is used potentially several times. Nevertheless, in the general case, this approach is not applicable.

In generic EAV attribute values should be of the same data type. Often, these are just strings. This produces several difficulties when constructing complex queries. There are two options to avoid this inefficiency.

The first option is to use several columns for different data types (Figure 4). Only one of the column values is allowed not to be null.

The second option is to store attributes of different types in different tables (Figure 5). This approach eliminates the redundancy of the first one, but the queries should be done over all these tables.

VID	EID	AID	ValueStr	ValueInt	ValueBool
1	1	1	NULL	NULL	Yes
2	2	2	NULL	80	NULL
3	1	3	NULL	NULL	Yes

Figure 4. One table for all attribute values

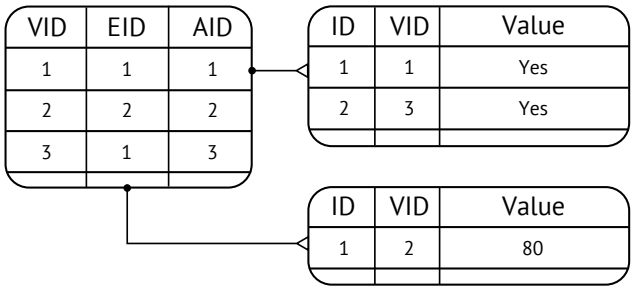


Figure 5. Separate tables for attributes of different types

The second option leads to more complex queries and increases the processing delays. Nevertheless, it helps to get rid of the numerous empty cells. Hence, the choice should be made depending of the nature of the data.

IV. EAV Approach in CardiaCare Database

A. Overview of CardiaCare schema

CardiaCare data model has hybrid design. Part of the tables have conventional relational column-based structure. Other tables are row-modeled and introduce the EAV approach.

Recordings of clinical trials heavily rely on predefined values of attributes that are organized as dictionaries. CardiaCare model supports compound attributes. Thus, an entity (e.g., medical trial) can be described not only by a set of attribute-value pairs, but also by a tree of attributes. Attribute dictionaries may also have hierarchical structure, as shown in Figure 6. As an example, one can consider some geo points dictionary with two-level (country-city) hierarchy.

There are two types of entities in CardiaCare data model:

- profiles of patients may be extended with arbitrary fields from contact details to billing information;
- clinical trials may be extended with arbitrary parameters and new types of trials may be incorporated as well.

Clinical trials can be considered not only as a hierarchy of parameters describing the event, but also as a form that consists of a hierarchy of fields to fill in to describe the event.

Attributes are described by metadata tables, as shown in Figure 7. The list of all applicable attributes is stored in ’attribute’ table. The names of attributes are listed here and their belonging to the entities is defined. Elsewhere, a machine-generated identifier is used to refer to the attribute.

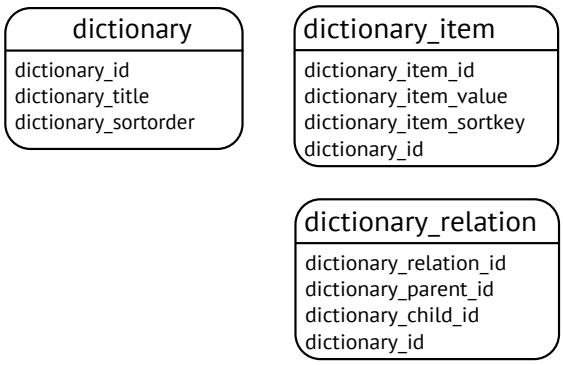


Figure 6. Hirarchical dictionaries in CardiaCare data model

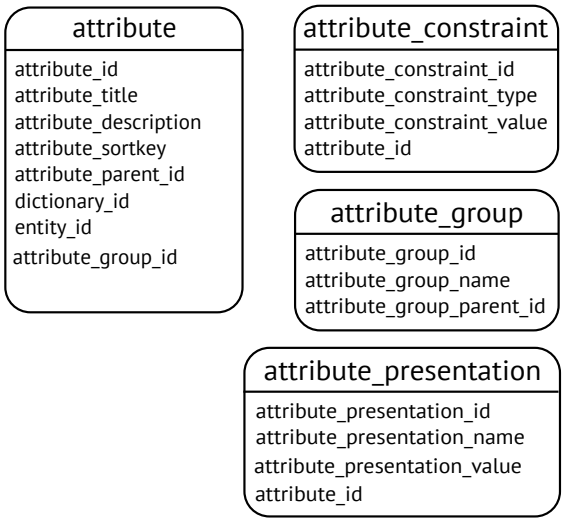


Figure 7. Attribute representation in CardiaCare data model

There are several other metadata tables, with the important ones being listed below.

- The 'attribute_constraint' table contains various metadata used for validation purposes, including data type checks, minimal and maximal allowed length of string values, allowed range of integer values, regular expressions, etc.
- The 'attribute_group' table contains attribute grouping for presentation to the user. Attribute groups can be nested.
- The 'attribute_presentation' table contains visual parameters of form fields for input forms and reports.

The aim of this auxiliary metadata is to provide an opportunity to construct user interfaces and REST APIs automatically.

V. Conclusion

Data model for CardiaCare service was conducted based on open schema design principles. This will allow to extend a set of supported medical measurements without the need of making changes to the database schema.

The schema was designed taking into account a possibility of automatic user and machine interface construction. Development of such framework is the purpose of the next stage of this research.

Acknowledgment

This research is financially supported by the Ministry of Education and Science of the Russian Federation: 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"

References

[1] Y. Apanasik, I. Shabalina, and L. Kuznetsova, "The Research Platform for the Medical Diagnostic Services Building," in Proceedings of the 14th Conference of Finnish-Russian University Cooperation in Telecommunications Program, Helsinki, Finland. 11.11.2013-15.11.2013, 2013, pp. 9–15.

[2] A. Borodin, A. Pogorelov, and Y. Zavyalova, "The Cross-platform Application for Arrhythmia Detection," in Proceedings of the 12th Conference of Finnish-Russian University Cooperation in Telecommunications Program, Tampere, Finland. 5.11.2012-9.11.2012, 2012, pp. 26–30.

[3] V. Dinu and P. M. Nadkarni, "Guidelines for the effective use of entity-attribute-value modeling for biomedical databases." I. J. Medical Informatics, vol. 76, no. 11-12, 2007, pp. 769–779.

[4] R. Paul and A. S. M. L. Hoque, "Search efficient representation of healthcare data based on the hl7 rim." Journal of Computers, vol. 5, no. 12, 2010, pp. 1810–1818.

 ISBN: 978-1-61208-418-3

Understanding Individual's Behaviors in Urban Environments

Claudia Liliana Zúñiga-Cañón[1,2] and Juan Carlos Burguillo[1]

[1] Information Technologies Group GTI, Department of Telematics Engineering
University of Vigo, Vigo, Spain.
[2] Research Group COMBA R & D, Department of Engineering
University of Santiago de Cali, Cali, Colombia.
Email:clzuniga@ieee.org,(clzuniga, J.C.Burguillo)@uvigo.es

***Abstract*—UrbanContext is an urban computing context abstraction model that follows an individual centered approach and validates the use of the Theory of Roles to understand the behavior of the individual within a social environment. The roles defined in UrbanContext allow the interpretation of the states of the individual, facilitating its interaction with the environment and offering services without damaging its privacy. In this paper, we mainly describe a real evaluation scenario for UrbanContext at the UBI platform in Oulu (Finland).**

Keywords–Urban Computing; Urban Context; Models; Theory of Roles; Smart Cities.

I. INTRODUCTION

Several projects have focused their works on improving cities, with the objective of making them more intelligent and ubiquitous. The ubiquitous city [1][2] is defined as a city with high technological interaction that has as goal to offer services and information at any place and time to its inhabitants.

These urban environments become spaces where persons, places and technologies converge. These three aspects form the so called triad [3] in urban computing. People become dynamic individuals and the main subject of study, if we want to offer people-centric services within an urban environment.

To model the context in the urban environment we face a big complexity due to the great number of variables involved in such spaces. This complexity demands techniques that allow the modeling and the representation of the individuals' behaviors in the cities.

Our contribution to face this problem is UrbanContext, a model for urban computing systems that uses the Theory of Roles [4] to manage context. UrbanContext facilitates the interpretation of the states of the individual, the development of adapted services and the generation of positive relationships. We propose a validation scenario in an adapted real environment, and evaluate our roles model approach.

The paper is organized as follows: The first section is an introduction about the context and the theory of roles in urban computing. The second section presents the roles model used in UrbanContext and describes the dynamic of the component. The third section presents an evaluation scenario for UrbanContext at Oulu (Finland), and finally, we conclude and present some future work.

II. THE CONTEXT AND THE THEORY OF ROLES

To model the urban environments it is necessary to consider the context. Schilit and Theimer [5] introduced the term *"context aware"*, and they consider the *"context"* as the location, identities of nearby people or objects, and the changes happening to these objects. Later, this definition was complemented in [6], where it was stated that the important aspects of context were: where you are, who you are, and what resources are near you.

One of the broader definitions was made by Dey [7], who defined *"context"* as any information that can be used to characterize the situation of an entity. These definitions were widely discussed and subsequently improved by Dourish [8], who indicated that to understand and model the context it was also necessary to involve social issues. Based on the latter idea, in UrbanContext we need a user-oriented approach that allows the identification of the individual's behavior and interaction within the urban environment.

The Academic Community has done big efforts to understand the different points of view present in urban environments. A representative example from the beginnings could be the project "Familiar Strangers" presented by Paulos in Intel Research [9], while recently we could mention the project "Urban Computing Middleware" funded by the South Korean Government [10], that tries to identify people, present in an environment, sharing the same likes in order to provide them with common services.

One of the challenges found in urban environments is the need to generate interacting spaces that allow to establish strong relationships among individuals. To achieve that it is necessary to understand the roles that people play in an urban environment in order to satisfy their real needs [11].

To face all these challenges, we use the Theory of Roles proposed by Erving Goffman [4]. This theory considers an individual who determines its behavior according to the role he/she plays in a certain situation. The individual's behaviors are influenced by the interactions he/she experiences, and is constantly changing the roles played within the social environment.

Therefore, the objectives of the UrbanContext, are to improve the interaction of the individual within the urban context, to identify its needs in every interaction and to provide services for it.

III. ROLES MODEL IN AN URBAN ENVIRONMENT

Understanding human behavior generates technical challenges for managing and classifying data with the required quality [12]. These challenges are related also with the issues to understand how the individual socially interacts, and how this interaction can be powered.

 ISBN: 978-1-61208-418-3

The individual interacts in physical and social environments that influence directly its behaviors, so we need to understand what he/she is doing, and what he/she really wants to do in specific physical places to estimate past behaviors in order to predict future ones. The roles model used in UrbanContext focuses on those challenges, and characterizes several individual states, defining if the user wants to be disconnected from the system or if it wants to play a certain role in a particular spatial or temporal scenario within the urban environment.

A. *The Management of the Context in the UrbanContext Model*

As we said before, we have created UrbanContext [13], which is a general-purpose model with a set of components, used for the design of urban computing platforms that applies the Theory of Roles to manage the individual's context in urban environments.

UrbanContext was designed from the result of previous projects and experiments developed in controlled and open environments of urban interaction. Our previous experiences have been performed gradually and were restricted to three major projects [14][15][16] (See Figure 1).

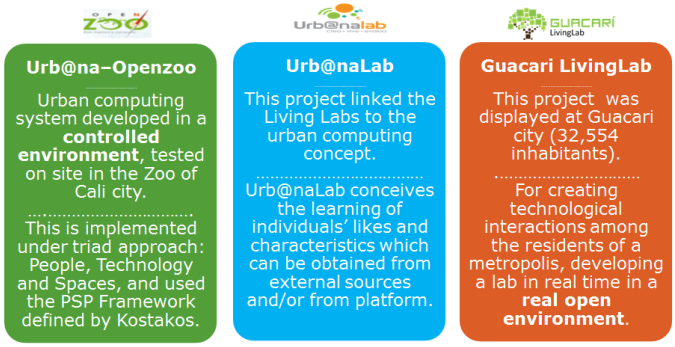


Figure 1. Previous Projects Developed in Urban Computing.

Openzoo [14] was the first project, and it was developed in an urban tourist controlled environment. Here, we worked along with anthropologists and sociologists analyzing a group of 400 individuals. After that, we had seven work sessions with focus groups of 45 users of different categories in a zoo. This allowed an ethnographic analysis of the individuals, the identification of their perception about the urban environment and the availability of technological resources.

Based on this first experience, we developed Urb@naLab [15], an urban computing platform that mixes the living lab concept in its design to allow the creation of collaborative spaces where the individual could participate actively in the co-creation of services. This originated a third project called Guacarí LivingLab [16], an Urb@naLab adaptation on an open real environment. LivingLab was considered as an ideal space to experiment, where we carried out three sessions with control groups of 45 people to observe and measure the interaction among individuals.

These previous experiments allowed us to identify the constant needs when designing urban computing platforms and led us to propose Goffman's Theory of Roles [4] to model the context of the individual. In order to describe it, we outlined our proposal with five components that we consider relevant to develop urban platforms (See Figure 2): interface, roles, semantic, cloud and services. The components of the UrbanContext model can be described as:

Interface Component, is the main external interaction point, uses several devices and technologies present in urban environments like mobile devices, augmented reality, etc.

Roles Component, is in charge of understanding human behaviors for modelling human states. This component allows the system to manage context information, to identify user interactions and to provide personalized services.

Semantic Component, is in charge of classifying all the information obtained, in order to be processed afterwards by algorithms to introduce a semantic level for reasoning.

Cloud Component stores all the data obtained from the individual and the urban atmosphere.

Service Component provides a set of adapted services to the individual needs.

In UrbanContext, the management of the context and the roles assumed by the individual are modeled through the Roles Component. The Roles Component is composed of four sub-components: urban agent, urban atmosphere, context and context management. Next, we describe the main aim of all those sub-components:

Urban Agent SubComponent, which is focused on the characterization of the user. It should collect all the information of the user directly from its interaction, as well as from different social sources he/she is related to.

Urban Atmosphere SubComponent, that allows filtering through all the data about the place, the environment, as well as the persons attending the space.

Context SubComponent, which is associated with the identification of the contexts in which the individual participates within the urban atmosphere. It considers that the individual can be in three different contexts: personal, social and global. This sub-component collects and identifies the roles of the individuals within every selected context.

Context Management SubComponent, that focuses on the semantic processing of all the data collected by the other components. It also achieves discovering services, establishes the logic of the individual's behaviors in the atmosphere and builds effective relationships.

B. *Roles Component Flow*

In UrbanContext we use a multi-tier approach to manage the context, which is divided in three main parts: global, social and personal. Thus, the Roles Component provides:

The ***global context*** is fixed to an individual in an open space of the urban atmosphere. This context manages what people share through a set of services around a public place: a concert, the meeting in a square or a spontaneous congregation.

The ***social context*** corresponds to what is obtained from socializing with other individuals; for example, friends, acquaintances and people the individual gets in contact with.

The ***personal context*** is based on the individual's own world, the one that is only available to it. In this space, the user is represented as a big bubble with some needs, fears, concerns, ideas and tastes.

 ISBN: 978-1-61208-418-3

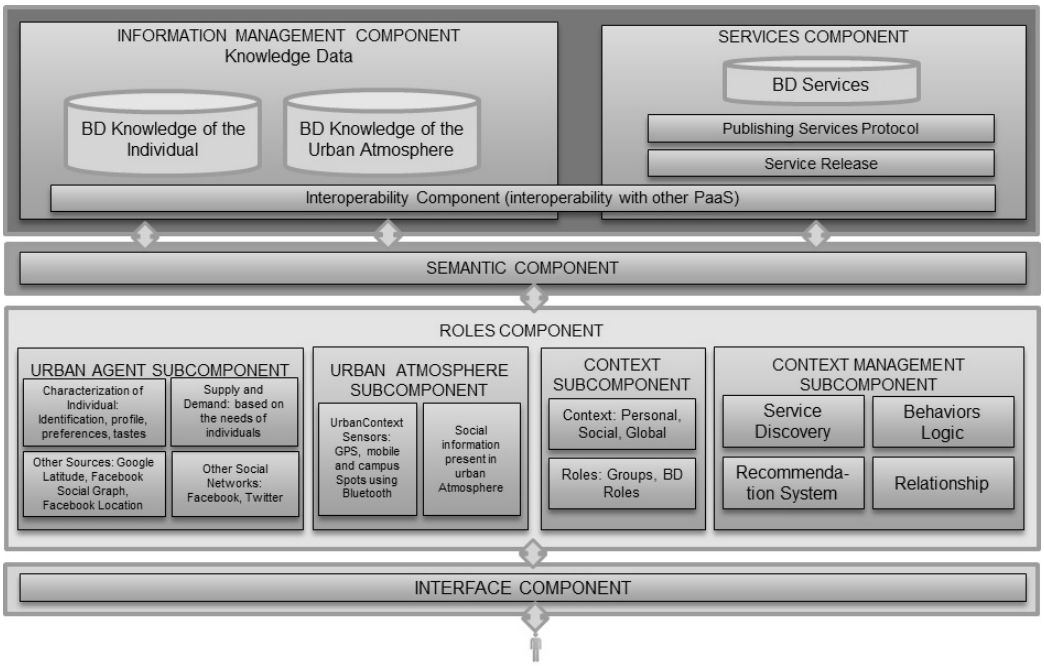


Figure 2. UrbanContext Model.

At this level of abstraction, we use a second ranking that allows to offer customization. For this approach, we used the Theory of Roles, which considers individuals always playing roles according to their situation. By applying this concept in UrbanContext, we establish that for every context an individual can play different roles that we can structure hierarchically as:

Role Groups: Personal, Family, Professional or Academic, Social and Rest of the World Roles.

Role Categories: Within each Role Group we have created several categories that identify the role the individual is playing, i.e., father, uncle, teacher, friend, or just himself.

Individuals can release public information and needs according to the role and the context they are at a certain moment. This information is fed into a knowledge base that identifies the role of individuals in an urban environment to be able to provide the right services adapted to their needs.

IV. Evaluation Process of UrbanContext

The evaluation process of the UrbanContext model considers three main steps. The first one identifies ***who is the individual***, the second one identifies ***what role is playing the individual*** and the third one predicts ***which services are relevant for the individual.***

The evaluation of the UrbanContext takes place at Oulu (Finland), where citizens have available a smart city platform, ***Open UBI(quitous) Oulu*** [17], which is accessible through displays scattered over the city. The UBI platform offers services like games, transport information, event information, maps, etc. (See Figure 3).

As the interaction with the UBI platform is done through smart displays and not with mobile devices, and taking into account that to identify the individual's roles it is necessary that participants are available online, we have designed a mobile application in Android to support all these issues.

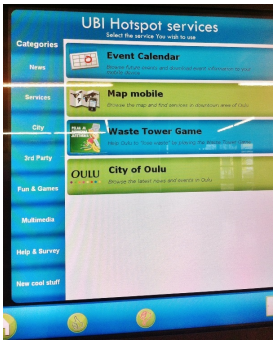


Figure 3. UBI Platform Interface.

The development of this application has taken into account technical and sociological factors. The application shall be able to consider different scenarios. For example, in many cases we need different services for a person when he/she is 20 years old, than when he/she is 35 years old.

The Android application can be freely downloaded by the participants in the evaluation. The application identifies the individuals and their devices. The application also stores a set of individuals' interaction records during a certain amount of time. Finally, the mobile application also includes an algorithm to manage the knowledge needed for the predicting future services (See Figure 4).

Concerning the three evaluation steps described above, they are performed as described next:

Individual and urban atmosphere information: the information is obtained through *UrbanContext Application*, uploading the Urban Agent SubComponent (name, age, sex, etc), and the Urban Atmosphere SubComponent (location in real time, place, etc).

 ISBN: 978-1-61208-418-3

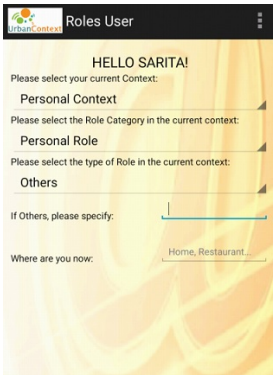


Figure 4. UrbanContext Android Application.

In Oulu, there exists demographic repositories with data coming from the users of the UBI platform. These data are combined with the data coming from the *UrbanContext Application* to obtain context enriched individual information.

Individuals' Roles: the individuals' roles are also captured through *UrbanContext Application*. The application was designed to start using a predetermined roles repository, which were obtained according by a previous study introduced in [18]. Using this predefined set of roles, an individual can choose at any time the role he/she plays from the default list or add new roles to the knowledge base.

Regarding the use of the mobile application, the user provides information about the actual context and the role it is playing when logging in for the first time. After that, the application enters a hibernation mode, appearing as an active icon on the screen of the mobile device. In hibernation mode, the user may at any time change its role by a simple touch. The application will use data mining algorithms to predict individual services based on the roles played by the individual.

Services prediction: the database obtained by the roles model, and the user interaction with the mobile device, allows to predict different services for resting or working time. For instance, if the user is resting at home at 6pm, the application can suggest a social network, but if the user is working at 9am, other work related services can be suggested instead.

While the knowledge roles' database is initially tuned by the user, the next times it adapts progressively to each user, and by means of data mining techniques (supervised classification techniques through decision trees), it is able to suggest the recommended services.

V. Conclusion and Future Work

In this paper, we present UrbanContext, a model for urban platforms that follows an individual centered approach. We consider the Theory of Roles to understand the individual's behavior within a social environment. We also define an evaluation environment in a real scenario to validate UrbanContext.

We can conclude that UrbanContext, and its roles model proposal, aim to understand the interpretation of the states of the individual, as well as its interaction with the environment. We consider that the urban atmosphere and the individuals' context directly influence the individuals' needs.

We have also realized, through the characterization of the individual and the urban atmosphere, that the roles model provides a knowledge base that facilitates the interaction and promotes positive relationships.

As future work, we plan to evaluate the UrbanContext model, at the UBI platform in Oulu (Finland), by means of our new Android mobile application. We also plan to collect enough data to measure the level of interaction of the individuals, in order to provide the adequate services according to the roles that individuals are playing at a certain time.

References

[1] J. Hwang, "u-city: The next paradigm of urban development," Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City, 2009, pp. 367–378.

[2] M. Foth, "From social butterfly to engaged citizen," Urban Informatics, Social Media, Ubiquitous Computing, and Mobile Technology to Support Citizen Engagement, 2011, p. Chapter 17.

[3] M. Foth, J. H.-j. Choi, and C. Satchell, "Urban informatics," in Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, ser. CSCW '11. NY, USA: ACM, 2011, pp. 1–8.

[4] E. Goffman, The presentation of self in everyday life, ser. Doubleday anchor books. Doubleday, 1959.

[5] B. Schilit and M. Theimer, "Disseminating active map information to mobile hosts," Network, IEEE, vol. 8, no. 5, Sept 1994, pp. 22–32.

[6] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications," in Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on, Dec 1994, pp. 85–90.

[7] A. K. Dey, "Understanding and using context," Personal Ubiquitous Comput., vol. 5, no. 1, Jan. 2001, pp. 4–7.

[8] P. Dourish, "What we talk about when we talk about context," Personal Ubiquitous Comput., vol. 8, no. 1, Feb. 2004, pp. 19–30.

[9] E. Paulos and E. Goodman, "The familiar stranger: Anxiety, comfort, and play in public places," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ser. CHI '04. NY, USA: ACM, 2004, pp. 223–230.

[10] J. Lertlakkhanakul, S. Hwang, and J. Choi, "Developing spatial information framework for urban computing environment: A socio-spatial computing framework for smart urban environment," in Management and Service Science, 2009. MASS '09, Sept 2009, pp. 1–5.

[11] T. Kindberg, M. Chalmers, and E. Paulos, "Guest editors' introduction: Urban computing," Pervasive Computing, IEEE, vol. 6, no. 3, July 2007, pp. 18–20.

[12] N. Oliver, "Urban computing and smart cities: Opportunities and challenges in modelling large-scale aggregated human behavior," in Human Behavior Understanding, ser. Lecture Notes in Computer Science, A. Salah and B. Lepri, Eds. Springer Berlin Heidelberg, 2011, vol. 7065, pp. 16–17.

[13] C. L. Zuniga-Canon and J. Burguillo, "Urbancontext: A management model for pervasive environments in user-oriented urban computing," Computer Science, vol. 15, no. 1, 2014, pp. 75–88.

[14] C. L. Zuñiga and et al., "Design of service-oriented pervasive system for urban computing in cali zoo (openzoo)," World Academy of Science, Engineering and Technology, vol. 4, no. 3, 2010, pp. 990 – 995.

[15] C. Zuniga and et al., "Software platform for services in colombian cities using the living labs approach," in GLOBECOM Workshops (GC Wkshps), 2011 IEEE, Dec 2011, pp. 1258–1262.

[16] "CO-T1199: Wireless Networks and Digital Inclusion Services in the Municipality of Guacarí, Sponsored by the IDB (Interamerican Development Bank) and the Italian Trust Fund of Information and Communication Technology for Development," 2011, URL: http://www.iadb.org/en/projects/project-description-title,1303.html?id=CO-T1199 [accessed: June, 2015].

[17] "UBI, Open Ubiquitous Oulu-Finland," URL: http://www.ubioulu.fi/en/ [accessed: June, 2015].

[18] C. Zúñiga-Cañón and J. Burguillo, "Applying data mining in urban environments using the roles model approach," in Advances in Artificial Intelligence – IBERAMIA 2014, ser. Lecture Notes in Computer Science, A. L. Bazzan and K. Pichara, Eds. Springer International Publishing, 2014, vol. 8864, pp. 698–709.

 ISBN: 978-1-61208-418-3

Adaptive Streaming Scheme for Improving Quality of Virtualization Service

Sunghee Lee and Kwangsue Chung
Department of Electronics and Communications Engineering
Kwangwoon University
Seoul, Korea
e-mail: shlee@cclab.kw.ac.kr, kchung@kw.ac.kr

Abstract— **The streaming-based virtualization services require Quality of Service (QoS) support for achieving a seamless display and low latency. Dynamic Adaptive Streaming over HTTP (DASH) has been proposed to support QoS of multimedia transmission. Existing bitrate adaptation schemes based on DASH are unsuitable for virtualization services due to latency problem. This paper proposes a DASH based adaptive streaming scheme to improve the QoS of virtualization service. The proposed scheme provides seamless display by adjusting the quality of the segment based on the segment throughput and the buffer status. It also reduces latency by using the server push mechanism of HTTP 2.0. The simulation results show that the proposed scheme has a better performance.**

Keywords- virtualization service; DASH; QoS.

I. INTRODUCTION

With the introduction of fast and reliable core networks and wide-spread availability of Internet access, a trend towards moving more and more services away from the end devices to remote data centers has established itself. This is widely referred to as streaming-based virtualization service. This results in greatly increased requirements on Quality of Service (QoS) to achieve a seamless display and low latency.

Meanwhile, Dynamic Adaptive Streaming over HTTP (DASH) has been proposed to support QoS of multimedia transmission. DASH can support seamless display by handling the bitrate of contents based on the time varying bandwidth conditions. In DASH, the server response depends on the client's requests when the server is otherwise idle or blocked for that client. To adapt the bitrate of the content according to the network status, the content is divided into short-duration media segments, each of which is encoded at various bitrates and can be decoded independently. During download, the client dynamically picks the segment with the right encoding bitrate that matches or is below the bandwidth, and requests that segment from the server.

Several bitrate adaptation schemes, such as Rate Adaptation for Adaptive HTTP Streaming (RAHS) and Adaptive Streaming of Audiovisual Content (ASAC) have been proposed for improving QoS of DASH. These schemes adjust bitrate of segment based on the ratio of Media Segment Duration (MSD) to Segment Fetch Time (SFT) [1][2]. However, these schemes cannot guarantee the quality of virtualization service because ratio of MSD to SFT cannot estimate the network status precisely due to VBR characteristics of contents. Moreover, DASH clients receive a manifest file, request and download the referred segments over HTTP, and play them back. This procedure introduces additional latency making HTTP streaming unsuitable for virtualization service that requires low latencies [3].

To reduce the latency of DASH, the server push based streaming scheme has been proposed. The server push mechanism pushes a resource directly to the client without the client request. However, server push is not suitable for DASH as the DASH adjusts the bitrate of contents based on the client request. In this paper, we propose a DASH based adaptive streaming scheme to improve the QoS of virtualization service. The proposed scheme provides seamless display by adjusting the quality of the segment based on the segment throughput and the buffer status. It also reduces latency by using the server push mechanism.

The rest of this paper is organized as follows: in Section II, we describe the concepts and algorithms introduced in the proposed scheme. In Section III, we show the simulation results. Conclusions and future works are presented in Section IV.

II. ADAPTIVE STREAMING FOR VIRTUALIZATION SERVICE

A. Overview of adaptive streaming scheme

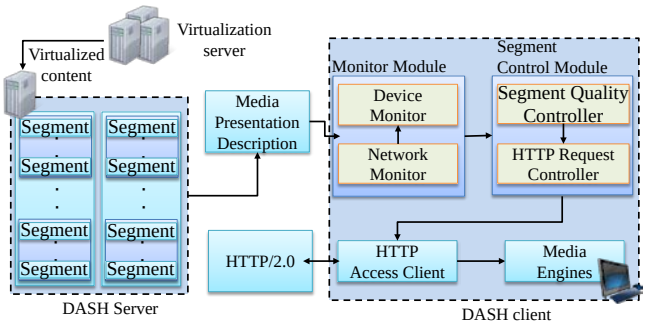


Figure 1. Architecture of adaptive streaming for virtualization service

Figure 1 illustrates the architecture of adaptive streaming for a virtualization service. In this architecture, HTTP 2.0 is used for server push. The proposed system performs segment quality control and HTTP request control algorithm to

 ISBN: 978-1-61208-418-3

reduce playback discontinuity and latency. In order to achieve this, information related to network and device status is monitored at the client. *Device Monitor* measures the remaining frames in a playback buffer, and *Network Monitor* calculates the segment throughput. The network and device information are forwarded to *Segment Quality Controller*. *Segment Quality Controller* decides the bitrate of next segment and *HTTP Request Controller* decides when to send HTTP request to server for preventing playback discontinuity.

Figure 2 depicts the behavior of the proposed adaptive streaming scheme. At the beginning of the service, the client requests virtualized content which has the minimum bitrate because there is no information about the network status and the playback buffer is empty. Then, the server pushes the segments until the client requests another bitrate of virtualized content. Based on this approach, the proposed scheme is better able to reduce latency and prevent playback discontinuity.

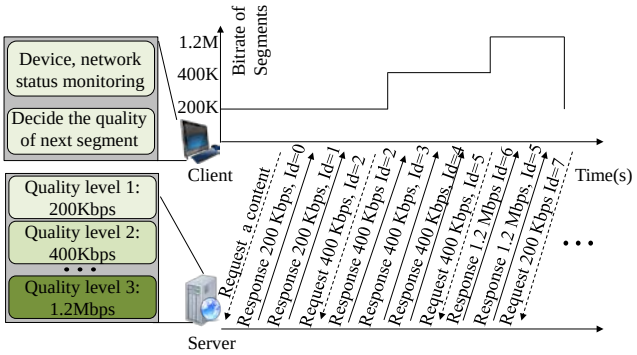


Figure 2. Behavior of the proposed adaptive streaming sceme

B. *Quality adaptation algorithm*

To provide seamless service, the proposed scheme adjusts timing to send a HTTP request and bitrate segments according to the network and device status. After receiving a segment, the client estimates segment throughput as follows:

$$Th_{seg} = \frac{S_{seg}}{SFT} \quad (1)$$

where Th_{seg} is the segment throughput, S_{seg} is the size of a segment, and SFT is the segment fetch time. Therefore, the video bitrate of the next segment is decided on the basis of the SFT of the most recently downloaded segment. To reduce the latency caused by the HTTP request from a client, the proposed scheme sends HTTP request only when the quality change is required. The quality level increases when the Th_{seg} is higher than the bitrate of currently downloaded segment to improve the quality of the video. On the other hand, the quality level decreases when Th_{seg} is lower than the bitrate of currently downloaded segment and there is not enough data in the playback buffer.

The condition of quality increment is as follows:

$$Th_{seg} > R_{cur+1} \quad (2)$$

where R_{cur+1} is the bitrate of the next higher quality level. If segment throughput is larger than the bitrate of next higher quality level, the proposed scheme increases a quality level as follows:

$$R_{cur} = R_{cur+1} \quad (3)$$

where R_{cur} is the bitrate of current quality level. We use stepwise increment of quality level to prevent buffer underflow when the available bandwidth is drastically decreased after increasing the quality level. The proposed scheme involves switch down the quality level when the amount of playback buffer is not enough to play the video until receiving next segment in the current quality level.

The condition of quality decrement is as follows:

$$T_{buf} < SFT \quad (4)$$

where T_{buf} is the amount of playback buffer in time. If the SFT is larger than the amount of playback buffer, the buffer underflow will occur before receiving next segment. Therefore, the proposed scheme selects the segment that has the maximum video bitrate among video qualities which can be downloaded during the remaining playback buffer time as follows:

$$R_{cur} = \max\{R_n \mid R_n \leq \frac{T_{buf} \times Th_{seg}}{MSD}\} \quad (5)$$

where R_{next} is the bitrate of next segment, R_n is the bitrate of n-th quality level.

III. Simulation Results

To evaluate the performance, we have implemented the proposed scheme in a DASH reference player developed by DASH Industry Forum (IF). We compared the amount of playback buffer among the proposed scheme, RAHS, ASAC, and default adaptation algorithm of DASH IF player. Figure 3 shows that the proposed scheme stably maintains the buffer level without buffer underflow because the quality is adjusted on the basis of the segment throughput and buffer status.

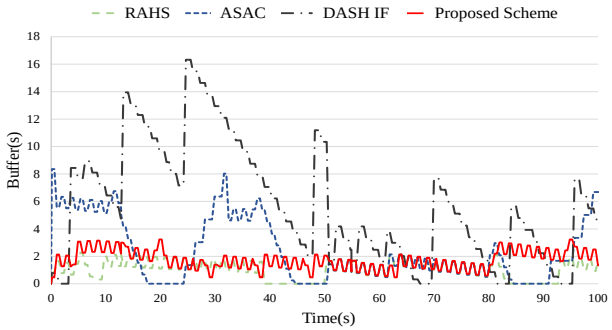


Figure 3. Comparison of the amount of playback buffer

We compared the latency between the proposed scheme with server push and one without server push mechanism. Figure 4 shows that the latency is evidently decreased when the server push mechanism is deployed.

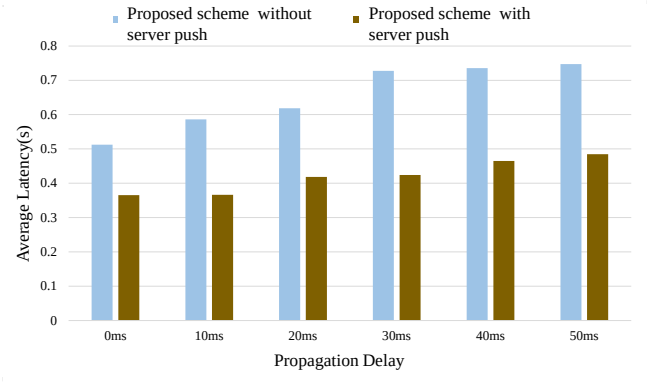


Figure 4. Comparison of latency

IV. Conclusion and Future Work

DASH can improve the QoS of streaming-based virtualization service by adjusting the quality of content according to the network condition. However, HTTP request/response procedure of DASH introduces additional latency which degrades the quality of virtualization service. In this paper, we propose an adaptive streaming scheme for DASH to improve the QoS of streaming-based virtualization service. The simulation results show that the proposed scheme provides seamless playback and low latency by adjusting the quality of content based on the server push mechanism of HTTP 2.0. In the future work, we will enhance the proposed adaptive streaming scheme to reduce the frequent quality change due to the varying network condition.

Acknowledgment

This work was supported by the ICT R&D program of MSIP/IITP. [R0112-14-1059, Development of Reducing Power Consumption Technology for Pay Broadcast Set-top Box]

References

[1] T. Thang, Q. Ho, J. Kang, and A. Pham, “Adaptive streaming of audiovisual content using MPEG DASH,” IEEE Trans. Cons. Elec. Fukushima, vol. 58, no. 1, pp. 78-85, February 2012.

[2] S. Garacia, J. Cabera, and N. Garcia, “Quality-optimization algorithm based on stocastic dynamic programming for MPEG DASH video streaming,” IEEE Int. Conf. Cons. Elec. Madrid, pp. 574-575, January 2014.

[3] S. Wei and V. Swaminathan, “Low latency live video streaming over HTTP 2.0,” ACM Work. Net. Oper. Sys. Sup. Digi. Aud. Vid. San Jose, pp. 37-42, March 2014.

 ISBN: 978-1-61208-418-3

Pervasive Social Network

Alexiei Dingli

University of Malta
Msida, Malta
Email: alexiei.dingli@um.edu.mt

Daniel Tanti

University of Malta
Msida, Malta
Email: daniel.tanti.11@um.edu.mt

***Abstract*—Social Networking sites like Facebook and Twitter have become extremely important and are used by millions of people worldwide. In addition, the advent of mobile technology coupled with advances on the communications front, means that technology has started to move away from the traditional desktop setting and is becoming more pervasive. This research aims to show how a big screen setup in a public space, displaying a stream of comments from an online social network, can be utilised by the general public. The goal is to find whether such a system is able to instigate discussions between people - both physically and virtually, on the social network. The evaluation of privacy concerns, related to such a system in comparison with traditional social networks, will also be an important focus of this research. Similar work has already been done in particular contexts such as a classroom or a conference, however we aim in finding specific uses for such a system where the context is not as clearly defined. In order to achieve this, we created a social network called Occupy. The study described in this paper took place at the University of Malta, where a big screen projecting the stream of comments from our online social network was setup for discussion among those people on campus and those from outside. 66% of the users of our system believe that a pervasive social network adds value to traditional social networks, mainly by merging virtual discussions happening on the social network with physical discussions between groups of people. Through the use of a survey, analysis of the collected data and a focus group, benefits regarding the use of a pervasive social network can be presented.**

Keywords–Pervasive Technology; Social Network; Comments Stream; Big Screen; Privacy

I. INTRODUCTION

In the last few years, technology has started to shift from the traditional desktop setting and is becoming more pervasive. The rapid development of mobile technology combined with advancements in communication capabilities meant that people can have access to technology wherever they may be, at all times. This new kind of technology is referred to as Pervasive Technology and examples of this can be clearly seen in modern devices such as smart phones and tablets. These devices have excellent computational capabilities and are network-enabled. This means that their users are constantly connected to the internet and to each other. Social Networking Websites like Facebook, Twitter and Google+, have become extremely important over the years and are used by millions of people worldwide. In addition, Social Networking Websites are now starting to exploit the pervasiveness aspect of technology by developing mobile applications [1]. These applications allow users to be constantly connected with the social network through their pervasive devices.

A pervasive social network is an extension of the traditional social network. The most important aspect borrowed from the traditional social network is the recent intrusion in the field of mobile technology - mobile social networks. In addition, the basic structure of posting, retrieving and rating of comments is also common to both types of social networks. A pervasive social network also provides a mechanism for the creation of connections with other users as well as the browsing of these connections.

On the other hand, there are a variety of differences between the two types of social networks. The first of these differences is the fact that comments are displayed on a physical big screen placed in a specific location. This means that the discussions are not restricted to those users who are participating online. Instead, anyone who happens to be near the big screen may follow the discussions and participate.

In addition, the fact that the screen is tied to a particular location, means that the topics of the discussions might be "*hijacked*" by the context of the screen. Furthermore, the pervasive social network will post specific comments itself to try and instigate discussions between its users. In addition, the pervasive social network will try to suggest users in the vicinity who may be interested in starting a particular discussion by providing the location of the particular users.

The research that will be presented in this paper aims at finding new ways of extending social networks, so that they exploit the functionality offered by pervasive technology. We aim at identifying ways in which a big screen set up in a public space, displaying a stream of comments, can be used by the general public. Moreover, we aim to describe the reaction of the general public to such a system, in comparison to the way they normally use traditional social networks. The evaluation of the issues related to privacy as mentioned in [2] and [3] is also one of the main focuses of this research.

The reminder of the paper is structured as follows. Section II presents the Aims and Objectives, followed by the Literature Review in section III. In section IV we explain the Methodology with the Evaluation in Section V. Finally we present the conclusion and future work.

II. AIMS AND OBJECTIVES

The research question for this project is the following: "How can a big screen set up in a public place, displaying a stream of comments, be utilised by the general public?". The following is a list of goals that must be achieved in order for this project to be successful.

 ISBN: 978-1-61208-418-3

1) **Identify Uses:** The main aim of our research is to identify ways in which such a pervasive social network can be utilised by the general public. Moreover, we aim to describe the reaction of the general public to such a system, in comparison to the way they normally use traditional social networks.
2) **Merging Virtual and Physical Interactions:** Another important goal is to assess whether a pervasive social network can be successfully used to merge virtual interactions happening on the social network, with physical interactions between groups of people. Furthermore, the social network we add further information to the social graph such as location, thus allowing nearby users to find each other.
3) **An Active Social Network:** To further enhance the quality of the discussions, the proposed social network must be able to play an active role in the users' discussions, and so we will assess the users' reaction to such interactions.
4) **Privacy Concerns:** The evaluation of the issues related to privacy is also one of the main focuses of this research. We will compare the users' perception of our social network in comparison to other social networks.

III. Literature Review

A. Social Networking

A social network is a web-based service that enables a user to hold a public (with optional limitations) or semi-public profile. It also allows the user to maintain a set of connections with other users with the added possibility of viewing and managing his/her list of connections and connections made by others [4]. The way this web service is handled varies from one social network to another [5] [6] [7]. The main goal behind social networks, apart from allowing individuals to meet other people, is to provide the users a way to create and make public their own social networks.

There are many different types of social networks available on the internet, some of which are built for a specific purpose, while others are built for general connections. The most popular general purpose social networks are Facebook[1], Twitter[2] and Google+[3].

The normal procedure of interacting with a social networking website, is to first create an account. This can be done in a number of ways, but the most popular format is to provide answers to specific questions posed by the website. This helps in creating a web-based profile of the individual for other users to see and react to. Some social networks even allow the user to upload a profile picture or an avatar of themselves [8]. The users can then create links with other users which can be bi-directional on some websites or unidirectional on others. Moreover, users are also able to post content on the website which may be viewed by other users on the social network. Additional features such as the posting of photos and tagging of users may be present in certain types of social networks.

For social networks to be useful to our purpose, we must find ways of interacting with them and gaining access to as much information as possible from these valuable sources. The reason why we need to interact with these social networks, is that the users provide information about themselves and the connections that they have with others. More importantly, we can make use of existing, tried-and-tested technologies for developing a social network, without having to reinvent the wheel [9]. Several of the most popular social networking web sites are now launching what are called social network connect services [9]. Some examples of these connect services are Facebook Platform[4], Google+ API[5] as well as Twitter's API[6].

B. Pervasive Technology

The main idea behind the pervasiveness of technology is that as the years go by, technology and communication capabilities would be found in every environment imaginable, while at the same time, they are able to integrate seamlessly into the human users' everyday life [10]. It is now extremely common to integrate computing devices into anything electronically-based: we have programmable fridges and washing machines, smart phones and even smart TVs. Furthermore, social networks are also becoming pervasive through the development of mobile applications.

In Pervasive Computing, we can outline the accomplishments and the remaining challenges in these main areas: Context-Awareness, Automated Capture and Access to Live Experiences, Privacy, Time and Natural Interfaces [11].

A phenomenon that has happened in most of the major cities in the world and is now also spreading to Malta is what is known as *Digital Signage*. This is a fast growing market which aims at replacing the traditional poster billboards with electronic public displays [12]. This can already be seen in urban areas such as Shibuya Crossing, Tokyo and Times Square, New York, where the landscape is filled with large displays showing adverts from major companies [13]. Until now, these screens have been used for marketing purposes; however in this project we will propose a social network that can make use of such public displays.

C. Related Work

The literature we have found is mainly related to education, more specifically, the context of a conference [14]. Our system will not be tied to a particular context and as such, it will be designed differently and we expect that it will also behave differently.

The first interesting aspect that we noted out of the various research papers that we studied [15] [16] was the fact that most of them used a microblogging type of social network such as Twitter. However, we believe that by developing a custom social network, we will have more control over the discussions. Moreover, we will also be able to include more features that may be of interest to our research, but are not present in existing tools.

The first of these experiments was conducted at the ED MEDIA 2008 Conference in Vienna [17]. The focus of this study was to establish whether a microblogging website can be used to enhance a live event. To address this issue, a Twitter stream was set up during the ED MEDIA 2008 conference,

[1] https://www.facebook.com
[2] https://twitter.com
[3] https://plus.google.com
[4] https://developers.facebook.com
[5] https://developers.google.com/+/
[6] https://dev.twitter.com

 ISBN: 978-1-61208-418-3

and all the conference participants were invited to follow this channel and participate. In addition, this Twitter stream was projected during the conference's keynote session and breaks, using an application called TwitterCamp [18], so that the online conversation among members of the audience using Twitter, could be displayed to the rest of the audience that were not using Twitter. The participants of this experiment, were asked to append the #edmedia08 hashtag in order to group the comments related to the conference together.

This experiment was redone a year later at the ED MEDIA 2009 Conference with a similar setting to the one done in 2008. However, an attempt to engage the audience more by keeping the Twitter Feed on for the entire duration of the conference was done [19]. In addition, the system was given more publicity than the year before in order to attract a larger audience. Another change from the previous year was the use of the #edmedia tag instead of the #edmedia08.

The following is a list of categories of comments identified by various studies [17] [20] [19] conducted on such experiments:

1) **Concerning the presentation**: Comments directly related to the presentation or any of the presenters.
2) **Discussion**: Interaction between two or more users.
3) **Links**: These comments contain links to online content that may or may not be relevant to the presentation.
4) **Comments**: This category encompasses feelings, thoughts and opinions of the members of the audience that are not necessarily related to the presentation.
5) **Establish Online Presence**: This is a very interesting point which can be defined as posting for the sake of posting.
6) **Pose Organisational Questions**: These are questions related to the logistics of the conference and its proceedings but not directly linked to the conference's topic as such.
7) **Exchange of Social Activities**: This category entails the setting up of social activities outside of the conference with other members of the audience. An example given by Ebner and Reinhardt [19] is that of inviting another person to go sightseeing.
8) **Arrange Short Meetings**: This means that users used Twitter to arrange meetings amongst themselves to perhaps continue their discussion about the conference privately.
9) **Documentation of Conference Activities**: Linked to the sharing of resources highlighted by [17] [20] but is related to the actual resources used for the conference.

All of these experiments tested the activity on the Twitter channel, before, during and after the conference. It is interesting to note that all experiments show an increase in participation when the conference starts in comparison to the activity before, and a decline in participation after it ends. This greatly suggests that such a system is indeed effective in creating discussion during conferences and live events.

IV. METHODOLOGY

For the purpose of this research, we created a pervasive social network called Occupy which is made up of two main parts. The first part is a traditional online social network that enables users to communicate through the use of comments. Furthermore, this social network enables users to create connections with other individuals, most importantly, those users that are within their vicinity. The second part of this social network is another website, which we refer to as the *Interactive Wall*. This website is responsible for extracting a stream of comments posted by the users of our system. This stream of comments is then displayed onto a big screen which is placed in a public space. Apart from displaying the stream of comments, the *Interactive Wall* will also try to play an active role in the discussions by displaying articles relevant to the current discussion on the wall (which are selected from news sources using artificial intelligence techniques).

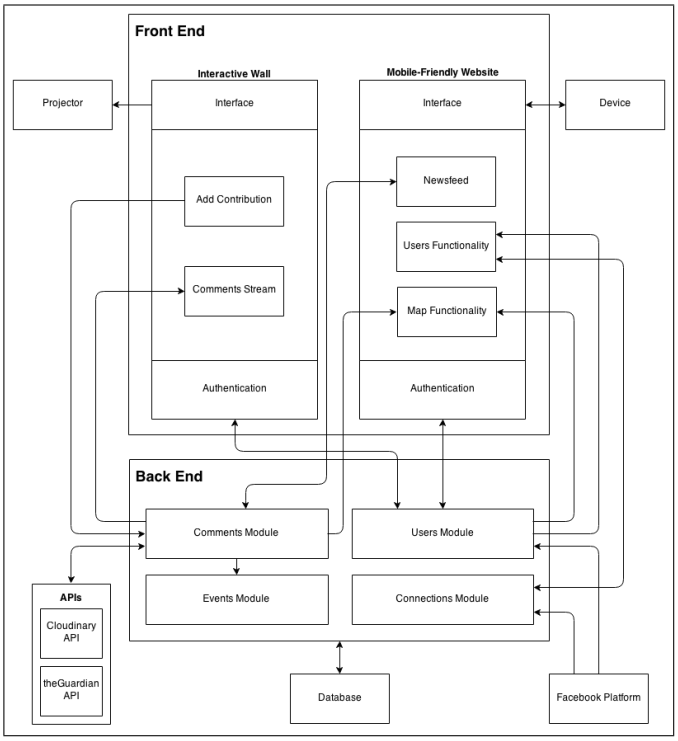


Figure 1. High-Level Block Diagram of the system

This system's design is split into two main partitions - a back end and a front end - as shown in Figure 1. These two partitions, which are further subdivided into modules, communicate with each other using the REST architecture[7]. This type of architecture enables the separation of specific roles between the server and the clients, by restricting the communication to standard methods.

The back end of the system - also referred to as the server - is responsible for handling all the functionality related to data. This is done in order to encapsulate the inner complexities of the system from the front end applications, so that they can then focus entirely on the graphical user interface and general client-side functionality. In addition, it also allows cross-platform compatibility since the client-side applications do not have to use a particular programming language to access and

[7]Representational State Transfer (REST) is a programming architectural ideology that works on the principle of sharing references to the data rather than a copy of the data itself

 ISBN: 978-1-61208-418-3

manipulate data. The front end of the system - also referred to as the client - is responsible for exposing the system's feature in a user-readable format, namely the two websites mentioned. The front end does not perform any complex functions and more importantly it does not communicate directly with the database. It communicates with the back end in order to delegate the necessary functions requested by the users. Web sockets were also used to facilitate the exchange of certain types of real-time messages between the two parts.

The social network website includes three main features. The first feature is that of a *Newsfeed* where all the comments posted by the users of this system are displayed, sorted by a custom ranking algorithm that takes into consideration the number of *likes*, *dislikes* and comments as well as the time elapsed since its creation. This was done in order to keep the discussions flowing. The *Newsfeed* allows the users to contribute to the discussions by posting either a completely new comment or a reply to a previous comment. These contributions can also include an image. To give the users some measure of control over their content, the possibility of deleting comments was also included. Similar to the way other social networks operate, the comments are saved with the Global Positioning System (GPS) coordinates of the location from where the comment was created. However, to protect the privacy of the users, this feature can be switched off.

The second feature is that of a *Users* page. This allows users to see a list of people that have joined this social network, highlighting those that are currently *online*. To further protect the privacy of the users, only basic public information, such as the name, gender and locality, are displayed. Apart from giving the ability to see *Users* and *Friends* (and also creating new friendships), this system includes a feature that shows an individual what users are currently in his/her physical vicinity (*Nearby Users*).

The third feature is a map that is able to show the locations of some of the latest comments posted. Of course, since the users can opt not to share their location, only those comments originating from users giving the system their consent to track them are shown. Apart from the comments, the map also displays the location of the users that have agreed to share their location. This feature, together with the *Nearby Users* feature, were created in an attempt to investigate whether such a system is able to merge virtual discussions on the social network with physical discussions. The hypothesis is that by utilising this feature, people may confront a person (whom they may not know) posting on the social network, to continue the discussion privately.

The next part of our pervasive social network is the *Interactive Wall*, which is a website containing a stream of comments from our social network, that is displayed on a big screen. This wall is the main focus of our research, as it is the enabler of discussions, both physically and virtually. To further enhance this social network, we decided to make the Wall active in the discussions. This is done by allowing the social network to scan the different keywords attached with the comments posted on the system and retrieve a related article from an online source. This article is then posted to the social network as a contribution to the discussions. We will study the effect of such a system over the general public and what, if any, their reactions will be to such contributions coming from the system itself.

V. Results and Evaluation

A. Evaluation Methodology

An experiment was carried out on the University of Malta campus, where a big screen showing our social network's interface was set up in a prominent place on campus. This experiment was held during the second week of January over a stretch of five days. It is important to note that this was during an exam period and so it might have introduced some bias to our study. Data obtained from this experiment was used as a valuable source of information for the purpose of this research, however we believe that this data alone is not enough to be able to draw conclusions. Similar to the previous work done in this area of research [19] [17] [16], we decided to conduct an online survey with the users of our system. Questions relating to the usage of this system as well as any concerns for privacy were put forward so that we would gather a general understanding of the public's view of our system and be able to draw conclusions based on this information. In total, the survey consisted of 21 questions that were split into three main sections – questions relating to demographics, to the usage of our system, and to privacy concerns. Apart from the survey, we also conducted a number of interviews during the experiment, so as to gather a better understanding of the users' first impressions in relation to our system. The same questions used for the online survey were asked during these interviews.

The last part of our research consists of a focus group, where a number of questions related to our system and how it compares to other similar systems were posed. The discussion generated in this focus group is also a major part of the evaluation of our research. The participants of this focus group were chosen based on their activity during the experiment and a total of six individuals took part.

B. Results

The response to our university experiment was highly satisfactory. A total of 425 unique users registered to our social network, 34% of which were females. Interestingly, only 35% of these users were active participants in this experiment and provided a total of 422 comments. In comparison to Facebook, our social network was outperformed in the percentage of users that are active in the discussions. In fact, 60% of all the users registered with Facebook are active participants in discussions [21]. We believe that the reason for this, is that people might feel apprehended when trying to post a comment to this social network. Other social networks may give a false sense of security and so people express their opinions freely without thinking that other users will see their comment. On the other hand, our system makes it explicit that whatever you will post, will inevitably be seen by a large number of people. Another point raised during the focus group is the fact that our social network is public. In other words, whatever a user posts will be seen by all the other users indiscriminately, while also including those individuals that are not part of our system, but can simply view the comment on the Interactive Wall. This further adds to the apprehension that users may feel, since they are sharing their information not only with their circle of friends but with virtually anyone who is either using the social network or is standing near the screen.

276 of the comments posted, introduced a unique topic to the social network, while the rest were a continuation

 ISBN: 978-1-61208-418-3

of previous discussions. The vast majority of the comments posted (94%) were text-only comments, while the remaining 6% included images. Figure 2 shows the usage of our system over a period of time. This includes a day before the actual start of the experiment, and a period of three days after it had ended.

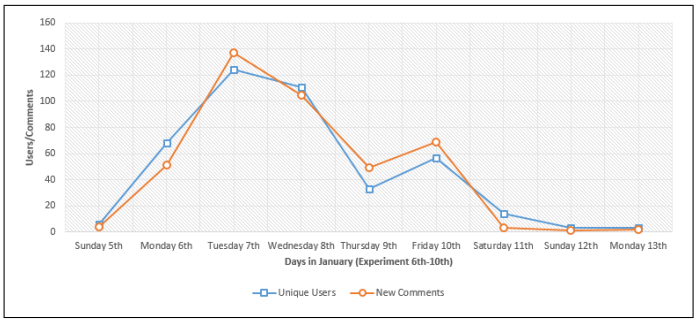


Figure 2. Usage of the system over time

From this graph, it can be clearly seen that the users of the system were most active during the experiment, at which time the screen with the comment stream was on. There was an immediate decline in usage after the screen was removed from campus, suggesting that this social network really does engage more with its users in its pervasive form rather than in its traditional form.

The topics discussed by the users varied, ranging from events related to the university, news from around the world, sports and other more general topics. Figure 3 shows the categories of comments that we identified during the experiment. Due to the fact that this was a new concept, a large portion of the comments posted were related to the system itself. When compared to the content uploaded to other social networks, the comments posted to our system seem to be heavily influenced by events taking place near the *Interactive Wall*. In fact, 68% of the comments posted are directly related to events that happened near the wall. Furthermore, despite the fact that the majority of the comments were posted by people who were near the wall (72 %), there was a small number of comments that were posted from other locations, including three comments from foreign countries. While the comments generated near the wall mostly had to do with events that happened near the wall, the other comments were largely personal advertisements. However, it is interesting to note that some of these comments were indeed linked to events happening near the wall, but in a different way. Instead of describing the events, these comments asked questions about them, and thus they created a real-time connection between people near the wall, and those at other locations.

As was expected, most of the participants of this experiment (48%) got to know of our system through word of mouth as well as through online media (31%), mainly Facebook. The poster set up to attract the users' attention and instruct them on how to interact with the system was not as effective as we had hoped, since only 12% of the users got to know of the system through it. It is interesting to note that 66% of the people who used our system believe that a pervasive social network adds value to traditional social networks. The most popular reason given for why it adds value, is the fact that it

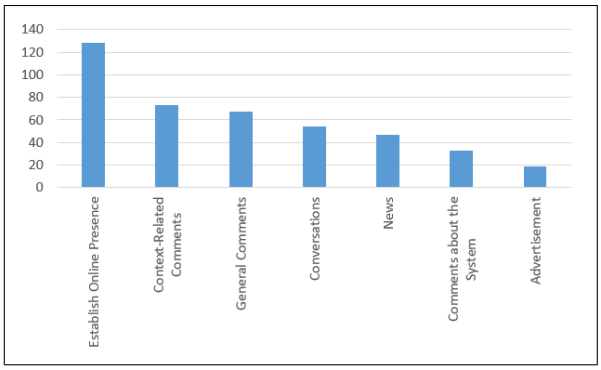


Figure 3. Purposes of using Occupy

enhances communication between those users who are online on the social network, to those users who are physically near the *Interactive Wall*. In fact, 73% of these users, believe that this social network is effective in merging virtual interactions with physical interactions.

A number of particular uses for such a system were identified during the focus group. The first of these uses is that it might be a tool for uniting people about a specific topic E.g. discussing the forthcoming exams (since the system was installed in a University during the exam period). Other social networks display content and comments that are relevant to you or your circle of friends, however, our system displays the same content to each one of its users. This may lead to a system that is able to reach a very wide audience with a single comment and so it can then be used to organise protests or similar gatherings. Another purpose identified during the focus group, is that you can immediately gather feedback about a particular topic, from a targeted sector of the population (within the context of the screen), simply by posting a comment. In addition, some users noted the fact that this system would be an ideal tool during academic conferences or cultural debates. In fact, the literature presented in this research shows evidence of the effectiveness of such a system in these contexts. This idea was developed even further and some people suggested that it might be a useful tool during concerts or political rallies.

Moreover, contrary to our hypothesis, the contributions made by the system (32 in total), went largely unnoticed. In fact, only one reply was made to a comment posted by the system, meaning that this feature was largely ineffective. In addition, 65% of the respondents answered that the discussions initiated by the pervasive social network were not effective in creating discussions. The reason we identified for this, is the fact that the source chosen for the harvesting of online information is not a local website, meaning that some of the articles extracted from this source were not relevant to the University of Malta's context.

Another question that we asked our respondents was whether our system's Nearby Users feature and the Map functionality were effective in creating physical discussions with people in the vicinity. 66% of the users of our system believe that they were indeed very useful features in that they allow you to continue discussions privately with people that are commenting on the social network. These users argued that these features extend traditional social networks in the sense that they make the communication on them more natural.

 ISBN: 978-1-61208-418-3

79% of the users of our system were concerned about their privacy when using any of the social networks. Furthermore, 74% of these users were aware that they have some measure of control over their own privacy. Interestingly, 52% of our users believed that our pervasive social network further invades their privacy, mainly because it constantly tracks their location and that they have no control over who sees their comments. This is a very interesting point, because although a large percentage of our users claimed that they were aware that they have control over their own privacy (switching off location-tracking and deleting comments), they still believed that our pervasive social network poses a greater risk to their privacy than other social networks. On the other hand, some users identified the fact that traditional social networks offer a false sense of security to their users. They claimed, that through the use of a pervasive social network, this false sense of security is not present, and so people are more careful of what they post.

VI. Conclusions and Future Work

A. Future Work

Despite the fact that the evaluation carried out proved that this system has met most of the goals set at the beginning of this final year project, some improvements can always be made. In this section we will be describing some of the improvements that were identified by ourselves or else suggested by the respondents of the survey or from the discussion generated during the focus group. The following is a list of future improvements that can be applied to this project:

1) **Automated Moderation:** Since this system is setup in a public place, there is no way of controlling who is actually viewing the content on the Interactive Wall. Some of this content may not be appropriate for minors within the audience and so some sort of moderation is required. Despite the fact that we employed manual moderation over the content being posted we believe that automated moderation would be far more effective.
2) **More Interactive Walls:** Another interesting suggestion that emerged during the focus group, was to include more than one Interactive Wall. Each wall would have its own URL and the comments are grouped together based on the URL they are originating from. This would be very interesting because we can then compare the different discussions originating in different contexts.
3) **More Informative Posters:** Based on the low percentage of users who got to know of our system through the informative poster, we think that more of these posters would help to attract the people's attention even more and as a consequence the activity rate increases.
4) **Hash-tag Functionality:** The users of our system are able to reply to existing comments and the topic of the original comment is assigned to the replying comment. However, this functionality was not clearly understood by the users. Instead of replying to comments through our reply functionality, the users created new comments and attached the original comment's topic manually. By removing this functionality, and employing a system of Hash-tags similar to that employed by Twitter, these comments can then be grouped together based on the topic given, and the discussions could then flow much better. In other words, the users would then be able to view comments that are related based on the topic chosen.

B. Conclusion

The purpose of this research was to analyse the ways in which a big screen, set up in a public place, displaying a stream of comments can be utilised by the general public. For this purpose, we created a pervasive social network that is embedded in the environment and which is also context-aware and active in the discussions being held on the social network. We have thus attempted to find ways in which such a system could be used and assess its value in comparison to other social networks. Although not all of the goals set at the beginning of this paper were met, a number of positive points emerged, mainly the fact that it successfully merges virtual discussions happening on the social network to those happening physically between groups of people. In addition, this social network helped in raising awareness on privacy concerns related to social networks. Through the use of our system, the users paid more attention to the content that they post as they were constantly aware that whatever they will post will inevitably be seen by a large number of people. The main result obtained from this research is that this system can be a very useful extension to traditional social networks, given that the content being uploaded is moderated in some way. Furthermore, the research question proposed for this project was answered successfully as we identified a number of uses for such a social network.

References

[1] W. Dong, V. Dave, L. Qiu, and Y. Zhang, "Secure Friend Discovery in Mobile Social Networks," in INFOCOM, 2011 Proceedings IEEE. IEEE, 2011, pp. 1647–1655.

[2] O. Mabrouki, A. Chibani, and Y. Amirat, "Privacy in pervasive social networks," in Constructing Ambient Intelligence. Springer, 2012, pp. 296–301.

[3] M. Jadliwala, J. Freudiger, I. Aad, J.-P. Hubaux, and V. Niemi, "Privacy-triggered communications in pervasive social networks," in World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2011 IEEE International Symposium on a. IEEE, 2011, pp. 1–6.

[4] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," Journal of Computer-Mediated Communication, vol. 13, no. 1, 2007, pp. 210–230. [Online]. Available: http://dx.doi.org/10.1111/j.1083-6101.2007.00393.x

[5] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, "Understanding Online Social Network Usage from a Network Perspective," in Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference, ser. IMC '09. New York, NY, USA: ACM, 2009, pp. 35–48. [Online]. Available: http://doi.acm.org/10.1145/1644893.1644899

[6] A. Yamada, T. H.-J. Kim, and A. Perrig, "Exploiting Privacy Policy Conflicts in Online Social Networks," Technical Report, Carnegie Mellon University, Tech. Rep., 2012.

[7] S. B. Mokhtar, L. McNamara, and L. Capra, "A middleware service for pervasive social networking," in Proceedings of the International Workshop on Middleware for Pervasive Mobile and Embedded Computing. ACM, 2009, p. 2.

[8] J. Sunden, "Material Virtualities : Approaching Online Textual Embodiment," Ph.D. dissertation, Linköping University, Department of Communications Studies, Faculty of Arts and Sciences, 2002.

[9] M. N. Ko, G. P. Cheek, M. Shehab, and R. Sandhu, "Social-Networks Connect Services," Computer, vol. 43, no. 8, 2010, pp. 37–43.

 ISBN: 978-1-61208-418-3

[10] F. Michahelles, S. Karpischek, and A. Schmidt, “What can the internet of things do for the citizen? workshop at pervasive 2010,” Pervasive Computing, IEEE, vol. 9, no. 4, 2010, pp. 102–104.

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

[12] M. Strohbach, E. Kovacs, and M. Martin, “Pervasive display newtorks–real-world internet service for public displays,” in Proc. of 4th European Conference on Smart Sensing and Context, 2009, pp. 35–38.

[13] J. Müller, J. Exeler, M. Buzeck, and A. Krüger, “Reflectivesigns: Digital signs that adapt to audience attention,” in Pervasive computing. Springer, 2009, pp. 17–24.

[14] S. Johnson, “How twitter will change the way we live,” Time Magazine, vol. 173, 2009, pp. 23–32.

[15] J. Letierce, A. Passant, J. Breslin, and S. Decker, “Understanding how twitter is used to spread scientific messages,” 2010.

[16] J. Letierce, A. Passant, J. G. Breslin, and S. Decker, “Using twitter during an academic conference: The# iswc2009 use-case.” in ICWSM, 2010.

[17] M. Ebner, “Introducing Live Microblogging: How Single Presentations Can be Enhanced by the Mass.” Journal of Research in Innovative Teaching, vol. 2, no. 1, 2009.

[18] D. Dura, “TwitterCamp at Daniel Dura,” Nov. 2008. [Online]. Available: http://www.danieldura.com/code/twittercamp

[19] M. Ebner and W. Reinhardt, “Social Networking in Scientific Conferences–Twitter as Tool for Strengthen a Scientific Community,” in Proceedings of the 1st International Workshop on Science, vol. 2, 2009, pp. 1–8.

[20] M. Ebner, G. Beham, C. Costa, and W. Reinhardt, “How People are Using Twitter during Conferences,” Creativity and Innovation Competencies on the Web, 2009, p. 145.

[21] C. Smith, “By the Numbers: 89 Amazing Facebook User Statistics (updated march 2014),” Mar. 2014, retrieved from: http://expandedramblings.com/index.php/by-the-numbers-17-amazing-facebook-stats/.

 ISBN: 978-1-61208-418-3

A Robust Model for Person Re-Identification in Multimodal Person Localization

Thi Thanh Thuy Pham
Faculty of Information Technology
University of Technology and Logistics
Bacninh, Vietnam
Email: thanh-thuy.pham@mica.edu.vn

Thi Lan Le
and Trung Kien Dao
and Duy Hung Le
International Research Institute MICA
University of Science and Technology
Hanoi, Vietnam

Van Toi Nguyen
The University of Information
and Communication Technology
under Thai Nguyen University
ThaiNguyen, Vietnam

***Abstract*—Determining person ID (Identity) is one of the crucial steps in indoor human localization system. It is more exactly stated as person Re-ID (Re-Identification) problem because for each user's position, the user ID at the first occurrence needs to be shown correspondingly at the later times of localization. In this paper, a multimodal person localization system of WiFi and camera is proposed, with the analysis for the key role of appearance-based person Re-ID in fusing different information sources. A new model for person appearance representation based on kernel descriptor is proposed to tackle the challenges of person Re-ID in camera network. Additionally, a dataset for the real scenario of the proposed multimodal person localization is also established, in which we set a database for vision-based human Re-ID evaluation. The experiments on the benchmark datasets and our dataset show the outperforming results in comparison with other state-of-the-art approaches.**

Keywords–Multimodal person localization; Person Re-ID; KDES descriptor; Camera; WiFi.

I. INTRODUCTION

Person Re-ID and positioning are two key problems in a typical human localization system. In case of multi-object localization, we need to identify the person who is localized, therefore we know the determined positions belong to which objects. Person Re-ID in camera network is a hard problem and increasingly attracted many researchers. Three basic steps need to be done for vision-based person Re-ID problem. People detection in consecutive frames is firstly executed, then feature extraction within the detected regions and feature descriptor is generated, finally object matching is done for Re-ID. Each step has its own challenges and these affect strongly to the system performance. In general, they include (1) illumination conditions that are different by time and space; (2) pose, scale and appearance variation of people at distinctive camera FOVs (Fields of View). This is considered as the most challenging, because the human appearance features are mainly used in the human re-identification system; (3) occlusions in which people are obscured by each other or obstacles in the environment; (4) re-identification scenarios involving closed set Re-ID (the identified objects are included in both gallery and probe sets) or open set Re-ID (the objects may not be contained in the gallery set).

Many approaches are proposed for vision-based person Re-ID problem, however most of them are oriented to (1) build a distinctive feature descriptor for each object and then apply an effective object classifier for that or (2) design potential distance metrics from data. In this paper, we concentrate on establishing a robust feature descriptor which improves the original KDES (Kernel Descriptor) of [1], and applying multi-class SVM as relative ranking for person Re-ID in camera network. This is proven to be more robust than original KDES or other state-of-the-art methods in solving vision-based person Re-ID problem.

The rest of the paper is organized as follow. In Section II, the related works on vision-based human Re-ID are presented. Section III indicates a combined system of WiFi and visual signals for human localization, in which appearance-based person Re-ID problem in camera network is solved by improved KDES. Some experimental results on benchmark datasets and our dataset are shown in Section IV. Conclusion and future directions will be finally denoted.

II. RELATED WORK

Design of a robust person descriptor is the most decisive step for vision-based person Re-ID problem. Many kinds of features are utilized for this, in which human body appearance is the simplest and the most popular one. Color, texture, and shape are features that can be extracted for human appearance. In [2][3], color histogram is used for feature descriptor. There are two ways to represent the image of detected people with color histogram: global color histogram and local color histogram. A single histogram is used in the first method for the whole image, while in the second way, image is divided into some parts and concatenating the part-based color histograms is done to give a final result. Most reported person Re-ID works pay attention on the second solution, such as in [4], a weighted color histogram derived from MSCR (Maximally Stable Colour Regions) and structured patches are combined for visual description. In [5][6], color histogram on different color models is calculated and syndicated with texture features to make person descriptor more robust. Shape features are also extracted for appearance model. However, they are unstable because of non-rigid objects as people; so, in [7], color and texture features are associated with shape feature to enhance the effectiveness of person descriptor. Local region descriptors, such as SIFT (Scale-Invariant Feature Transform), SURF (Speeded Up Robust Features) and GLOH (Gradient Location and Orientation Histogram) are evaluated in [8] for person Re-ID in image sequences. The results show that GLOH and SIFT outperform both shape context and SURF descriptor. Additionally, a large number of visual features are exploited

 ISBN: 978-1-61208-418-3

for person Re-ID problem, such as Haar-like features, HOG (Histogram of Oriented Gradients), edges, covariance, interest points, etc.

The next step in human Re-ID process is classification, with two scenarios of single-shot and multi-shot being reported. The first case is more simple with one-to-one matching between a pair of probe and gallery image for each person, whereas in the second scenario, each object has multiple images, either in the gallery or the probe set. In general, the purpose of classification in person Re-ID is finding out the most similar candidate for a target or ranking the candidates based on a standard distance minimization strategy, which is known as distance metric. This metric can be chosen independently (non-learning based method) [9] or learned from the data (learning-based method) [10] in order to minimize intra-class variation whilst maximize extra-class variation. They typically include histogram-based Bhattacharyya distance, K Nearest Neighbor classifiers, L1-Norm, diffusion distance [11]. Additionally, some later proposed methods, such as LMNN-R (Large Margin Nearest Neighbor) distance metric in [12] or PRDL (Probabilistic Relative Distance Learning) in [13] are more robust.

To get an ID ranking list, distance scores between true and wrong matches can be compared directly or relatively (ranking the scores that show the correspondence of each likely match to the probe image). The relative ranking treated by either Boosting as RankBoost in [14] or kernel-based learning, such as RankSVM [15], primal-based RankSVM [16] or Ensemble RankSVM [6].

III. PROPOSED SYSTEM

A. Overview of multimodal person localization system

Object localization is known as a problem of determining the object position in the environment. For each user in multi-user localization system, two problems of positioning (where the user is) and identifying (who the user is) must be solved simultaneously. A general diagram for object localization is illustrated in Figure 1. In this figure, the input cues can come from different sensors, such as optical, radio frequency, ultra sound, inertia, DC Electromagnetic sensors, etc. From the input cues, localization and Re-ID are executed simultaneously to give the output for object position and ID. Multimodal object localization is defined as a problem of multi-cue combination for input or fusion of different positioning methods. As proven by Vinyals et al. [17] and Dao et al. [18], compounding of different models gives better positioning results than applying a single model. Teixeira et al. [19] proposed to use the motion signature taken from wearable accelerometer for identifying people in camera network.

Figure 1. Flowchart of object localization system.

Our research aims at developing a multimodal person localization system by using both WiFi and camera systems. This offers some benefits in comparison with single-method systems. (1) System setting cost is limited because of available WiFi infrastructure and uncrowded-deployed cameras. (2) Positioning range is easily broaden by simply adding more APs (Access Point) in the environment. (3) Computational expense is much lower for WiFi-based than vision-based positioning system. (4) The positioning accuracy is provided in accordance with the application-specific demands. Although the camera-based system brings more impressed positioning results, but not every where in building needs high localizing accuracy. (5) Sampling frequency is improved for the WiFi-based system, because it has lower sampling rate (about one signal measure per second) than vision-based system (approximately 15 fps). (6) The information for person Re-ID becomes richer. One object can be identified simultaneously by both WiFi and camera systems. These ID cues can be used in the way of supporting one another in multi-model object localization system. For example, at a certain time, one object is localized and identified by WiFi system, with the position of P_{WiFi} and the identity of ID_{WiFi} (the MAC address of mobile device) respectively. At the same time, this object is also determined by P_{cam} and ID_{cam} from camera system. However, P_{WiFi} is not as accurate as P_{cam}, whilst ID_{WiFi} is clearer than ID_{cam}. Therefore, by using both of these systems, the object can be localized by P_{cam} and identified by ID_{WiFi}.

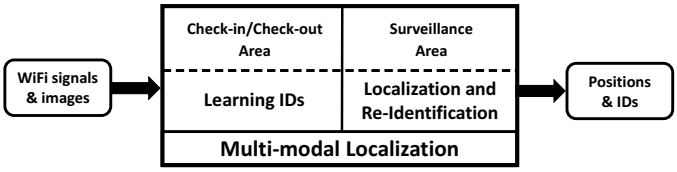


Figure 2. Multimodal localization system fusing WiFi signals and images.

Figure 2 shows a framework for our multi-model human localization system using both WiFi signals and camera network. The framework indicates that the proposed system is implemented in two subregions of the whole positioning area: check-in/check-out region and surveillance region. In the first region, learning ID cues is executed. Person holding a WiFi-integrated device will one by one come in and come out of the first region. At the entrance of the first region, the person's ID will be learned individually by the images captured from cameras and MAC address of WiFi-enable equipment held by that person. One camera, which is in front door of check-in gate, captures human face and then a face recognition program is executed. Another camera acquires human body images at different poses and learning phase of appearance-based ID is done for each person. In short, in the first region, we get three types of signature for each person (N_i): face-based ID (ID^i_F), WiFi-based ID (ID^i_{WF}), and appearance-based ID (ID^i_{Apr}). Depending on different circumstances, we can map among signatures of (ID^i_F, ID^i_{WF}), (ID^i_F, ID^i_{Apr}), (ID^i_{Apr}, ID^i_{WF}) and utilize them for person localization and identification in the surveillance region. The user will end up his route at the exit gate and he will be checked out by other camera. This camera acquires human face for person Re-ID, and based on this, the user will be removed from the localization system. By using check-in/check-out region, we can (1) control the human appearance changes (the difference in cloth colors) at each time people come in the positioning area, (2) decrease the computing cost by eliminating the checked-out users from the system, (3) map between different ID cues for the same person.

 ISBN: 978-1-61208-418-3

In the surveillance region, two problems of person localization and Re-ID will be solved concurrently by combining visual and WiFi information. Figure 3 demonstrates a surveillance region which contains WiFi range and camera FOVs. In this region, the WiFi range covers some visual ranges (the camera FOVs: FOV of C_1, FOV of C_2,.., FOV of C_n). This means the user always move within WiFi range but switch from one camera FOV to others.

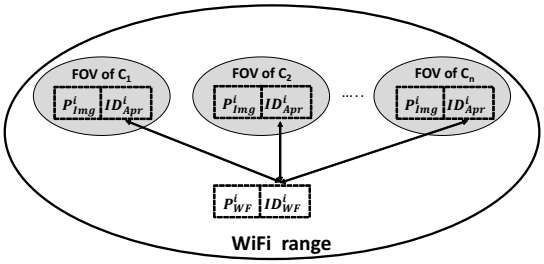


Figure 3. Surveillance region with WiFi range and disjoint cameras' FOVs.

In each camera FOV and for an individual, we calculate image-based and WiFi-based positions (P^i_{img}, P^i_{WF}) and ID^i_{Apr}. From ID^i_{Apr}, we know ID^i_{WF} correspondingly by ID mapping result taken from the first region. Outside the camera FOV, there only exits the information of P^i_{WiFi}, ID^i_{WF}, and ID^i_{Apr} respectively. When people switch from one camera FOV to others, their positions and IDs will be updated in the WiFi-available region. The localization accuracy then be tuned by combination of WiFi-based and vision-based systems.

From the above analysis, we see that finding ID^i_{Apr} plays a key role in the proposed multimodal person localization system. It is used to link the object trajectories from one camera range to others through the intermediate positioning range of WiFi. Therefore, ID^i_{Apr} must be shown at each frame captured from different cameras in the surveillance area. That means the appearance-based person Re-ID problem needs to be solved. In this circumstance, it belongs to multi-shot person Re-ID problem, with multiple images for each detected person at different resolutions, lighting conditions, and poses are processed.

B. Vision-based person re-identification

1) The system overview: The flowchart of vision-based person Re-ID system is illustrated in Figure 4. It includes three stages of (1) person detection, (2) feature extraction, and (3) classification. In the first stage, from the input frames, the ROI (Region of Interest) of person can be determined by using the state-of-the-art methods. The features are then extracted from these regions and feature descriptors are created in the second stage. Finally, a classifier is applied to learn the person model and predict the corresponding ID.

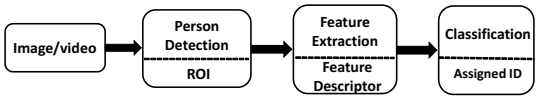


Figure 4. A diagram of vision-based person Re-ID system.

In this section, we present in detail the second stage since it is the main contribution of our paper. For this, we propose a new person appearance representation model based on KDES. This descriptor is firstly proposed by Bo et al. [1] and has been proved to be robust for recognition of different objects, such as hand pose recognition [20].

2) KDES-based person representation: The basic idea of the representation based on kernel methods is to compute the approximate explicit feature map for kernel match function (see Figure 5). In other words, the kernel match functions are approximated by explicit feature maps. This enables efficient learning methods for linear kernels to be applied to the non-linear kernels. This approach was introduced in [1][21]. Given a match kernel function $k(x, y)$, the feature map $\varphi(.)$ for the kernel $k(x, y)$ is a function mapping a vector x into a feature space so as $k(x, y) = \varphi(x)^\top \varphi(y)$. Suppose that we have a set of basis vectors $B = \{\varphi(v_i)\}_{i=1}^{D}$, the approximation of feature map $\varphi(x)$ can be:

$$\phi(x) = Gk_B(x) \tag{1}$$

where G is defined by: $G^\top G = K_{BB}^{-1}$ and K_{BB} is $D \times D$ matrix with $\{K_{BB}\}_{ij} = k(v_i, v_j)$. k_B is a $D \times 1$ vector with $\{k_B\}_i = k(x, v_i)$.

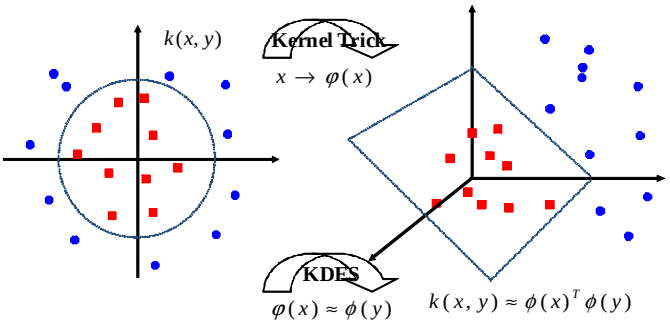


Figure 5. The basic idea of representation based on kernel methods.

Feature extraction is then done at three levels of pixel, patch and the whole image of detected person. At pixel level, a normalized gradient vector is computed for each pixel of the image. The normalized gradient vector at a pixel z is defined by its magnitude $m(z)$ and normalized orientation $\omega(z) = \theta(z) - \overline{\theta}(P)$, where $\theta(z)$ is orientation of gradient vector at the pixel z, and $\overline{\theta}(P)$ is the dominant orientation of the patch P that is the vector sum of all the gradient vectors in the patch. This normalization will make patch-level features invariant to rotation. In practice, the normalized orientation of a gradient vector will be:

$$\widetilde{\omega}(z) = [sin(\omega(z))\ cos(\omega(z))] \tag{2}$$

At the second level, the image with different resolutions will be divided into a grid of a fix number of cells as in [20], instead of size-fixed cells as in [1]. A patch is then set by 2×2 cells and two adjacent patches along x-axis or y-axis are overlapped at two cells. This division results to size-adaptive patches to the different image resolutions, and nearly the same feature vectors for the scale-varied images of intraclass are created (see Figure 6). In our work, this technique is utilized for KDES extraction because of a large variation of person size caused by different distances from pedestrian to the stationary camera. For each patch, we compute patch features based on

 ISBN: 978-1-61208-418-3

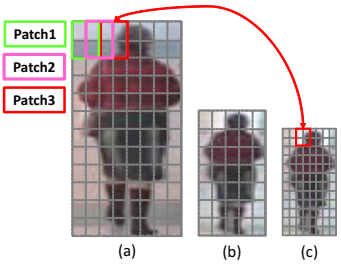


Figure 6. Illustration of size-adaptive patches (a, c) and size-fixed patches (a, b) which is mentioned in [1].

a given definition of match kernel. The gradient match kernel is constructed from three kernels: gradient magnitude kernel $k_{\widetilde{m}}$, orientation kernel k_o, and position kernel k_p.

$$K_{gradient}(P,Q) = \sum_{z\in P}\sum_{z'\in Q} k_{\widetilde{m}}(z,z')k_o(\widetilde{\omega}(z),\widetilde{\omega}(z'))k_p(z,z') \tag{3}$$

where P and Q are patches of two different images need to measure the similarity. z and z' denote the 2D positions of a pixel in the image patch P and Q respectively. $\varphi_o(.)$ and $\varphi_p(.)$ are the feature maps for the gradient orientation kernel k_o and position kernel k_p respectively. Then, the approximate feature over the image patch P is constructed as:

$$\overline{F}_{gradient}(P) = \sum_{z\in P} \widetilde{m}(z)\phi_o(\widetilde{\omega}(z)) \otimes \phi_p(z) \tag{4}$$

where $\otimes$ is a Kronecker product, $\phi_o(\widetilde{\omega}(z))$ and $\phi_p(z)$ are approximate feature maps (1) for the kernel k_o and k_p respectively.

The last level is finished by creating a complete descriptor for the whole image. As in [22], a pyramid structure is used to combine patch features. Given an image, the final representation is built based on features extracted from lower levels using EMK (Efficient Match Kernels) proposed in [1]. First, the feature vector for each cell of the pyramid structure is computed. The final descriptor is the concatenation of feature vectors of all cells.

Let C be a cell that has a set of patch-level features $X = \{x_1, ..., x_p\}$, then the feature map on this set of vectors is defined as:

$$\overline{\phi}_S(X) = \frac{1}{|X|}\sum_{x\in X}\phi(x) \tag{5}$$

where $\phi(x)$ is approximate feature map (1) for the kernel $k(x,y)$. The feature vector on the set of patches, $\overline{\phi}_S(X)$, is extracted explicitly.

Given an image, let L be the number of spatial layers to be considered. In this case, $L = 3$. The number of cells in layer l-th is (n_l). $X(l,t)$ is a set of patch-level features that fall within the spatial cell (l,t) (cell t-th in the l-th level). A patch is fallen in a cell when its centroid belongs to the cell. The feature map on the pyramid structure is:

$$\begin{aligned}\overline{\phi}_P(X) \quad = \quad & [w^{(1)}\overline{\phi}_S(X^{(1,1)}); ...; w^{(l)}\overline{\phi}_S(X^{(l,t)}); \\ & ...; w^{(L)}\overline{\phi}_S(X^{(L,n_L)})]\end{aligned} \tag{6}$$

In (6), $w^{(l)} = \frac{\frac{1}{n_l}}{\sum_{l=1}^{L}\frac{1}{n_l}}$ is the weight associated with level l.

Once the KDES computed, multiclass SVM is applied to train the model for each person. For each detected instance, a list of ranked objects will be generated based on the probability of SVM.

IV. Experimental Results

This section will present the testing datasets and the results obtained for vision-based person Re-ID. The CMC (Cumulative Match Curve) is employed as the performance evaluation method for person Re-ID problem. The CMC curve represents the expectation of finding correct match in the top n matches.

A. Testing datasets

In our experiments, two multi-shot benchmark datasets of CAVIAR4REID and i-LIDS are used. We also build our own dataset in the context of multimodal person localization. The CAVIAR4REID dataset includes 72 pedestrians, in which 50 of them are captured from two camera views and the remaining 22 from one camera view. i-LIDS dataset contains 119 individuals, with the images captured from multi-camera network. Both of them, especially CAVIAR4REID, are challenging because of broad changes in resolution, lighting condition, occlusion, and human pose. Concerning to our dataset, we build it for multimodal person localization evaluation. A database for testing appearance-base person Re-ID is also established in this.

Figure 7 shows the 8th floor plan of our office building. It is set as the testing environment for our combined person localization system. At the entrance, people hold smart phones or tablets go one by one through the check-in gate, then move inside the surveillance area, and finish their routes by going out check-out gate.

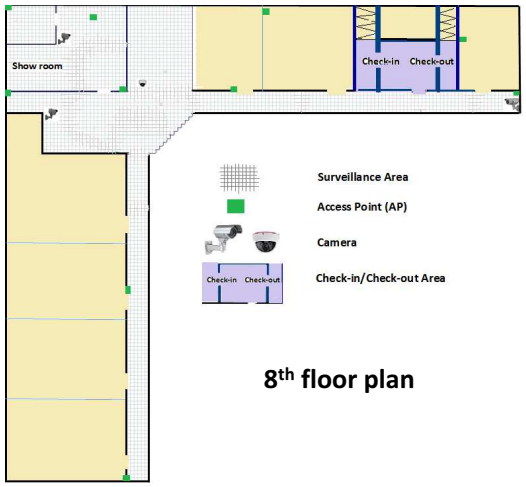


Figure 7. Testing environment.

In the check-in and check-out area, we set three cameras. Two of them are used at the entrance. One camera captures human face in order to check-in user by face recognition. The remaining camera acquires human body images at different poses. This will help the system learn appearance-based signature of the checked-in user. The third camera is used to capture human face at the exit, and based on this, the system will check out or release the user from its process. In the surveillance

 ISBN: 978-1-61208-418-3

area, four cameras with non-overlapping FOVs are deployed along the hallway and in a room. People are detected, localized, and re-identified at each frame captured from these cameras. Besides this, 11 APs are established throughout the testing environment. RSSIs (Received Signal Strength Indicators) and the MAC address are consecutively scanned and sent from mobile device to the server to calculate the position and ID of the device holder.

In short, a total of seven AXIS IP cameras and eleven APs are deployed throughout the testing environment of the 8^{th} storey floor plan. These cameras and APs are fixed at certain distances from the floor ground (about 1.6m-2.2m for cameras and 2m-2.8m for APs). They are configured with static IP addresses. The camera frame rate is set to 20 fps and image resolution is 640x480.

The dataset for human Re-ID includes 25 people with different routes in the testing environment. Each person spends from 3 to 5 minutes for his route. An approximation of 800 values of RSSIs are scanned, about 2000 frames are captured for each camera in the surveillance area. All captured frames are processed as real Re-ID scenario of multimodal pedestrian localization system. Firstly, the images of person body at different poses are extracted from video sequence of the entrance camera. They are later used for training phase of appearance-based person identification. In the surveillance area, each frame from the sequences of four cameras is processed for human detection, so each of 25 pedestrians will have the appearance images (the bounding boxes of each person) at different views. The image filename identifies the video sequence to which it belongs, the camera ID, frame number, time (hour, minute, second), and the person ID: `VVV_WW_XXXXX_YYYYYY_ZZZ.jpg` (E.g., `025_01_01260_153702_012.jpg` means the appearance image belongs to video sequence 25; it is captured from the camera 01; frame number 1260; the time is 15h:37m:02s; the person ID is 12). These images are utilized for testing phase of person Re-ID system. An example in the dataset for person Re-ID in camera network is shown in Figure 8. The images on the top are used for training phase of appearance-based person identification. They are captured by a camera at the check-in/check-out region. The images for testing phase are shown at the bottom. They come from four different cameras in the surveillance region.

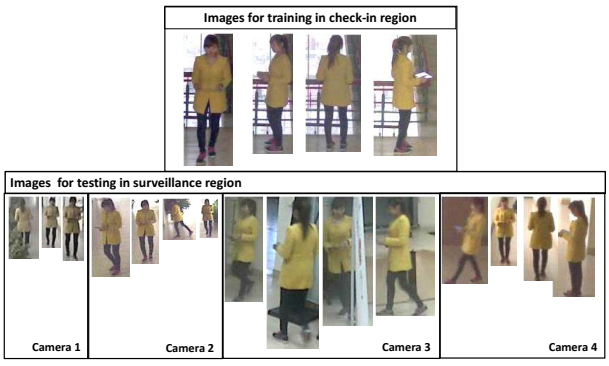


Figure 8. The instances in dataset for vision-based person Re-ID.

In comparison with other person Re-ID benchmark datasets, such as iLIDS, ETHZ, PRID 2011, CAVIAR4REID and VIPeR, our dataset contains multiple images for each person. These images are captured from many cameras (4 cameras) at different FOVs. This makes more variations for intraclass images in terms of resolution, illumination, pose and scale. In addition, it is set for real scenario of our proposed multimodal person localization system.

B. Person re-identification results

We compare the results of our proposed method with original KDES [1] and other state-of-the-art approaches on multi-shot datasets of CAVIAR4REID and iLIDs. The experimental settings are kept the same as in [23], with modified version of iLIDS dataset is selected (including only 69 individuals, with at least 4 images for each) and 72 pedestrians for CAVIAR4REID dataset, in which 50 different individuals are captured under both views. Each view has 10 images for each pedestrian. The outperforming results of the proposed method are shown in Figure 9. For CAVIAR4REID dataset, rank-1 recognition rate of AHPE (Asymmetry-based Histogram Plus Epitome) [23] is much lower than our method. It is only about 8 %, compared with 67.76 % of the original KDES [1] and 73.81 % for our method. However, both KDES and our method gain nearly the same figures from rank-13 and backward. For iLIDS dataset, the gap between rank-1 recognition rates of the proposed method with AHPE [23] or SDALF (Symmetry-Driven Accumulation of Local Features) [24] is approximately 20 %, and about 7 % with the original KDES [1]. This gap is slightly decreased for KDES but significantly reduced for AHPE and SDALF after rank-15.

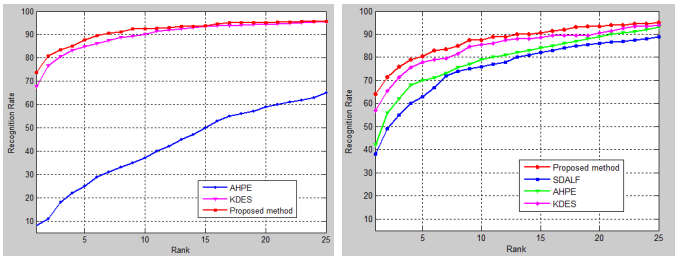


Figure 9. The results of proposed method against AHPE [23], SDALF [24] and KDES [1] on (a) CAVIAR4REID dataset and (b) iLIDS dataset.

Other experiments with iLIDs dataset are presented in Figure 10-a in comparison with other methods reported in [25]. The highest result for rank-1 belongs to RDC (Relational Divergence Classification) as mentioned in [25], but it is roughly 14% lower than our proposed method (66.18%). KDES [1] is tested on this dataset with 61.76% for rank-1, which is approximately 5% smaller than our method at the first 7 ranks.

The state-of-the-art SDALF [26] and the proposed method for person Re-ID are also tested on our dataset, with the gallery images (about 50 images for each class) from the entrance camera and the probe images (from 60 to 193 images for each person) from four cameras in the surveillance area (see Figure 10-b). The testing result is 73.13% at rank-1, compared with the original KDES of 67.16% and 30% for SDALF. The deviation between two recognition rates of our method and KDES gradually declines and almost reaches to the same value as SDALF after rank-21.

 ISBN: 978-1-61208-418-3

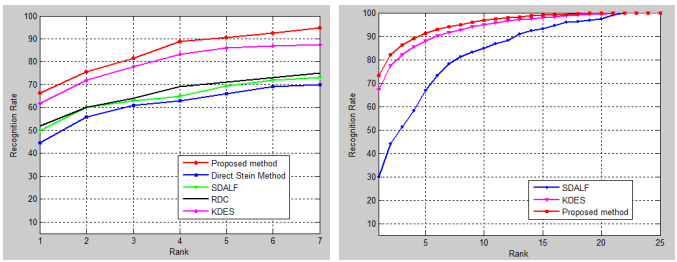


Figure 10. The comparative results with (a) reported methods in [25] and (b) the results are tested on our dataset.

The experimental results obtained with three different datasets have proved the better performance of the proposed method in comparison with the original KDES and other state-of-the-art methods. Based on these results, we will use the proposed method for person Re-ID in our multimodal human localization system.

V. Conclusion and Future Work

In this paper, person Re-ID problem in camera network achieves state-of-the-art performance on the benchmark datasets and our dataset by applying a robust person appearance representation based on KDES. The visual person ID can be used in connective and complementing manner of different types of information in the proposed multimodal pedestrian localization system of WiFi and camera. The experimental results are promising, and based on this, a multimodal method, which uses particle filter and integrated data association algorithm, will be promoted in the future work to increase the performance of the combined person Re-ID and localization system.

Acknowledgement

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2013.32.

References

[1] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada, 2010, pp. 244–252.

[2] L. F. Teixeira and L. Corte-Real, "Video object matching across multiple independent views using local descriptors and adaptive learning," Pattern Recognition Letters, vol. 30, no. 2, 2009, pp. 157–167.

[3] D. N. T. Cong, C. Achard, L. Khoudour, and L. Douadi, "Video sequences association for people re-identification across multiple non-overlapping cameras," in Image Analysis and Processing (ICIAP). Springer, 2009, pp. 179–189.

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in BMVC, vol. 2, no. 5. Citeseer, 2011, p. 6.

[5] D. Figueira, L. Bazzani, H. Q. Minh, M. Cristani, A. Bernardino, and V. Murino, "Semi-supervised multi-feature learning for person re-identification," in Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on. IEEE, 2013, pp. 111–116.

[6] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking." in BMVC, vol. 2, no. 5, 2010, p. 6.

[7] N. Martinel, C. Micheloni, and C. Piciarelli, "Learning pairwise feature dissimilarities for person re-identification," in Distributed Smart Cameras (ICDSC), 2013 Seventh International Conference on. IEEE, 2013, pp. 1–6.

[8] M. Bauml and R. Stiefelhagen, "Evaluation of local features for person re-identification in image sequences," in Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on. IEEE, 2011, pp. 291–296.

[9] S. Bak, E. Corvee, F. Brémond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," in Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on. IEEE, 2010, pp. 435–440.

[10] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, pp. 649–656.

[11] D. Figueira and A. Bernardino, "Re-identification of visual targets in camera networks: A comparison of techniques," in Image Analysis and Recognition. Springer, 2011, pp. 294–303.

[12] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja, "Pedestrian recognition with a learned metric," in Computer Vision–ACCV 2010. Springer, 2011, pp. 501–512.

[13] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 3, 2013, pp. 653–668.

[14] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," The Journal of machine learning research, vol. 4, 2003, pp. 933–969.

[15] D. Simonnet, M. Lewandowski, S. A. Velastin, J. Orwell, and E. Turkbeyler, "Re-identification of pedestrians in crowds using dynamic time warping," in Computer Vision–ECCV 2012. Workshops and Demonstrations. Springer, 2012, pp. 423–432.

[16] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," Information Retrieval, vol. 13, no. 3, 2010, pp. 201–215.

[17] O. Vinyals, E. Martin, and G. Friedland, "Multimodal indoor localization: An audio-wireless-based approach," in Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on. IEEE, 2010, pp. 120–125.

[18] T. K. Dao, H. L. Nguyen, T. T. Pham, E. Castelli, V. T. Nguyen, and D. V. Nguyen, "User localization in complex environments by multimodal combination of gps, wifi, rfid, and pedometer technologies," The Scientific World Journal, vol. 2014, 2014.

[19] T. Teixeira, D. Jung, G. Dublon, and A. Savvides, "Identifying people in camera networks using wearable accelerometers," in Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments. ACM, 2009, p. 20.

[20] V. T. Nguyen, T. L. Le, T. H. Tran, R. Mullot, and V. Courboulay, "A New Hand Representation Based on Kernels for Hand Posture Recognition," in The 11th IEEE International Conference on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 2015.

[21] S. Maji, A. C. Berg, and J. Malik, "Efficient classification for additive kernel svms," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 35, no. 1, 2013, pp. 66–77.

[22] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in Advances in neural information processing systems, 2009, pp. 135–143.

[23] L. Bazzani, M. Cristani, A. Perina, and V. Murino, "Multiple-shot person re-identification by chromatic and epitomic analyses," Pattern Recognition Letters, vol. 33, no. 7, 2012, pp. 898–903.

[24] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 2360–2367.

[25] A. Alavi, Y. Yang, M. Harandi, and C. Sanderson, "Multi-shot person re-identification via relational stein divergence," arXiv preprint arXiv:1403.0699, 2014.

[26] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," Computer Vision and Image Understanding, vol. 117, no. 2, 2013, pp. 130–144.

 ISBN: 978-1-61208-418-3

Analyzing Consumer Loyalty of Mobile Advertising: A View of Involvement, Content, and Interactivity and the Mediator of Advertising Value

Wei-Hung Hsiao
Institute for Information Industry,
Taipei, Taiwan
e-mail: miswhhsiao@gmail.com
Shwu-Ming Wu
Department of Human Resource Management, National Kaohsiung University of Applied Science
Kaohsiung, Taiwan
e-mail:mingwu@cc.kuas.edu.tw
Ing-Long Wu
Department of Information Management, National Chung Cheng University
Chia-Yi, Taiwan
e-mail: ilwu@mis.ccu.edu.tw

Abstract—**Mobile commerce has been growing popular recently and mobile advertising is one of the important aspects in marketing. Advertising value is an important criterion for measuring its success. Previous studies have focused on message content or technology use. The purpose of mobile advertising is to provide personalized information for consumers. Individual beliefs thus play an important role in identifying advertising value. Three major concerns arise for advertising value, personal involvement, message content, and user interactivity. However, advertising value may be temporary to define advertising effectiveness. Consumer loyalty is defined as the target for rapid growth and proliferation of advertisements. Based on these issues, this study thus proposes a novel research model for defining the relationship structure for the key drivers, advertising value and customer loyalty, in a m-commerce. The empirical results show important links among these components and advertising value acts as a critical mediator in realizing consumer loyalty.**

Keywords-Mobile advertising; Consumer loyalty; Advertising value; Involvement; Message content; Interactivity.

I. INTRODUCTION

Mobile advertising has become increasingly important in marketing practice. According to Gartner research, mobile advertising revenue reached 3.3 billion dollars in 2011, more than double the 1.6 billion dollars of revenue in 2010. Existing research for exploring its effectiveness has been fragmentally focused on two perspectives, technology use and message content. The first examines it from a technology use perspective according to relevant theories, such as the technology acceptance model (TAM) [13], the theory of planned behavior (TPB) [1] or their extensions, to investigate advertising value of mobile advertising in a national or cross-cultural context [20]. Other studies investigated it from content perspective in terms of identifying message characteristics to study advertising value, attitude, or intention toward mobile advertising [26].

However, to study how consumers process mobile advertising/message in their minds, it is necessary to gain an understanding of the individual's involvement state when experiencing this type of advertising [4]. For example, different people perceive the same product differently and have different levels of involvement with the same product. Recently, interactivity has been identified as an important concern for technology use of mobile devices, such as in the case of the Web 2.0 platform, as it is the major feature which differentiates new media from traditional ones [14]. However, most technology-use based theories have focused more on the attributes of IT itself, such as perceived usefulness, computer self-efficacy, and other technology related factors [40]. Next, previous studies have confined research targets to only examining advertising value or initial acceptance of mobile advertising, which is just an initial step to defining the success of mobile advertising. Bhattacherjee [6] contended that long-term viability of an IT/IS and its eventual success depends on its continued use rather than initial acceptance. There has been a lack of the studies of consumer loyalty or continued use in relation to mobile advertising [3].

In sum, this study first attempts to define the antecedents of advertising value by integrating three perspectives, involvement, advertising content, and interactivity in the particular context of mobile advertising. Furthermore, this study explores customer loyalty in relation to mobile advertising, as loyalty issue has been considered as the main purpose from effectively realizing advertising value with the rapid growth and proliferation in the mobile advertising [18]. Accordingly, this study thus proposes a novel research model to define a relationship structure for the key drivers, advertising value, and customer loyalty, in an m-commerce context.

Specifically, based on a comprehensive review of the literature, we further define five attributes of message characteristics, entertainment, informativeness, irritation, credibility, and personalization. When it comes to the meaning structure of mobile devices, both involvement and interactivity are complex factors. We thus define both as a second-order construct with three indicators for each. The former includes personal, stimulus, and situation indicators and the latter contains communication, synchronicity, and user control indicators. Finally, the structure of this paper is as below, Section II for literature review, Section III with research design, Section IV for hypothesis testing, Section V with findings and discussions, and Section VI for conclusions and suggestions.

II. LITERATURE REVIEW

Based on the above discussion, Figure 1 provides a pictorial depiction of this research model. The proposed model is also

 ISBN: 978-1-61208-418-3

consistent with the basic relationship structure, including stimulus, satisfaction/value, and loyalty/retention, when often applied in consumer research [5]. As follows, we discuss the relevant literature and the development of the hypotheses.

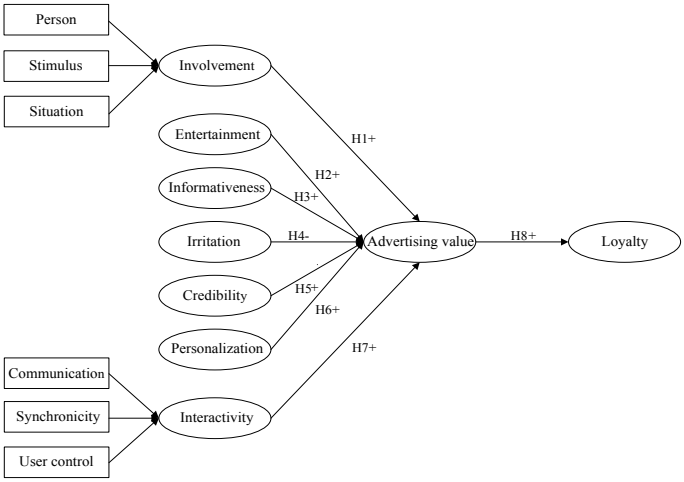


Figure 1. Research model
Node: Variable, Edge: Link for two variables, Rectangle: First order indicator

A. Mobile Advertising

Mobile advertising focuses on communication by mobile media with the intent to influence consumers' cognitive and emotional states. Mobile advertising employs two basic publicization methods: push-based and pull-based. The push-based strategy signifies that marketers send information directly to recipients without the latter's request. In contrast, the pull-based strategy implies that the specific information received by consumers is sent as a result of their own request [21]. Although the two publicization methods make no distinctions in traditional advertising or flyer promotions, however, they are different with mobile devices for each user. In sum, mobile advertising has some distinctive features, that is, interactivity and personalization.

Advertising value refers to a subjective evaluation of the relative worth or the utility of advertising to consumers [27]. Advertising value may serve as an index of customer satisfaction in the communication with products purchased from a vendor; moreover, it also represents a general cognitive belief or an attitude toward advertising. Studies of advertising value have demonstrated it to be a useful performance criterion for evaluating advertising effectiveness and further to help advertisers develop their advertising strategy more effectively [20]. When mobile advertising does not provide certain values for consumers, it may be perceived by consumers as coercive and unwelcome and fail to capture the attention of consumers [29].

B. Involvement

Involvement generally refers to the mediating role in determining if a cognitive experience of external stimuli is effectively relevant to the receiver [29]. In the advertising domain, involvement is manipulated by making the advertisement relevant to the consumers in terms of being personally affected and hence motivated to respond to the advertisement [26]. This study is mainly concerned with the impact of involvement when a consumer experiences mobile advertising messages. Although there are no agreements on what clusters constitute a good taxonomy of involvement, most researchers agree that involvement should include at least three dimensions: personal, stimulus, and situational [19]. The personal dimension defines a person's inherent value system along with her/his unique experiences, which determines whether the person is affected by the advertisement. The stimulus dimension relates to the physical characteristic of the advertisement. The situational dimension refers to the varying situations in presenting advertisement that temporarily affect the state of involvement with the advertisement. Involvement is conceptualized as a formative construct with the three indicators as involvement is clearly a composite of three indicators that may be very different [23]. Reflective indicators, however, are interchangeable and share a common theme.

As follows, we develop a hypothetical link. Researchers noted that consumers' involvement with online advertising has a direct impact on their attitude toward web advertising and their perceived value of the advertised products or services [28]. As such, involvement has several influences on personal perceived value in relation to stimuli in different types of advertising messages [28]. Following this argument, we can propose the first hypothesis.

H1: Involvement with mobile advertising positively affects advertising value.

C. Message Content

Information objects include a storage format (physical or digital) and one or more forms of human usable expressions (e.g., visual, aural, and tactile) [22]. Mobile advertising is an information object in a digital form for human use and the information content of advertisement is usually related to products or services to be sold to consumers. Several studies have proposed similar classifications for indentifying message content. Ducoffe [11] defined three major message attributes forming personal attitudes toward web advertising, including entertainment, informativeness, and irritation.

Haghirian et al. [15] discussed the advertising value of mobile marketing in Austria, and suggested that media characteristics, including entertainment, informativeness, irritation, and credibility, significantly affect the perceived advertising value of consumers. Brackett and Carr [7] discussed similar message characteristics, including, entertainment, informativeness, irritation, and credibility, for understanding how college students perceive the advertising value of web advertising.

Personalized mobile advertising aims precisely at target customers and accommodates their needs according to their preference profiles and shopping experiences. This can further ensure that customers find the most appropriate and appealing mobile advertising and create positive benefits, ranging from improved advertising value or attitude toward the advertising to purchasing the advertised products [17]. Taken together, we identify five attributes of message content, entertainment, informativeness, irritation, credibility, and personalization, based on a comprehensive classification discussed above. By the same token, the five message attributes are important antecedents to determine advertising value. Accordingly, we thus propose the following hypotheses.

H2: Entertainment positively affects the advertising value of mobile messages.

 ISBN: 978-1-61208-418-3

H3: Informativeness positively affects advertising value of mobile message.
H4: Irritation negatively affects the advertising value of mobile messages.
H5: Credibility positively affects the advertising value of mobile messages.
H6: Personalization positively affects the advertising value of mobile messages.

D. Interactivity

Interactivity is the most salient feature of mobile communications [5] and is considered as an important factor in differentiating between new forms of media and traditional ones [23]. Advertisers can get immediate and direct feedback from the consumers through the mechanism of mobile advertising, which is based on interactivity [15]. In contrast, consumers could have the ability to choose and respond to particular advertisement of their liking [29].

Interactivity is generally believed to be a multidimensional construct [14]. Most previous studies have agreed that there are three major components in the exchange of communication between various parties, user control, two-way communication, and synchronicity [19]. A user, as a receiver, has more power to control the advertisement as mobile devices allow prompt, two-way communication online. Accordingly, interactivity was conceptualized as a formative construct of the three indicators. The reasons are similar to the above arguments.

In this study, we argue that interactivity is a critical factor affecting consumers' perceptions of advertising value in a mobile environment. One study attempted to examine the effect of perceived interactivity on attitude toward mobile advertising experimentally by manipulating the design of various interactive features to customers in order to understand their differences; the results indicated that interactivity was a strong predictor of attitude toward mobile advertising [14]. Moreover, a study of e-tailing investigated the role of interactivity, showing that consumers' interactivity with commercial messages plays an important role in creating perceived value of these messages and increasing customer satisfaction [28]. Based on these arguments, we propose the following hypothesis.

H7: Interactivity offered by mobile devices positively affects advertising value.

E. Consumer Loyalty

Consumer loyalty has long been recognized as an important issue in marketing [12]. Oliver [22] defined loyalty as "a deeply held commitment to re-buy or re-patronize a preferred product/service consistently in the future, thereby causing repetitive same-brand or same brand-set purchasing, despite situational influences and marketing efforts having the potentials to cause switching behavior." In particular, with rapid growth and proliferation of mobile services, it is necessary to know what factors mainly determine consumers' attitudinal commitment and behavioral intention continue using these services [18]. This study defines loyalty as consumer' behavioral intention to continuously use mobile advertising, as well as their inclinations to recommend mobile advertising to other people.

Advertising value refers to a subjective evaluation of the relative value or the usefulness of advertising by consumers [12]. It is a form of perceived value regarding mobile advertising. Several studies of consumer loyalty have suggested that when the perceived value of mobile services is low, customers are more inclined to switch to competitors in order to increase perceived value, thus contributing to a decline in loyalty. Turel et al. [25] looked into the usage of hedonic digital artifacts (i.e. mobile phone ringtones) from the perceived value perspective. Their results showed that consumers' perceived value can successfully predict behavioral usage in the future and leads to positive word-of-month intentions. Harris and Goode [16] indicated that perceived value of online services is a critical factor affecting customer loyalty. Accordingly, this study proposes the following hypothesis.

H8: Advertising value of mobile advertising positively affects consumer loyalty.

III. RESEARCH DESIGN

A survey study was conducted to collect empirical data. The research design is describe below.

A. Instrument

The instrument includes a three-part questionnaire. The first part uses a nominal scale and the rest uses a 7-point Likert scale.

1) Basic Information: We collected basic information about respondent characteristics including gender, age, education, occupation, type of mobile advertising (push or pull-based), and volume of mobile advertising received daily.

2) Antecedents of advertising value: Here, we measure the seven antecedents of advertising value. The items used for measuring involvement were adapted from the instruments developed by Huang et al. [19]. Involvement consists of three sub-constructs, person, stimulus, and situation, each including four items.

3) There are three items used for measuring entertainment, informativeness, irritaton, and credibility, adapted from the instruments developed by Brackett and Carr [7]. The three items used for measuring personalization were adapted from the instrument developed by Xu [40]. The three items used for measuring interactivity were adapted from the instrument developed by Gao et al. [14]. Interactivity is defined as three subconstructs, each including three items.

4) Advertising Value and Loyalty: This part measures the perceived value of mobile advertising and consumer loyalty. The three items used for measuring advertising value were adapted and revised from the instrument developed by [11]. The three items used for measuring loyalty were adapted from the instrument developed by [10].

B. Sample Design

Consumers qualified for this study require previous experience using mobile advertising. An online survey was placed in online communities to seek users as potential respondents. The online survey was placed in several larger online communities simultaneously. Website users (potential subjects) can reach the questionnaire with a hyperlink to its website when they access these online communities. A wider variety of data sources in terms of these different larger communities, were covered for the survey, allowing the responses to be more representative for the population. This

 ISBN: 978-1-61208-418-3

survey was carried out during the period of April-June, 2012. A reward system was also provided for the respondents. At least 30 participants were drawn from the response sample, with a reward of 10 USdollars.

C. Scale Validation

Initially, a pretest was conducted for the scale. The scale was carefully examined by selected practitioners and scholars in this area, including translation, wording, structure, and content. Their comments were used to modify the scale, in order to guarantee acceptable initial reliability and validity. When the questionnaire had been finalized, the online survey was performed, using the previously mentioned sampling procedure. A total of 533 respondents were received with a certain level of experience using mobile advertising. After 57 invalid responses were deleted, including 23 with no experience mobile advertising, there was a response sample of 476. The demographics showed that a high proportion of respondents (more than 80%) had the experience of receiving push-based mobile advertising and received at least 3 mobile messages daily.

In addition, common method bias was examined with Harman's single factor test [24]. We included all items for a factor analysis to determine whether the majority of the variance could be accounted for by one general factor. The result reported no single factor accounted for the bulk of covariance, leading to the conclusion that common method bias was not present.

D. Measurement Model

Partial Lease Square (PLS) allows latent variables to be modeled as either formative or reflective constructs and places minimal demands on sample size [8]. In this study, involvement and interactivity variables were formulated as a second-order structure with formative indicators. A measurement model was conducted for reliability and validity.

Reliability is assessed by the criterion that Cronbach's α is larger than 0.7. Convergent validity is assessed by three criteria: (1) item loading (λ) is larger than 0.70 (2) composite construct reliability is larger than 0.80, and (3) the average variance extracted (AVE) is larger than 0.50 [13]. Next, discriminant validity between constructs is assessed using the criterion that the square root of AVE for each construct should be larger than its correlations with all other constructs [13]. Item loadings range from 0.83 to 0.98, composite construct reliabilities range from 0.91 to 0.98, and average variances extracted (AVE) range from 0.75 to 0.95. The results indicate that all the constructs are highly acceptable for reliability, and convergent and discriminant validity.

IV. HYPOTHESIS TESTING

PLS was used for analyzing the structural model. This study uses bootstrapping analysis with 1000 subsamples to estimate path coefficients and relevant parameters, including means, standard errors, item loadings, and item weights. Next, we need to compute the coefficient of determination (R^2) for endogenous variables to assess the predictive power in the structural model. Figure 2 presents the results of the structural model.

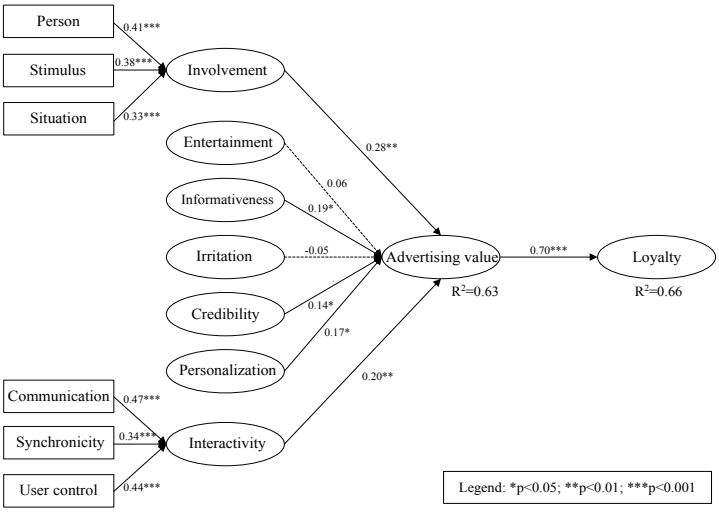


Figure 2. Result of the structure model

Hypotheses testing is reported in Table 1. For the psychological state, involvement was found to be a notable predictor of advertising value at the level of 0.01 (β=0.28). Hypothesis 1 is supported. Among the five message attributes, informativeness, credibility, and personalization were important in affecting advertising value at the level of 0.05 (β=0.19, 0.14, and 0.17), but entertainment and irritation were not important (β=0.06 and -0.05). Hypothesis 3, 5, and 6 are supported. However, Hypothesis 2 and 4 are not supported. For the technology use, interactivity had a significant effect on advertising value at the level of 0.01 (β=0.20). Hypothesis 7 is supported. The three issues, involvement, message attributes, and interactivity, jointly explain 63% of variance in advertising value. In turn, advertising value was critical in determining customer loyalty at the level of 0.001 (β=0.70). Hypothesis 8 is supported. It further explains a large of proportion of variance (R^2 =66%) in loyalty.

Table 1. Hypotheses testing

Hypotheses	Testing results
H1	Supported
H2	Not supported
H3	Supported
H4	Not Supported
H5	Supported
H6	Supported
H7	Supported
H8	Supported

This gives rise to a thinking of an important mediating role of advertising value in the realization of loyalty from the initial drivers. We based on a comparison between the original model and a competing model with extra paths for the antecedents directly to loyalty. Accordingly, the f^2 statistic is based on the difference between the R^2 in the two models, was used to assess their effect size [9]. This indicates a full mediating role of advertising value in realizing loyalty. Moreover, personal, stimulus, and situation are three important indicators in explaining involvement (W=0.41, 0.38 and 0.33, weight score). Interactivity is significantly explained by three key indicators, communication, synchronicity, and user control (W=0.47, 0.34, and 0.44).

 ISBN: 978-1-61208-418-3

V. FINDINGS AND DISCUSSIONS

Involvement with mobile advertising is the underlying basis for explaining a consumer's psychological state during the subjective evaluation of advertisement in terms of being personally affected by affective and cognitive relevance, such as personal, stimulus, and situational attributes. In order to understand the effect of involvement, researchers commonly manipulated involvement by leading subjects to believe one of these stimuli or situations. In their studies, subjects in both high and low-involvement groups receive the same communication messages. However, high-involvement subjects are led to believe the communicated message affects them, while low-involvement subjects do not believe the communicated message has a personally relevant effect. Practitioners therefore place their emphasis on the issue of involvement to increase advertising value.

Informativeness, credibility, and personalization remain as important predictors of advertising value and in turn, consumer loyalty, as in prior studies of mobile advertising [17]. Informativeness seems to be the primary predictor among three key factors. While mobile-based mechanisms are highly penetrable for their users to send or receive timely and important information, such as mobile advertising, consumers are often find it easy to understand the new products or services offered by marketers. Credibility indicates that the content of mobile advertising can be trusted or believed by consumers for their futures decisions. Marketers should take notice of the building of the initial trust of consumers as an important precursor for further being able to nurture and recognize the effectiveness of advertising in their purchasing decision process. Personalized mobile advertising is a major concern of consumers in relation to their willingness to read/receive messages as it targets the real needs of consumers. For marketers, the importance of personalization makes one-to-one mobile marketing a better marketing strategy. Marketers could take the initiative in sending personalized advertising by means of the collection and analysis of consumers' basic information, purchase records, or a combination with location-based service/awareness, to increase the return on marketing investment.

In contrast, entertainment and irritation show a non-significant impact on advertising value. The findings should be quite interesting to marketers. The reasons behind this can be explained as below. First, the most common way forms of mobile advertising are short message services (SMS) and multimedia messaging services (MMS). These advertisements are usually presented with texts, pictures, or together with hyperlinks. All of these look similar and monotonous. However, many marketers have started using the APPs (Applications) to increase the degree of pleasure and enjoyment during interaction with mobile advertising. In this study, a high proportion of respondents (more than 80%) have received pushed-based mobile advertising on a daily basis. Mobile advertising offers new opportunities to marketers to deliver messages in a personalized manner to meet the needs of consumers. It is thus recognized as a friendly type of service in terms of both meeting the needs and time restraints of individual consumers. As a result, consumers may gradually get accustomed to mobile messages and may no longer feel irritated about them.

Interactivity has positive relationship with advertising value, which is consistent with some previous studies [14]. In general, this study indicates the importance of considering the interactivity perspective in the mobile context. Marketers can take full advantage of the features of mobile devices (e.g., mobility, synchronicity, and two-way communication), and design a better form of interactive interface. System interfaces should be presented in a user-friendly way to enable good communication with consumers. Therefore, marketers can quickly get feedback from consumers and further provide them with the capability to adjust sale forecasts and marketing strategies in real time. In contrast, if consumers have the experience of interaction failures when using mobile advertising, they may not feel that mobile advertising is valuable, thus decreasing advertising value. Finally, advertising value is useful in evaluating the effectiveness of mobile advertising and is an important mediator for achieving customer loyalty when using mobile advertising. Advertising value should be considered as an important criterion when designing an effective mobile advertising mechanism for marketers.

VI. CONCLUSIONS AND SUGGESTIONS

Several important practical implications arise from our findings. A high proportion of respondents (83.6%) have the experience of receiving push-based mobile advertising. The primary work, in general, focuses on improving the value of mobile advertising so that consumers do not think mobile advertising as equivalent to spam.

For the involvement, marketers may need to look into consumer behavior to understand the decision-making process, including personal characteristics and external stimuli. For example, personal income and job would be considered by marketers concerned with delivering mobile advertising to the appropriate person in order to effectively induce purchase. External stimuli, such as price, may be sensitive to a variety of factors, such as preference for a particular brand, relative importance or perceived differences of product attributes, and promotions to target consumers. After such considerations, mobile marketing could be an important mechanism to effectively communicate with consumers and increase their level of involvement.

Next, message content are also important for advertising value. Efforts by marketers may focus on the following directions. In terms of informativeness and personalization, marketers should use a new mobile advertising approach, location-based service (LBS), which is pull-based mobile advertising. This service could facilitate timely delivery of customized marketing information to target consumers. In addition, marketers may often need to conduct surveys to understand the real requirements of consumers in order to reduce the gap between marketers/vendors and consumers. Regarding credibility, as building the initial trust belief is important for the effectiveness of mobile advertising, marketers may initially apply both traditional (print, radio, and TV) and mobile advertising approaches for promoting consumer recognition of marketers/vendors. Finally, while interactivity

 ISBN: 978-1-61208-418-3

implies that the mechanism of mobile advertising needs to be further enhanced for improving two-way communication channels and consumers' self-efficacy in using the advertising, marketers can create an effective conversation between marketers/vendors and consumers through better accessible hardware, higher speed networking systems, and user-friendly interfaces.

Some theoretical implications are also noted from the findings. We approach this research as an integration of three unique features. This will provide a new way of thinking for the future research. Based on this study, future research could focus on a variety of issues. As cross culture issues may result in important effects in the fields of consumer behavior and commercial advertising, future research can target consumers from different countries, to understand their differences and similarities. IT-based features, such as interactivity, are usually an important concern of users of mobile advertising.

Although this research has produced some interesting results, it may have a number of limitations. For example, push-based advertising occupies a high proportion (more than 80%) in the survey as compared to pull-based advertising. However, this is a major stream of current mobile advertising. The result does properly reflect the regular population distribution of subjects.

REFERENCES

[1] I. Ajzen, "The theory of planned behavior," Organizational Behavior and Human Decision Processes, vol. 50, issue 2, 1991, pp. 179-211.

[2] K. Amoako-Gyampah, "Perceived usefulness, user involvement and behavioral intention: an empirical study of ERP implementation," Computers in Human Behavior, vol. 23, issue 3, 2007, pp. 1232-1248.

[3] R. E. Anderson and S. S. Srinivasan, "E-satisfaction and e-loyalty: a contingency framework," Psychology and Marketing, vol. 20, issue 2, 2003, pp. 123-138.

[4] H. Barki and J. Hartwick, "Rethinking the concept of user involvement," MIS Quarterly, vol. 13, issue 1, 1989, pp. 53-63.

[5] S. J. Barnes, "Wireless digital advertising: nature and implications," International Journal of Advertising, vol. 21 issue 3, 2002, pp. 399-420.

[6] A. Bhattacherjee, "Understanding information systems continuance: an expectation-confirmation model," MIS Quarterly, vol. 25, isssue 3, 2001, pp. 351-370.

[7] L. K. Brackett and B. N. Carr, "Cyberspace advertising vs. other media: consumer vs. mature student attitudes," Journal of Advertising Research, vol. 41, issue 5, 2001, pp. 23-32.

[8] W. W. Chin, B. L. Marcolin, and P. R. Newstead, "A partial least squares latent variable modeling approach for measuring interaction effects: results from a Monte Carlo simulation study and an electronic-mail emotion/adoption study," Information Systems Research, vol. 14, issue 2, 2003, pp. 189-217.

[9] J. Cohen, "Statistical Power Analysis for the Behavioral Sciences," Lawrence Erlbaum, NJ, 1988.

[10] Z. Y. Deng, Lu, K. K. Wei, and J. Zhang, "Understanding customer satisfaction and loyalty: an empirical study of mobile instant messages in China," International Journal of Information Management, vol. 30, issue 4, 2010, pp. 289-300.

[11] R. H. Ducoffe, "How consumers assess the value of advertising," Journal of Current Issues and Research in Advertising, vol. 17, issue 1, 1995, pp. 1-18.

[12] R. H. Ducoffe, "Advertising value and advertising on the web," Journal of Advertising Research, vol. 36, issue 5, 1996, pp. 21-35.

[13] C. Fornell and D. F. Larcker, "Structural equation models with unobservable variables and measurement error: algebra and statistics," Journal of Marketing Research, vol. 18, issue 3, 1981, pp. 382-388.

[14] Q. Gao, P. L. Rau, and G. Salvendy, "Perception of interactivity: affects of four key variables in mobile advertising," International Journal of Human-Computer Interaction, vol. 25, issue 6, 2009, pp. 479-505.

[15] P. Haghirian, M. Madlberger, and A. Tanuskova, "Increasing advertising value of mobile marketing: an empirical study of antecedents," In: Proceedings of the 38th Hawaii International Conference on System Sciences, 2005, pp. 1-10.

[16] L. C. Harris and M. M. H. Goode, "The four levels of loyalty and the pivotal role of trust: a study of online service dynamics," Journal of Retailing, vol. 80, issue 2, 2004, pp. 139-158.

[17] G. S. Kim, S. B. Park, and J. Oh, "An examination of factors influencing consumer adoption of short message service (SMS)," Psychology and Marketing, vol. 25, issue 8, 2008, pp. 769-786.

[18] H. H. Lin and Y. S. Wang, "An examination of the determinants of customer loyalty in mobile commerce contexts," Information and Management, vol. 43, issue 3, 2006, pp. 271-282.

[19] P. B. Lowry, N. C. J. Romano, J. L. Jenkins, and R. W. Guthrie, "The CMC interactivity model: how interactivity enhances communication quality and process satisfaction in lean-media groups," Journal of Management Information Systems, vol. 26, issue 1, 2009, pp. 155-195.

[20] Y. A. A. Megdadi and T. T. Nusair, "Factors influencing advertising message value by mobile marketing among Jordanian users: empirical study," European Journal of Economics, Finance and Administrative Sciences, vol. 31, 2011, pp. 87-98.

[21] S. A. Nasco and G. C. Bruner, "Comparing consumer responses to advertising and non-advertising mobile communications," Psychology and Marketing, vol. 25, issue 8, 2008, pp. 821-837.

[22] R. L. Oliver, "Whence consumer loyalty?" Journal of Marketing, vol. 63, 1999, pp. 33-44.

[23] S. Petter, D. Straub, and A. Rai, "Specifying formative constructs in information systems research," MIS Quarterly, vol. 31, issue 4, 2007, pp. 623-656.

[24] P. M. Podsakoff, S. B. Mackenzie, J. Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: a critical review of the literature and recommended remedies," Journal of Applied Psychology, vol. 88, issue 5, 2003, pp. 879-903.

[25] O. Turel, A. Serenko, and N. Bontis, "User acceptance of hedonic digital artifacts: a theory of consumption values perspective," Information and Management, vol. 47, issue 1, 2010, pp. 53-59.

[26] C. S. Wan, S. H. Tsaur, Y. L. Chiu, and W. B. Chiou, "Is the advertising effect of virtual experience always better or contingent on different travel destinations?" Journal of Information Technology and Tourism, vol. 9, issue 1, 2007, pp. 45-54.

[27] E. Wang, "Forming positive advertising and product attitude: the role of product involvement," International Journal of Technology Marketing, vol. 6, issue 3, 2011, pp. 259-271.

[28] W. S. Yoo, Y. Lee, and J. Park, "The role of interactivity in e-tailing: creating value and increasing satisfaction," Journal of Retailing and Consumer Services, vol. 17, issue 2, 2010, pp. 89-96.

[29] J. L. Zaichkowsky, "Conceptualizing involvement," Journal of Advertising, vol. 15, issue 2, 1986, pp. 4-34.

 ISBN: 978-1-61208-418-3

SUMMIT: Supporting Rural Tourism with Motivational Intelligent Technologies

Mei Yii Lim, Nicholas K Taylor
School of Mathematical and Computer Sciences
Heriot-Watt University
Edinburgh, UK
e-mail: M.Lim@hw.ac.uk, N.K.Taylor@hw.ac.uk

Sarah M Gallacher
Department of Computer Science
University College London
London, UK
e-mail: S.Gallacher@ucl.ac.uk

***Abstract*— SUMMIT is a mobile app that aims to gamify the experience of walkers and hikers and benefit the local communities through which they perambulate. It encourages physical activity through gamification of the user experience by adding additional elements of social fun and motivation to walking and hiking activities. It rewards users for their physical effort by offering access to local resources, hence increasing awareness and appreciation of the local assets and heritage and contributing to the local economy. The evaluation results show that both businesses and walkers were very receptive to the idea.**

Keywords; location-based; gamification; personalisation; rewards; tourism.

I. INTRODUCTION

Romanticism era in the 18th century brought forth a shift in attitudes to the landscape and nature leading to the manifestation of the idea of walking through the countryside for pleasure [1]. An explosion of long distance walking routes occurred in the late 20th century with the Appalachian Trail in the USA [2] and the Pennine Way in Britain [3] as early examples. In Scotland, tourism figures from 2012 show that walking and hiking was the second most popular tourist activity among domestic visitors [4]. However, for many of these visitors, the walk can be the sole purpose of their trip and they may not access any other local attractions or local businesses.

SUMMIT is a location-based mobile app that encourages the walking and hiking community to avail themselves of local resources including hospitality businesses, product vendors, tourist attractions and local information. The key goal is to "gamify" the user experience by adding additional elements of social fun, motivation and rewards to walking activities whilst increasing cultural appreciation through promotion of the local amenities and services to the benefit of the local economy.

The idea behind SUMMIT is to challenge walkers and hikers to reach checkpoints (geo-fenced areas) that are located along popular walking and hiking routes. When walkers reach a checkpoint they are presented with a list of rewards on their mobile app from which they can choose their favourite. The rewards are provided by local businesses in the area and may include things like a free muffin or a 20% discount on a product. For example, if the walker decides to choose a free muffin as his reward at some checkpoint, he selects this in his app and a virtual muffin is added to his "reward knapsack". He then takes this virtual muffin to the local shop that offered this reward to exchange his virtual muffin for a real one. While he is there he may also buy a coffee or take friends with him who may also make some purchases.

In this way, SUMMIT benefits both walkers and local businesses. It encourages physical activity by making such activities more fun and rewarding but also introduces walkers and hikers to new local resources in the area that they might not have visited otherwise.

The rest of this paper is organised as follows. Section II presents the related work. The design of SUMMIT is presented in Section III along with an evaluation of the first prototype that was carried out in the wild on Arthur's Seat in Edinburgh, Scotland in Section IV. Section V details and discusses the results of the evaluation while Section VI concludes the paper.

II. RELATED WORK

Pervasive gaming takes the gaming experience into the real world, focusing on introducing game elements into the everyday life of players. It exploits interaction devices such as handhelds to display virtual elements [5], generates location-sensitive responses to interaction [6], employs technology through which human game-masters can exercise control of the game experience [7] and involves interactive actors to perform non-player characters [8].

Pirates! [9] was one of the first successful attempts to port the computer game into the real world and the IPerG project [10] has successfully executed a number of pervasive games in real spaces such as Epidemic Menace [11] and Day of the Figurines [12]. Other groups have produced educational pervasive games such as Virus [13] and Paranoia Syndrome [14]. Artistically oriented pervasive games such as Can You See Me Now? [5] used whole cities as the game environment. Ludocity [15], a collection of pervasive games inspired by theatre, painting, dance and other art forms also exploits public places such as city streets and parks for social play. All Ludocity games are released under creative common license which allows everyone to run the games for free. Ingress by Google [16] is a near real-time augmented reality massively multiplayer online pervasive game with a

 ISBN: 978-1-61208-418-3

complex science fiction back story and continuous open narrative.

On the other hand, SUMMIT is a real-world outdoor treasure hunt game using Global Positioning System (GPS)-enabled devices inspired by geocaching [17]. Analogous to geocaching, SUMMIT "hides" rewards of different categories at different places along a popular route for users to find and collect. These rewards reflect the distinctive resources offered by the local area and community encouraging users to appreciate and take advantage of the local amenities on offer. SUMMIT also logs users' achievements and allows them to perform social comparison of their performance against others, thus introducing a competitive element to the overall walking/hiking experience.

To date, quite a few treasure hunt based pervasive applications aiming at increasing cultural heritage appreciation have emerged including the Regensburg REXplorer game [18] the Global Treasure Apps [19], the National Museum of Scotland Apps [20] and Huntzz [21]. The main difference between these applications and SUMMIT is that these applications do not reward users based on physical achievements but on solving puzzles and clues. Only the Global Treasure Apps include real-world rewards but the focus is on promoting artifacts and attractions rather than local businesses and communities.

Although other stamping schemes for tourist checkpoints exist [22], these schemes usually require all checkpoints to be reached to validate the completion of a tour with the aim of collecting badges or similar rewards. On the other hand, SUMMIT users do not need to reach all checkpoints to collect rewards and have the flexibility of choosing their desired rewards. Instead of automated checkpoint verification, the stamping schemes involve manually dating and stamping of a personal completion brochure or manually entering codes collected from checkpoints on the respective websites for electronic validation.

III. The SUMMIT System

The SUMMIT system consists of two main components: a web app which allows business users to manage the rewards that they provide, and a mobile app, which is used by the walkers and hikers. Fig. 1 illustrates the SUMMIT system deployment including the server where information about business users and app users are stored.

The web app was developed to enable easy sign-up of local businesses as reward providers. Once registered as business users, they can perform the actions depicted in the workflow diagram in Fig. 2. The supplier can add, edit or delete a business. They can add, deactivate, re-activate, delete and edit a specific reward item. They can also approve claims from the mobile app users. Fig. 3 shows the web app dashboard which displays the list of businesses and rewards own by a provider as well as the available actions. Alert icons will appear beside reward items that reach zero count so that the provider can decide to add more of the reward or delete it.

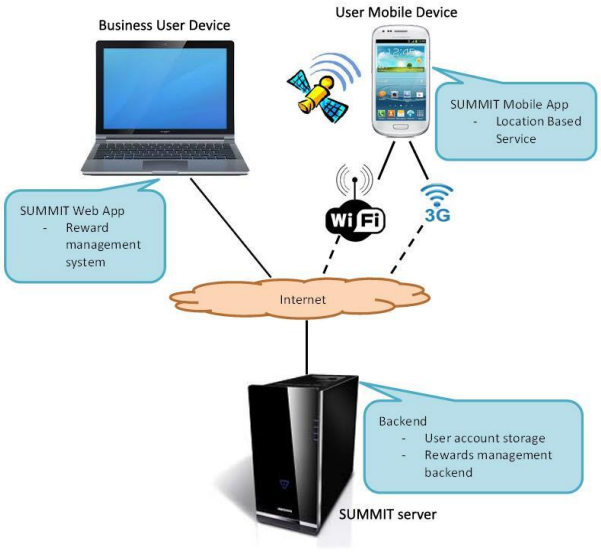


Figure 1. SUMMIT system deployment.

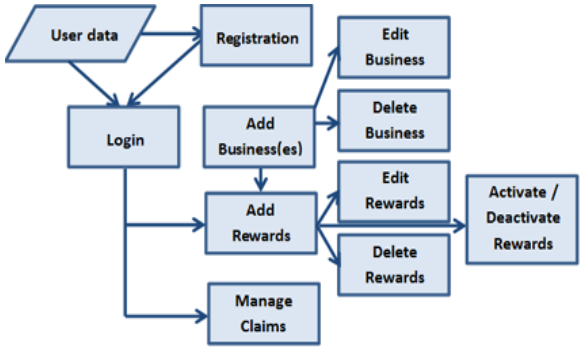


Figure 2. SUMMIT web app workflow.

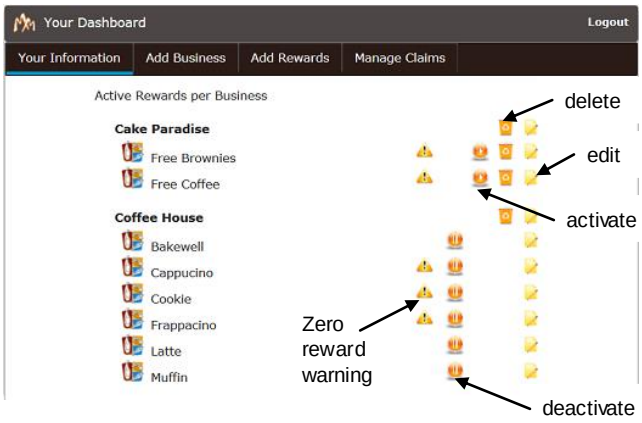


Figure 3. Web app dashboard.

The mobile app was developed for the Android platform. It aims to enhance the walking activities by supporting the users' personal achievement element through a reward scheme and the social competition element through comparisons of their progress against others via the social network site, Facebook. The mobile app monitors the users'

outdoor locations while they are en route using GPS. Each route has several pre-defined checkpoints, usually selected based on their touristic values to the respective region that are geo-fenced areas. The app does not provide real-time navigation but as users reach checkpoints, the phone will start vibrating and notifications will appear on the system bar. When this happens, users will unlock new virtual reward items provided by local businesses which they can exchange into real rewards. They can also post their achievement onto Facebook if they wish.

Prior to the game, users can check out different routes and rewards associated with each of the routes. They can then select a route that provides the rewards they desire and suits their constraints in terms of time and distance. This flexibility enables users to customise their gaming experience based on their needs at any particular time. Fig. 4 shows the workflow of the mobile app. When users select a route, the route information will be downloaded onto their phone assuming Internet connection is available. By pre-loading the routes, the issue of unreliable 3G signal is avoided as the route information is now locally stored, hence will always be available to users when en route. During the hike, only GPS signal is required to track users' position. Since each checkpoint covers an area of 50-metre radius, a short lost of GPS signals will not affect the performance of the app. These approaches give users the Virtual "Always-On" Connectivity impression allowing them to have an undisrupted interaction experience. The problem of draining the battery power is also minimised as the phone is not constantly connected to the network. Synchronisation with the server occurs the next time network connectivity is available and activated by the user when all logged data on the mobile device is uploaded. To help users locate the rewards, a map that shows the locations of the different checkpoints is provided as illustrated in Figure 5(a). Fig. 5(b) shows a rewards selection dialog box.

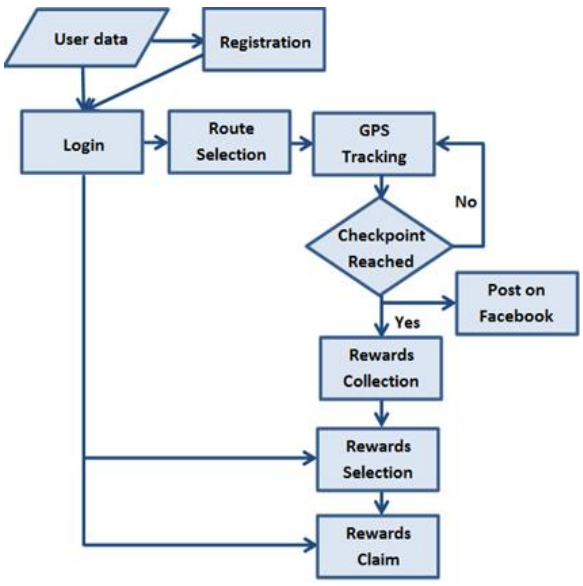


Figure 4. SUMMIT mobile app workflow

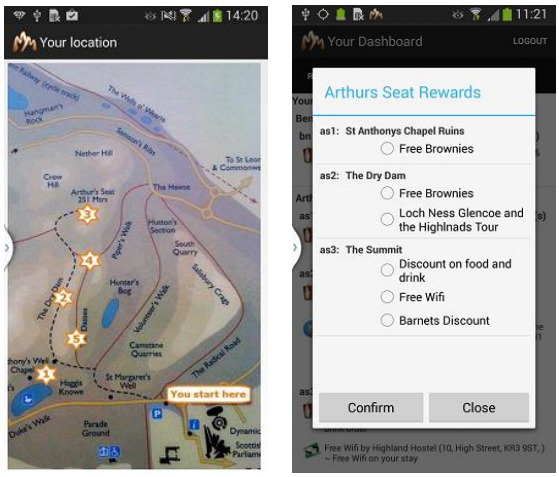


Figure 5. (a) Map with checkpoints, (b) Rewards selection dialog box

IV. EVALUATION

A trial of the SUMMIT mobile app has been carried out with 24 participants; 18 males and 6 females. Participants were volunteers who are either interested in mobile applications or hikers and walkers. They were issued with Samsung Galaxy SIII phones with the game pre-installed and were asked to hike up Arthur's Seat, a popular rural area within the City of Edinburgh. After participants had played the game, they were asked to complete a questionnaire. They were asked to rate different features of the game on a 5-point Likert scale. These features included – S1: route information, S2: map, S3: rewards motivation, S4: advance knowledge of rewards, S5: rewards selection, S6: claim system, S7: rewards choices, S8: claim intention, S9: Facebook functionality and S10: Ease of use of the app. Additionally, they were given the freedom to provide further comments about any part of the game or their experience of using the mobile app. Please refer to Appendix I for the list of questions.

A total of 7 businesses signed up to the rewards scheme. A week after the trial ended, they were contacted to gather their feedback on the web app. The questions included the number of customers the mobile app brought into the shops, other desired features for the web app and free comments on the web app and their experience in using it. Please refer to Appendix II for the list of questions.

V. RESULTS AND DISCUSSION

The chart in Fig. 6 shows the overall average rating of all 24 participants. On average participants were neutral on the usefulness of the route information (S1). Taking the level of significance, $\alpha = 0.05$, a Mann-Whitney test on this variable between the younger (less than 40 years old, n=17, M=3.418, SD=0.425) and the older (more than 40 years old, n=7, M=2.286, SD=0.694) users showed a significant difference with U(24)=13, Z=-3.050, p=0.002 (see Fig. 7).

 ISBN: 978-1-61208-418-3

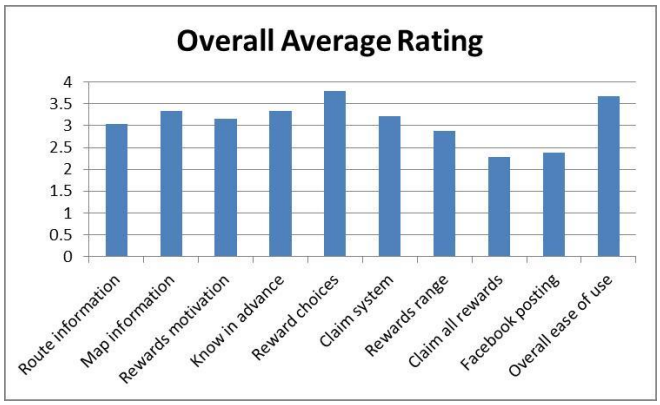


Figure 6. Overall average rating of all 24 participants.

The older generation found the route information not useful while the younger generation found it useful. This might be because the older users were used to using guidebooks when walking and were expecting directional information such as descriptions of terrain and photographs of each checkpoint which was not provided via the app and might have led to some of them getting lost along the way.

On average the participants found the map informative (S2). However, they would have preferred an interactive map. The participants found the rewards motivated them to go on the hike (S3). Some participants mentioned in their questionnaire that the rewards served as an initial motivation. As they moved from one checkpoint to another, the fact that there were rewards attached to each of the checkpoints became less important to them and, instead, their ultimate goal was to reach all the checkpoints and complete the route. This interesting finding suggests that the gamification aspect of numbering the checkpoints itself provided enough motivation for the user to carry on once they had started.

Overall, participants thought that it was important to know the rewards in advance (S4) and to be given the option to choose from a selection of rewards (S5). They also found the claim system easy to use (S6).

In terms of the range of rewards provided (S7), there was again a significant difference between younger (n=17) and older (n=7) users. This was revealed by applying the Mann-Whitney test, with U(24)=17, Z=-2.842, p=0.005, to the results in Fig. 7. The younger users seemed to be satisfied with the type of rewards provided which included discounts on food, drinks, shoes, sweets, clothes, souvenirs and tours, while the older users were not. The older users would have liked some rewards that they could redeem immediately after the hike, for example, refreshments, discount at a local hotel or B&B and rewards targeted at kids.

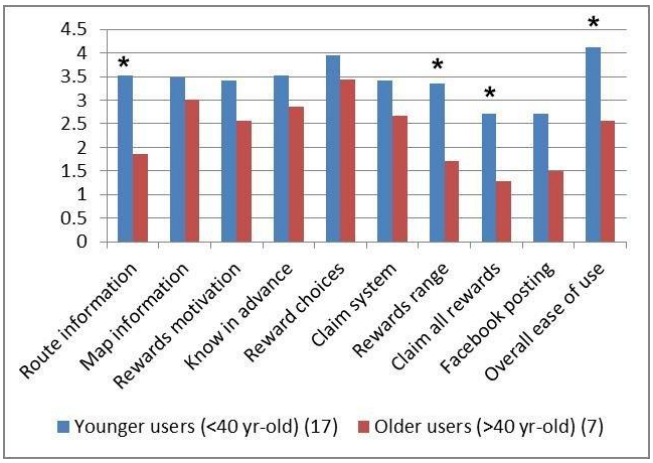


Figure 7. Average rating comparison between younger and older users. (* denotes variables with significant differences, $\alpha = 0.05$)

The intention to claim all rewards (S8) also revealed a significant difference between the two age groups. Many of the participants were exhausted after the hike and selected their rewards only after they were home and once they were not in the vicinity of the shops. They were therefore less keen to make the effort to return to the area to collect their rewards at a later date. Again, the Mann-Whitney test showed a significant difference between the older and younger users, U(24)=21.5, Z=-2.529, p=0.013. Since the older users were less interested in the rewards, they were also less inclined to claim them. The participants rated the ability to post their achievement onto Facebook (S9) fairly low. One reason for this might be because, as the game was only a prototype, the users had to use test user accounts instead of their own accounts. As a result the achievement posts did not appear on their own Facebook wall or timeline. Finally, the average rating for ease of use of the app (S10) was good. However, there was again a significant difference between the older and younger users, as confirmed by the Mann-Whitney test, U(24)=17.5, Z=-2.801, p=0.005. This could be due to the fact that the younger users were more accustomed to gaming and more familiar with mobile apps and thus had a better idea about the flow of control and operations of the mobile app and phone in general.

The feedback on subjective questions revealed that some participants would have liked the mobile app to provide more interesting information about the route and checkpoints. One of the experienced hikers suggested that it would be useful if the app could show real-time progress such as the time he took to go from one checkpoint to another and the overall time he took to complete the route. This would allow users to compare their real-time progress with each other, hence increasing the competitive element of the game. The game has also been found to provide motivation for a second time visitor to hike a hill/mountain that they have conquered before, as one of the participants stated:

> "Thanks for giving me a reason to walk up Arthur's Seat. I am feeling revitalised and refreshed now I'm

 ISBN: 978-1-61208-418-3

home! This serves as an excellent reason to walk up hills/mountains that you have already conquered (I've been up Arthur's Seat twice)."

From the perspective of the suppliers, overall they were very satisfied with the usability of the web app. Unfortunately they were not very meticulous in recording the actual rewards that were redeemed so we are unable to report actual numbers but we were assured that rewards were indeed claimed.

In order to encourage claims after the trial, one of the suppliers offered an additional deal on top of those provided on the mobile app if participants claimed within a particular period of time. The suppliers remained very enthusiastic about the SUMMIT system following the trial and one of the suppliers suggested that it might be useful to include an online claim facility which might encourage more claims as the participants would be able to redeem their rewards anywhere at their own convenience.

VI. CONCLUSION

SUMMIT has successfully added the elements of social fun and motivation to walking and hiking activities. It helps to promote local resources around a route by making users aware of their existence through its rewards scheme and checkpoints assignment. Business users were satisfied with its ease of use and appreciated its potential as a useful medium for advertising and delivering their wares and services. The authors are now investigating other ways in which the concepts and business model underlying SUMMIT can be deployed to support the mobile, rural and tourist sectors.

ACKNOWLEDGMENTS

This work was supported by dot.rural, University of Aberdeen and the SICSA Smart Tourism Programme. The authors are solely responsible for the content of this publication.

REFERENCES

[1] *The Norton Anthology of English Literature*, ed. M. H, Abrams, 7th ed., vol. 2, pp. 9-10, 2000.

[2] Appalachian Trail FAQ, http://www.outdoors.org/conservation/trails/appalachian-trail/at-faq.cfm, available online, accessed May 09, 2015.

[3] National Trails: Pennine Way, http://www.nationaltrail.co.uk/pennine-way, available online, accessed May 09, 2015.

[4] Visit Scotland – Key Facts on Tourism 2012, http://www.visitscotland.org/pdf/VS%20Insights%20Key%20Facts%202012%20%282%29.pdf, available online, accessed May 09, 2015.

[5] S. Benford et al., "Can you see me now?," ACM Transactions on Computer-Human Interaction vol. 13, no. 1, pp. 100–133, 2006.

[6] G. Chen and D. Kotz, "Solar : A pervasive computing infrastructure for context-aware mobile applications," Technical Report TR2002-421, Department of Computer Science, Dartmouth College, 2002.

[7] J. Soderberg, A. Waern, K. P. Akesson, S. Bojrk, and J. Falk, "Enhanced reality live role playing," Workshop on Gaming Applications in Pervasive Computing Environments, Second International Conference on Pervasive Computing, Vienna, Austria, 2004.

[8] J. Stenros, "Between Game Facilitation and Performance," International Journal of Role-Playing, Special Issue: Role Playing in Games, Issue 4, pp. 78-95, 2013.

[9] S. Bjork, J. Falk, R. Hansson, and P. Ljungstrand, "Pirates! Using the physical world as a game board," Proceedings of Interact '01, 2001, pp. 9-13.

[10] IPerG: Integrated Project on Pervasive Gaming. http://iperg.sics.se, available online, accessed March 17, 2015.

[11] I. Lindt, J. Ohlenburg, U. Pankoke-Babatz, and S. Ghellal, "A report on the crossmedia game Epidemic Menace," Computers in Entertainment, vol. 5, no. 1, 2007.

[12] M. Flintham, G. Giannachi, S. Benford, and M. Adams, "Day of the Figurines: Supporting Episodic Storytelling on Mobile Phones," Virtual Storytelling, Using Virtual Reality Technologies for Storytelling, Lecture Notes in Computer Science, Vol. 4871, 2007, pp. 167-175.

[13] V. Collella, R. Bororvoy, and M. Resnick, "Participatory simulations: Using computational objects to learn about dynamic systems," SIGCHI conference on Human factors in computing systems (CHI'98), Los Angeles, USA, 1998, pp. 9-10.

[14] G. Heumer et al., "Paranoia Syndrome - A pervasive multiplayer game using PDAs, RFID, and tangible objects," Third International Workshop on Pervasive Gaming Applications on Pervasive Computing 2006, Dublin, Ireland, 2006.

[15] Ludocity, http://ludocity.org/wiki/Main_Page, available online, accessed May 09, 2015

[16] Ingress: Niantic Labs, https://www.ingress.com/, available online, accessed May 09, 2015.

[17] Geocaching: http://www.geocaching.com, available online, accessed May 09, 2015.

[18] R. Ballagas, A. Kuntze, and S. P. Walz, "Gaming tourism: Lessons from evaluating REXplorer, a pervasive game for tourists," Pervasive Computing, Lecture Notes in Computer Science, Vol 5013, pp.244-261, 2008.

[19] Global Treasure Apps: http://globaltreasureapps.com/, available online, accessed May 09, 2015.

[20] National Museum of Scotland: Museum Apps. http://www.nms.ac.uk/our_museums/national_museum/museum_explorer_app.aspx, available online, accessed May 09, 2015.

[21] Huntzz Treasure Everywhere, 2011, http://www.huntzz.com/about-us.html#.VU4sjZNWIhQ, available online, accessed May 09, 2015

[22] Proposed features/Checkpoint for Tourism, http://wiki.openstreetmap.org/wiki/Proposed_features/Checkpoint_for_Tourism, available online, accessed May 09, 2015

 ISBN: 978-1-61208-418-3

Appendix I: SUMMIT Android App User Trial Questionnaire

About You

Username : _____

Age : _____

Gender : male female

Prior experience with mobile apps:
novice intermediate experienced

Hiking experience :
novice intermediate experienced

How often do you go on hiking trips? _____

About SUMMIT Android App

Please rate your degree of agreement with the following statements: From Disagree (1) to Agree (5)

1) The route information was useful

2) The map was informative

3) The rewards motivate me to continue hiking

4) It is important to know what rewards are available in advance

5) It is important to be given some choices of rewards to select from

6) The reward claim system was easy to use

7) I found the rewards useful

8) I intend to claim all the rewards I have chosen

9) I found the ability to post my achievements onto Facebook useful

10) Overall, the SUMMIT Android App was easy to use

_____

What other type of reward would you like to be included?

_____

Other comments

_____

Appendix II: SUMMIT Web App User Trial Questionnaire

About You

Age : _____

Type of business : _____

Have you used any app for advertising purposes before? :
Yes No

About SUMMIT Web App

Please rate your degree of agreement with the following statements: From Disagree (1) to Agree (5)

The registration process was straightforward

It is easy to add business(es)

It is easy to add reward(s)

The claim management system is easy to use

The SUMMIT App is a useful advertising medium

_____

Did the SUMMIT app bring you customers? If yes, how many?

_____

What other features would you like the app to provide?

_____

Other comments

_____

 ISBN: 978-1-61208-418-3

Smart Spaces Approach to Development of Recommendation Services for Historical e-Tourism

Aleksey G. Varfolomeyev
Faculty of Mathematics
Petrozavodsk State University
Petrozavodsk, Russia
e-mail: avarf@petrsu.ru

Dmitry G. Korzun
Department of Computer Science
Petrozavodsk State University
Petrozavodsk, Russia
e-mail: dkorzun@cs.karelia.ru

Aleksandrs Ivanovs
Department of History
Daugavpils University
Daugavpils, Latvia
e-mail: aleksandrs.ivanovs@du.lv
Research Institute for Regional Studies (REGI)
Rezekne University
Rezekne, Latvia

Oksana B. Petrina
Faculty of Mathematics
Petrozavodsk State University
Petrozavodsk, Russia
e-mail: petrina@cs.karelia.ru

***Abstract*—Recommender systems support advanced services in many domains. In this paper, we study personalized recommendation services for historical e-Tourism. The considered recommendation-making is based on a corpus of historical data distributed over multiple sources and requires taking into account various semantic relations between historical objects. We adopt a multi-agent architecture to develop services using the smart spaces approach. In contrast to the existing web-based solutions and mobile standalone applications, we propose to provide a tourist with a smart space. It integrates and self-generates historical and other relevant information. Based on this information, personalized recommendations with quantitative and qualitative estimates are constructed and then visualized to assist the tourist in history-aware analysis of points of interests.**

Keywords–Historical e-Tourism; Recommender Systems; Point of Interest; Smart Spaces; Personalized Services.

I. INTRODUCTION

Currently, the role of e-Tourism recommender systems is growing. A lot of research has been done, especially in the direction of making services of such systems mobile and intelligent [1][2]. A topical subdomain is historical tourism [3][4], which we distinguish from a more general cultural heritage tourism. In particular, the historical tourism focuses on visiting historical Points of Interest (POI) and on studying their relation with other historical objects (POIs, events, persons, etc.).

For a historian tourist, a POI is recommended not only if it is nearby and within the user's interests. Such a tourist would like to see a spectrum of historically related POIs; some are closely located and some can be faraway. Clustering important POIs for recommendation needs semantic analysis of their relation with historical objects and can be performed by means of ontologies [5]. The content for reasoning about, however, must be extracted from some historical databases or archives. Moreover, some historical relations are subjective, e.g., depend on context or personal vision of historical facts.

The study represented in this paper is motivated by lack of "smart" assistants for historian tourists, although there is a lot of them for mobile e-Tourism in general [1]. Based on our previous work [6], we expect that practical development of recommendation services with built-in semantic analysis of historical data can be implemented using ontology-based technologies of Semantic Web. Furthermore, traditional web-based architectures and mobile standalone applications seem insufficient for this development. We focus on the emerging approach of smart spaces [7][8]. A ubiquitous computing environment is created where mobile users, multisource data, and various services constructed over these data are intelligently connected based on ontology-driven information sharing and self-generation. Services can be personalized by means of augmentation of personal data to the shared content and customization of required reasoning about the content.

We continue our research [6] on the historical POI recommendation problem. The scope of this paper is limited with such important parts of the service development as concept definition and design. We provide a reference scenario of recommendation services for historical e-Tourism. We present a smart space based architectural design solution for the studied class of recommender systems. The contribution consists of the concept definition and system design for creating a personalized smart space to assist the historian tourist by means of recommendations.

The rest of the paper is organized as follows. Section II discusses the related work motivating development of recommendation services for historical tourism. Section III introduces our reference scenario for historical e-Tourism services. Section IV describes our system design based on the smart spaces approach. Section V analyses our proposal contrasting it with the previous work. Section VI concludes the paper.

 ISBN: 978-1-61208-418-3

II. Motivation and Related Work

Historical tourism has distinctive features [3] compared with a more general application domain of cultural heritage tourism. The latter embraces both historical and present-day cultural phenomena. Its main focus is not on historical events and persons, but on artifacts (e.g., artworks, architectural monuments) and cultural traditions (e.g., festivals, cuisine). According to Nora [4], historical tourism addresses the so-called "sites of memory". They present any material traces of historical events, which sometimes coincide with cultural heritage artifacts. For instance, an architectural monument is directly "involved" in historical developments related to its construction. The other example is any place or a spot associated with a historical event. Traces of historical facts are presented in the multitude of historical sources, including open sources in the Internet.

In general, a point of interest (or attraction) is an actual spot with precise localization on the geographical map (e.g., geo-position coordinates or postal address). Nowadays, POI recommender systems form an important services class in e-Tourism [1]. In addition to POIs, historical tourism takes into consideration a lot of other historical-valued objects such as persons, events, and data sources (written records and narratives, alternative information sources, data and knowledge bases available on the Web). Relations between historical objects contain important semantics [9], localized in space and time. For instance, an event might be conditionally defined as a semantic relation between several historical objects [10].

Ontologies become of high application interest for knowledge representation and reasoning in historical research [5] and e-Tourism [2]. In historical tourism, we expect that the introduced semantic relations can be effectively represented and manipulated using technologies of the Semantic Web. To the best of our knowledge, no specialized knowledge base that comprise semantically enriched information about historical objects has been created yet, e.g., see [11]. To a certain extent, a corpus of historical information is represented in ontological form in such knowledge bases as DBpedia, Freebase, or YAGO. Additional information can be extracted from web publications of historical sources [6][12]. In these settings, the methods of web-based systems, mobile programming, and multi-agent systems provide means for implementation of data search, access, and reasoning [1][2].

There are mobile services for cultural heritage e-Tourism developed using semantic technologies, see survey [13]. For instance, an intelligent tourist guide [14] utilizes cultural heritage information. Nevertheless, the present-day developments do not take into account the principal peculiarity of historical tourism—semantic relations among historical data.

Methods of ubiquitous computing and, in particular, the recent progress in communication technologies of the Internet of Things make possible creating environments where diverse devices and computer systems cooperatively construct services surrounding the user [8][15]. New programming paradigms emerge, such as smart spaces [7]. A smart space supports cooperation by establishing a shared view of resources in the environment. The shared view is ontology-based, applying the technologies of Semantic Web. For e-Tourism services, a smart space is mobile and personalized, i.e., created around a traveling tourist, attracting appropriate web services and other data sources from the Internet [6][16][17].

The discussion above motivates our research focus on POI recommendation services for historical tourism. First, semantic relations between historical objects cannot be bypassed in recommendation making. Second, there is no single source of needed information. The latter is distributed within multiple sources, each represents the information either in ontological form or requires an extraction procedure. Third, a historian tourist needs personalized services, i.e., source information and the result are subject to her/his preferences and context. Last but not least, a recommendation service is "ubiquitous", i.e., the service intelligently accompanies its mobile tourist.

III. Recommendation Scenario

Consider a historical POI recommender service that provides personal assistance for a tourist during her/his journey. Table I summarizes our formal symbol notation for the reference recommendation scenario.

Let P consist of POIs the service accesses from multiple sources. Typical examples of historical POIs are buildings, monuments, fountains, bridges, squares, etc. Spacious objects, such as streets or rivers, are not POIs. As a rule, a historical POI has a particular name. There may be several names associated with a given $p \in P$, e.g., due to historical developments or due to the use of different languages. Each POI has distinctive properties: coordinates and/or address, date, architect, etc. In the service, POIs and other historical objects form the set H.

Historical POI recommendation is essentially based on relations between the elements of H. First, "direct" links exist between historical objects. For instance, a person x is the architect of a building y. Second, links can appear between objects due to similarity. For instance, two buildings have been constructed by the same architect or they are located on the same street. Third, links are a result of involving diverse objects and POIs in a common historical event. Therefore, a semantic network G with nodes H can be constructed.

The historical relation semantics can be personified: some relations are treated differently by different historians or dependently on a context. For instance, there can be several visions of the role of a person for a certain POI. An important context correpsonds to an initial POI; a tourist selects $p \in P$ to consider semantic graph G_p. The first type of links mentioned above typically represents some stable and widely accepted historical relations. The last two types of links are the result of inference, and can be clear subject to context and personalization.

TABLE I. SYMBOL NOTATION

Symbol	Description
H	The set of all historical objects $H = H_1 \cup \ldots \cup H_n$ derived for the consideration from n data sources. Overlapping $H_i \cap H_j \neq \varnothing$ is possible.
$P \subset H$	The set of all POIs $P = P_1 \cup \ldots \cup P_n$, where $P_i \subset H_i$ for any data source i.
G, G_p	Semantic network G where nodes are from H and links are historical relations. In G_p, an initial POI $p \in P$ is fixed.
O	Ontology O describes the historical domain: possible classes and properties of historical objects as well as relations and restrictions for them.
R_p	Star graph R_p is a POI recommendation, where the internal node is the initial POI $p \in P$ and leaves are recommended POIs $q \in P$.
$r_q > 0$	Real-valued rank r_q shows the recommendation degree of $q \in P$ in respect to the initial POI p.
t_q	Annotation t_q summarizes (in a human-readable form) the reason of recommending q if the initial POI is p.

 ISBN: 978-1-61208-418-3

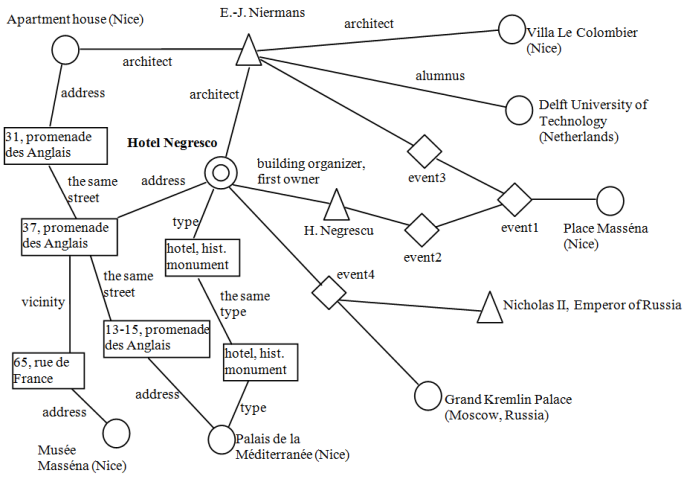


Figure 1. Sample semantic network: historical relations built around Hotel Negresco.

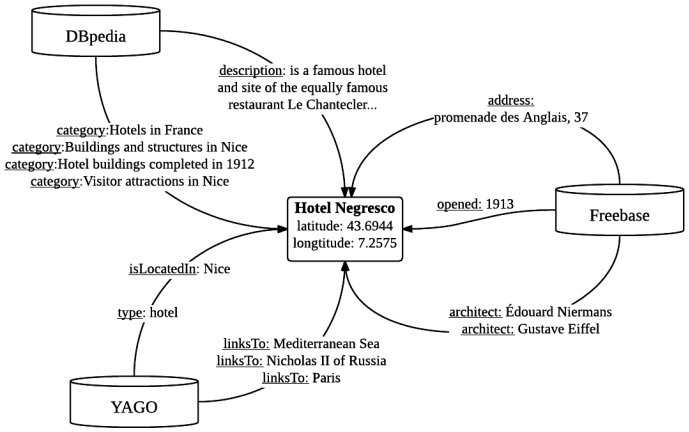


Figure 2. Hotel Negresco: information extraction from different sources

Figure 1 shows a sample semantic network that is built around Hotel Negresco, one of the most famous buildings of Nice. Small circles are POIs, triangles denote historical persons, text rectangles describe POI properties, and rhombuses represent historical events. The initial POI—Hotel Negresco—is linked with seven other POIs (five of them are located in Nice). The links are based on different properties: one and the same architect, close location, involvement in common historical events, etc.

Hotel information is extracted from different sources such as DBpedia, Freebase, or YAGO. Figure 2 shows an example of extracting facts from such sources. The search for POIs is usually performed by their coordinates. In DBpedia, historical description of each POI and categories the POI has can be found. Freebase provides formal attributes, such as POI address, the date of opening of the hotel, and architects. This information allows relating a POI with other entities.

Based on G_p, a tourist would like to understand which POIs are interesting for her/his personal consideration from historical perspective. An important context for this understanding is that she/he starts from p (e.g., being actually or virtually in this POI). The recommendation result can be represented as a

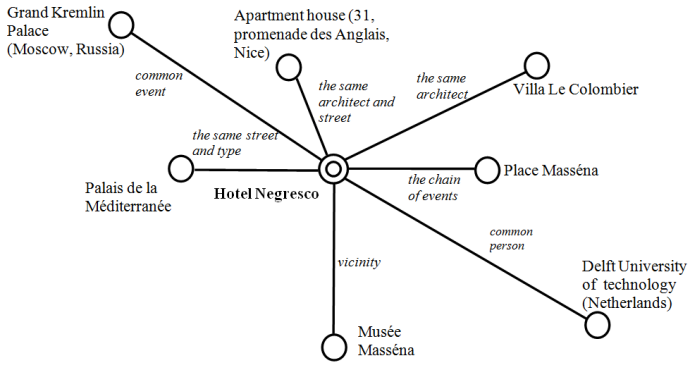


Figure 3. Star graph example: Hotel Negresco and recommendations on its historical surrounding.

star graph R_p. Its internal node is p and leaves represent all recommended POIs $q \in P$. Ranks $r_q > 0$ can be associated with the POIs to describe the recommendation degree of q (the higher rank the more recommendable). In a visual representation of G_p the length of edge (p, q) is proportional to the rank. Additional annotations t_q, which describe the reason of recommendation (in an aggregative form), can be also associated. Visual layout of R_p can also take into account the geographical position of q in respect to p (e.g., when q is on the North-East of p).

An example of recommendation is shown in Figure 3. The star graph is derived from the semantic network of Figure 1. POI geographical positions are not reflected.

Our reference scenario of a recommendation service consists of the following steps.

Step 1: Initial POI selection. It can be made either manually (e.g., pointing out coordinates, a spot on the map, or POI's name) or automatically (e.g., within a definite area pointing out the nearest POI). The area can be either set by a tourist (for instance, on the map), or determined automatically taking into consideration the current tourist's location.

Step 2: Semantic network around the initial POI. The sets H and $P \subset H$ as well as semantic relations among them are searched and retrieved from available knowledge bases and other data sources. Since the network G_p is potentially infinite, the process is limited. For instance, if the construction reaches another POI q from p, then the search for additional historical objects interconnected with q is terminated. Note that the path from p to q is subject to analysis in order to derive the reason of recommending q (construction of an annotation t_q). This example of limiting the construction process straightforwardly leads to a star graph R_p.

Step 3: POI ranking. Differentiation of recommended POIs can be based on ranks r_q. They are computed based on tourist preferences. For instance, he/she wants to find a building constructed by the same architect, an edifice built in the same architectural style, or another historical building located on the same street. Such preferences can be defined in the user's profile. They can be manually defined for the initial POI (before the implementation of the second step of this scenario). A significant component of the user's profile is the history of the choices of previous initial POIs (e.g., history of visits). For instance, the previously chosen POIs acquire lower ranks, since these POIs should not be repeatedly recommended.

 ISBN: 978-1-61208-418-3

Step 4: Visualization. The recommendation results achieved on Steps 2 and 3 are visually presented in a user friendly way, i.e., by means of a star graph possibly augmented with a map and textual/visual descriptions. For instance, annotations t_q show the reasons of the provided POI recommendations.

Step 5: Feedback. The recommendation process is iterative. Based on the presented results (the star graph with ranks and annotations), the tourist supplements this information by expanding the semantic network G (additional data retrieved from historical sources). The process—supplementing and expanding network—is represented in G: new historical events appear in which both the user and the objects are involved. The user becomes a historical person—a network node in G.

IV. Smart Space based Design

The smart spaces paradigm considers computing networked environments equipped with a variety of devices and with access to the Internet [7][8][15]. Software agents—knowledge processors (KPs)—run on the devices and interact via information sharing. A semantic information broker (SIB) is a mediator for information collection and exchange. Each KP produces its share of information and makes it available to others via the SIB. Similarly via the SIB, a KP consumes information of its own interest. The information storage employs RDF (Resource Description Framework) [18]. Agents can apply such advanced Semantic Web technologies as SPARQL Protocol and RDF Query Language or Web Ontology Language (OWL) for shared information maintenance, search, and reasoning [19].

This programming paradigm suits well for the development of e-Tourism services, as recent works [6][16][17] indicated. Figure 4 shows a high-level architecture that we adopted from [6] for the case of historical recommendation services.

The recommendation service is constructed by cooperative work of multiple KPs on historical data and other information. Consider properties of the proposed architecture to analyze the advantages that the smart spaces approach provides to the development of recommendation services for historical tourism. Some advantages are valid for the more general e-Tourism case.

Popular architectural styles for e-Tourism recommender systems are web-based, agent-based, and mobile [1][2]. Smart spaces support them to be applied in a composition. SIB is deployed on a host machine in the Internet, similarly as it happens with web services now. Each user (tourist) is mobile, acts using her/his client KP (e.g., on smartphone or tablet), and consumes the service anywhere and anytime. Other KPs produce the information collecting it in the smart space for the use by the service. They can be hosted on the same machine with SIB (web-like solution) or on other computers (agent-based solution). The latter property leads to higher flexibility for system deployment. For instance, some KPs are provided by a travel agency and some KPs are from the user side in order to augment the system for personalized operation.

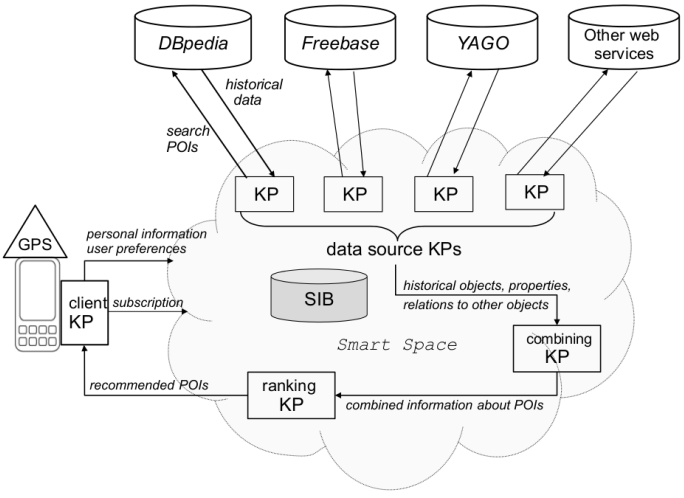


Figure 4. Multi-agent architecture: historical data from various sources and other information are semantically related and analyzed in the smart space.

The recommendation service becomes not attached to a fixed source of historical information. A wide pool of available sources is used, where a data source KP is assigned per source (DBpedia, Freebase, YAGO, etc.). Configuration of the pool is flexible and subject to dynamic inclusion/exclusion. Some data source KPs are set up by system administrators. Some KPs can be attached by the users if the appropriate rights are delegated. Each data source KP has to implement its source-specific interface to access and search for information. Note that a client KP can also provide historical information to the smart space, in addition to her/his preferences, context, and control. The information is further used for personalization.

The function of data source KPs is to extract historical information from two key types of data sources. The first type is tourism-oriented or universal knowledge bases (e.g., DBpedia). They store many POIs and associated information. POI search is primarily based on coordinates, similar to popular location-based systems. The other type is historical publications (as a rule, in HTML—HyperText Markup Language or in PDF—Resource Description Framework) or archival databases records (e.g., in XML—eXtensible Markup Language). XML-files can be mapped into OWL [6]. HTML sources can be processed by means of NLP (Natural Language Processing) tools. In treating data sources of the second type, the main difficulty is that historical objects are usually identified by their names only (e.g., by means of record linking techniques).

As a result, the smart space contains a representation of the semantic network G, integrating the information extracted from multiple data sources. Their parallel activity is coordinated by combining KPs. The common ontology O is used to represent G in the smart space. The ontology provides a system of classes, relations, and restrictions that collected historical information must confirm. As a result, they constitute a historical semantic network for the reference recommendation scenario. A combining KP reasons over the extracted historical information and establishes semantic relations between historical objects. There can be several combining KPs, and the consistency of G is ensured by O. A combining KP may represent interests of a given tourist, act on behalf of a group of tourists, or perform generic context-aware construction.

Based on semantic network G in the smart space, a ranking KP constructs recommendations. Each recommendation is represented as star graph R_p for a given tourist, initial POI p, and context. Visualization on the client KP can utilize additional information such as ranks r_q and annotations t_q for all recommended POIs from R_p. Importantly that there can be many ranking KPs, each employs own computational method of POI selection for the recommendation. For instance, in POI selection method [6], values of r_q reflect the closeness of q's categories to the categories the initial POI p has. Then an

 ISBN: 978-1-61208-418-3

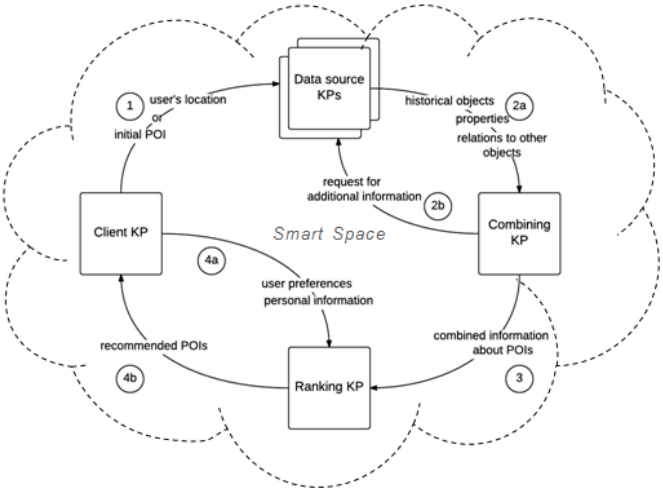


Figure 5. Many KPs interact in a smart space to construct a service.

annotation t_q can describe the common categories of p and q.

The proposed architecture reduces the service construction to interactions of KPs. It follows the principle that a smart space service is knowledge reasoning over the shared content and delivering the result to the users [20]. In our case of historical recommendation, the proposed model of KPs interaction is presented in Figure 5.

Smart space content is shared forming a subject to self-generation. That is, the steps in Figure 5 are performed simultaneously, with event-driven synchronization. An important event to activate data extraction is specifying the initial POI of tourist (step 1). Content self-generation also supports the service personalization. New historical objects can be found and new semantic relations can be associated with this given POI (the iteration in steps 2a and 2b). Request for additional information about historical objects (step 2b) occur until semantic network around tourist is being formed. The POI recommendation and ranking are further personalized (the iteration in steps 4a and 4b) when personal information is directly integrated into the rank computation, e.g., the POIs the tourist has already visited. Updating the user's personal information or preferences (step 4a) requires the rank conversion. This conversion may occur until the user likes the recommended POIs (step 4b).

Table II illustrates this content self-generation model showing the construction of the semantic network from Figure 1 and consequently the star graph from Figure 3. Note that in this paper we do not focus on particular ranking criteria. Intuitively, the closer and richer relations the higher rank. For instance, Apartment house receives the highest rank since the POI is a) located on the same street and b) designed by the same architect.

V. Comparison and Applicability Analysis

The proposed service scenario differs from other proposals. The most close to our work is [14]. The authors explore the use of location aware mobile devices for searching for and browsing a large number of general and cultural heritage information repositories. The application—Mobile Cultural Heritage Guide— searches for POIs in the current tourist's physical location and constructs a "mental map" of nearby POIs within a circular shape. Next semantic crawling is applied

TABLE II. EXAMPLE OF CONTENT GENERATION

Step	Interaction	Generated content
1	Initial POI	Hotel Negresco
2	Data source KPs retrieve facts about the initial POI from different data sources.	Hotel Negresco is located at Promenade des Anglais, 37. The architect of Hotel Negresco is E.-J. Niermans. The first owner of Hotel Negresco is H. Negrescu.
3	Combining KP advance the information description on POIs to create a semantic network.	POIs on the same street: Apartment house, Palais de la Méditerranée. POIs by the same architect: Place Masséna, Villa le Colombier, Apartment house. POIs related with H. Negrescu: Place Masséna.
4	Ranking KPs differentiate the POIs in the semantic network, selecting the most attracting for the user.	Rank-sorted list: Apartment house, Place Masséna, Palais de la Méditerranée, Villa le Colombier.

to resemble the process of a human using the Web to find other information relevant to these POIs. Finally, augmented reality is used in combination with facet selection to present this POI-related information to an active tourist on her/his mobile device. Similarly to our scenario, this application aims at dynamic provision of semantically-enriched information in favor of a classical travel guide. In contrast, our scenario introduces both nearby and faraway POIs, which are semantically related within a variety of historical objects, including common historical persons and events. Our scenario supports automation of semantic crawling with POI ranking; the ranks are then used for visual representation of POI recommendations to the user.

Paper [21] considers a prototype application with POI ranking. It supports content-based recommendations for generating personalized routes along cultural heritage assets in outdoor environments (e.g., city tours). The case of indoor environments, such as museums, is studied in recent work [22]. The mobile application helps the visitors to access information concerning exhibits that are of primary interest to them during pre-visit planning, to provide the visitors with relevant information during the visit, and to follow up with post visit memories and reflections. In contrast, our scenario resembles the process of a historian studying historical facts.

Based on recent studies of the Smart-M3 performance, we expect the proposed system design is applicable in real-life setting and has advantages over the other approaches to recommender system development. The applicability of the smart space based architecture, similar to the one shown in Figure 4, is discussed in [7]. A realistic case is a smart space [16][23] where the number of large data sources, such as web-based databases and repositories, is of the 10-order magnitude and the number of mobile users is of the 10^3-order magnitude.

In contrast to mobile standalone applications, the workload is delegated from a personal mobile device to smart space infrastructure deployed on powerful hosts. Experiments in [24] confirm that this design solution additionally improves reliability and fault tolerance, essentially in wireless network settings. In our scenario, the delegated workload includes the construction of semantic networks for many users and POI ranking over these networks. A personal mobile device visualizes aggregated fragments (e.g., a star graph) of the whole semantic network enriched with derived ranks.

In comparison with web-based recommender system, the proposed smart space based system design provides flexibility in selection of 1) data sources, 2) semantic network construction, 3) POI ranking, and 4) personalization. Although the cost is performance, Semantic Web technologies are now capable

 ISBN: 978-1-61208-418-3

to create and maintain relatively large RDF triple stores [18], where the number of RDF triples is of the 10^5-order magnitude and more. In particular, Smart-M3 SIB employs the Redland library for RDF triple store and SPARQL support [8][15].

VI. CONCLUSION

This paper addressed recommendation services development for historical e-Tourism. We studied the problem of historical POI recommendation, the necessity of using semantic relations between historical objects, and the personalized (subjective, contextual) aspect of services. We proposed the smart space based system design for implementing such a recommender system. The proposal provides a concept definition and design solutions for creating a smart space to accompany a historian tourist. Multiple external sources of historical data can be attached to the smart space and used for provision of information relevant to the situation and user's interests. The information is integrated using Semantic Web technologies and analyzed to produce personalized recommendations. The result is visually presented with quantitative (POI ranks) and qualitative (reason annotations) estimates.

Our study makes a step towards concept development for historical e-Tourism. Feasibility study of the proposed concept, including ontology engineering for integrated representation of historical objects, analysis of POI ranking methods over a semantic network of historical objects, and experimental evaluation, is subject to our further research.

ACKNOWLEDGMENT

This work is financially supported by the Ministry of Education and Science of the Russian Federation within project # 2.2336.2014/K from the project part of state research assignment and 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 presented results are part of the research carried out within project # 14-07-00252 of the Russian Foundation for Basic Research.

REFERENCES

[1] D. Gavalas, C. Konstantopoulos, K. Mastakas, and G. Pantziou, "Mobile recommender systems in tourism," J. Netw. Comput. Appl., vol. 39, Mar. 2014, pp. 319–333.

[2] J. Borrás, A. Moreno, and A. Valls, "Intelligent tourism recommender systems: A survey," Expert Syst. Appl., vol. 41, no. 16, Nov. 2014, pp. 7370–7389.

[3] V. L. Smith, Hosts and Guests: The Anthropology of Tourism. University of Pennsylvania Press, 1989.

[4] P. Nora, "Between memory and history: Les lieux de mémoire," Representations, no. 26, 1989, pp. 7–24, Special Issue: Memory and Counter-Memory.

[5] N. Ide and D. Woolner, "Historical ontologies," in Words and Intelligence II: Essays in Honor of Yorick Wilks, ser. Text, Speech and Language Technology, K. Ahmad, C. Brewster, and M. Stevenson, Eds. Springer, 2007, vol. 36, pp. 137–152.

[6] A. Varfolomeyev, D. Korzun, A. Ivanovs, and O. Petrina, "Smart personal assistant for historical tourism," in Recent Advances in Environmental Sciences and Financial Development. Proc. 2nd Int'l Conf. on Environment, Energy, Ecosystems and Development (EEEAD 2014), C. Arapatsakos, M. Razeghi, and V. Gekas, Eds., Nov. 2014, pp. 9–15.

[7] 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, Aug. 2013, pp. 48–59.

[8] J. Kiljander, A. D'Elia, F. Morandi, P. Hyttinen, J. Takalo-Mattila, A. Ylisaukko-oja, J.-P. Soininen, and T. S. Cinotti, "Semantic interoperability architecture for pervasive computing and Internet of Things," IEEE Access, vol. 2, Aug. 2014, pp. 856–874.

[9] M. Kalus, "Semantic networks and historical knowledge management: Introducing new methods of computer-based research," The Journal of the Association for History and Computing, vol. 10, Dec. 2007, Retrieved: May, 2015. [Online]. Available: http://hdl.handle.net/2027/spo.3310410.0010.301

[10] A. Ivanovs and A. Varfolomeyev, "Computer technologies in local history studies: Towards a new model of region research," Acta Humanitarica Universitatis Saulensis, vol. 19, 2014, pp. 97–107.

[11] A. Meroño-Peñuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. van Harmelen, "Semantic technologies for historical research: A survey," Semantic Web journal, 2015, to appear.

[12] E. Ahonen and E. Hyvonen, "Publishing historical texts on the semantic web - a case study," in Proc. 2009 IEEE Int'l Conf. on Semantic Computing (ICSC '09). IEEE Computer Society, 2009, pp. 167–173.

[13] L. Ardissono, T. Kuflik, and D. Petrelli, "Personalization in cultural heritage: The road travelled and the one ahead," User Modeling and User-Adapted Interaction, vol. 22, no. 1-2, Apr. 2012, pp. 73–99.

[14] C. van Aart, B. Wielinga, and W. R. van Hage, "Mobile cultural heritage guide: Location-aware semantic search," in Proc. 17th Int'l Conf. on Knowledge Engineering and Management by the Masses (EKAW'10), LNCS 6317, P. Cimiano and H. S. Pinto, Eds. Berlin, Heidelberg: Springer, 2010, pp. 257–271.

[15] I. Galov and D. Korzun, "Design of semantic information broker for localized computing environments in the Internet of Things," in Proc. 17th Conf. of Open Innovations Association FRUCT. ITMO Univeristy, Apr. 2015, pp. 36–43.

[16] A. Smirnov, A. Kashevnik, A. Ponomarev, N. Teslya, M. Shchekotov, and S. Balandin, "Smart space-based tourist recommendation system," in Proc. 14th Int'l Conf. Next Generation Wired/Wireless Networking and 7th Conf. on Internet of Things and Smart Spaces (NEW2AN/ruSMART 2014), LNCS 8638, S. Balandin, S. Andreev, and Y. Koucheryavy, Eds. Springer, Aug. 2014, pp. 40–51.

[17] K. Kulakov and A. Shabaev, "An approach to creation of smart space-based trip planning service," in Proc. 16th Conf. of Open Innovations Association FRUCT. ITMO Univeristy, Oct. 2014, pp. 38–44.

[18] C. Gutierrez, C. A. Hurtado, A. O. Mendelzon, and J. Pérez, "Foundations of semantic web databases," J. Comput. Syst. Sci., vol. 77, no. 3, May 2011, pp. 520–541.

[19] D. G. Korzun, A. A. Lomov, P. I. Vanag, J. Honkola, and S. I. Balandin, "Multilingual ontology library generator for Smart-M3 information sharing platform," International Journal on Advances in Intelligent Systems, vol. 4, no. 3&4, 2011, pp. 68–81.

[20] D. Korzun and S. Balandin, "A peer-to-peer model for virtualization and knowledge sharing in smart spaces," in Proc. 8th Int'l Conf. on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2014). IARIA XPS Press, Aug. 2014, pp. 87–92.

[21] N. Stash, L. Veldpaus, P. D. Bra, and A. P. Rodeirs, "Creating personalized city tours using the CHIP prototype," in Late-Breaking Results, Project Papers and Workshop Proceedings of the 21st Conf. on User Modeling, Adaptation, and Personalization (UMAP 2013): Proc. Workshop on Personal Access to Cultural Heritage (PATCH 2013), ser. CEUR Workshop Proceedings, S. Berkovsky, E. Herder, P. Lops, and O. C. Santos, Eds., vol. 997, Jun. 2013, pp. 68–79.

[22] T. Kuflik, A. Wecker, J. Lanir, and O. Stock, "An integrative framework for extending the boundaries of the museum visit experience: linking the pre, during and post visit phases," Information Technology & Tourism, vol. 15, no. 1, 2015, pp. 17–47.

[23] P. Vanag and D. Korzun, "SmartSlog knowledge patterns: Initial experimental performance evaluation," in Proc. 11th Conf. of Open Innovations Association FRUCT and Seminar on e-Tourism, S. Balandin and A. Ovchinnikov, Eds. SUAI, Apr. 2012, pp. 176–180.

[24] I. Galov and D. Korzun, "Fault tolerance support of Smart-M3 application on the software infrastructure level," in Proc. 16th Conf. of Open Innovations Association FRUCT. ITMO Univeristy, Oct. 2014, pp. 16–23.

 ISBN: 978-1-61208-418-3

Integrating Application-Oriented Middleware into the Android Operating System

Julian Kalinowski and Lars Braubach

Computer Science Department, University of Hamburg
Distributed Systems and Information Systems
Hamburg, Germany
Email: {kalinowski|braubach}@informatik.uni-hamburg.de

***Abstract*—As mobile devices are becoming more advanced in technology, the type of software they are able to process develops from simple apps to complex applications. Fortunately, a main area in software engineering research is dedicated to examining the handling of complexity. The common approach of adding abstraction layers is embodied in various middleware solutions, including application-oriented middleware that feature generic abstractions for decomposition and distribution as well as support for non-functional criteria and higher-level concepts in programming. However, embedding middleware into a mobile operating system environment bears many challenges. The several attempts of porting a middleware to Android have only been partially successful, as they either require developers to use an uncommon programming language or abandon the well-proven Android design principles. We propose a universal architecture for integrating middleware into the Android operating system while maintaining the core features of the Android application framework. The presented architecture provides the shared use of middleware libraries during runtime as well as a middleware execution platform for shared use of different apps and an event-based mechanism for middleware/android component coupling.**

Keywords–Android; Middleware; Mobile Applications; Software Agents.

I. INTRODUCTION

The possibilities of mobile applications increase with the advent of faster hardware, bigger screens and more stable broadband connections. This ongoing evolution of mobile computing leads to larger applications and increases the need for methods reducing software complexity [1]. While reducing complexity has played a central role in software engineering right from the beginning, there is still no silver bullet; complexity can only be coped with abstraction [2]. Several types of middleware can aid software developers by providing some of the abstractions that are needed to create nowadays programs.

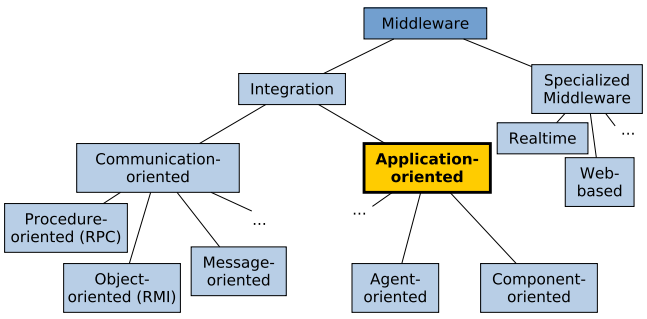


Figure 1. Middleware taxonomy, based on [3].

A helpful categorization of middleware is shown in Figure 1. This paper focuses on a subset which we call *application-oriented middleware*. This subset includes component and agent-oriented middleware. In addition to communication, they provide support for decomposition or other generic programming abstractions; thus supporting application development in multiple aspects.

Android, as well as other mobile operating systems, has been developed with limited resources in mind. Processing performance, available memory, and battery capacity have been considered at system level. In consequence, the system was designed to run rather small, self-contained applications or *apps*, which are executed in separate virtual machine (VM) environments for security and safety reasons [4]. App developers are restricted to fixed design principles to make sure their application integrates properly with the operating system. For example, apps have to be split up in activities and services, depending on whether the given part represents user interface (UI) or executes background tasks. Activities and services are subject to a specific life cycle, which is executed by Android. Acting as a framework, Android is allowed to start, stop, pause and resume apps if it needs to for various reasons such as limited resources, user interaction or even an incoming phone call.

The strict requirements for application developers lead to limitations in software architecture design and the integration of middleware in particular. As applications are executed in different VM processes and the developer cannot influence application loading, it is not possible to share libraries in a convenient and secure way [4]. Nevertheless, as apps get bigger and use more libraries, the possibility of two apps sharing some code increases. Middleware, in contrast, is built to handle more than one running application and thus supports access to common functionality by design.

Furthermore, since the Android system determines the way applications are loaded, instantiated and started, there is little chance for the application to influence mechanisms like class and resource loading. Application-oriented middleware, however, form an abstraction layer between operating system and applications as shown in Figure 2. This usually requires the use of specific classloading or startup mechanisms, as middleware provides a runtime environment called *platform* to control the life-cycle execution of runtime application components [5][6].

This paper presents an architecture that deals with these challenges and integrates a middleware platform within an Android application; to be used and accessed by *client applications* and achieving a higher level of abstraction during application development. As the presented architecture is independent

 ISBN: 978-1-61208-418-3

of a specific middleware implementation, it is conceptually applicable to various middleware.

The article is structured as follows: Section II introduces the requirements we would like to fulfill. Section III provides an overview on related work. Section IV explains the challenges of the Android operating system regarding middleware integration. Section V describes the architecture we propose to cope with these challenges. Section VI presents a prototypical implementation of the architecture, which is then evaluated in Section VII. Finally, Section VIII concludes the paper and gives an outlook on future work.

II. Requirements

In contrast to embedding a middleware into an Android *application* [7][8], the goal of this paper is the integration of middleware into the Android operating system as an abstraction layer to be used by *other applications* as shown in Figure 2.

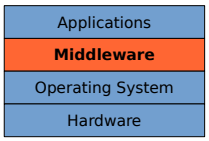


Figure 2. Middleware layer between applications and operating system, based on [9, p. 23].

This goal leads to the following functional requirements:

1) *Independent deployment:* The middleware is deployed independently of other applications and can be used by *client applications* (which in turn do not need to include middleware libraries).
2) *Multi-client capable:* Multiple applications can run on a single instance of the middleware.

As outlined in the introduction, the Android application framework introduces several design principles to simplify application development in the context of mobile devices. These principles should obviously still be applicable when developing applications using an integrated middleware. Application-oriented middleware can provide additional programming approaches, such as agent-orientation, which should be available, too. Furthermore, components developed with middleware concepts should be able to interoperate with Android application components. This leads to the following criteria:

3) *Concept integration:* Concepts of both the integrated middleware and the Android operating system should be available to the developer.
4) *Component coupling:* Components running on the middleware and components based on Android paradigms must be able to communicate easily.

As non-functional requirements cannot generally be fulfilled by an integration architecture, but are determined by the concrete middleware and application, they will not be considered here.

III. Related Work

In this section, other work regarding the use of middleware on mobile operating systems is reviewed with respect to the requirements given in the previous section.

A. Component-Oriented middleware

Component-oriented middleware realizes the idea of interchangeable and reusable software components. They implement a component model, which defines syntax and semantics of component definitions and their relations [10]. Several approaches, all based on OSGi [11], have been proposed for the use on Android devices.

Equinox was originally developed to provide a plugin-based architecture for the Eclipse IDE. In the progress of evaluating the application of Equinox on Android devices, necessary changes were added by Hargrave and Bartlett in 2008 [12]. For the time being, Equinox does not provide a concept for integrating UI. In consequence, it is uninteresting for many real-world scenarios.

The Apache OSGi implementation *Felix* supports execution on Android since version 1.0.3. It is possible to use the Felix command line shell to add bundles and run console applications, just as with Equinox. Furthermore, Felix can be embedded in Android Apps and executed during the initialization of an app [7]. Based on this approach, Escoffier showed how to create Android apps that dynamically load *.jar* bundles. Felix uses a special *ViewFactory* Interface for creating application UI from within any bundle [13].

The commercial OSGi implementation *ProSyst mBS* was designed for embedded hardware, and features explicit support for Android devices. As in Felix, application components are deployed as *.jar* bundles and can contain UI, which has to be implemented using the interface *ApplicationFactory* instead of activities. As opposed to the previously described OSGi implementations, the ProSyst platform is deployed inside a standalone Android application. To launch an individual application, a dummy app is installed on the device, instructing the platform application to load a specific application bundle and display its UI [14]. This execution model enables sharing of the middleware platform between applications, while keeping the original user experience.

B. Agent-Oriented middleware

Software agents provide a high-level approach to implement complex and concurrent software systems. In order to use such an abstraction, a runtime environment (*platform*) is required to provide services, e.g., for executing or discovering agents.

JaCa-Android [8] was specifically developed for Android and combines the *Agents & Artifacts* paradigm with an agent runtime called *CArtAgO* [15]. Agents are implemented using *Jason*, an AgentSpeak implementation [16]. The runtime model is based on embedding the runtime platform into applications, including a central *JaCa-Middleware* application, which provides several artifacts to enable using services like contact management, localization or SMS from within agents. User interfaces are developed using default Android activities and are represented to the agents as *artifacts* to enable communication between them. The Agents & Artifacts approach allows for an elaborated integration of agent and Android design principles, but introduces an implementation language that is very different from traditional languages.

Another agent-oriented middleware is *JADE*, which also features an Android version. *Jade-Android* can either integrate with a back-end or be executed as standalone platform. In any case, the runtime platform is included in applications;

 ISBN: 978-1-61208-418-3

increasing the application sizes and loading times. Agents can communicate with Android activities using the *Object-to-Agent Interface (O2A)*. O2A utilizes Android intents sent by agents and received by activities as well as Java interfaces, which are used by activities to call agent methods [17].

TABLE I. FEATURE OVERVIEW.

	Independent deployment	Single instance	Concept integration	Component coupling
Equinox	+	+	-	-
Felix	-	-	o	-
ProSyst	+	+	o	o
JaCa	-	-	+	+
JADE	-	-	+	+

Legend: +: supported, o: partly supported, -: not supported.

In Table I, all previously described approaches are compared in respect to the requirements stated in Section II. It can be seen that no approach is able to fulfill all requirements to a satisfactory degree.

IV. CHALLENGES OF THE ANDROID OPERATING SYSTEM

Several properties of Android prevent middleware developers from simply porting and using a Java SE middleware. In order to fulfill all the requirements mentioned in Section II, integrating the middleware into the Android operating system has to be done with great care. We will focus on three of the most critical characteristics of Android that have to be considered.

First, Android implements an effective way to run every application on its own virtual machine. To avoid loading core libraries twice, a central VM process called *Zygote* is used to load them into memory. This process is forked each time a new VM is needed, providing every running VM access to the previously loaded core libraries [18]. This works fine for sharing Android core libraries, but the process separation prevents applications from sharing common libraries at runtime. For the integration of middleware, this is turning into a problem: As per requirement #1, we want middleware and client applications to be deployed independently; but at the same time run multiple applications on one middleware instance (#2). To achieve this, every running client application must have access to classes which are included in the middleware application.

Second, Android applications, specifically their activities and services, utilize a preset life cycle. While dynamic library sharing is impossible due to the above-mentioned process separation, this also complicates static sharing, i.e., using the same filesystem copy of a library. An Android application's entry point is an activity or application object [19, p. 75], which is instantiated before developers gain control of execution. In particular, replacing or modifying the class loader that is used to load the application's activities and services is not possible, which renders statical sharing of libraries a hard problem.

In Figure 3, the control flow of an application with two activities is shown. The first activity requests startup of the second activity in its `onCreate()` method; passing control to the system. This is why some implementations from Section III don't allow the use of activities to implement UI, but rather provide special interfaces to be implemented by the application

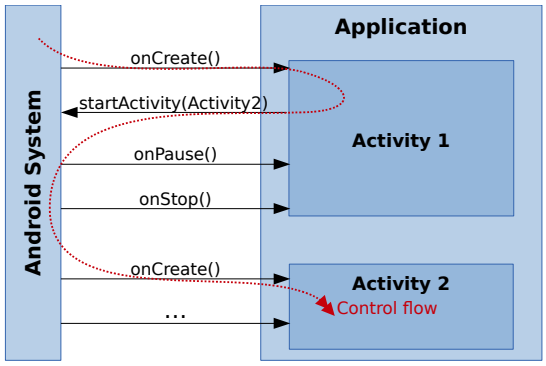


Figure 3. Control flow during application startup and activity change.

developer. This way, they can ensure the right class loader is used to call the interface methods.

Third, Android introduces a simple way to access *Resources* from code. Resources can be images, UI layouts, string values and arbitrary binary files such as sounds. At compile time, every resource is assigned an id, which is stored in the generated class `R`, pointing to the correspondent file. For resolving resources at runtime, Android automatically uses the `R` class belonging to the current application context. In consequence, loading an application-external class that contains resource ids (e.g., UI), which is what middleware has to do to show application-specific UI, will usually fail. This is mirrored by the fact that in current implementations which allow sharing of middleware libraries, it is not possible to use resources such as XML-UI layouts (see Section III).

V. ARCHITECTURE

Our middleware-embedding architecture is based on separate Android applications with no additional .jar deployment, as this would differ from default Android concepts (requirement #3). One application provides the middleware runtime and contains all middleware-specific classes and libraries. This *middleware app* can then be called from *client apps*, providing access to middleware libraries and functionality. While keeping both application types separated at deployment time, our architecture executes client apps in the middleware process to allow for sharing of middleware libraries at runtime. To preserve Android and middleware concepts, the proposed architecture does not allow to create or modify UI inside of middleware components. Instead, middleware and Android components are created separately, while a convenient way to communicate between them is part of the architecture.

The following subsections will describe the startup procedure as well as three important architectural details: *UI instantiation*, *service binding* and *communication* between Android and middleware components.

A. Startup phase

The interaction on startup of a client app is illustrated in Figure 4. First, the client app has to send a special startup *intent* adressed to the middleware app, which is started on demand. This intent must contain information about the client app: the full path to the installed android application package (APK) file and the name of the main activity to launch. This behavior can be extracted to a base activity class that can be extended by the application developer.

 ISBN: 978-1-61208-418-3

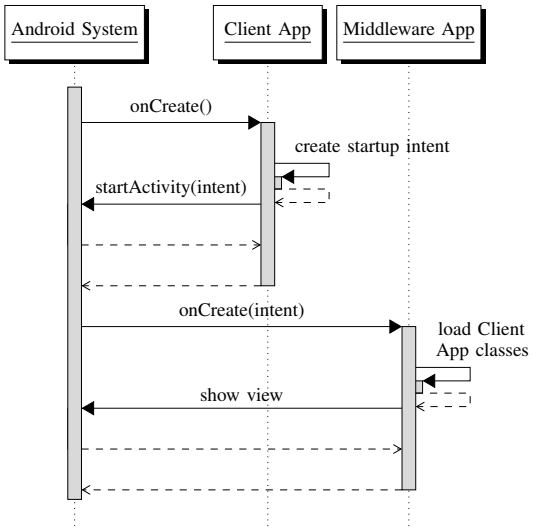


Figure 4. Interaction on application startup.

Upon reception of the intent, the middleware app will access the client app's classes by creating a class loader that points to the correspondent APK. Interestingly, it is possible for applications to read other applications' compiled code, if it is not explicitly forward-locked [20, p. 79]. After loading the classes, the middleware app will instantiate the main activity, which has been specified in the startup intent. The mechanism that is used to display the client app's UI will be introduced in the next section.

B. UI instantiation and management

As discussed earlier, app developers are not able to influence activity instantiation and presentation themselves. Fortunately, since the release of Android 3.0, there is a way to manage sub-views, which are called *Fragments*, from inside an application. Fragments implement their own life cycle, which is derived from the system-executed activity life cycle, but executed by a *Fragment Manager* instead [21]. Contrary to activities, this enables the use of fragments that are instantiated by application-own code, which in turn is the requirement for using a custom class loader.

In consequence, we are not allowing activities inside a client app, but instead use fragments as top-level replacement. As fragments were introduced to allow the partition of user interfaces, they also provide all capabilities to implement Android user interfaces (including fragments in fragments) and thus provide a viable replacement for activities. Three essential elements are needed in order to make the replacement work:

- A basic `FragmentActivity` inside the middleware app which will contain the client application's layout.
- A mechanism that implements switching between shown fragments and used resource contexts, depending on active client application.
- A base class extending `Fragment` that will (possibly transparently) replace activities in the client app. This class will be called `ClientAppFragment` and its instances are referred to as *client fragments*.

To focus on architectural design, we omit discussion of these elements here. They are discussed in detail in [22].

C. Service binding

When executing a client app inside the middleware process, the Android service binding mechanism needs special attention. Generally, binding an Android service returns the control flow to the Android system, which then determines which service to call and whether it has to be started or is already running. As this procedure is similar to starting an activity, it does not allow the developer to influence class loaders beforehand. This makes it impossible to build services that in turn use middleware libraries or communicate with the middleware. Since this is obviously undesirable, we propose an alternative way services are bound by client apps.

The used approach was inspired by the way Android handles fragments. Instead of letting the system create, bind and destroy services, a single service is always bound to the middleware app's main activity. All service calls originating from client fragments are handled by this *universal service*, which manages the (quite simple) service life cycle and maintains all existing service instances and connections.

The binding process is shown in Figure 5. In step (1), a binding is requested by a client fragment. The request, containing the classname of the targeted service, is handled by the universal service. In step (2), the service is instantiated and its life cycle is executed until `onBind()` is called, returning a binder object in step (3). The binder object is passed to the corresponding client fragment by calling `onServiceConnected()` in step (4), finally establishing the connection between client fragment and client service.

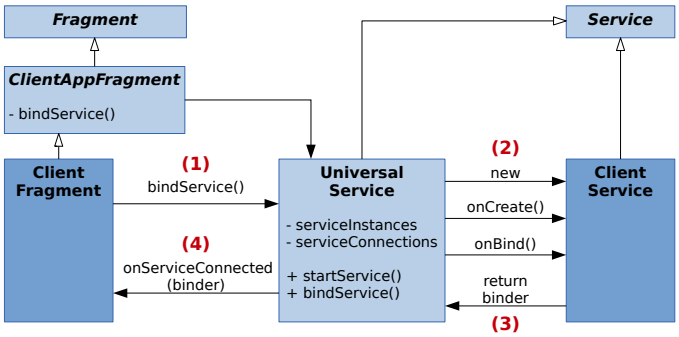


Figure 5. Universal service handles binding of client services.

Calls to external services, e.g., from other third-party apps, are still handled by the original service binding mechanism. Its possible to differentiate between internal and external services by the type of intent used: *Explicit* intents always refer to application-internal classes, while *implicit* intents can address external services [21].

D. Communication/Coupling

Besides displaying UI and binding services, at some point, the client app has to interact with the middleware in terms of starting the platform or using an already running instance, starting or stopping a middleware-supported runtime element, or looking up middleware services and components. Since interactions of this kind generally take place during the whole application lifetime, they should not be executed inside short-living activities/fragments, but rather in long-living Android services. For the sake of convenience, we only allow one service inside an application to communicate with the middleware directly; we will call this service *platform service*.

 ISBN: 978-1-61208-418-3

In the requirements section, we further demanded easy communication between relevant Android components, which we cut down to services and middleware components. The latter can be software components, agents, or any other runtime elements that are supported by the used middleware. For coupling between the platform service and middleware components, we use the observer pattern to allow for loose coupling of the components: a middleware component can be called from the platform service through interface methods or using the middleware's service model if available. The platform service can in turn register for typed events that are thrown inside middleware components; either by calling a static method or by integrating an appropriate event service into the middleware itself, which can be used by runtime components.

E. Overview

Figure 6 shows an overview of the architecture containing the elements described above. Green dots represent connection endpoints, embodied on Android by implementations of the classes `ServiceConnection` and `Binder`.

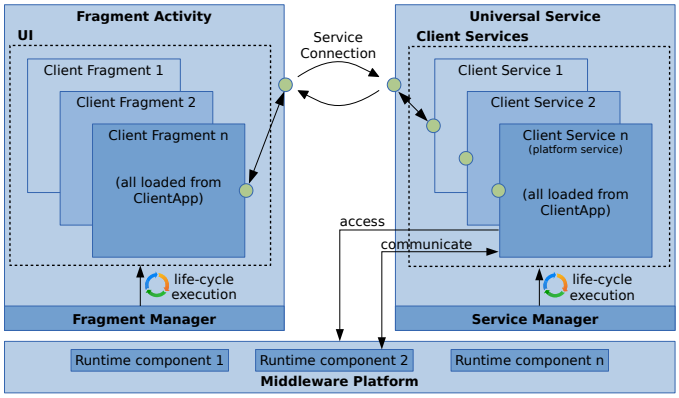


Figure 6. Overview of the middleware integration architecture.

On the left side, the *fragment activity* provided by the middleware app is shown. It loads all needed client fragments and displays them according to their life cycle, which is executed by the fragment manager. Client fragments communicate with client services on the right side, which are contained in the *universal service*. Their life cycles are executed by a service manager and the *platform service* is able to access the underlying *middleware platform* for managing and communicating with runtime components. Everything that is shown runs inside the same process and application context.

VI. Implementation

While the Android-specific implementation is rather universal and could be generalized, the used middleware might need some extensions. For example, it must provide a class loading mechanism that supports dynamic adding of class loaders. This is required for running a shared middleware platform, as it is unknown in advance which client application classes will be needed in the future. Upon requests by applications, the middleware must to be able to distinguish between individual applications and their classes to make sure classes can be loaded and the correct runtime components are managed. Since middleware may already handle dynamic class loading, this is a modest requirement. In addition, the aforementioned class loading mechanism has to support the Android class loaders; supporting the android-specific class formats. The presented integration architecture was implemented using the active component middleware *Jadex* [23].

To use the middleware platform, developers can extend the a special Android service class providing methods for configuring platform options, such as platform name, sharing of the platform, and other Jadex-specific options. After the platform was started in shared mode, other client apps can access it using their own implementations of the platform service. Each app-specific platform service can now start Jadex components on the shared platform and register for their events. The platform takes care of correlating apps with their Jadex components by using the corresponding class loader for each client app. Also, loading resources, layouts and assets from the right client app package is done transparently.

On the UI side, an app has to define a class inheriting from `JadexClientLauncherActivity` and implement a method returning the class name of the app's default fragment. After initiatiing the startup procedure described in Section V, this activity will show the default fragment. When the developer would like to implement further activities, he has to write a new Fragment instead. To launch these, `startActivity()` is enhanced to support Fragments and display them as if they are contained in a new Activity. Similarly, `startService()` is modified to use the universal service from Section V. Enhancing the methods specified by the Android framework keeps differences in the programming model as small as possible. In the case where fragments are required to split application layouts, implementing `ClientAppFragment` enables them to gain access to the top-level fragment for layouting purposes, such as adding/removing fragments.

VII. Evaluation

For evaluation, two versions of a demo application were compared against each other regarding startup time and application size. The first app is a client app and thus capable of using a shared platform, while the second embeds the Jadex platform, as it would have to without the presented architecture. For this evaluation, the applications just contain one fragment and bind itself to a platform service which starts up a platform component. In consequence, they model the smallest possible Jadex applications in each setup and the total startup time heavily depends on the platform startup time.

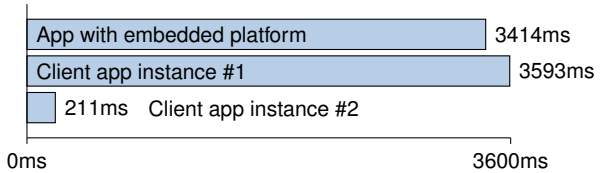


Figure 7. Application startup time comparison.

Figure 7 compares application startup times, showing the average of ten measurements on a Samsung Galaxy Nexus i9250. We differentiated between the first client app starting a shared platform and the case of a second app instance using an already running platform. As the chart indicates, the second client app starts up much faster, as the platform is already running. In comparison to embedding the platform, starting the first app using the integration architecture does

 ISBN: 978-1-61208-418-3

not significally influence startup time. Additionally, Table II shows corresponding application sizes. As splitting application and middleware results in only 132 KB overhead compared to embedding the platform, the benefit of using the integration architecture is confirmed again.

TABLE II. APPLICATION SIZE COMPARISON.

	Own classes	*DEX size*	*APK size*
Embedded	64 KB	3.516 KB	1.548 KB
Clientapp	58 KB	274 KB	136 KB
Middleware app	49 KB	3.532 KB	1.544 KB

Nevertheless, the proposed architecture comes with some limitations, which are mostly induced by circumventing the process separation of Android and running multiple apps in one process. First, client-side manifest declarations are ineffective, as the client app is only started by itself to initiate the startup phase. Afterwards, only the manifest of the middleware app is respected. This is especially critical for permissions, while intent receivers, or styles, could be handled by passing them through for the middleware app to handle. For evaluation purposes, our middleware app was given all necessary permissions. While app internal services are already handled by the architecture, externally visible services cannot be declared.

Second, access to the Android storage options, such as databases, preferences and private files is shared between all client apps running on the middleware. Since they use the same application context, applications also have access to the same storage resources. This could be prevented to a certain extent by enhancing the middleware with access control features.

Last, running multiple apps in one process affects runtime security and safety. Running in the same process means sharing memory with each other, possibly causing sensible data to be exposed. With big apps, one might also reach resource limits such as heap space sooner, which in the worst case might lead to termination of the middleware including all client apps.

VIII. CONCLUSION AND FUTURE WORK

This paper presented an approach that enables the integration of middleware into the Android system. In particular, the architecture allows different apps to use one middleware platform jointly at runtime. It further provides full access to the sophisticated Android design principles, which are unavailable on most current middleware implementations for Android. An event-based mechanism ensures smooth interaction between application components running on the middleware platform and Android application components. Most of the problems handled in the presented architecture arise from the fact that Android itself provides an extensive framework and runtime environment, complicating the integration of middleware.

With the presented architecture, other middleware can be integrated into Android, allowing developers to program using alternative programming principles, decomposition features and non-functional criteria of modern application-oriented middleware. Future work may include removing limitations where possible, evaluating the architecture using other middleware, as well as implementing a generic integration solution that abstracts from the concrete middleware platform.

REFERENCES

[1] J. Dehlinger and J. Dixon, "Mobile application software engineering: Challenges and research directions," in Workshop on Mobile Software Engineering, 2011, pp. 27–30.

[2] F. P. Brooks, Jr., "No silver bullet essence and accidents of software engineering," Computer, vol. 20, no. 4, Apr. 1987, pp. 10–19.

[3] T. A. Bishop and R. K. Karne, "A survey of middleware," in Procs. 18th Int. Conf. Computers and Their Applications, 2003, pp. 254–258.

[4] D. Ehringer, "The dalvik virtual machine architecture," Tech. Rep., 03 2010.

[5] R. H. Bordini et al., "A survey of programming languages and platforms for multi-agent systems," Informatica (Slovenia), vol. 30, no. 1, 2006, pp. 33–44.

[6] W. Emmerich, "Software engineering and middleware: a roadmap," in Proceedings of the Conference on The future of Software engineering. ACM, 2000, pp. 117–129.

[7] The Apache Software Foundation, "Apache felix framework and google android," 05 2009, [retrieved: 2015-06-10]. [Online]. Available: http://felix.apache.org/documentation/subprojects/apache-felix-framework/apache-felix-framework-and-google-android.html

[8] A. Santi, M. Guidi, and A. Ricci, "Jaca-android: an agent-based platform for building smart mobile applications," in Languages, Methodologies, and Development Tools for Multi-Agent Systems. Springer Berlin Heidelberg, 2011, pp. 95–114.

[9] P. Naur and B. Randell, Eds., Software Engineering: Report of a Conference Sponsored by the NATO Science Committee, Garmisch, Germany, 7-11 Oct. 1968, 1969.

[10] G. T. Heineman and W. T. Councill, Component-based software engineering: putting the pieces together. Addison-Wesley Longman Publishing Co., Inc., 2001.

[11] The OSGi Alliance, OSGi Service Platform Core Specification, Release 5, Jun. 2012.

[12] B. Hargrave and N. Bartlett, "Android and osgi: Can they work together?" [retrieved: 2015-06-10]. [Online]. Available: http://www.eclipsecon.org/2008/sub/attachments/Android_and_OSGi_Can_they_work_together.pdf

[13] C. Escoffier, "ipojo on android," 10 2008, [retrieved: 2015-06-10]. [Online]. Available: http://ipojo-dark-side.blogspot.de/2008/10/ipojo-on-android.html

[14] ProSyst Software GmbH, "Osgi runtime on android," [retrieved: 2015-06-10]. [Online]. Available: http://dz.prosyst.com/pdoc/mBS_SDK_7.3.0/modules/framework/common/android/introduction.html

[15] A. Ricci, M. Viroli, and A. Omicini, "Carta go: A framework for prototyping artifact-based environments in mas," in Environments for Multi-Agent Systems III, ser. Lecture Notes in Computer Science, D. Weyns, H. Parunak, and F. Michel, Eds. Springer Berlin Heidelberg, 2007, vol. 4389, pp. 67–86.

[16] R. H. Bordini, J. F. Hübner, and M. Wooldridge, Programming multi-agent systems in AgentSpeak using Jason. John Wiley & Sons, 2007, vol. 8.

[17] F. Bergenti, G. Caire, and D. Gotta, "Agents on the move: Jade for android devices," in Procs. Workshop From Objects to Agents, Sep. 2014.

[18] D. Bornstein, "Dalvik vm internals," Google I/O conference presentation video and slides, 2008, [retrieved: 2015-06-10]. [Online]. Available: http://sites.google.com/site/io/dalvik-vm-internals

[19] C. Collins, M. Galpin, and M. Käppler, Android in Practice. Manning Publications Company, 2011.

[20] N. Elenkov, Android Security Internals: An In-depth Guide to Android's Security Architecture. No Starch Press, 2014.

[21] Open Handset Alliance, "Android Developer Docs," [retrieved: 2015-06-10]. [Online]. Available: http://developer.android.com/

[22] J. Kalinowski, "Analysis and integration of application-oriented middleware into mobile devices in context of the android operating system," Master's thesis, Universität Hamburg, Fachbereich Informatik, Vogt-Kölln-Str. 30, 22527 Hamburg, Germany, Apr. 2014, in German.

[23] A. Pokahr and L. Braubach, "The active components approach for distributed systems development," International Journal of Parallel, Emergent and Distributed Systems, vol. 28, no. 4, 2013, pp. 321–369.

 ISBN: 978-1-61208-418-3

Fortifying Android Patterns using Persuasive Security Framework

Hossein Siadati
New York University
New York, Brooklyn
email: hossein@nyu.edu

Payas Gupta
New York University Abu Dhabi
Abu Dhabi
email: payasgupta@nyu.edu

Sarah Smith, Nasir Memon
New York University
New York, Brooklyn
email: sesmith325@gmail.com
nasir@nyu.edu

Mustaque Ahamad
Georgia Institute of Technology
New York University Abu Dhabi
Atlanta, Georgia
email: mustaq@cc.gatech.edu

***Abstract*—Android Pattern, form of graphical passwords used on Android smartphones, is widely adopted by users. In theory, Android Pattern is more secure than a 5-digit PIN scheme. Users' graphical passwords, however, are known to be very skewed. They often include predictable shapes (e.g., Z and N), biases in selection of starting point, and predictable sequences of the points that make them easy to guess. In practice, this *decreases* the security of Android Pattern to that of a 3-digit PIN scheme for at least *half* of the users. In this paper, we effectively *increase* the strength of Android Patterns by using a persuasive security framework, a set of principles to get users to behave more securely. Using these principles, we have designed two user interfaces that *persuade* users to choose *stronger* patterns. One of the user interfaces is called BLINK, where the starting point of the pattern is suggested to user, effectively *nudging* her to create a pattern with a significantly *less predictable* starting point. The other user interface is called EPSM, where the system gives *continuous* feedback to user while she is creating a new pattern, effectively persuading her to create a complex pattern. Security and usability of our proposed designs evaluated by conducting a user study on 270 participants recruited from Amazon MTurk demonstrated that while only 49% of subjects choose *strong* patterns in Android Pattern user interface, our suggested designs increase it to 60% in BLINK and 77% in EPSM version.**

Keywords–Android; nudging; persuasive security; blinking.

I

I. INTRODUCTION

The rising trend of smartphones in our daily lives and the amount of personal information being carried on these devices call for stronger authentication measures than ever. Smartphones are used to perform sensitive personal and financial tasks including online banking, messaging, and used as a two-factor authentication. PINs have been the traditional way of locking a phone and securing critical data on it [1]. However, *Android Pattern*, a graphical password scheme, has seen a tremendous increase in adoption due to its perceived user-friendliness [2]. According to a recent study, 40% of Android users are using Android Pattern to unlock their devices instead of a PIN [3].

A pattern can be denoted by a sequence of numbers indicating the position of points on the screen (see Figure 1). In the Android Pattern scheme, enrollment and verification works as in a typical password-based user authentication, where a user chooses a secret (i.e., a pattern) in an enrollment phase and recalls it at the time of verification. Whereas the *theoretical password-space* of Android Pattern is larger than that of a 5-digit PIN scheme, Uellenbeck et al. [4] demonstrated biases in starting points (i.e., some points are more frequently chosen than others) and n-grams (i.e., frequent subsequences of patterns) that make user patterns guessable. Based on these findings, they were able to guess about 50% of the patterns with only 1000 guesses. In other words, the *effective password-space* of Android Pattern is equivalent to that of just a 3-digit PIN scheme for 50% of Android users! The problem is that users do not appear to effectively use the large password-space of Android Pattern.

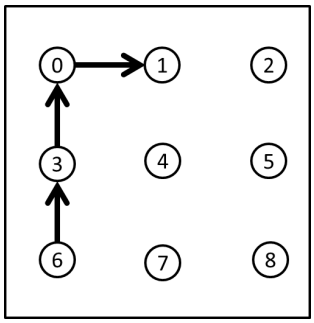


Figure 1. Path from point o_6 to point o_1. $o_6 - o_3 - o_0 - o_1$ indicates that the pattern is started from o_6, then moved to o_3, then to o_0 and then finally ended at o_1.

Several intuitive solutions appear promising but unfortunately fail to address the problem. For example, with *blacklisting*, the authentication system forbids frequently-chosen patterns, but this only shifts the distribution to a new set of frequently-chosen patterns, and does not hinder a resourceful attacker. With *random assignment*, the authentication system chooses a random pattern for the user, but this comes with a significant cost on usability and memorability. With *rearrangement*, Uellenbeck et al. [4] removed frequently-chosen starting points and rearranged all points, but found that this approach by itself does not expand the *effective* password-space. With *user education*, users are taught the differences between weak and strong passwords so that they may prefer the latter over the former, but we found in a survey (described later in this paper) that users are already aware of these differences, and yet this awareness does not translate into choosing stronger passwords.

In this paper, rather than overwhelming users with instructions and cumbersome security measures, or forcing them to choose certain patterns, we use a *persuasive security authentication framework* to *nudge* or *persuade* users to behave more securely [5].

We present two persuasive mechanisms that nudge users to choose *strong* patterns, thereby expanding the *effective password-space*, and reducing the advantage the adversary may have from a priori knowledge of pattern distribution.

 ISBN: 978-1-61208-418-3

Specifically, these user interfaces are:

1) Embedded Pattern Strength Meter (EPSM): A mechanism that provides realtime visual feedback based on the pattern's strength while it is being drawn.
2) BLINK: A mechanism that provides recommendations and nudges users at appropriate points — by blinking — to eliminate the problem of starting point bias, and to persuade users to create stronger patterns.

In summary, this paper makes the following research contributions:

- We show that EPSM helps users to create stronger, more complex patterns compared to Android Pattern. EPSM dramatically reduces the success of a guessing attack: a hypothetical attacker is able to guess 50% of Android patterns by only 1000 guesses, whereas the same attacker is able to guess only 22.6% of the patterns in EPSM.
- With BLINK, we show that we can eliminate the starting point bias. Consequently, the probability that a point is chosen as a starting point is diffused across all points in the unlock pattern grid. This makes patterns stronger against guessing attacks (40% of the patterns can be guess by 1000 guesses, 10% less than the Android patterns).
- To derive these results, we tested BLINK and EPSM with 270 participants recruited from the Amazon Mechanical Turk. We show that BLINK and EPSM improve the security of Android Pattern with only a negligible usability and memorability impact.

The rest of the paper is organized as follows. In Section II, we discuss background details on state-of-the-art approaches for graphical passwords and persuasive authentication. We also provide details from the past work which may seem plausible options at first, however, may not help to improve the security of Android Pattern. In Section III, we provide our persuasive security mechanism design choices. In Section IV, we present experimental details of our user study followed by results in Section V. We conclude in Section VI.

II. BACKGROUND

The first type of grid-based graphical passwords called "Draw a Secret" (DAS) was proposed by Jermyn et al. [6]. In DAS, user creates a password by drawing a pattern that connects cells of a grid on a screen. Followup works have proposed variations of DAS to improve on its security and usability. Most notably, Tao et al. introduced Pass-Go [7] that uses intersections of the cell in a grid (instead of the cells) and improved its usability. Android Pattern is a type of Pass-Go system and is widely adopted by Android users [3].

In this section, we briefly describe the problem of *bias* in users-choices of patterns, and its effect on the security of Android users (see Section II-A). We also list the previous efforts taken to fortify the security of Android Pattern and describe why they have insignificant effect (see Section II-B). Thereafter, we describe how the *persuasive security framework* can help to address the problem (see Section II-C). Finally, we show how to calculate the strength of patterns using Markov model (see Section II-D).

A. Biases of user-chosen patterns

Thorpe et al. [8][9] demonstrated the limitations of user selected patterns and effective password space of DAS by analyzing the memorable space of graphical passwords where patterns are partially or completely symmetric. They conclude that the *effective password-space* of DAS is much smaller than *theoretical password-space*. Andriotis et al. [10] have studied the biases of patterns chosen by users. They have found that 50% of the users choose the top-left point as the starting point of their patterns. Uellenbeck et al. [4] analyzed the bias of choosing the sequence of the points in their patterns as well. Exploiting these biases, they have estimated that 50% of users choose a pattern weaker than that of a 3-digit PIN.

B. Efforts to fortify the patterns

Previous efforts to increase the security of grid-based graphical passwords can be categorized into two different classes. The first class is focused on increasing the theoretical password-space (and implicitly increasing the effective password space used by users), either by increasing the size of the grid or introducing new degrees of freedom such as rotation and layering to the user interface. The second class is focused on developing approaches that explicitly expand the effective password space used by user. In this section, we enumerate significant works in both classes aimed to improve the security of free-form graphical passwords.

1) Background: Dunphy et al. [11] suggested "Background Draw-a-Secret" (BDAS) to improve the security of patterns by adding a background to the grid of the DAS scheme. Using a usability test, authors showed an increase in the length of patterns [12]. However, Gao et al. [13] and Zhao et al. [14] demonstrated the ease of guessing patterns based on the detectable hot-spots in the background images in the *Window 8 graphical password*, an approach similar to BDAS.

2) Rotation: Chakrabarti et al. [15] proposed a scheme called R-DAS adding rotation as a degree of freedom to DAS. This intuitively increases the theoretical password-space and may increase the effective password space. By drawing the same pattern but using rotation between several strokes, users hypothetically can achieve a stronger pattern. However, authors did not study the usability and effect of rotation on what users generate as their patterns. Applying rotation on Android Pattern is not practical because it is a single-stroke scheme. In addition, Android Pattern is used for frequent authentication and rotation possibly hampers its usability to a great extent.

3) Layering: Chiang et al. [16] proposed an extension of DAS called Touch-screen Multi-layered Drawing (TMD) where they add "wrap cells" that allow users to continuously draw their passwords across multiple layers. This improves the theoretical password-space. However, the usability study shows that biases of starting point and shape of the patterns remain pertinent.

4) Blacklisting: An intuitive approach to strengthen the security of patterns is to blacklist certain patterns that are used frequently (e.g., a pattern like "Z" or "N", or any pattern starting from the top-left point) and do not allow users to choose them as their pattern. Uellenbeck et al. [4] have experimented such an approach by removing the most frequently used starting point, o_0, from the Android unlock screen (i.e., blacklisting all patterns that start from that point).

 ISBN: 978-1-61208-418-3

They noticed that this resulted in a *new* frequently used starting point (o_1). Indeed, the blacklisting approach only shifts the distribution of patterns, and does not transform the skewed distribution to a uniform one.

5) Random assignment: Another option to strengthen the security of patterns is to assign a random pattern to the user. This resolves the problem of skewed distribution of patterns. Nonetheless, random assignment will suffer from practical weaknesses including usability [17] and memorability.

6) Rearrangement: Some believe that the shape of the grid and the arrangement of cells create some inherent biases on what users choose as their patterns (e.g., choosing straight vertical or horizontal lines, instead of a cross line, because of the visual effects of the grid). Therefore, rearranging of the points of a grid in a shape other than square (e.g., circle or random) is studied as a potential technique to remove such biases. However, it has been observed that the biases only shift to a new set of points and results in a new set of frequently used sub-sequences [4]. These modifications do not help to remove the biases of the patterns, and do not increase the security of scheme.

C. Persuasive password security as an alternative approach

Persuasive Technology is a psychological framework which can be defined as "interactive computing systems designed to change people's attitudes and behaviours" [18]. Built on top of persuasive technology, there has been prior work on persuasive password security which persuades users to choose strong passwords by creating suitable user-interfaces [19][20]. A persuasive user-interface guides user to choose options that are desirable from the perspective of the designer of the system. In the same way, a persuasive password user-interface guides user to choose passwords that are strong. Chiasson et al. [21] have proposed and studied the "cued click point" a variant of PassPoints [22] which employs persuasive password security techniques to reduce the biases and reduces the predictable hotspots from 40% to 8%.

Forget et al. have proposed a persuasive authentication framework [5] that enumerates possible techniques for persuasion. These include simplification, personalization, monitoring, conditioning, and social interaction, as applied to a user-interface. For example, personalization includes suggestions of secure options to the user, and monitoring includes feedback to users about the security of their choices.

D. Pattern strength

In this section, we discuss guessing attacks on the Android Pattern system. Assuming the attacker has a perfect knowledge of the system and the distribution of all Android pattern used by users, an attacker can build a probabilistic model for computing the probability $P(X)$ of every possible pattern X. A pattern X_1 is considered stronger than pattern X_2 if $P(X_2) > P(X_1)$; resulting in an attacker tries X_2 before trying X_1 to guess someone else's pattern. In summary, more likely patterns are guessed before less likely ones. Therefore, to evaluate the strength of a pattern, we develop a score function $f(X)$ based on the probability of a given pattern X, in which, a more likely pattern gets a low score, and a less likely one gets a higher score.

Uellenbeck et al. [4] demonstrated that a Markov probabilistic model can effectively estimate the probability of patterns as:

$$P(X = o_1 o_2 \dots o_m) = P(o_1 o_2 \dots o_{n-1}) * \prod_{i=n}^{m} P(o_i | o_{i-n+1} o_{i-n+2} \dots o_{i-1}) \quad (1)$$

To compute this probability, we need an appropriate training dataset to compute the conditional probability

$$P(o_i | o_{i-n+1} o_{i-n+2} \dots o_{i-1})$$

For a 3-gram Markov model, we should compute the probability $P(o_i|o_{i-2}o_{i-1})$ for all different combinations of the nodes. We use an estimation of this probability instead, by collecting enough sample of patterns, using an appropriately designed experiment. If a sequence does not occur in our dataset, the probability of *zero* is assigned to that n-gram, leading to estimation of zero as the probability of a rare pattern. To fix this issue, we use Kneser-Ney smoothing (an advanced form of absolute-discounting interpolation) [23]. It is considered as the most effective method of smoothing.

We use a simple score function based on the Markov probabilistic model. MM-score(X) = $-log(p(X))$, where probability of X, $P(X)$ is computed by (1). We refer to this as MM-score function in the rest of the paper. A pattern X_1 is stronger than pattern X_2 iff:

$$MM - score(X_1) > MM - score(X_2)$$

We categorize all patterns of Android Pattern into 3 different levels of security: *weak*, *medium*, and *strong*. We compute the strength of all possible Android Pattern based on MM-score defined above. Defining an interval for the score of the patterns in each of these levels is subjective. A minimum security of a 4-digit PIN system is considered appropriate for authentication in ATMs [24][25] and smartphones. Therefore, we classify the patterns which offers the security of a 2-digit PIN as *weak* patterns accounting to a total of 100 patterns. The next 900 patterns are labeled as *medium* security, as they provide a maximum security of a 3-digit PIN, and all other patterns are labeled as *strong* patterns, since they provide the minimum security of a 4 to 5 digits PIN.

III. Proposed Approach

In this section, we first present our findings on what users are aware and where they lack understanding of strength of patterns (see Section III-A). Instead of assigning random pattern to a user, imposing unreasonable restrictions or enforcing them to not use certain blacklisted patterns/points, we propose persuasive security mechanisms to nudge/persuade users to choose secure patterns, without potentially hampering the usability or security of the system. In this paper, we propose a) EPSM (using self-monitoring) and b) BLINK (using personalization) to help users create stronger patterns by infusing knowledge of the global pattern distribution to the system (see Section III-B and Section III-C, respectively).

 ISBN: 978-1-61208-418-3

A. Plausible awareness and obliviousness

To provide better security suggestions or instructions, it is first important to understand what users are aware of and where they are lacking. To understand this, we conducted a short online survey using Amazon Mechanical Turk (MTurk). From 336 total participants, we analyzed the responses from only 266 participants. We eliminated a) anyone who provided contradicting answers to the same question asked multiple times with different wording. b) who completed the survey in less than 30 seconds or took more than 5 minutes. All participants, even those we eliminated, were paid $0.50. Participation for this survey was restricted to only Android users who have either used or are using Android unlock pattern as authentication mechanism on their phones. This survey and all the experiments reported in this paper were approved by the Institutional Review Board of the New York University (IRB approval reference IRB-13-9674) and Institutional Review Board of the New York University Abu Dhabi. Data collected from the participants was anonymized and protected according to the procedures described in the corresponding IRB submission documents.

We used a within-subjects design and asked ("*How strong is the following pattern?*") the participants to rate the strength of six patterns as shown in Figure 2 on a 5-point Likert scale. We chose patterns from three different security levels based on their strength, *weak*, *medium*, and *strong*, as is defined in section II-D. Patterns 2(a) and 2(b) are weak patterns, patterns 2(c) and 2(d) are of medium security level, and patterns 2(e) and 2(f) are strong patterns (refer to Figure 2). To avoid biases, we randomized the order that the patterns were shown to the participants.

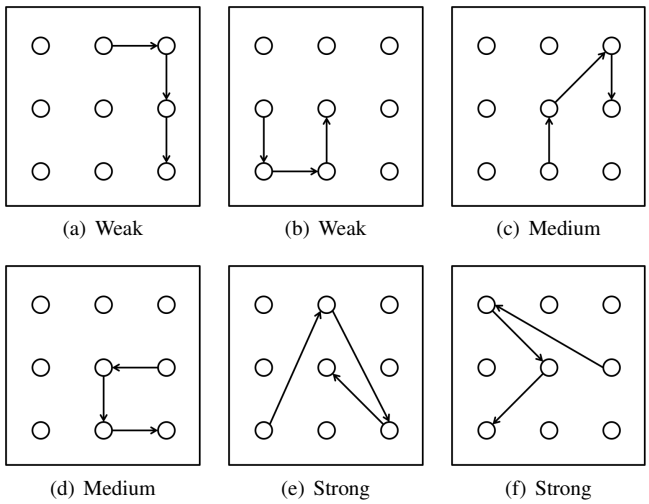


Figure 2. Choice of different proposed patterns (based on pattern strengths)

As it can be observed from result of the survey in Figure 3, users *are aware* of the relative strength of patterns and can distinguish between complex (Figures 2(e) and 2(f)) and easy to guess (Figures 2(a) and 2(b)) patterns.

However, this knowledge is not translated into selection of strong patterns by a large number of users and many still choose weak patterns. However, we exploit this *awareness* and design a simple but effective feedback mechanism called EPSM which provides feedback to users about the security

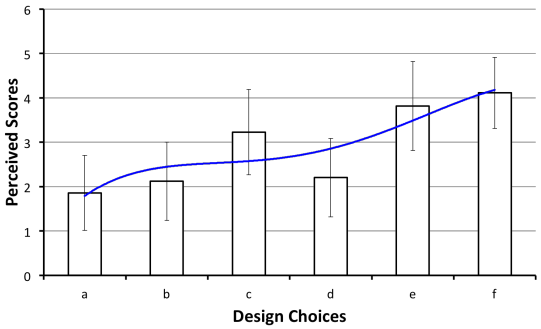


Figure 3. Perceived strength of patterns by users. It is mostly consistent with computed strength class for patterns.

of their pattern (by updating the color of the pattern). EPSM does not provide any hint on how to create a better pattern because users are already *aware* of which patterns are more secure.Therefor, users will be able to update their patterns to a stronger one.

Moreover, based on this survey, we observed that the perceived security strength of patterns 2(b) and 2(d) is almost same (as shown in Figure 2), where their strength is not the same in reality (because of the different starting points and n-grams they use). We believe that this could be because of the similar shapes of the two patterns. This suggests that even if user chooses a pattern of a shape similar to a weaker pattern but with a different starting point, it can increase the strength of the pattern. For this purpose we propose BLINK in which the system suggests a different starting point to the users.

B. EPSM: Embedded pattern strength meter

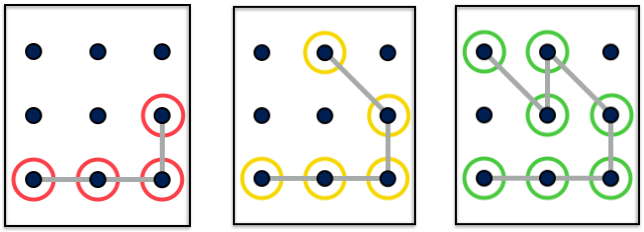
Self-monitoring is one of the persuasive security principles helping users to adjust their security behavior [20] and was used to design EPSM. Andriotis et al. [26] showed the promise of this approach by providing a text-based feedback to users about the strength of their patterns after they complete drawing it. In their experiment, one out of five subjects changed their patterns after knowing their patterns are not strong.

In EPSM, instead of giving a delayed feedback after users generate their patterns and using a separate user interface element (e.g., text-based feedback), we provide a *continuous* and *embedded* feedback while the user is drawing a pattern. This design choice was made because a) it is more effective to provide continuos feedback influences the user's decision of what to choose as her pattern. b) smartphones have relatively small screen size which demands a compact representation of information and feedback. This is helpful for users to adjust the strength of their patterns as they create the patterns.

EPSM provides a fine-grained continuous embedded feedback by coloring the user's pattern according to pattern's strength level. The *red* color alarms a *weak* pattern, *yellow* indicates *moderate*, and *green* represents a *strong* pattern (see Figure 4). The system also pops-up a message describing the meaning of each color "*As you draw your pattern, the color of your pattern changes from red to green. Red one is bad (others can guess your pattern), yellow is good, and green is perfect.*" Regardless of the strength of the pattern the color of the pattern remains *red*, until the pattern satisfies the minimum required length (i.e., four). Thereafter, based on the

 ISBN: 978-1-61208-418-3

strength of the pattern, the color of the pattern gets updated (see Section II-D for details on pattern strength). Note that change in the color usually goes from red to yellow to green, however in certain cases it can sometimes goes the other way around, i.e., from green to red. For example, a half-drawn Z ($o_0 - o_1 - o_2 - o_4 - o_6 - o_7$) is a more secure pattern than a full drawn Z ($o_0 - o_1 - o_2 - o_4 - o_6 - o_7 - o_8$).



(a) Red (b) Yellow (c) Green

Figure 4. Sequence showing the change in colors for a single pattern, as user draw it

C. BLINK- Nudging

Nudging, a concept in behavioral science, argues that positive reinforcement and suggestions can influence the motives, incentives and decision making of groups and individuals [27]. Nudging can be used to suggest stronger patterns to users. However, suggesting a random pattern hampers usability and memorability of the Android Pattern. A more practical option is to provide partial suggestions. For example, suggesting users where to start their patterns is helpful to remove the bias of starting points (e.g., more than 40% of the users use upper most left point to start their patterns [4][10]) that can be used by attackers to guess the patterns easily.

In a pilot study, we examined a number of techniques to suggest a starting point to the users without hampering the usability of Android Patterns. Based on our observations and users suggestions, we concluded that a) suggestion by blinking a point is very effective, and b) users need to be told what is expected be done with the suggested point without hampering the usability of the system. Our final design uses a blinking point (see Figure 5) and similar to EPSM it pops-up a recommendation message stating "*It is recommended to start your pattern with the blinking point but NOT MANDATORY*", as it helps user to understand what to do and how to proceed with the blinking point.

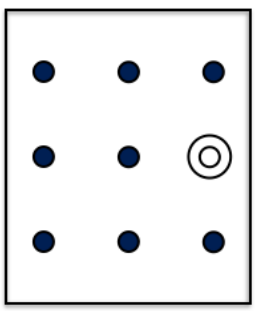


Figure 5. Suggesting a start point by blinking

During the registration phase, the system randomly recommends one of the 9 points in the screen to the user by creating an additional circle around the recommended point and blinking indefinitely. The circle stops blinking when the user starts drawing a pattern.

IV. Experimental Setup

To measure the efficacy of the designed persuasive security schemes, we conducted a between-subjects usability study on Amazon Mechanical Turk (MTurk). Subjects were assigned randomly to three different user interfaces: a) the control group, where users were assigned to work with the normal Android Patterns user interface (NORMAL), b) the BLINK group, where users were assigned to the BLINK user interface, and c) the EPSM group, where users were assigned to the EPSM user interface. We compare the security, memorability and usability of our proposed user interfaces (EPSM and BLINK) with that of the normal Android Patterns (NORMAL).

Because of Amazon's MTurk policy, we could not ask our participants to install an app on their smartphone to participate in our user study, therefore, we implemented the NORMAL, EPSM and BLINK using web technologies (HTML, CSS, and Javascript) accessible by visiting a link on participants' smartphones. In order to avoid the possibility of doing the experiment on a desktop or any other device beside an Android phone, we checked the "Browser Agent" field of the HTTP requests and only permitted those requests issued by an Android phone. Web pages are rendered differently on different devices and browsers, and there is a possibility that users do not see the user interface as we expected. To detect any such distortions in users' experience, we asked the participants about the quality of the user interface, at the end of the user study in the post user study survey. Any data from those who reported a distortion in the main page is excluded from our analysis.

The user study procedure involved two main steps a) *Registration* - Participants were assigned randomly to one of the groups, i.e., NORMAL, EPSM or BLINK. After that, they were asked to choose a pattern and then to verify it immediately. All participants were instructed to imagine that they have received a new phone and would like to set an Android Pattern on it. They were asked to choose a pattern of a minimum length four. b) *Survey* - After creating a pattern, participants were asked to complete a survey. We paid $0.40 to each participant upon the completion of our user study.

TABLE I. GROUP DEMOGRAPHICS.

Group	Total Participants
NORMAL	92 M (28); F (64)
EPSM	72 M(25); F(47)
BLINK	106 M(31); F(75)

We recruited a total of 270 US-based workers to participate in our experiment. Note, that these are different participants than the participants described in Section III-A. We also confined our user study to those who are familiar with the Android Patterns. Demographics of the participants and number of participants are given in Table I.

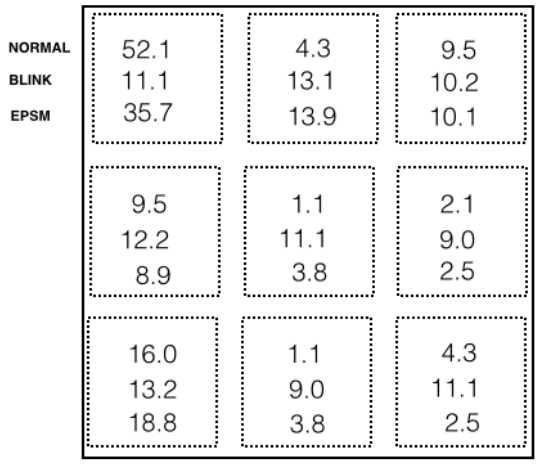
V. Results

In this section, we analyze the patterns obtained during our user study, and compare the security, usability and memorability of each design schemes proposed.

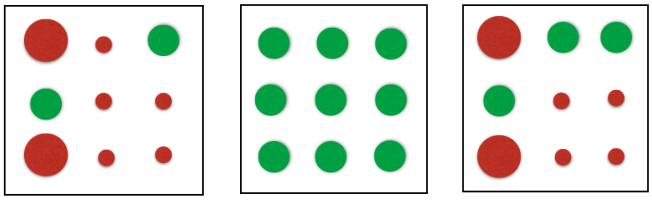
 ISBN: 978-1-61208-418-3

A. Starting point distribution

Figure 6(a) shows the percentage of patterns starting from each of the nine points in all three schemes. As expected in NORMAL and EPSM, we can observe that the starting point probabilities of the top two corner points (52.1% and 35.7%) are much higher than the other points. This is because there was no recommendation provided for eliminating the starting point bias in NORMAL and EPSM.



(a) Start point distribution(%)



(b) NORMAL (c) BLINK (d) EPSM

Figure 6. Distribution of starting point of the patterns in the NORMAL, BLINK and EPSM. BLINK removes the bias of starting points.

On the other hand, the bias of starting point probability in BLINK is eliminated and it is distributed almost uniformly. This happens because BLINK suggests the starting points randomly, and 85% of users use the suggested point. For some others, this has apparently nudged them to choose the start point of their patterns wisely. Table II shows the percentage of the suggestions used by users for each of the nine points in the BLINK.

TABLE II. PERCENTAGE OF USED SUGGESTION FOR EACH POINT IN THE BLINK USER INTERFACE. POINTS ARE SUGGESTED UNIFORMLY AND RANDOMLY.

Point	0	1	2	3	4	5	6	7	8
Used	80%	90%	77%	100%	84%	73%	100%	81%	85%

B. Pattern strength

Theoretical password-space of the normal Android Pattern is higher than that of a 5-digit PIN, but a large number of users (51%) use patterns with strength level (see Section II-D) of a 3-digit PIN (*weak* and *medium* strength patterns). We designed EPSM and BLINK to persuade users to use patterns of greater strength. Figure 7 shows the strength level of the patterns generated in our user study. The percentage of users that create strong patterns is increased to 60% in the BLINK, and 77.4% in the EPSM. This shows that BLINK and EPSM are able to persuade users to choose stronger patterns.

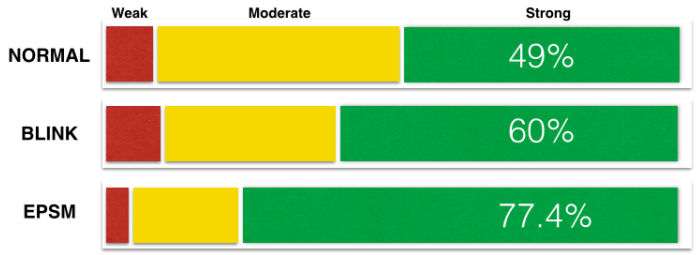


Figure 7. Strength level of the patterns created by users of NORMAL, BLINK and EPSM. Red, yellow, and green bars indicate weak, moderate, and strong patterns respectively.

C. Security against partial guessing attack

In this section, we study the resilient of patterns generated by each user interface against guessing attacks. In a guessing attack, an attacker has access to an oracle (a blackbox) that gets a password and answers yes if it is correct. An oracle may answer an unlimited number of queries or apply some limitations on the amount or speed of returning the results. For example, on Android phones, a maximum number of 20 guesses are granted before the attacker get locked out completely. An oracle may also appear in the form of a secure hardware module that is rate-limited for the purpose of deterrence of attackers, and answer the queries very slowly (e.g., iPhone [28]).

An optimal attacker tries the weak patterns before the strong ones because they are more likely to be used by users. Indeed, such an adversary builds a dictionary

$$D_{pattern} = \{pt_1, pt_2, ..., pt_n\}$$

as a set of all possible patterns. Then, he computes the probability of occurrence of each pattern based on a probabilistic model (see equation 1), and then sorts the patterns based on their probability to compute the ordered list $G = (g_1, g_2, ..., g_n)$ where

- $g_i \in D_{pattern} \quad \forall \quad 1 \leq i \leq n$
- $P(g_i) \leq P(g_j) \quad \forall \quad i \leq j$

The attacker tries patterns in the order in which they appear in the ordered list G. Since guessing very strong patterns is time consuming, and is not cost-effective in many cases, an attacker usually guesses a portion of passwords with a reasonably small effort. This is called *partial guessing attack* and is used to evaluate the security of text-based passwords [29] and Android Patterns [4]. Accordingly, we evaluate the security of our proposed user interfaces against *partial guessing attack*.

Figure 8 shows the success rate of guessing attack against NORMAL, EPSM and BLINK. As we can observe, it is evident that patterns in EPSM and BLINK are stronger than NORMAL, and an attacker needs more effort (i.e., number of guesses) to guess a portion of them in comparison with patterns in NORMAL. Specifically, an attacker needs only 886 guesses to guess 50% of the patterns in NORMAL, whereas he

 ISBN: 978-1-61208-418-3

needs 3344 and 1918 guesses to guess 50% of the patterns in EPSM and BLINK, respectively. In terms of *partial entropy* of patterns, this translates to 9.79 bits of entropy for the patterns used by 50% of NORMAL, and 11.7 and 10.9 bits of entropy for the same proportion of users in EPSM and BLINK.

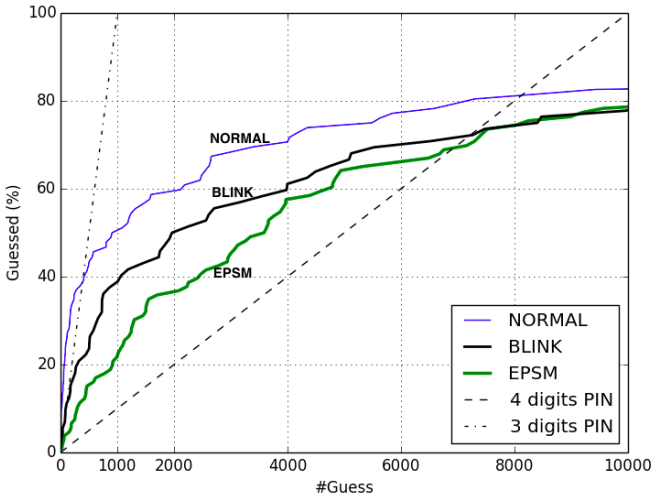


Figure 8. Guessing attack against NORMAL, BLINK and EPSM. Guessing model is built over the NORMAL version.

D. Pattern Length

Table III shows statistics about the length of patterns for each design schemes. Even though there are minor differences between the length of patterns in EPSM as compared to NORMAL and BLINK, Kruskal Wallis test (This is a non-parametric equivalent of the one-way analysis of variance (ANOVA)) does not show any significant difference between length of the patterns (χ^2=2.5, p>0.28). This emphasizes that the increased security offered by EPSM is not resulted by longer patterns, but is because of using more complex patterns with higher strength levels.

TABLE III. LENGTH OF PATTERNS.

Group	Average length	Std. Dev
NORMAL	6.13	1.61
BLINK	6.12	1.78
EPSM	6.5	1.72

E. Short-term recall rate

One of the design considerations for our new variations is to create a strong yet easy to use pattern scheme without hampering the users' recall rate. In the frequent authentication schemes (e.g. unlocking phone that is done several times a day), the repetition of password entry helps users to recall the pattern over long intervals. Consequently, it is reasonable to measure the short-term memorability of the patterns [30].

Since the maximum idle timeout before the Android phone gets locked down in normal Android Patterns is 30 minutes, we tested the memorability of the patterns after 20 minutes. It was not possible to guaranty that the subjects we recruited from the Amazon Mechanical Turk will take the recall task within 20 minutes. Therefore, we ran a between-subjects in-lab study and recruited 60 students to evaluate the recall rate of

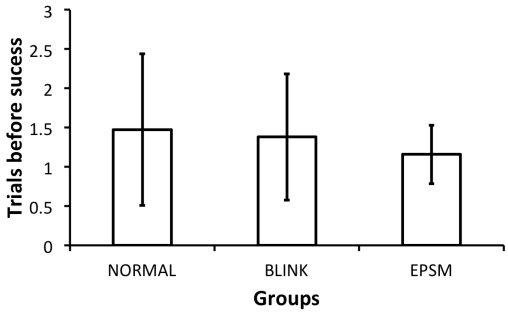


Figure 9. Recall accuracy (in terms of trials before success)

the proposed user interfaces. When they created their patterns using NORMAL, BLINK, and EPSM, we asked them to return back after 20 minutes to do a follow-up test where we asked them to re-enter their pattern. Figure 9 shows the recall rate of the patterns for each version after 20 minutes. We compute the recall rate accuracy in terms of how many trials participants took on average to verify themselves against the system. Based on the pairwise *t-test* conducted we found that there is no statistically significant difference between the the recall rate of NORMAL and BLINK; and NORMAL and EPSM.

VI. Conclusion

In this paper, we proposed two Android Patterns schemes with the goal of improving the security of patterns chosen by users. We used the principles from *persuasive security framework* to nudge users to choose starting points uniformly and to use more complex sequence of points in their patterns. We recruited 270 participants from Amazon Mechanical Turk and conducted a usability user study to measure the effect of our proposed schemes on security and usability of the system.

While only 49% of subjects choose *strong* patterns in standard Android Patterns, our suggested schemes increase it to 60% in BLINK and 77.4% in EPSM version. Accordingly, the partial entropy of the patterns is increased from 9.79 in NORMAL to 10.9 in BLINK and 11.7 in EPSM. These improvements are achieved without hampering the usability in term of the length of the pattern and short-term recall rate.

VII. Acknowledgement

The first, third, and fourth authors were supported by NSF grant 1228842. We would like to thanks Markus Jakobsson for suggestions of the design for the experiment, and all the anonymous reviewers for their comments and feedbacks towards this work.

References

[1] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner, "Are you ready to lock?" in Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, ser. CCS '14. New York, NY, USA: ACM, 2014, pp. 750–761.

[2] D. Van Bruggen, S. Liu, M. Kajzer, A. Striegel, C. R. Crowell, and J. D'Arcy, "Modifying smartphone user locking behavior," in Proceedings of the Ninth Symposium on Usable Privacy and Security. ACM, 2013, p. 10.

[3] D. C. Van Bruggen, "Studying the impact of security awareness efforts on user behavior." Ph.D. Thesis, University of Notre Dame, 2014.

 ISBN: 978-1-61208-418-3

[4] S. Uellenbeck, M. Dürmuth, C. Wolf, and T. Holz, "Quantifying the security of graphical passwords: The case of android unlock patterns," in Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security, ser. CCS '13. New York, NY, USA: ACM, 2013, pp. 161–172.

[5] A. Forget, S. Chiasson, and R. Biddle, "Persuasion as education for computer security," in World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, vol. 2007, no. 1, 2007, pp. 822–829.

[6] I. Jermyn et al., "The design and analysis of graphical passwords," in Proceedings of the 8th USENIX Security Symposium, vol. 8. Washington DC, 1999, pp. 1–1.

[7] H. Tao and C. Adams, "Pass-go: A proposal to improve the usability of graphical passwords." IJ Network Security, vol. 7, no. 2, 2008, pp. 273–292.

[8] J. Thorpe and P. Van Oorschot, "Towards secure design choices for implementing graphical passwords," in Computer Security Applications Conference, 2004. 20th Annual. IEEE, 2004, pp. 50–60.

[9] P. C. van Oorschot and J. Thorpe, "On predictive models and user-drawn graphical passwords," ACM Transactions on Information and system Security (TISSEC), vol. 10, no. 4, 2008, p. 5.

[10] P. Andriotis, T. Tryfonas, and G. Oikonomou, "Complexity metrics and user strength perceptions of the pattern-lock graphical authentication method," in Human Aspects of Information Security, Privacy, and Trust. Springer, 2014, pp. 115–126.

[11] P. Dunphy and J. Yan, "Do background images improve draw a secret graphical passwords?" in Proceedings of the 14th ACM conference on Computer and communications security. ACM, 2007, pp. 36–47.

[12] Z. Zhao, G.-J. Ahn, J.-J. Seo, and H. Hu, "On the security of picture gesture authentication," in Proceedings of the 22Nd USENIX Conference on Security, ser. SEC'13. Berkeley, CA, USA: USENIX Association, 2013, pp. 383–398.

[13] H. Gao, W. Jia, N. Liu, and K. Li, "The hot-spots problem in windows 8 graphical password scheme," in Cyberspace Safety and Security. Springer, 2013, pp. 349–362.

[14] Z. Zhao, G.-J. Ahn, J.-J. Seo, and H. Hu, "On the security of picture gesture authentication." in USENIX Security, 2013, pp. 383–398.

[15] S. Chakrabarti, G. V. Landon, and M. Singhal, "Graphical passwords: drawing a secret with rotation as a new degree of freedom," in Proceedings of the Fourth IASTED Asian Conference on Communication Systems and Networks. ACTA Press, 2007, pp. 561–173.

[16] H.-Y. Chiang and S. Chiasson, "Improving user authentication on mobile devices: A touchscreen graphical password," in Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services. ACM, 2013, pp. 251–260.

[17] E. A. Stobert, "Memorability of assigned random graphical passwords," Ph.D. dissertation, Carleton University, 2011.

[18] B. J. Fogg, "Persuasive technology: using computers to change what we think and do," in Ubiquity, vol. 2002, no. December. ACM, 2002, p. 5.

[19] D. Weirich and M. A. Sasse, "Persuasive password security," in CHI'01 Extended Abstracts on Human Factors in Computing Systems. ACM, 2001, pp. 139–140.

[20] A. Forget, S. Chiasson, P. C. van Oorschot, and R. Biddle, "Persuasion for stronger passwords: Motivation and pilot study," in Persuasive Technology. Springer, 2008, pp. 140–150.

[21] S. Chiasson, P. C. van Oorschot, and R. Biddle, "Graphical password authentication using cued click points," in Computer Security–ESORICS 2007. Springer, 2007, pp. 359–374.

[22] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon, "Passpoints: Design and longitudinal evaluation of a graphical password system," Int. J. Hum.-Comput. Stud., vol. 63, no. 1-2, Jul. 2005, pp. 102–127.

[23] J. M. Gawron, "Discounting," http://www-rohan.sdsu.edu/~gawron/compling/course_core/lectures/kneser_ney.pdf, accessed: 2015-03-19.

[24] I. O. for Standardization, "Iso 9564," http://en.wikipedia.org/wiki/ISO_9564, accessed: 2015-03-19.

[25] B. Milligan, "The man who invented the cash machine," http://news.bbc.co.uk/2/hi/business/6230194.stm, accessed: 2015-03-19.

[26] P. Andriotis, T. Tryfonas, G. Oikonomou, and C. Yildiz, "A pilot study on the security of pattern screen-lock methods and soft side channel attacks," in Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks, ser. WiSec '13. New York, NY, USA: ACM, 2013, pp. 1–6.

[27] R. Thaler and C. Sunstein, Nudge. Yale University Press, 2008.

[28] S. Garfinkel, "The iphone has passed a key security threshold," http://www.technologyreview.com/news/428477/the-iphone-has-passed-a-key-security-threshold/, accessed: 2015-03-19.

[29] J. Bonneau, "The science of guessing: analyzing an anonymized corpus of 70 million passwords," in Security and Privacy (SP), 2012 IEEE Symposium on. IEEE, 2012, pp. 538–552.

[30] A. Beautement and A. Sasse, "Gathering realistic authentication performance data through field trials," in SOUPS USER Workshop, 2010.

 ISBN: 978-1-61208-418-3

An Architecture for Self-healing in Internet of Things

Fernando Mendonça de Almeida, Admilson de Ribamar Lima Ribeiro, Edward David Moreno
Department of Computing
Federal University of Sergipe – UFS
São Cristóvão, Brazil
e-mails: fernando.m.al.91@gmail.com, admilson@ufs.br, edwdavid@gmail.com

***Abstract*—Security in Internet of Things has an important role into mass adoption of the technology. There are some research works in this area, but most of them are related to Wireless Sensor Networks or Internet Networks. The association of a security mechanism and the self-* properties is very beneficial for the Internet of Things, considering the growth of connected devices. This paper proposes an architecture that uses the Dendritic Cell Algorithm for a security system with self-healing property for the Internet of Things.**

Keywords-Internet of Things; Dendritic Cells Algorithm; Self-healing.

I. INTRODUCTION

The Internet of Things (IoT) is a novel paradigm whose concept is based on ubiquitous presence of objects - sensors, actuators, Radio-Frequency IDentification (RFID) tags, mobile devices etc - that interact with each other using unique addresses to achieve common goals [1]. The IoT is extremely vulnerable to attacks, most of communication is wireless based, most of the components have constrained resources and it is possible to physically attack the IoT components [1][2]. Considering that the IoT will have information about almost everything, security and privacy are key concerns in IoT research [3][4].

Xu, He and Li [3] say that the research about security in IoT is necessary for the massive adoption of this technology in Industry. Gubbi, Buyya, Marusic and Palaniswami [4] highlight the need of self-protection in domestic applications, arguing that actuators will be connected to the system and they will need protection from intruders.

According to Roman, Zhou and Lopez [5], fault tolerance will be essential in the IoT. The number of vulnerable systems and attacks will increase, so there is a need to develop intrusion detection and prevention systems to protect the components of the IoT.

The growth of connected devices in IoT make the human intervention less effective. Autonomic computing aims to reduce the human intervention in complex systems, like the human nervous system controls autonomically some functions of body, like the digestive system [6].

The proposed architecture defines five components, distributed between the monitoring phase, analysis phase and knowledge component, to add the property of self-healing into the Internet of Things. The Dendritic Cell Algorithm (DCA) is combined with an Artificial Intelligence component to allow the DCA learning the role of each information sensed. The architecture described in this paper can implement a self-healing system in IoT nodes distributing roles between them in order to ease design of self-healing systems for IoT.

The remainder of this paper has five more sections: Section II introduces some aspects of this subject and the self-* properties; Section III presents four kinds of attacks in the Internet of Things that will be mitigated with the proposed architecture; Section IV presents three works related to security in IoT, self-protected system and use of dendritic cell algorithm; Section V describes the architecture to mitigate four kinds of attacks in IoT and Section VI presents the conclusion.

II. AUTONOMIC COMPUTING

The first utilization of Autonomic Computing term was made by the International Business Machine (IBM) in 2001 to describe self-managing systems [10]. The Autonomic term comes from biology, where, for example, the autonomic nervous system from the human body takes care of most bodily functions, by removing from the consciousness the need to coordinate all the bodily functions.

In IBM's manifesto, they suggested that complex systems should have autonomic properties and distilled the four properties of self-managing systems: self-configuration, self-optimization, self-healing and self-protecting.

A. Self-configuration

The self-configuration property is found in systems that are capable to self-install and self-set to achieve the user goals.

B. Self-optimization

The self-optimization property is found in systems that can make some changes proactively to improve the performance of the system.

C. Self-healing

The self-healing property is found in systems that detect and diagnose problems. It is important that the self-healing systems have fault tolerance.

D. Self-protecting

The self-protecting property is found in systems that protect themselves from malicious attacks. The autonomic

 ISBN: 978-1-61208-418-3

system adjusts itself to offer security, privacy and data protection.

E. MAPE-K Autonomic Loop

In the IBM's manifesto, they presented a reference model, the MAPE-K autonomic control loop, where the responsibilities are shared between the components of the MAPE-K loop: Monitor, Analyze, Plan, Execute and Knowledge, Sensors and Effectors.

The monitor phase is responsible to collect data from sensors and process that data through the analysis phase. The analysis phase is responsible to receive the processed data and detect possible problems in the system. The plan phase organizes the necessary actions to fix the detected problems in analysis phase. The execution phase implements the actions planned.

III. Dendritic Cell Algorithm

The Dendritic Cell Algorithm (DCA) was introduced by Greensmith, Aickelin and Cayzer [11] and is inspired by the Danger Theory of mammalian immune system. The main elements of DCA are: Dendritic Cells (DC), Lymph nodes and antigens. The input signals of DCs are: danger signal, safe signal, PAMP (pathogenic associated molecular patterns) and inflammatory signal. The output signals of DCs are: Costimulatory Molecules (CSM), semi-mature signal and mature signal.

The antigens are the input of DC and they are presented iteratively to dendritic cells. Each antigen increments the CSM. When the CSM pass the migration threshold, the DC migrate to lymph node. The danger signal and PAMP increments the mature signal of DC and the safe signal increments the semi-mature signal. The inflammatory signal raises all other signal increments.

When the DC achieves the migration threshold, it will move to lymph node and the DC will be labeled as mature or semi-mature, comparing the mature and semi-mature signals. After receiving a defined number of DCs, the lymph node will calculate the Mature Context Antigen Value (MCAV), that is the percentage of mature DCs per all DCs received. The Dendritic Cell Algorithm detects an attack if the MCAV surpass a defined threshold.

IV. Security Threats in Internet of Things

Ashraf and Habaebi [2] proposed a taxonomy for security threat mitigation techniques. In this taxonomy, there are fifteen threats classified between actors (Managed Resources and Autonomic Managers), layer (Machine to Machine, Network and Cloud) and approach (Self-Protecting, Self-Healing and Hybrid). This paper discusses four security threats: Jamming, Sinkhole, Hello Flood and Flooding.

A. Jamming

The Jamming threat affects the managed resource and is classified in the Machine to Machine (M2M) layer. It is an attack that occupies the wireless spectrum blocking the communication between the IoT devices. The attacker uses noise signals to interference the wireless communication. To detect a jamming attack, the device monitors the Received Signal Strength Indicator (RSSI) values. It signal is abnormally high when a jamming attack occurs. That technique was used by Salmon et al. [9] to detect jamming attack in a Wireless Sensor Network.

B. Sinkhole

A sinkhole attacker announces a beneficial routing path to receive route traffic through it. It is classified in the Network layer and affects the managed resources and autonomic manager. The Intrusion Detection System proposed by Raza, Wallgren and Voigt [7] uses the representation of the network to find inconsistence and detect a sinkhole attack.

C. Hello Flood

The Hello Flood attack can occur when the routing protocol prompts a node to send hello messages to announce its presence to the neighbors. The Routing Protocol for Low power and lossy networks (RPL) needs to build the routing paths with some kind of hello messages, which makes the RPL vulnerable to Hello Flood attack. The Hello Flood attack is classified in the Network layer and affects both managed resources and autonomic managers.

D. Flooding

In a flooding attack, the attacker tries to run out the victim's resources, e.g. battery, sending many connection establishment requests. The Flooding attack is classified in the Cloud layer and affects the Autonomic Manager. Considering that most part of IoT communication with IP protocol uses UDP, this attack can be mitigated by setting traditional connection barriers [2].

V. Related Work

A. SVELTE

Raza, Wallgren and Voigt [7] designed, implemented and evaluated SVELTE, an Intrusion Detection System (IDS) for the Internet of Things. The SVELTE detects sinkhole and selective-forwarding attacks in IPV6 over Low power Wireless Personal Area Networks (6LoWPAN) wireless network that uses RPL routing protocol.

Their IDS has a hybrid approach, it has distributed and centralized modules. The three main modules are: 6Mapper (6LoWPAN Mapper), Intrusion Detection Component and a mini-firewall.

The 6Mapper builds the network topology of RPL in the border router. Each node needs to have an 6Mapper client. The Intrusion Detection Component uses four algorithms to detect sinkhole and selective-forwarding attacks, the first algorithm detects inconsistence in network topology, the second algorithm detects nodes that may have messages filtered, the third algorithm verifies the network topology validity and the last algorithm verifies the end-to-end losses. The mini-firewall module has a server, in border router, and a client, in each node. Each node, when necessary,aks the border router to block messages from an external attacker.

The authors concluded that SVELTE can be used in the context of RPL, 6LoWPAN and IoT. Detecting sinkhole attacks with nearly 90% true positive rate in a small lossy network and almost 100% true positive rate in a lossless network configuration.

 ISBN: 978-1-61208-418-3

B. Dai, Hinchey, Qi and Zou

Dai, Hinchey, Qi and Zou [8] proposed a self-protected system based on feature recognition using virtual neurons. The virtual neurons have three components: information collector, neighbor communicator and feature recognizer. The information collector senses useful data for the self-protecting mechanisms, like processing use, memory, processing status, message size, transmission direction etc.

The virtual neurons communicate with each other in Peer to Peer (P2P) model and hierarchical model. The hierarchical communication allows a fast message propagation between clusters, each cluster having a head-neuron that have a fast communication with each other head-neurons.

The authors propose five self-protecting mechanisms, each one with an associate algorithm to detect and prevent one type of attack. The attack types are: eavesdropping, replay, masquerading, spoofing and denial of service.

Each mechanism senses the environment's data and, if an attack is detected, the connection is finished and all nodes involved are alerted.

In the paper, they present three use cases and the results. The use cases test the self-protection with eavesdropping, replay and denial of service attack. In the eavesdropping use case, with 15% of node coverage and with one node per three seconds of monitor frequency, it is possible to effectually prevent the attack. In the replay use case, it is necessary to use a buffer and the prevention grows when the buffer size and frequency inspection grows too. In denial of service use case, the authors verify that the proposed mechanisms increase the server availability.

C. Salmon et al.

Salmon et al. [9] proposed an anomaly based IDS for Wireless Sensor Networks using the Dendritic Cell Algorithm. The proposed IDS architecture has five elements: Monitoring, responsible for sensing the environment's values, Context Manager, responsible for managing the monitoring and parameter base, Intrusion Detection Manager, responsible for organizing the tasks and coordinate the responses and actions to other managers, Decision Manager, responsible for executing the dendritic cell algorithm, detect an attacker and manage the rules base, and Countermeasures, responsible for executing the actions to combat the identified attacks.

In their proposal, the authors divide two roles to be represented by the nodes: Dendritic Cells (sensor-dc) and Lymph node (sensor-lymph). The sensor-lymph have the Decision Manager and Countermeasures components, while sensor-dc have all the other components.

In the experiment, Salmon et al. used MICAz mote [16] with TinyOS [17]. The scenarios were simulated with TOSSIM (TinyOS Simulator) and they try to identify jamming attacks. The environment data used by dendritic cell algorithm included RSSI level, representing the PAMP signal, the received messages rate, as danger signal, and the inverse of received messages rate, as safe signal.

Several experiments were done, including changing configuration, time of attack, number of sensor-dc. Through the tests, the authors concluded that the IDS proposed is efficient for Wireless Sensor Networks saving energy from the nodes while there is a jamming attacker.

D. Comparison of Related work

The related work listed in this paper has some common goals, but not all of them. The SVELTE IDS [7] is designed for Internet of Things, but it was not designed considering the autonomic properties. Similar to SVELTE IDS, Salmon et al. [9] work was designed with resource constrains, but did not have the same network topology. Dai, Hinchey, Qi and Zou [8] designed their system with autonomic properties, but it is not possible to know if their approach fits into Internet of Things constrains. There is a summary of features of each work in Table I.

VI. SELF-HEALING ARCHITECTURE

The proposed architecture is based on MAPE-K loop. The phases of MAPE-K loop are divided into components and distributed on the nodes. The architecture is based on RPL Destination-Oriented Directed Acyclic Graph (DODAG), that is the default topology of a 6LoWPAN network that uses the RPL protocol. A typical RPL DODAG is depicted in Figure 1, it has a root and the other nodes. All nodes have a node ID, i.e., an IPv6 address, and all nodes, except the root, have one or more parents and a rank, that is relative to the distance between the node and the root.

TABLE I. RELATED WORK FEATURES COMPARISON

Feature	Related Work		
	SVELTE [7]	***Dai, Hinchey, Qi and Zou [8]***	***Salmon et al. [9]***
Has Autonomic Property		x	
Designed for IoT	x		
Detect Jamming			x
Detect sinkhole	x		
Detect DoS		x	
Detect Selective-Forwarding	x		
Detect Eavesdropping, Replay, Masquerading and Spoofing		x	

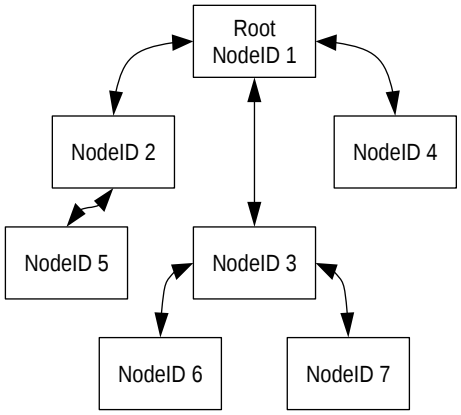


Figure 1. A typical RPL DODAG with one root and six other nodes.

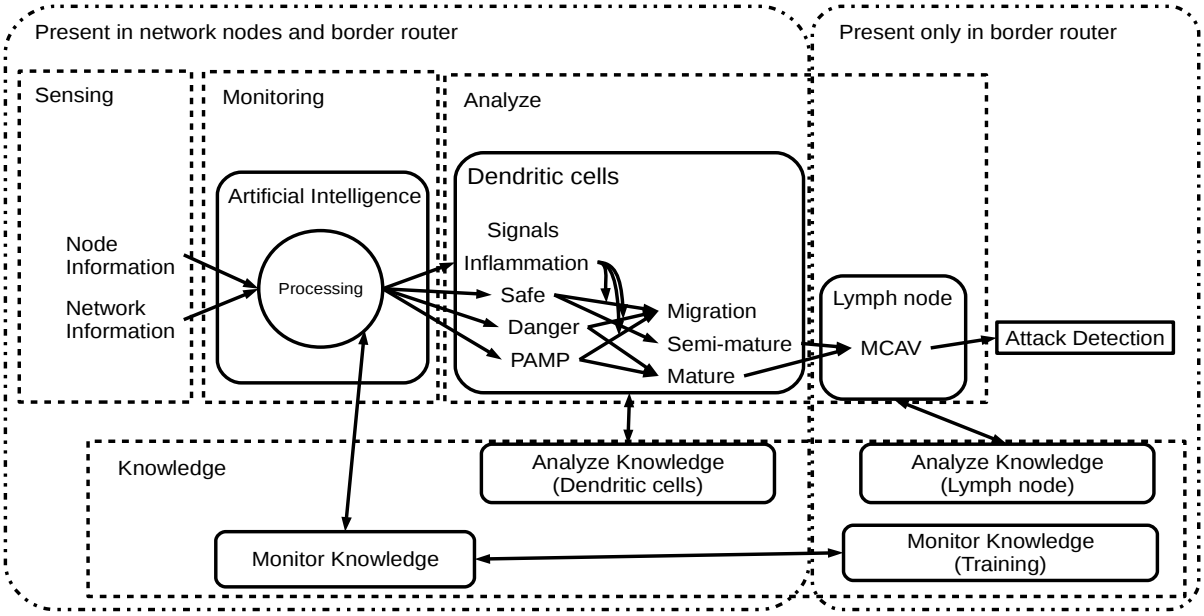


Figure 2. The proposed Architecture

There are five components described in the proposed architecture: Artificial Intelligence, Dendritic Cell, Lymph node, Monitor Knowledge and Analyze Knowledge. The other components of MAPE-K loop, Planning and Execution Phases, are not defined yet in the proposed architecture, they will be analyzed in future works.

The current components of the proposed architecture are distributed between Sensing, Monitoring, Analyzing and Knowledge elements of MAPE-K loop. The distribution of the proposed architecture components in the MAPE-K loop elements is depicted in Figure 2. Some components are not present on all network nodes, the Lymph node, Analyze Knowledge and part of Monitor Knowledge are present only in border router. Sensing, Monitoring, Analyzing, Monitor Knowledge and Analyze Knowledge components can be present in any network node and are present in border router.

A. Sensing Phase

The sensing phase is present in all nodes. In this phase the network and node information will be sent to monitoring phase. Node information will be the rate of successfully sent packets, total sent packets, RSSI level and more. Network information will be the network packets information, the DODAG routing tree information and more. As the root node is connected to the Internet, it will get more information about the network.

B. Monitoring Phase

The monitoring phase also is present in all nodes of the local network. In this phase there is the Artificial Intelligence component responsible to get the information from the sensing phase, the node and network information, process it and generate useful information for the analyzing phase.

The Artificial Intelligence proposed in this paper is an Artificial Neural Network Multi-Layer Perceptron (MLP). The MLP will be used to get useful information from the sensing phase, e.g. the network packets processed have some information such as total time to process and respond, and the MLP can try to predict this information without processing the packet, in order to provide this information quickly to the Analyzing phase.

The MLP will give a mix of real and inferred information, to the Analyzing phase in order to detect a possible attack to the node and network.

C. Analyzing Phase

In analyzing phase, the information generated by the monitoring phase will be used as input to the dendritic cells DC component. The DC component is present in all nodes of the local network. The signals of dendritic cells are classified as safe, danger or inflammatory signal. The monitoring phase receives the information from the sensing phase and infers the signal levels to this phase.

When the DC have enough information, they will migrate their result to the lymph node, present only in border router. The lymph node processes the DC result and detects if there is an attack on the network. This attack detection information in passed to the Planning phase.

D. Knowledge

The Knowledge components described in this paper are the Analyzing and Monitor Knowledge. The Plan and Execute Knowledge components will be present in the architecture too.

The monitor knowledge is split in two components, the training component and the component itself. The monitor knowledge itself is present in all nodes of the local network while the training monitor knowledge is present only in border router. This split occurs because the training of the artificial intelligence may need more resources than the node can offer.

The analyze knowledge is split in two parts, the dendritic cell part and the lymph node part. Each one is present where it counterpart component in analyze phase is present.

 ISBN: 978-1-61208-418-3

E. Planning, Execution and Effection Phase

Planning, execution and effection phase are not described in this architecture yet. Most efforts to define the network are concentrated on define how the nodes of an IoT network will detect an attack.

The planning phase will receive the analyze phase warning about an attack in the network and plan how to mitigate the side effects of the attack. This phase should consult previous network packages to improve the plan. The planning phase will be split in two components, one present in border router, the node with more processing resources in the local network and the other present in all nodes of the local network. The planning phase in the border router will list actions to mitigate the side effects of the attack and will distribute these actions to the planning component inside all the nodes. When the planning component receives the actions, it will pass the actions to the Execution phase.

The Execution phase will receive the actions planned in the planning phase and deliver each order from each action to the effection phase. The effectors of the node will receive clear orders from the execution phase and they will perform it to mitigate the side effects of the attack detected in the planning phase.

F. Attack Detection

The early implementation of the proposed architecture will try to detect Jamming, Sinkhole, Hello Flood and Flooding attacks. To detect Jamming attacks, as in SALMON et al. [9], the RSSI level and rate of messages will be sensed in the sensing phase. To detect sinkhole attack, like SVELTE [7], the DODAG tree information will be processed and inconsistency, validity etc will be sensed in the sensing phase. To detect hello flood and flooding, the number of connection attempts and RPL's hello messages from the same address will be sensed. The proposed architecture will capture the mentioned information and use it to detect an attack.

VII. RESULTS

The initial results of this research are about the technique used in the monitoring phase. The chosen technique for the first efforts is an Artificial Neural Network, a Multi-Layer Perceptron with Limited Weights (MLPLW) based on the neural network with limited precision weights [12].

The MLPLW implemented has 10 neurons in the hidden layer and each weight is represented by a byte. The training technique of the MLPLW is the Quantized Back-Propagation Step-by-Step (QBPSS) [12], a modified version of Back-Propagation for neural network with limited weights.

To check the implementation, the KDD99 dataset [13], widely used in Intrusion Detection Systems, was used. Yan, Wang and Liu [14] used the KDD99 dataset to evaluate their hybrid technique that uses a rule-based decision and neural network. Their accuracy rate was 99.75% and the false positive rate was 0.57%.

The KDD99 dataset was used with our ANNLW implementation with a stream based training [15]. Each input is used once and the accuracy and false positive rates were measured every thousand inputs. After the first thousand inputs, the MLPLW achieved 97,65% accuracy, but oscillated until the thirty-fourth thousand input. The accuracy rate after the stabilization is close to the accuracy in Yan, Wang and Liu[13], considering the use of fewer neurons in the hidden layer. The oscillation of the accuracy rate of the MLPLW is depicted in Figure 3.

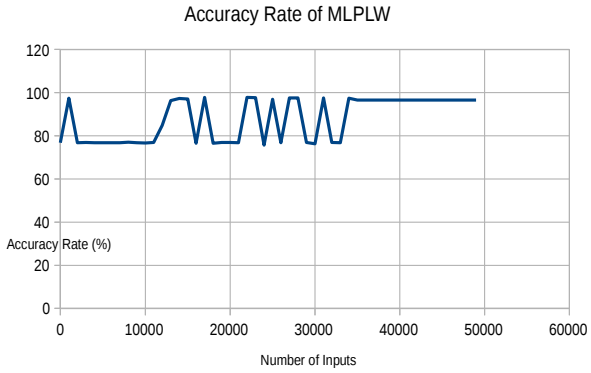


Figure 3. Accuracy rate of MLPLW over the number of inputs.

As our proposed architecture shall be used in the Internet of Things context, the used techniques should be in accordance to the use of resources. The memory used by MLPLW can be seen in Table II. It is possible to see in Table II that the MLPLW have a small impact in memory utilization in an ARM Cortex-M3 with 512 KB of Persistent memory and 64 KB of Volatile memory, 0.33% and 0.64% respectively.

VIII. CONCLUSION AND FUTURE WORK

The security in Internet of Things is a key property for the mass adoption of the technology. The research of security in Wireless Sensor Network and Internet itself can be used to show paths into security in IoT.

This work presents an architecture with self-healing property for the Internet of Things using ideas from Wireless Sensor Networks applied to a 6LoWPAN network.

The system will reuse components to detect multiple types of attacks, only by using additional information from the nodes and network, but without a dramatic algorithm change.

The proposed architecture will mitigate four different kinds of attacks of three different layers: Machine to Machine, Network and Cloud.

The MLPLW implemented shows that the monitor phase of the proposed architecture can use less than 1% of the memory of the embedded system and still have a high accuracy rate.

TABLE II. MLPLW MEMORY CONSUMPTION TABLE

Resource	***MLPLW consumption***
ROM memory	1716 bytes
RAM memory	420 bytes
Related ROM (related to 512 KB)	0.33%
Related RAM (related to 64 KB)	0.64%

 ISBN: 978-1-61208-418-3

For future work, there will be the implementation of the proposed architecture to validate it. The Artificial Intelligence component algorithm should be defined. The performance of the system should be evaluated to verify if the proposed architecture implementation has better results than related work.

ACKNOWLEDGMENT

This work was supported by CAPES and FAPITEC/SE

REFERENCES

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

[2] Q. M. Ashraf and M. H. Habaebi, "Autonomic schemes for threat mitigation in Internet of Things." Journal of Network and Computer Applications, vol. 49, 2015, pp. 112-127.

[3] L. D. Xu, W. He, and S. Li, "Internet of things in industries: A survey." Industrial Informatics, IEEE Transactions on, vol. 10, no. 4, 2014, pp. 2233-2243.

[4] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions." Future Generation Computer Systems, vol. 29, no. 7, 2013, pp. 1645-1660.

[5] R. Roman, J. Zhou, and J. Lopez, "On the features and challenges of security and privacy in distributed internet of things." Computer Networks, vol. 57, no. 10, 2013, pp. 2266-2279.

[6] M. C. Huebscher and J. A. McCann, "A survey of autonomic computing—degrees, models, and applications." ACM Computing Surveys (CSUR), vol. 40, no. 3, 2008, pp. 7.

[7] S. Raza, L. Wallgren, and T. Voigt, "SVELTE: Real-time intrusion detection in the Internet of Things." Ad hoc networks, vol. 11, no. 8, 2013, pp. 2661-2674.

[8] Y. S. Dai, M. Hinchey, M. Qi, and X. Zou, "Autonomic security and self-protection based on feature-recognition with virtual neurons." Dependable, Autonomic and Secure Computing, 2nd IEEE International Symposium on. IEEE, 2006.

[9] H. M. Salmon, et al.. "Intrusion detection system for wireless sensor networks using danger theory immune-inspired techniques." International journal of wireless information networks, vol. 20, no. 1, 2013, pp. 39-66.

[10] J. O. Kephart and D. M. Chess, "The vision of autonomic computing." Computer, vol. 36, no. 1, 2003, pp. 41-50.

[11] J. Greensmith, U. Aickelin, and S. Cayzer, "Introducint dendritic cells as a novel immune-inspired algorithm for anomaly detection." Artificial Immune Systems. Springer Berlin Heidelberg, 2005, pp. 153-167

[12] J. Bao, Y. Chen and J. Yu, "An optimized discrete neural network in embedded systems for road recognition." Engineering Applications of Artificial Intelligence, vol. 25, no. 4, 2012, pp. 775-782.

[13] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, June 2015.

[14] K. Q. Yan, S. C. Wang and C. W. Liu, "A hybrid intrusion detection system of cluster-based wireless sensor networks." Proceedings of the International MultiConference of Engineers and Computer Scientists, 2009, pp. 18-20.

[15] K. Faceli, A. C. Lorena, J. Gama and A. C. P. L. F. de Carvalho, "Aprendizado em Fluxos Contínuos de Dados." Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina, Grupo Gen-LTC, 2011, pp. 260-269.

[16] MEMSIC, "MICAz Datasheet". Available on: http://www.memsic.com/userfiles/files/Datasheets/WSN/micaz_datasheet-t.pdf, June 2015.

[17] P. Levis et al., "Tinyos: An operating system for sensor networks." Ambient intelligence. Springer Berlin Heidelberg, 2005.,pp. 115-148.

 ISBN: 978-1-61208-418-3

Object Location Estimation from a Single Flying Camera

Insu Kim and Kin Choong Yow
GIST College, Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
e-mail: ahinsutime@gist.ac.kr, kcyow@gist.ac.kr

***Abstract*—With the recent popularity and ubiquity of drones, there had been an increasing demand to deploy drones for the detection, localization and tracking of objects in a scene (e.g., pedestrians, cars, etc.). The problem with a single camera drone is that it is impossible to estimate distances from a single image. Although the drone can fly to another position to take a second image, the object that we are tracking may have moved during that time interval, rendering traditional stereo-vision algorithms useless. In this paper, we propose a novel system that instructs the drone to fly in a specific pattern so as to achieve a large baseline, and use three images (instead of the traditional two) to recover the distance to the object that is moving. The experimental results show that our algorithm can estimate depth with better or equal accuracy than other state-of-the-art methods. This algorithm would have great significance for small or low cost drones which are unable to carry additional devices (apart from the built-in camera), thus enhancing their ubiquity of use.**

Keywords - computer vision; drone; stereo-vision; distance extraction; object detection; location mapping.

I. INTRODUCTION

With the recent popularity and ubiquity of drones or Unmanned Aerial Vehicles (UAV), they have been increasingly deployed in various tasks such as object localization and tracking. Drones often carry a high-definition camera and it can be used for surveillance, expedition guidance, search and rescue, etc. However, with a single image of an object, it is usually impossible to obtain the distance of the object from the camera because we usually do not know the size of the object.

Traditionally, such a problem can be solved with stereo-vision, i.e., putting a second camera some distance away and computing the disparity of the object in the two images. However, existing stereo-vision algorithms require that the distance between the two cameras (i.e., the baseline) be fixed because the image disparity is a function of the object distance and the baseline.

Drones can be and have been fitted with two cameras for their mission. Carrillo et al. [1] showed the possibility of using stereo vision with inertial navigation system which can estimate the UAV's position accurately. They used two separated fixed camera on the drone. Knoppe [2] also proposed a system for a drone carrying a stereo camera to get ground surface scanning data. Schauwecker and Zell [3] introduced more sophisticated method for navigating Micro Aerial Vehicles (MAV) using four cameras. They performed stereo matching separately for the downward and forward of the MAV.

Certainly, the use of additional cameras could generate problems for a drone such as increased payload, insufficient power, reduced duration of flight, instability, etc. However, a more important problem is that the baseline between the two cameras is short relative to the distance of the object, resulting in very large errors in distance estimation. Since the drone can fly to any location with little or no restrictions, the obvious solution is to carry a single camera and fly as far as the environment allows (i.e., without losing sight of the object or crashing into an obstacle) so as to maximize the baseline. Traditional stereo-vision algorithms will still work, and it does not matter whether the two images are taken from two separate cameras, or from a single camera that had moved. This, however, works only if the target object remains stationary between the two views.

Hence, in this paper, we propose a novel algorithm that enables a drone carrying a single camera to produce an accurate estimate of the distance of a stationary or moving object. Our algorithm works by instructing the drone to fly in a specific pattern so as to achieve a large baseline, and use *three* images (instead of the traditional two) to recover the distance to the (moving) object.

Our proposed system is very effective for small or low cost drones (e.g., Parrot rolling spider [4]) which are unable to carry additional devices (apart from the built-in camera). This increases the ubiquity of using drones for object localization and tracking. Once the distance of the object from the drone is known (and angle, which can be obtained from a magnetic compass), we can map the object location to any global coordinate system (assuming that we know the drone's position e.g., through Global Positioning System (GPS)).

The rest of the paper is organized as follows: Section II discusses some related work, and Section III describes the proposed algorithm. Section IV discusses the implementation details and Section V provides the experimental results. Section VI concludes the paper.

II. RELATED WORK

Similar work has been done by Zhang and Liu [5]. They proposed a system where a drone estimates relative altitude from a ground object with (or without) movement if the size of the object is known. In our system, we do not make the assumption that we know the size of the object. This means that our algorithm is more generic and can be extended to locate and track any kind of objects. Sereewattana et al. [6]

 ISBN: 978-1-61208-418-3

did depth estimation of color markers for automatic landing control of UAV using stereo vision with a single camera. In their work, the color markers are static (i.e., not moving) and are close to the drone (below 3 meters). Kendall et al. [7] also proposed novel system for UAV which tracking an object of known color and size to get the depth information. For our case, our objects can be any size, moving and can be arbitrarily far (limited by the accuracy of the object detection algorithm).

A moving camera can also be compared to a Pan-Tilt-Zoom (PTZ) camera. Wan and Zhou [8] introduced a novel stereo rectification method for a dual-fixed-PTZ-camera system that can recover distance accurately. However, their PTZ camera pair is fixed so the object may disappear from the cameras' view. In comparison, our system can fly to different positions to avoid occlusion, as well as choosing the length of the baseline. Tran et al. [9] proposed a system that performed face detection using dual fixed PTZ cameras for large area security system. They showed good result for the detection rate (99.92%) with indoor detection range (5 to 20m). However, their work is concerned about object recognition rather than to estimate the distance of the object from the camera.

III. PROPOSED ALGORITHM

A. System Overview

Figure 1 shows an overview of our system. Our system consists of a drone mounted with a single camera at position (x_d, y_d). Using the algorithm described later in this section, the drone will compute the distance Z and angle θ_d of an object (e.g., pedestrian) relative to itself, and then use this distance to compute the position of the object (x, y) in global coordinates. We assume that the position of the drone is always known (e.g., through GPS, or WiFi positioning, or from Inertial Navigation System (INS), if available).

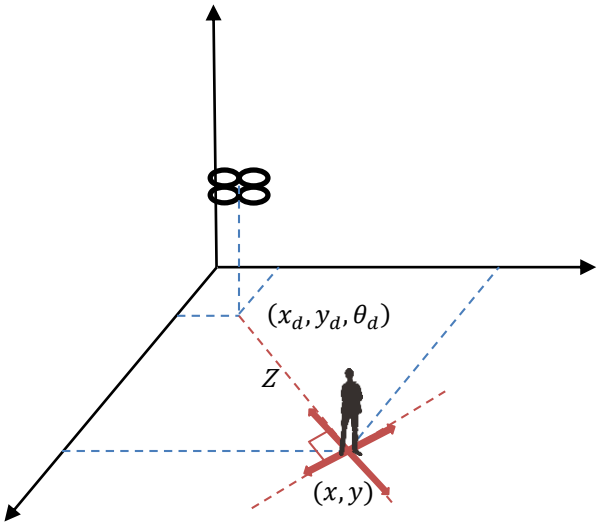


Figure 1. Overview of the localization system

After calculating the position of the object, the drone will compute a new position for it to fly to (if the object has moved) so that it will continue to have the object in its view. This process can be repeated as often as necessary. Figure 2 gives a flowchart of the system.

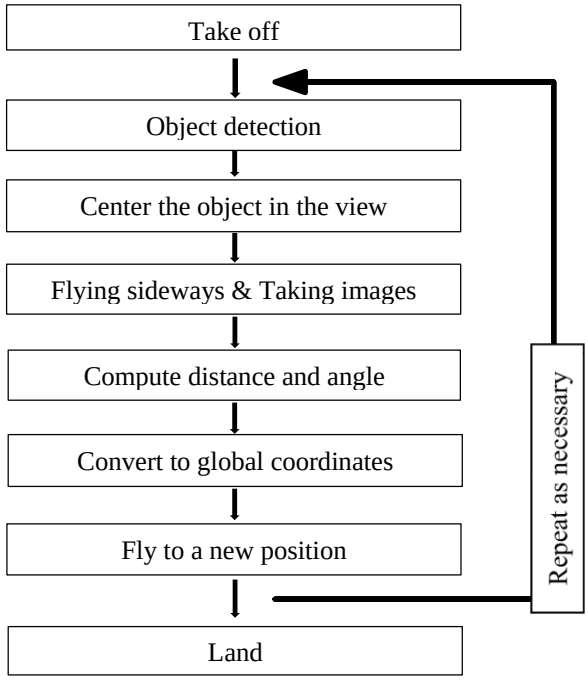


Figure 2. System flowchart

B. Object detection

Our system is not restricted to work on only one class of objects. It can be extended to work on any class of objects, but we need to have a reliable object detection algorithm for it. In this paper, we demonstrate our system on pedestrians, and we make use of the Histogram of Oriented Gradients (HOG) pedestrian descriptor from the OpenCV 2.4.10 library ("peopledetect.cpp"). Figure 3 shows an example of the pedestrian detection algorithm in OpenCV, which displays a bounding box over the detected pedestrian.

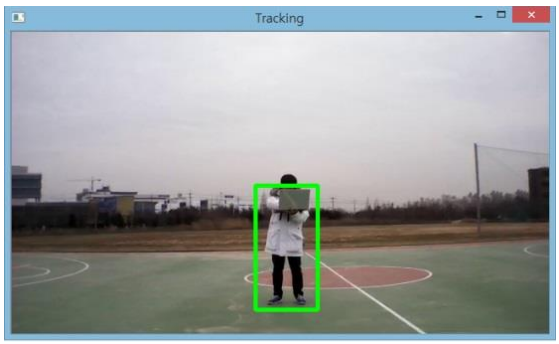


Figure 3. Example of pedestrian detection from the drone

 ISBN: 978-1-61208-418-3

The pedestrian detector from OpenCV can detect more than one person in a scene. Like with any other tracking algorithm, we need to make use of additional information (e.g., color, size, motion vectors, etc.) to correctly match the object in different scenes. However, our experiment shows that if the pedestrian is further than 11m from the drone's camera, the false detection rate becomes significantly increased. Thus in our experiments, we restrict our study to detect pedestrians at depths below 11 meters. The design of a more accurate pedestrian detector is not in the scope of our work.

C. Camera Calibration

In order to obtain an accurate estimation of the depth of an object, the camera needs to be calibrated. The basic concept of the classical structure-from-motion algorithm is from the geometry shown in Figure 4.

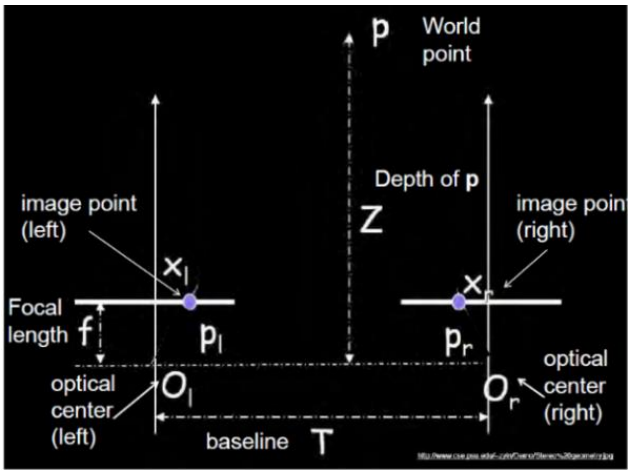


Figure 4. Classical disparity formula diagram

From Figure 4, using similar triangles, the following relation for depth, Z, can be derived easily:

$$Z = f\frac{T}{x_r - x_l} \quad (1)$$

where f is the focal length, T is the baseline, and $x_r - x_l$ is the image disparity. The camera needs to be calibrated to find f in order to be able to obtain Z based on the image disparity.

However, in our formulation (discussed in the next section), we need another important parameter which maps the image offset p (position of the object in the image from the image center) to the angle θ subtended by the object from the image center (see Figure 5). As the image offset p is dependent entirely on the focal length, we choose to calibrate for θ/p instead of the focal length f.

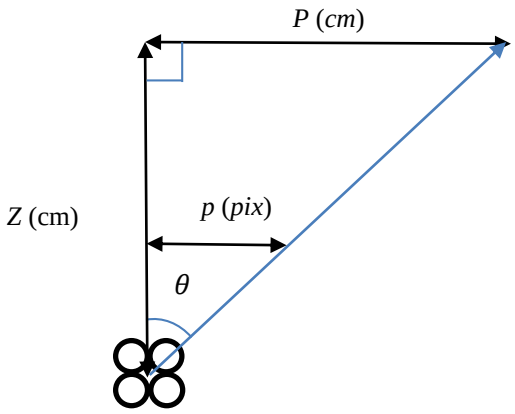


Figure 5. Geometry of view angle calibration

We can obtain this value by placing an object at various positions in the drone's field of view (Figure 6) and measuring the image offset p and the angle subtended θ.

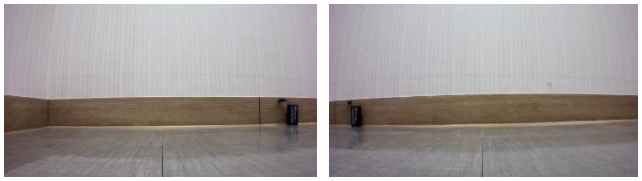


Figure 6. Images of an object taken at various positions

Table I shows the data measured from Figure 6. While the θ/p ratio is not constant throughout the entire image, the variation was found to be quite small (0.001°/pixel) between the two cases where the object is in the image center and at the image edge.

TABLE I. CALIBRATION OF THE DRONE'S CAMERA

Z (cm)	*P (cm)*	*arctan(Z/P) (°)*	*p (pixels)*	*θ/p (°/pixels)*
92	*57*	*31.78°*	*320*	*0.099*

At the image edge of 320 pixels (for a 640x480 pixel image), the θ/p ratio is found to be 0.099°/pixels.

D. Baseline calibration

Another important parameter to calibrate is the distance between the drone positions at each successive image capture. This represents the baseline between the images. To obtain these distances, the drone is instructed to take a sequence of images at constant rate while flying with a preset speed to the left (or right). To reduce the error in estimating depth, the length of the baseline should be as long as possible.

In our calibration experiments, we allow the drone to takeoff and hover for a few seconds until it has stabilized, and then we send an instruction to the drone to fly to the left (we choose 'left' for the ease of discussion) at the maximum speed (i.e., roll angle phi = -1.0, normalized), while taking images at 500ms interval for a total of 11 images. These

 ISBN: 978-1-61208-418-3

values were chosen after numerous tries to give the best tradeoff between distance flown and the time taken to complete, as the object (pedestrian) may have moved during that time interval.

To measure the length of the baselines, we placed numerous markers on a wall, and made the drone fly parallel to the wall, taking images as mentioned above. From the images, we can determine the positions of the drone where the images were taken. We observe that the first image (i.e., at position 0) was exactly the same as the image at hovering (i.e., it was taken before the drone has even started to rotate (roll)), and the second image (i.e., at position 1) was taken just after the drone has completed its rotation (roll) to the left (but before any horizontal translation takes place – so no change in its position). The rest of the images (at positions 2 to 10) were of increasing distances between each other, which is due to the inertial and the acceleration of the drone, which needed some time to accelerate to the desired speed.

We repeated the experiment three times, and averaged the results of the four experiments to produce the baselines shown in Table II.

TABLE II. ELINE CALIBRATION

position	*Average distance from previous position (cm)*
0	*- (Hover)*
1	*0 (Rotate)*
2	*47.50*
3	*51.75*
4	*75.75*
5	*91.25*
6	*100.5*
7	*108.75*
8	*126.75*
9	*146.50*
10	*189.25*

The baseline between any two images is simply the sum of the values between them. For example, the baseline between image 0 and image 5 is 47.50+51.75+75.75+91.25 = 266.25(cm).

E. Depth estimation

We now present our formulation for estimating the depth of an object from the drone's camera images and its position. We will divide our discussion into 3 parts (1) stationary object, (2) moving object in a direction parallel to the drone's flight, and (3) moving object in a direction perpendicular to the drone's flight. For a moving object in an arbitrary direction, a similar analysis has been performed but the result is not shown here due to the lack of space.

1) Stationary Object

If the object is stationary, then our problem reduces itself to the classical case of stereo-vision. Figure 7 illustrates the geometry needed for the computation of depth *Z* in this case.

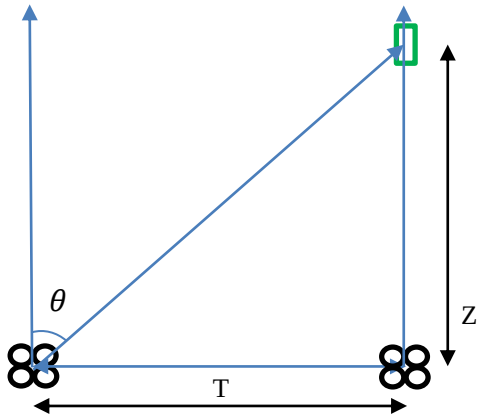


Figure 7. Classical stereo vision method for the stationary pedestrian

We will first instruct the drone to takeoff, and then hover for a few seconds for it to stabilize itself. Then, we will call the object detector function to find any objects within its view. For the ease of discussion, let us use the example of pedestrian detection. After the drone is stabilized, we call the HOG pedestrian detector from the OpenCV library. If more than one pedestrian is found, we need to choose which pedestrian is the one that we are trying to localize. After we have found the pedestrian, we rotate the drone (yaw) so that the pedestrian (i.e., the centroid of the bounding box) is in the center of the image (see Figure 2 for an overview of the system).

The next step is the most important step of the algorithm. The drone is instructed to fly left and take 11 images using the same parameters as in the baseline calibration. After the images are taken, the drone will return to its hovering state and then examine whether the pedestrian is detected in each image. For example, if the pedestrian disappears at image 6, then the drone will use image 0 and image 5 to calculate the baseline (otherwise use the last image so as to obtain the largest baseline). From the image offset *p* of the pedestrian in image 5, we can also use our calibrated θ/p to find the angle θ at image 5.

From Figure 7 we can obtain the equation:

$$Z = \frac{T}{\tan\theta} \qquad (2)$$

from which we can compute the depth *Z*.

2) Object moving parallel to drone's flight

Here, we assume that the object (pedestrian) is moving at a constant speed in a straight line. If he is not, we can approximate the pedestrian movement as piecewise linear. The loop in Figure 2 needs to be repeated for more accurate localization results. Figure 8 illustrates the geometry needed for the computation of depth *Z* in this case.

 ISBN: 978-1-61208-418-3

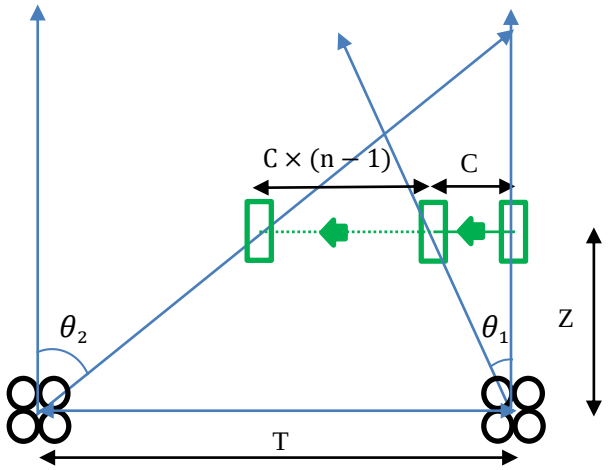


Figure 8. Geometry for pedestrian moving parallel to the baseline

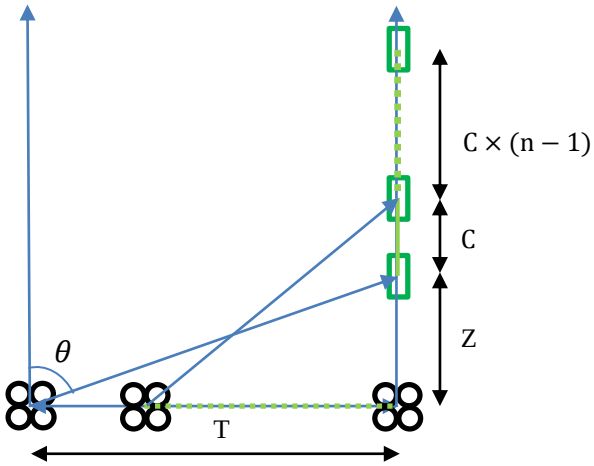


Figure 9. Geometry for pedestrian moving perpendicular to the baseline

The key idea here is that we now need **three** images, at position 0, 1 and *n* (the last image where the pedestrian can still be seen) to compute the depth *Z*. Due to the fact that at position 1 the drone has just completed its rotation (roll) to the left, the drone has not moved yet but the pedestrian had moved by a distance C. This gives us the very important image offset p_1 that allows us to calculate θ_1. At the *n*th position, at angle θ_2, we know that the pedestrian has moved by an additional distance of $C \times (n-1)$. So from the geometry, we can generate the following relation:

$$Z\tan\theta_2 = T - C \times n \quad (3)$$

$$C = Z\tan\theta_1 \quad (4)$$

Eliminating 'C', we obtain the following equation.

$$Z = \frac{T}{(tan\theta_2)+n(tan\theta_1)} \quad (5)$$

from which we obtain the depth Z.

If the pedestrian moves to the right instead, equation (5) is still valid. In that case, the pedestrian may disappear from the image much sooner. To overcome this, we can simply make the drone fly to the right for a second computation.

3) Object moving perpendicular to drone's flight

In the case of a pedestrian moving perpendicular to the baseline of the drone, the computation is very simple. Figure 9 illustrates the geometry needed for the computation of depth Z in this case. Assuming that the pedestrian starts from a distance of $C \times n + Z$ from position 0, after n images, he would be at a distance Z away from position 0. This is the same as the stationary case and we simply need equation (2) to obtain the depth *Z*.

F. Transformation to global coordinates

After we have obtained the depth *Z* and angle θ, we can use them together with the drone position to calculate the global coordinates of the pedestrian. Figure 10 shows the geometry for calculating the coordinate transformation for each of the 3 cases.

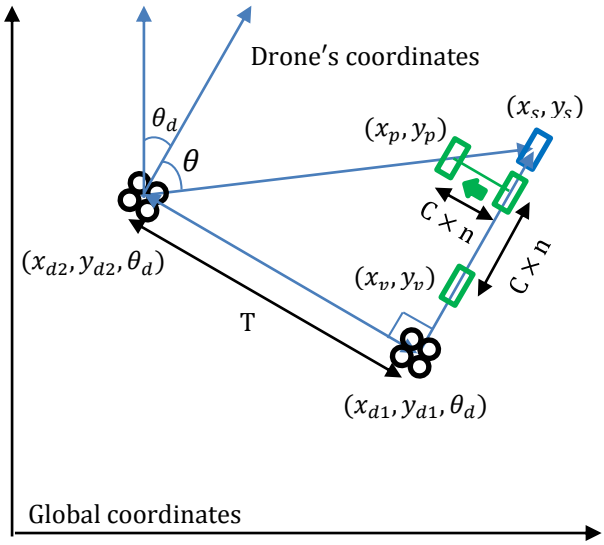


Figure 10. Coordinate transformation

In Figure 10, $(x_{d2}, y_{d2}, \theta_d)$ *and* $(x_{d1}, y_{d1}, \theta_d)$ refers the positions of the drone (which is assumed to be known). $(x_s, y_s), (x_p, y_p), (x_v, y_v)$ are the positions of the pedestrian in each of the three cases of stationary, moving parallel, or moving perpendicular (vertical) to the baseline, respectively. Note that the initial coordinate of the drone $(x_{d1}, y_{d1}, \theta_d)$ is vital which is used as boundary condition for coordinate transformation. From this geometry, we can directly

 ISBN: 978-1-61208-418-3

calculate the position of a pedestrian's position using one of the three following equations.

$$(x_s, y_s) = (x_{d1} + \frac{\tan\theta}{T}\sin\theta_d\ , y_{d1} + \frac{\tan\theta}{T}\cos\theta_d) \quad (6)$$

$$(x_p, y_p) = (x_{d1} + (T - \frac{T(tan\theta)}{(tan\theta_2)+n(tan\theta_1)})\cos\theta_d,$$
$$y_{d1} + (T - \frac{T(tan\theta)}{(tan\theta_2)+n(tan\theta_1)})\sin\theta_d) \quad (7)$$

$$(x_v, y_v) = (x_{d1} + (\frac{\tan\theta}{T} - \mathrm{C} \times \mathrm{n})\sin\theta_d,$$
$$y_{d1} + (\frac{\tan\theta}{T} - \mathrm{C} \times \mathrm{n})\cos\theta_d) \quad (8)$$

IV. IMPLEMENTATION

A. Hardware

The drone used in this study is the Parrot AR.Drone2 GPS edition [10]. It has a forward-looking 720p HD camera and a vertical QVGA camera. The drone is controlled from an Intel i5 laptop running Windows 8.1 with 4GB of RAM.

Figure 11. Parrot AR.Drone2.0

B. Software

The software we used to control the drone is the "CV Drone" package which is available from Github [11]. The image processing routines were from the OpenCV 2.4.10 library, and the entire system was developed in Microsoft Visual Studio 2013.

V. EXPERMENTAL RESULTS

A. Experiments

To evaluate our proposed algorithm, we conduct experiments for each of the three cases of stationary, moving parallel and moving perpendicular pedestrians. For each case, we perform the experiment four times with the pedestrian at 2.75, 5.5, 7.25, and 9 meters from the drone. As the HOG pedestrian detection from the OpenCV library did not work well from 11 meters and beyond, we stopped the experiment at 9 meters.

For the moving cases, the pedestrian was asked to move in the required direction at a speed of 0.6 m/s, covering a distance of about 3m in the 11 images that was captured. Examples of the images captured by the drone (at 5.5m, stationary) are shown in Figure 12. Notice that between the first and second image, there is only a rotation of the drone and there is no horizontal movement.

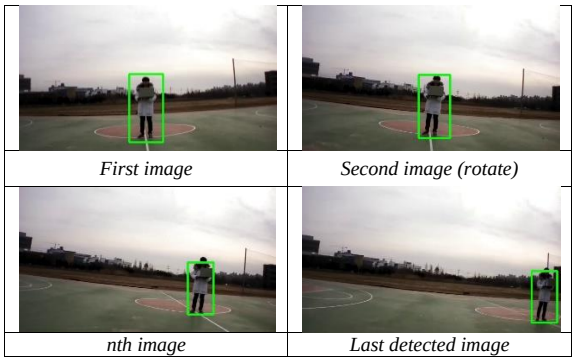


Figure 12. Examples of Image Sequence and pedestrian detection (5.5m, stationary pedestrian)

For each of the experiments, we compute depth error

$$\text{Depth error } (\%) = |Z - Z'|/Z \quad (9)$$

where Z refers to the actual depth while Z' indicates measured depth. Also, since we compute the pedestrian position in global coordinates, we also compute the position error using the following equation:

$$\text{Position error } (\%) = (\sqrt{(x - x')^2 + (y - y')^2})/Z \quad (10)$$

Here, (x, y) refers the actual position of a pedestrian while $(x',\ y')$ indicates the measured position. The results are shown in Tables III to V.

TABLE III. STATIONARY PEDESTRIAN

Actual depth, Z(m)	*Measured depth, Z'(m)*	*Depth error rate (%)*	*Position error rate (%)*
2.75	*2.35*	*14.5*	*16.5*
5.50	*6.01*	*9.3*	*18.0*
7.25	*7.51*	*3.6*	*12.2*
9.00	*8.96*	*0.4*	*7.2*

TABLE IV. PERPENDICULARLY MOVING PEDESTRIAN

Actual depth, Z(m)	*Measured depth, Z'(m)*	*Depth error rate (%)*	*Position error rate (%)*
4.50	*4.08*	*9.3*	*9.4*
7.60	*6.69*	*12.0*	*12.9*
5.45	*5.33*	*2.2*	*5.86*
6.90	*6.75*	*2.2*	*2.2*

TABLE V. PARALLELLY MOVING PEDESTRIANS

Actual depth, Z(m)	*Measured depth, Z'(m)*	*Depth error rate (%)*	*Position error rate (%)*
2.75	*3.43*	*24.7*	*27.7*
5.50	*6.20*	*12.3*	*12.7*
7.25	*8.28*	*14.2*	*22.8*
9.00	*9.82*	*9.1*	*11.7*

 ISBN: 978-1-61208-418-3

We can see from the tables that with the exception for the parallel case at the nearest distance (i.e., 2.75), all the experiments yield good results at < 15% depth error rate.

B. Comparison with other techniques

In this section, we compare our algorithm with the classical stereo-vision technique of computing image disparity and then calculating the distance using equation (1). Table VI shows the results.

TABLE VI. COMPARISON WITH THE CLASSICAL METHOD (STATIONARY PEDESTRIAN)

	Classical method		*Our method*	
Actual depth, Z(m)	***Measured depth, Z'(m)***	***Depth error rate (%)***	***Measured depth, Z'(m)***	***Depth error rate (%)***
2.75	*2.45*	*11.0*	*2.35*	*14.5*
5.50	*7.18*	*31.4*	*6.01*	*9.3*
7.25	*8.79*	*21.2*	*7.51*	*3.6*
9.00	*10.01*	*11.2*	*8.96*	*0.4*

While our method performs worse at short depth (2.75m), it actually works better than the classical method for the larger depths. We further compare our method with the method proposed by Sereewattana et al. [6], which is only evaluated for stationary objects less than 3m in depth. Table VII shows the results.

TABLE VII. COMPARISON WITH OTHER SYSTEMS (STATIONARY CASE)

System	***Accuracy (%)***
M. Sereewattana et al. [6] (Only for stationary object below 3m)	*3.9 ~ 12.4*
Classical stereo vision depth extraction	*11.0 ~ 31.4*
Our system	*0.4 ~ 14.5*

The results showed that our algorithm can estimate depth with better or equal accuracy than other state-of-the-art methods.

VI. CONCLUSION AND FUTURE WORKS

This paper proposed a novel system to estimate the location of an object from a single moving camera mounted on a drone. The proposed algorithm instructs the drone to fly in a specific pattern, which allows us to estimate the baselines between images so as to obtain depth. The algorithm is not restricted to any particular class of objects and can be easily extended to any class of objects. In addition, our formulation makes the novel use of **three** images, which allows us to extract depth even when the object is moving (with the assumption of constant speed and in a straight line). Experiments showed that our algorithm can estimate depth with better or equal accuracy than other state-of-the-art methods.

In this paper, we only reported the analysis and results of a pedestrian moving either parallel or perpendicular to the drone's flight, due to the lack of space. We already have the analysis of a pedestrian moving in an arbitrary direction, and our future work will be directed to complete the experiments for it. In addition, we will expand the capability of the system to cope with non-linear and non-constant pedestrian motion, and also occlusion (e.g., the pedestrian had turned round the corner of a building).

REFERENCES

[1] L. R. G. Carrillo, A. E. D. Lopez, R. Lozno, and C. pegard, "Combining Stereo Vision and Inertial Navigation System for a Quad-Rotor UAV," Journal of Intelligent & Robotic Systems, vol. 65(1-4), 2011, pp. 373-387.

[2] K. Knoppe, "A Lightweight Digital Stereoscopic Camera System," Institute for Geoinformatics Unsiversity of Munster, Matriculation No. 341901, 2013, pp. 1-58.

[3] K. Schauwecker and A. Zell, "On-Board Dual-Stereo-Vision for the Navigation of an Autonomous MAV," Journal of Intelligent & Robotic Systems, 2014, pp. 1-16.

[4] Parrot Rolling Spider. [Online]. Retreived from: http://www.parrot.com/usa/products/rolling-spider/ on 2015. 6. 9.

[5] R. Zhang and H. H. T. Liu, "Vision-Based Relative Altitude Estimation of Small Unmanned Aerial Vehicles in Target Localization," American Control Conference, June. 2011, pp. 4622-4627, ISBN: 978-1-4577-0080-4

[6] M. Sereewattana, M. Ruchanurucks, and S. Siddhichai, "Depth Estimation of Markers for UAV Automatic Landing Control Using Stereo Vision with a Single Camera," ICICTES, 2014.

[7] A. G. Kendall, N. N. Salvapantula, and K. A. Stol, "On-Board Object Tracking Control of a Quadcopter with Monocular Vision," International Conference on Unmanned Aircraft Systems, 2014, pp. 404-411.

[8] D. Wan and J. Zhou, "Stereo vision using to PTZ cameras," Computer Vision and Image Understanding, vol. 112(2), 2008, pp. 184–194.

[9] D. X. Tran et al, "Dual PTZ Cameras Approach for Security Face Dectection," Communications and Electronics, July. 2014, pp. 478-483, ISBN: 978-1-4799-5049-2

[10] Parrot AR.Drone2.0: Technical specifications. [Online]. Retrieved from: http://ardrone2.parrot.com/ardrone-2/specifications/ on 2015. 6. 9.

[11] CV Drone: OpenCV + AR.Drone. [Online]. Retrieved from: https://github.com/puku0x/cvdrone/ on 2015. 6. 9.

 ISBN: 978-1-61208-418-3

Generating Arbitrary View of Vehicles for Human-assisted Automated Vehicle Recognition in Intelligent CCTV Systems

Youri Ku, Insu Kim and Kin Choong Yow
GIST College, Gwangju Institute of Science and Technology
Gwangju, Republic of Korea
e-mail: kyr2234@gist.ac.kr, ahinsutime@gist.ac.kr, kcyow@gist.ac.kr

***Abstract*—Intelligent closed-circuit televisions (CCTV) are CCTV systems that can perform Video Content Analysis (VCA). However, in the area of Automatic Vehicle Recognition, there is still no good algorithm to recognize a car based on its description. In this paper, we propose a novel algorithm that will take an image (or several) of a car, extract special markings (if any) from it, and then texture map it to a 3D model of the same car. With the texture mapped 3D model, we can rotate the car to any arbitrary view point, especially to the view of another CCTV, so that non-sophisticated image matching algorithms can match the image to the actual CCTV feed. We performed experiments on three cars with different body markings and the results show that we can achieve quite realistic images of the car at any arbitrary viewpoint. This system will have significant impact on the use of intelligent CCTVs.**

Keywords - 3D model reconstruct; drone; generate unseen view of object; image warping; texture map UVW; arbitrary view .

I. INTRODUCTION

CCTV cameras, or closed circuit television cameras, are undoubtedly one of the most pervasive devices used in security systems all over the world today. In fact, over the years, these devices did not just help the authorities to pursue criminals, they were also used to view and monitor traffic incidents, estimate crowd density, and even detect suspicious activities within private businesses and residences. It can even be mounted on an Unmanned Aerial Vehicle (UAV), e.g., the Aeryon Scout [1] and be instructed to fly to a different incident location.

It is an ill-informed opinion that all CCTV does is to provide the ability to review an event after it has already taken place or to 'spy' on passers-by. Modern technologies are able to analyze the video data captured, alongside with the use of triggers, to prompt security actions for certain events or situations. The term "Intelligent CCTV" is now loosely used to describe systems that have such Video Content Analysis (VCA) ability.

In a study by Ju and Yi [2], the global video surveillance market is a huge market picking up 9 billion dollars in 2010, and it is expected to achieve 11.3 billion in 2012 and 14.4 billion in 2015. Among them, the intelligent CCTV market aggregated 0.2 billion dollars in 2010, and projected to hit 0.3 billion in 2012 and 0.6 billion in 2015.

The intelligent CCTV is touted to change the way we interact or react to people. The Intelligent CCTV has several advanced capabilities, some of which includes:

- Human Face Recognition
- Car License Plate Recognition
- Point of Sale (POS) / Shrinkage detection
- Object Tracking
- Unattended Object Detection
- Traffic Monitoring, and
- Behavior Recognition

We observe that there is a severe lack of technologies in the area of Automatic Vehicle Recognition in Intelligent CCTV systems. The only technology that is used here seems to be the Car Licensed Plate recognition technology. While it is true that the license plate may be the only unique "feature" to identify a car, there may be many situations where it is not possible to obtain a clear view of the license plate of a car that is leaving/entering a town/city, or a car that is simply parked in a crowded car park.

Suppose you have a suspicious car that you want to track. How would you convey to the Police the information about the car? Apart from the license plate information, you will also give a description of the car (e.g., "A White Audi Q7 with stripes on the hood", etc.). The Police will search the neighborhood first for a white Audi Q7, and after finding one, they will proceed to check if the car has stripes on its hood and whether the license plate matches.

This would be exactly the same for an Intelligent CCTV system. If we want the Intelligent CCTV system to automatically search an area for a car that matches our description, we need to provide it with more information than just the license plate number.

However, to communicate with the computer that you want a car "with stripes on the hood" is a notoriously difficult thing to do. The best option is to have an image of the car so that the computer can perform image matching on the CCTV feed. Another problem is that existing image matching algorithms do not perform well if the view of the object differs too greatly between the two images, or the background is drastically different.

This leads us to the motivation for our proposed solution. We want to develop an Intelligent CCTV system that takes at least one image of the vehicle that we want to find (additional images may provide views of the vehicle not seen in the first image), and with a very small amount of help from a human operator (to identify 5 to 9 points on the image), the system will be able to generate an arbitrary view of the car from a pre-defined 3D model that matches the view and background of the CCTV camera. With this new view, a non-sophisticated image

 ISBN: 978-1-61208-418-3

matching algorithm will be able to find a successful match of the car with low false positive rates.

The rest of the paper is organized as follows: Section II discusses some related work, and Section III describes the proposed algorithm. Section IV shows our reconstruction results and Section V concludes the paper.

II. RELATED WORK

Intelligent CCTV surveillance systems make use of a variety of image and video processing technologies to exact the information they need for their tasks. Foresti et al. [3] identify moving vehicles in video data streams by subtracting the current frame from an estimate of the background scene, based on the idea that anything 'new' in the current frame must be the mobile vehicle(s). To cope with occlusion, they made use of a Kalman filter to provide better estimates of the place and time of the vehicle's emergence from the occluded zone. Greenhill et al. [4] proposed a method which significantly improved the accuracy of object tracking by utilizing knowledge about the monitored scene. Such scene knowledge includes the homography between the camera and ground planes and the occlusion landscape identifying the depth map associated with the static occlusions in the scene.

Lotufo et al. [5] proposed the ANPR (Automatic Number Plate Recognition) system that is based on Computer Vision. The system performs a detailed matching process between the extracted character features and the reference features (contained in a database) using a statistical nearest neighbor classifier. Saravi and Edirisinghe [6] present an approach to Vehicle Make & Model Recognition (VMMR) in CCTV video footage that uses CPD (coherent Point Drift) to remove skew of vehicles detected. They also proposed a LESH (Local Energy Shape Histogram) feature based approach for vehicle make and model recognition that uses temporal processing to improve reliability.

III. PROPOSED ALGORITHM AND IMPLEMENTATION

A. Problem Definition

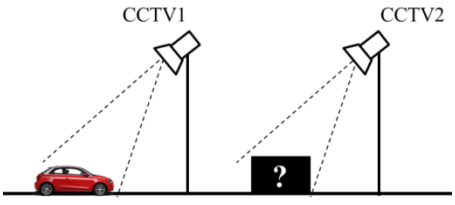


Figure 1. Problem Definition

Assume that we are living in a city that is equipped with Intelligent CCTV systems. In Figure 1, CCTV1 and CCTV2 are two arbitrary cameras located in the city. Each CCTV camera knows its own location and elevation as well as the inclination of the camera. A car is moving on the road and CCTV1 is tracking the car. However, since CCTV1 cannot move, the car will soon disappear from the field of view of CCTV1. Luckily, along the same road at some distance away there is another camera CCTV2, but the car is not yet in the field of view of CCTV2. Then how can CCTV1 communicate the car information to CCTV2 so that when the car comes into CCTV2's field of view, CCTV2 can continue to track it?

We identify that CCTV1 must provide at least one image of the car (with no restriction to its viewpoint) to CCTV2. Since CCTV2 knows its own position, elevation and camera inclination, it must generate a view of the car that matches its viewpoint of the road when the car begins to come into its field of view. Figures 2 and 3 illustrate two scenarios where the source image(s) can be obtained.

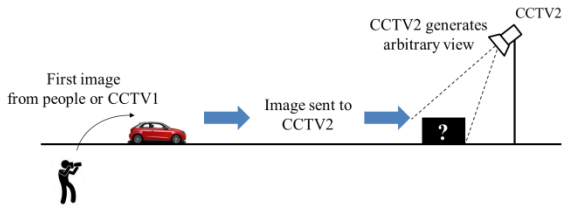


Figure 2. System with only one source image

In the first scenario, we have only one image of the car. The image may be generated by a person with a digital camera or mobile phone camera, or even by another intelligent CCTV camera (CCTV1). This image is then sent to CCTV2 where it will generate a new view of the car that matches its own viewpoint of the road.

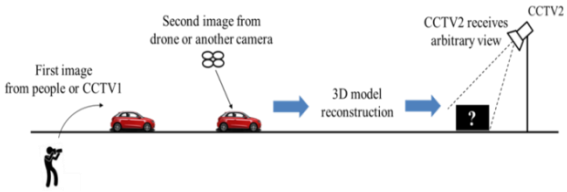


Figure 3. System with two or more source images

In the second scenario, we have two or more images. The first image may be generated by a person with a digital camera or mobile phone camera, or by another intelligent CCTV camera (CCTV1). The second image may be taken by a drone (i.e., UAV) or by another CCTV camera. If the drone is the source of the second image, it can offer not only the top view but also many views in different angles because it is a flying camera. These images are then sent to CCTV2 where it will generate a new view of the car that matches its own viewpoint of the road.

B. System overview

In this section, we describe our proposed system. We assume that the car to be tracked by CCTV2 has some special markings to differentiate it from another car of the same type (e.g., see Figure 5). If no special markings are available, then there may be many cars that look exactly the same and it may not be possible to correctly identify the right car to track. In addition, the intelligent CCTV camera (CCTV2) needs to receive a small amount of help from a human user (to input 5 – 9 corresponding points for texture mapping) as current Computer Vision correspondence algorithms are not robust enough yet to perform this task automatically. The overview of the system is shown in Figure 4.

 ISBN: 978-1-61208-418-3

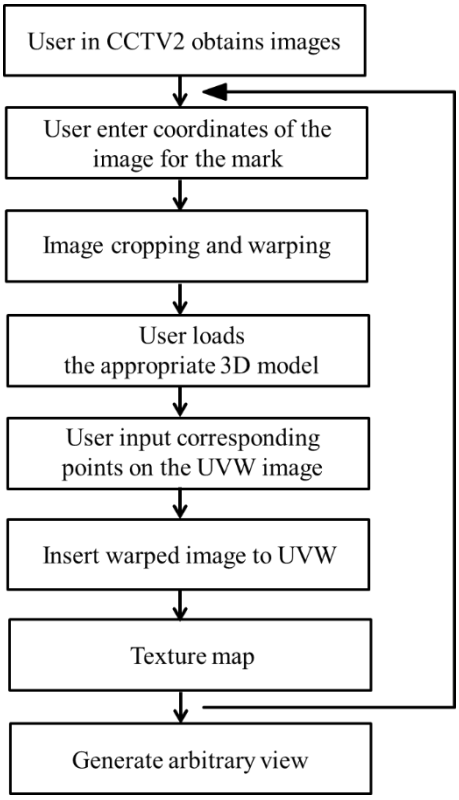


Figure 4. System overview

C. Special marking extraction

First, when the human user in CCTV2 receives the image(s), he needs to identify what make and model of the car it is. Although there are existing VMMR algorithms (e.g., Saravi and Edirisinghe [6]) for vehicle make and model recognition, they are not robust enough yet to perform this task automatically. Then, the human user will determine if there are any special markings on the car and on which panel they are (e.g., hood). If special markings exist, the human user will determine the coordinates of a 4-point polygon (representing a skewed rectangle) that covers the markings, and enters them into the system. The system will then automatically crop and warp the images into a rectangle.

Figure 5 shows an Audi Q7 that has stripes running down the hood. The four points in red (on the hood near to the stripes) in Figure 5 are the four points that the human user needs to enter into the system.

Figure 5. Audi Q7 with strips on the hood

Figure 6 shows an Opel Corsa that has yellow flame patterns on a blue plate. In this case, two images were available for the same car. If only one image was available (e.g., Figure 6(a)), then we could only extract the hood and driver-side door patterns and we will not know that there are also patterns on the trunk and the passenger-side door. This may lead to an incorrect 3D model reconstruction. To increase our chances for obtaining a correct 3D model, we can make certain assumptions (e.g., that the flame mark exists on the doors on both side of the car) and generate several candidate models for recognition.

(a) (b)

Figure 6. Opel Corsa with Flame patterns

If both images (Figure 6(a) and (b)) are available then we can completely reconstruct the 3D model of the Opel Corsa. The white circles in Figure 6 are the points of the polygon where the user has to identify and enter into the system.

(a) (b)

Figure 7. Opel Corsa with Green Stripes

Figure 7 shows another Opel Corsa that has a different pattern from Figure 6. This Opel Corsa has a black plate with a green stripe mark running from the hood to the trunk. In this case, there are also two images of the same car. The second image (Figure 7(b)) represents an image taken by a drone or from a CCTV camera that is mounted at the top of a building. If this image was not available, we will face the same problem as the Opel Corsa with Flame patterns. The white points in Figure 7(a) and (b) are the points of the polygon where the user has to identify and enter into the system.

D. Image warping

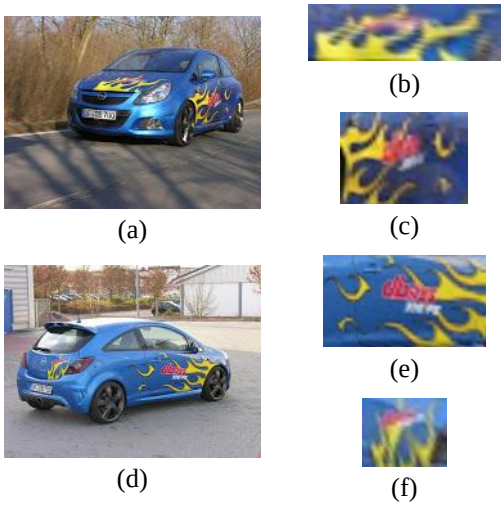
After the human user has identified the coordinates of the special markings, the system will perform image warping automatically using the OpenCV *warpPerspective* function and transform the polygon in the image into a rectangle. This is necessary for the next stage of texture mapping for the 3D model. Figures 8, 9 and 10 show the warped image of the pattern for the three cars in our experiments.

 ISBN: 978-1-61208-418-3

(a) (b)

Figure 8. Audi Q7 with warped image of pattern

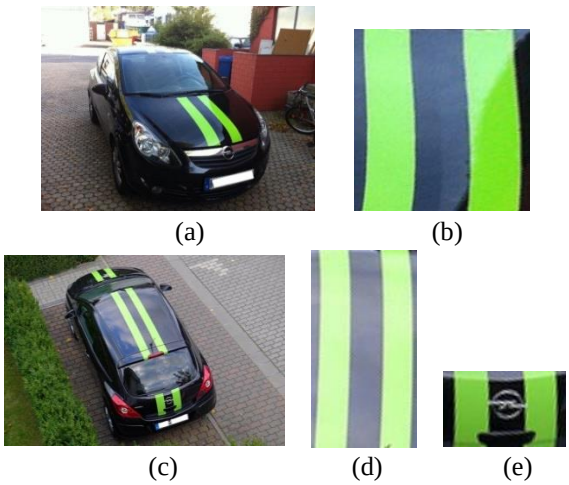
Figure 8(a) is original car image of Audi Q7 which is not modified at all. The Figure 8(b) is the warped image of extracted black stripe mark from Figure 8(a).



(a) (b) (c) (d) (e) (f)

Figure 9. Opel Corsa with warped image of flame pattern

Figures 9(a) and 9(d) are original car image of Opel Corsa with flame pattern. Figure 9(b) is the warped image of extracted flame mark on the front hood of the car from Figure 9(a). The Figure 9(c) is warped image of flame mark on the left door in Figure 9(a). The Figure 9(e) is warped image of mark extracted from right door of car in Figure 9(d), and Figure 9(f) is warped image of mark on car trunk in Figure 9(d).



(a) (b) (c) (d) (e)

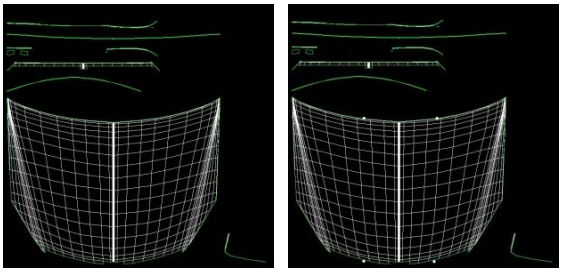
Figure 10. Opel Corsa with warped image of green stripes

Figures 10(a) and 10(c) are the raw images of car. Figure 10(b) shows the warped image of extracted mark which was on the front hood in Figure 10(a) originally. In the same way Figures 10(d) and 10(e) are warped image of extracted mark from Figure 10(c). Figure 10(d) is the mark which was on the top side of the car in Figure 10(c), and Figure 10(e) is the mark on the car trunk.

E. Texture Mapping

The next step is to apply the warped pattern images to the 3D model of the car (texture mapping). There are numerous websites that offer 3D car models for download (some are freeware and some are payware) and they come in a variety of formats (3DSmax, Maya, Wavefront, etc.). We obtain the 3D models for our three experiments from a website [7].

After we have downloaded the corresponding car model, we need to perform a UVW unwrap operation (from any 3D modeling software such as 3DSmax) to obtain the UVW map of the model. The UVW map allows us to perform texture mapping onto the 3D car model. An example of the UVW map of the Audi Q7 is showed in Figure 11(a).



(a) (b)

Figure 11. UVW map of the Audi Q7

To perform the texture mapping, we need to attach the warped image to the UVW map of the car. Here, the human user has to enter another 4 points on the UVW map to show where the warped image will go. Figure 11(b) shows the 4 points that the human user have entered.

After entering the 4 points, the warped image will be automatically resized and then “pasted” onto the UVW map. In addition, the human user needs to specify one more point in the image (that has the color of the body of the car) so that all the panels in the UVW can be filled with this color. This completes the texture mapping process on the UVW map and we are ready to reconstruct the 3D car. Figures 12, 13 and 14 show the completed UVW maps of all the three cars in our experiment.

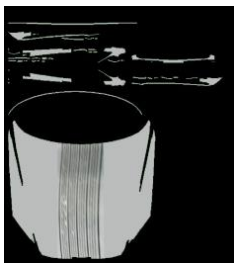
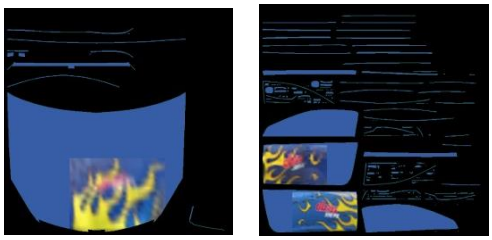


Figure 12. Completed UVW map of the Audi Q7

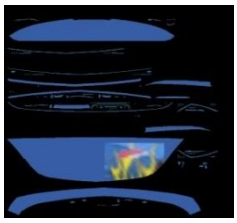
In Figure 12, the UVW map of the Audi Q7 is completed by pasting Figure 8(b) on the chosen part of raw UVW map which is Figure 11(b). Since the size of warped mark and chosen part of UVW map are different

 ISBN: 978-1-61208-418-3

each other, the algorithm performs resizing process before pasting.



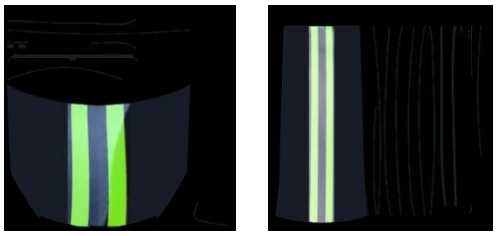
(a) (b)



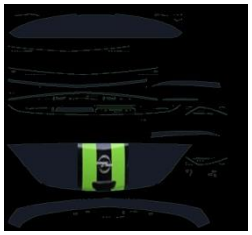
(c)

Figure 13. Completed UVW map of the Opel Corsa with Flame pattern

Figures 13(a), 13(b) and 13(c) are completed UVW map of Opel Corsa with flame pattern. Each of them is generated by attaching warped images of part D to the region of UVW which is chosen by user. Figure 9(b) is used for Figure 14(a), Figures 9(c) and 9(e) are used for Figure 14(b), and Figure 9(f) is used for Figure 13(c).



(a) (b)



(c)

Figure 14. Completed UVW map of the Opel Corsa with Green Stripes

Figures 14(a), 14(b) and 14(c) are the completed UVW map of the Opel Corsa with green stripes. The resized Figures 10(b), 10(d), and 10(e) are attached to corresponding position of UVW map of Opel Corsa.

F. Reconstruction of the 3D model in the Required View

We assume that the CCTV camera (CCTV2) position, elevation and inclination are already known. Let the CCTV camera position be (x, y, z). In addition, we also assume that the position of car in CCTV2's field of view is also known to be (x_c, y_c, z_c). Figure 15 shows the relationship between the two points.

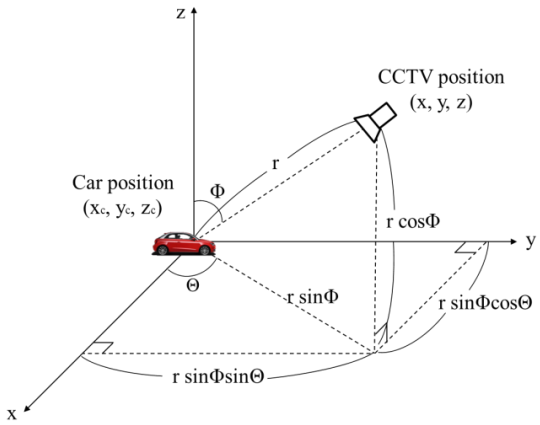


Figure 15. Construction of arbitrary views

With the information on the car position and the camera position, we can generate the expected view of the car from CCTV2 using the *gluLookat* function in the OpenGL library. By rotating the 3D reconstructed model, we can generate many arbitrary views of the car at different angles.

The final step is to merge the expected view of the car to the background image seen by CCTV2. This is a very simple process since the generated view of the car has a black background, so all we need to do is to replace the background image pixels by the non-zero car image pixels at the desired location.

IV. EXPERIMENTAL RESULTS

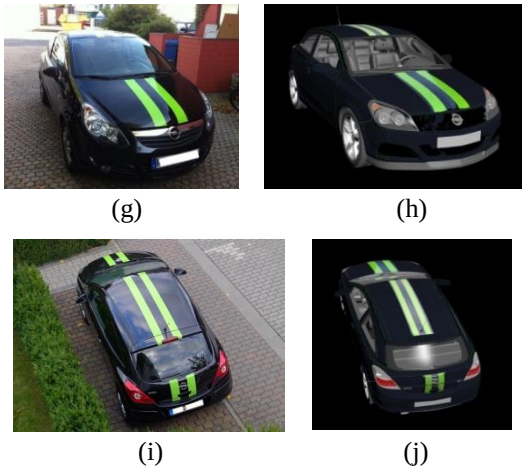
A. Result of reconstucting 3d model

Figure 16 shows the results of reconstructing the 3D model of the three cars to the same view as the original images so that we can verify that the result is correct.

(a) (b)

(c) (d)

(e) (f)



(g) (h)

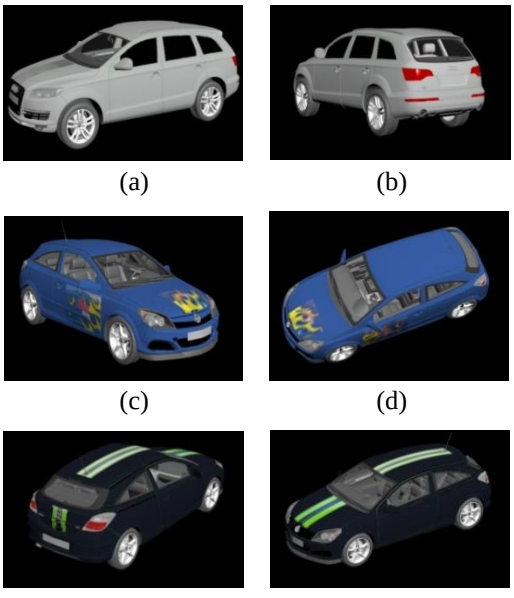
(i) (j)

Figure 16. Reconstruction results

Figures 16(a), 16(c), 16(e), 16(g) and 16(i) are the raw car images of Audi Q7 and Opel Corsa which are same with Figures 8(a), 9(d), 9(a), 10(a) and 10(c) each. Figures 16(b), 16(d), 16(f), 16(h) and 16(j) are the reconstructed 3D model of each car.

B. Result of generating new view

Figure 17 shows the results of generating new arbitrary views of the three cars.



(a) (b)

(c) (d)

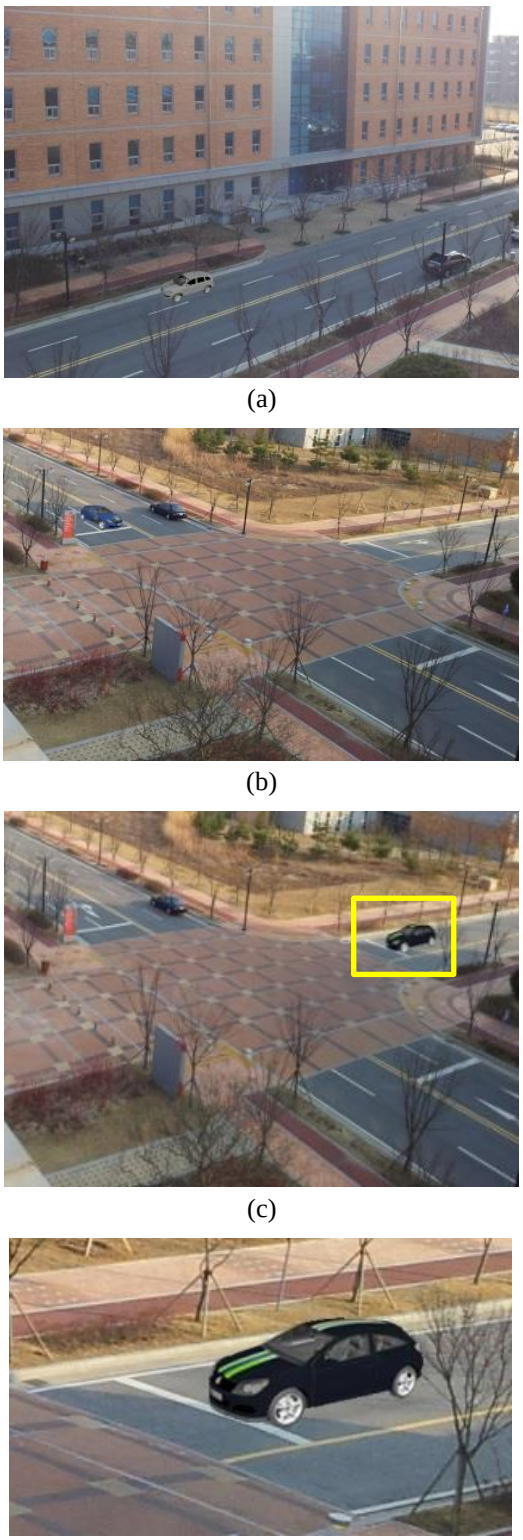
(e) (f)

Figure 17. New arbitrary views of the cars

Figures 17(a) and 17(b) are the arbitrary view of Audi Q7. Figures 17(c), 17(d), 17(e) and 17(f) are different view of Opel Corsa with flame pattern and green stripe. We can see that all the three cars in our experiment look realistic.

C. Merging with the CCTV background image

Figure 18 shows the result of merging the generated car image to the CCTV background image. The Audi Q7 is placed on the side of a road in Figure 18(a). Figures 18(b) and 18(c) show the Opel Corsa with flame patterns is placed at the top left corner of a junction, and the Opel Corsa Green Stripe is placed at the top right hand corner of the junction. Figure 18(d) is the zoom-in image of yellow box in Figure 18(c).



(a)

(b)

(c)

(d)

Figure 18. Merging with CCTV background image

We can see that the car images blend in very well to the CCTV background image and the resulting image looks quite realistic. We also showed a zoomed-in image of the Opel Corsa with green stripes to show that the

 ISBN: 978-1-61208-418-3

results look equally good when zoomed in.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new algorithm for generating an arbitrary view of a car from at least one image, taken from separate digital camera, mobile phone camera, or CCTV camera. Our system requires a little help from a human user (to identify the car's make and model, as well as to select and input 5 – 9 corresponding points for texture mapping). After the texture mapping is completed, we can generate any arbitrary view of the car, most importantly the view from the intelligent CCTV camera that will be tracking the car.

Although the resulting image looks realistic, this can be improved further by considering the position of the sun as well as the time of the day. At different times of the day, the position of the sun will cause different reflections as well as shadows on the car, causing the image to look different from what it was. This will be the focus of our future work in this project.

REFERENCES

[1] Aeryon Labs Inc. [online], available at http://www.aeryon.com/products/avs/aeryon-scout.html. Retrieved Mar 21, 2015.

[2] Y. W. Ju and S. J. Yi, " Implementing Database Methods for Increasing the Performance of Intelligent CCTV", International Journal of Security and Its Applications Vol.7, No.5 (2013), pp.113-120. http://dx.doi.org/10.14257/ijsia.2013.7.5.09.

[3] G. Foresti, C. Micheloni, L. Snidaro, P. Ramagnino, and T. Ellis, "Active Video-based Surveillance System," IEEE Signal Processing Magazine, March 2005, pp. 25-37.

[4] D. Greenhill, J. Renno, J. Orwell, and G. A. Jones, "Learning the Semantic Landscape: Embedding scene knowledge in object tracking," Realtime Imaging, Special Issue on Video Object Processing, Volume 11, Issue 3, June 2005, Pages 186–203. doi:10.1016/j.rti.2004.12.002

[5] R. A. Lotufo, A. D. Morgan, and A. S. Johnson, "Automatic number-plate recognition", IEE Colloquium on Image Analysis for Transport Applications, Feb 1990, pp.1-6.

[6] S. Savari and E. A. Edirisinghe, "Vehicle Make and Model Recognition in CCTV footage", 18th International Conference on Digital Signal Processing (DSP), 2013, July 2013, pp. 1-6.

[7] Crazy 3D Free.com, Retrieved from http://www.crazy3dfree.com on 2015.6.7.

 ISBN: 978-1-61208-418-3

Towards a Reliable and Personalized Disaster Warning System

Sungmin Hwang, Hiep Tuan Nguyen Tri, Kyungbaek Kim
Department of Electronics and Computer Engineering
Chonnam National University
Gwangju, South Korea
e-mail: hsmvirus@gmail.com, tuanhiep1232@gmail.com, kyungbaekkim@jnu.ac.kr

***Abstract*—The main goal of a disaster warning system is preventing loss of properties by providing suitable information to people who can be affected by disaster events. The challenges of a disaster warning system are geographically correlated characteristics of disaster events and user movements, multimodal communication channels of social connectivity, and unexpected large scale failures caused by disasters. To deal with the challenges, in this paper we propose a new design of a reliable and personalized disaster warning system to support city-wide mobile users. The proposed disaster warning system has three main components, namely, personality-aware location prediction, geo-social connectivity aware message generation, and failure-resilient geo-aware dissemination. Through the preliminary evaluation of each component, we show that the viability of the proposed disaster warning system.**

Keywords-Disaster Warning System; Mobile Users; Geo-Social Information; Geographical Failure; Personality; Location Prediction

I. INTRODUCTION

The main goal of a disaster warning system is preventing loss of properties by providing suitable information to people who can be affected by disaster events [2][7]. To achieve this goal, the system should analyze the impact of disaster events in detail, pick up all relevant people to the disaster, prepare proper messages and send the messages reliably in timely manner.

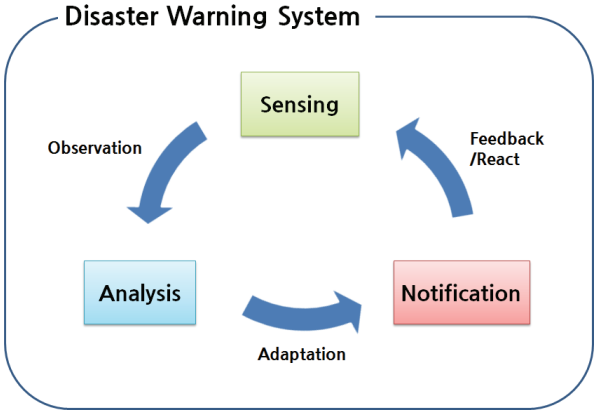


Figure 1. A conceptual view of a Disaster Warning System

Figure 1 illustrates a conceptual view of a disaster warning system. There are three basic components: sensing, analysis and notification. The sensing component continuously gathers information related to events, as well as users. The information related to events can be gathered by reading various sensors which are deployed on interesting locations [1][2] or getting the reports of events provided by local or government organizations. The information related users can be gathered through an interactive disaster portal or disaster warning notification applications. The interesting locations of users keep changing along with their current locations and the change of social connectivity. For example, if a user makes a new friend, the location of the new friend may be his new interesting location. Also, the information related to disaster event needs to be managed continuously. For example, the utility of hospitals, the traffic situation and failures of utility services need to be observed seamlessly in prior an event, and the system uses the information at the moment of the event.

The continuous observation, which is gathered by the sensing component, is used as an input of the analysis component. The role of analysis component is sorting out the most relevant users to the corresponding event and generating proper messages for each user. Because a disaster event is usually tightly coupled to geography, the location information of users and the interested location information of users are essentially required at this component [7]. If the system can predict the further location of a user based on the continuous observation, the analysis component also uses this location prediction for sorting out the relevant users. Beside the location information, the social connectivity is also very important for finding relevant users. For example, parents of elementary students desire to get the messages whenever their children can be affected by a disaster event.

Another role of analysis component is proper message generation. One of eventual goals of a disaster warning system is maximizing the coverage of warning message propagation. To achieve this, the warning system should pick better seeds of message propagation. Though the disaster warning system only uses the internet connection or SMS (Short Message Service), the message channel between end users are very diverse from phone call to human contact [6]. This multi-modal channel characteristic should be considered during selecting message seeds. In the aspect of effectiveness of the message, the message should have rich information for each user. By using the continuous observation, the system generates the proper message for

 ISBN: 978-1-61208-418-3

each group of user to react in proper way. For example, in the case of a building fire event, if the warning system observes damages of the building and generates messages which hold the clear evacuation plan, the users can immediately react to the disaster without any hesitation.

The output of analysis component is the list of relevant users and the proper messages for each of them. The notification component uses this output as an input, and disseminates the messages to the target users. At the time of the disaster event, some part of the network may be broken or end user host may not be reachable [8]. That is, unexpected failures happen in large scale, as well as in geography correlated manner. The goal of the notification component is how to keep the network more resilient to the unexpected geo-correlated large scale network failures. The action of notification component is considered as feedback and react of users.

In this paper, we propose a new design of a disaster warning system architecture which is mainly focus on considering personalized message generation (Analysis) and geo-correlated failure resilient message dissemination (Notification). The architecture of the proposed reliable and personalized disaster warning system has three main modules; personality aware location prediction module, geo-social connectivity aware message generation module, and failure-resilient geo-aware dissemination module. The personality aware location prediction module filters out targeted users who may be affected by the event, even though the users are not interesting in the event location. The geo-social connectivity aware message generation module generates messages for each set of users in order to provide personalized information of events to each individual user, as well as to maximize the coverage of message propagation through multi-modal communication channels on social connections. The failure-resilient geo-aware dissemination module manages the connectivity to users such as an overlay network of warning notification clients and supports the dissemination protocols which are highly resilient to unexpected geo-correlated severe failures caused by disaster events.

The rest of paper is organized as follows. In Section 2, the main idea of each module of the proposed system is described in detail. Also, the results of preliminary evaluations for each module are presented with the discussion of the viability of the proposed modules and system. Finally, Section 3 provides the conclusion of this paper.

II. DESIGN OF A RELIABLE AND PERSONALIZED WARNING SYSTEM

In this section, we describe each module of the proposed disaster warning system in detail. Figure 2 illustrates the overall architecture of the proposed disaster warning system. The observation, the user interface module and the database module are related to sensing component of a disaster warning system. The personality aware location prediction module and the geo-social connectivity aware message generation module are related to the analysis component of a disaster warning system. Finally, the failure-resilient geo-aware dissemination module is related to the dissemination component of a disaster warning system.

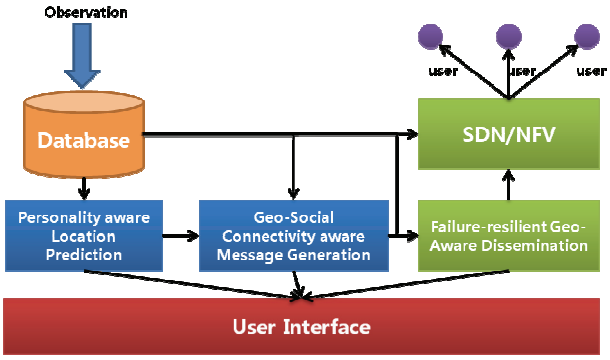


Figure 2. The architecture of the proposed disaster warnding system

A. Observation, Database and User Interface

The proposed disaster warning system manages interesting locations and social connectivity of users, as well as geography and other environmental information related to disaster events. To manage user information such as location subscriptions and social connectivity, a disaster warning system needs an interactive user interface. Recently we implemented a web-based interactive user interface for gathering and managing user information, as well as event information [3]. Beside the user information, other environmental information, such as geography and sensor readings can be gathered by using any IoT(Internet of Things)-based sensor network techniques [1][2] and using open API (Application Programming Interface) to other disaster portals such as USGS (United State Geological Survey). The gathered information is archived into the database module for further uses such as generating warning messages for target users.

B. Personality-aware Location Prediction

In the disaster warning system, the subscriptions of users are basically event type and interesting location, because a disaster event is highly coupled with geography. Also, the current location of a user and some locations where the user most likely visits are natural interesting locations of the user. To get the current location of a user, the system can gather the location of a user periodically by contacting warning notification clients which are running on machines of users.

With these user subscriptions and interesting locations, the disaster warning system filters out the relevant users to a disaster event. This process can be easily done by conducting a series of database queries. But, there is room for improvement of the filtering process. That is, even though the user does not have any subscription related to the event, the system can predict the relevant users to an event and improve the utility of the warning system. This prediction of interesting location of a user can be done by applying various machine learning techniques to user location history [4].

 ISBN: 978-1-61208-418-3

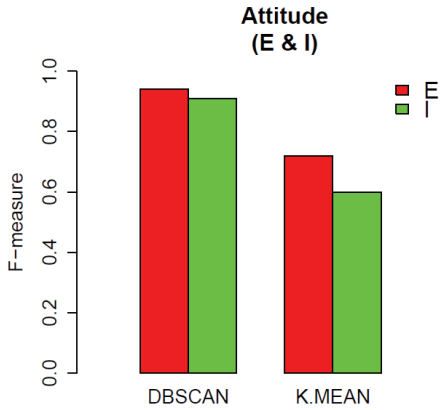


Figure 3. Performance of predicting user attitute (E: Extraverstion and I: Introversion) in MBTI (Myer-Briggs Type Indicator). SVM (Support Vector Machine) is used as a classifier and each location clustering technique is used to extract classification features.

As a novel feature of prediction of interesting locations of a user, we considered the personality of a user and its relationship to the movement pattern of the user. Recently, we researched on the relationship between user personality and their movement patterns, and we found the possibility of estimating user behavior based on the location clustering [5]. Figure 3 shows the performance of predicting user attitude personality of MBTI (Myer-Briggs Type Indicator) with different location clustering methods such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and K-means clustering. With DBSCAN, we achieved over 90% of F-measure which is the harmonic mean of precision and recall of prediction.

As results, we found that a user with outgoing personality travels more distance and visits more places. With this finding, the personality aware location prediction module leverages the personality information to improve the effectiveness of location prediction based on user location histories.

C. *Geo-Social Connectivity-aware Message Generation*

A warning under a disaster event requires not only sending a message to users in the event area but also letting people close to the users informed. In an ideal case, users of the warning system immediately get a message and try to evacuate from the disaster area. But, in real cases, users may not be able to check the message, or may not be able to evacuate within time limit. For example, let us imagine that a building is on fire and fire alarms start to ring. But there may be some disabled people, old people who cannot react fast, or young children who may not react properly. In this case, it is better to let others, who are physically close to them or who are family members of those improperly reacting users in the disaster area, and let them know about their detail condition. Then, informed people can help or let someone nearby help the users.

In this case, we should consider that the message for the target users inside the disaster region is different to the message for the other users who may help them. That is, message generation module should refer the user information with its current location, as well as the social connectivity in order to make proper warning messages for different type of users.

Additionally, we can imagine the information of warning messages spread out through social connections. For example, the well made personalized warning message also increases the possibility that the recipient carefully reads the message, and it improves the coverage of message propagation.

When the message propagation is considered, we can consider the multi-modal channels for the message propagation, because users usually use various channels to propagate messages through social connections [6]. In an emergency situation, the message propagation may enhance the possibility that a message reaches to the target user within a given time limit.

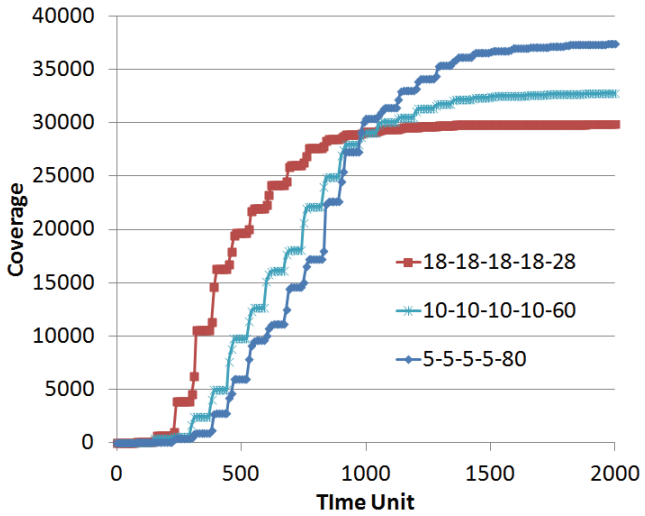


Figure 4. Propagation process with different channel preferences

Recently, we have researched on the relationship between channel preferences and the coverage of message propagation. To evaluate the relationship, the users are grouped in different five groups, and each group has different channel preferences. Figure 4 shows the coverage and the speed of message propagation under different distribution of user groups. In this figure, the portion of the fifth group, whose members prefer long delayed channel such as phone call, varies. As the portion of the fifth group increases, the speed of early stage of propagation becomes slow, but the eventual coverage of message propagation increases. That is, it is shown that the different channel preferences affect the eventual coverage of the message propagation and the speed of the propagation.

According to these consideration of message generation and propagation, the geo-social connectivity aware message generation module leverages location of users, social-connectivity of users, the multi-modal channel preferences, and possible time limits in order to maximize the performance of message propagation and the utility of a message.

 ISBN: 978-1-61208-418-3

D. Failure-Resilient Geo-aware Dissemination

In a disaster warning system, quickly and reliably sending warning messages to all of relevant users is the upmost important job. The users who are in the range of a disaster usually have to make a quick decision under the lack of information. For example, a man who is inside a building on fire need to quickly chooses an exit. A wrong choice at this time can be irredeemable. Therefore, giving them quick information to help them making decision is very important.

In a disaster event, there is a high probability that the network devices are ruined. In this situation, how to provide a reliable sending method is also important. For example, a mobile phone is a reliable method to send the message to user in the normal case. But in a disaster, the mobile phone can be ruined or the user cannot get the message. In this case, we need to choose another way to send the message to the user. Also, it is possible that the network devices such as switches or routers are destroyed. To send message to the user, we might have to reroute the message to another router.

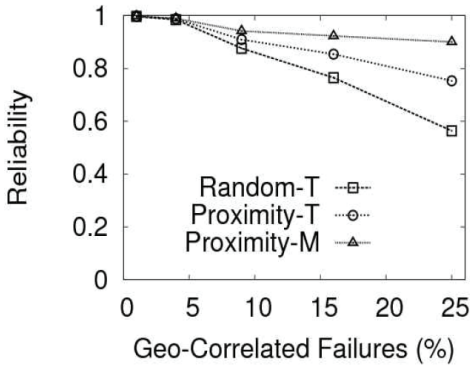


Figure 5. Reliability of message dissemination under large failures

Recently, we have researched on reliable overlay network construction which is resilient to geographical failures [7][8]. The constructed overlay network can be used as a communication mean for disseminating warning messages. Figure 5 shows the reliability of message dissemination under various scales of geo-correlated failures with different overlay networks. The reliability is measured by calculating the rate of successful message delivery to the total number of message receivers. The "Random-T" means the random multiple tree overlays and the "Proximity-M" and the "Proximity-T" mean the mesh network and tree network with proximity aware technique to mitigate the impact of geo-correlated large scale failure [8]. As a result, the proximity aware technique improves reliability of message dissemination substantially. However, this proximity aware technique requires the router information and it is hard to realize in the current internet.

In these days, SDN (Software Defined Networking) is promised as future of internet. SDN provides centralized network management, and SDN controller can provide detail information of underlay routers and make decision based on the global view of the current network condition. With help of the SDN technology, the failure-resilient geo-aware dissemination module manages a highly reliable overlay network, which quickly and reliably disseminates a warning message to all the target users. To do that, we need a system that can analyze the network condition and adapt the sending method based on the network condition.

III. Conclusion and Future Works

The main goal of a disaster warning system is preventing loss of life or properties by providing detail information to all relevant users to a disaster event in timely manner. To achieve this ideal disaster warning system, we propose a new architecture of a disaster warning system with three important components: personality aware location prediction, geo-social connectivity aware message generation, and failure-resilient geo-aware dissemination.

Currently, we have primitive versions of each module with convincible initial findings. The natural future work is improving all the modules and combining them properly. After we build a complete system, we expand this system by applying big data sensing and analyzing techniques.

Acknowledgements

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014R1A1A1007734).

References

[1] Z. Qin, G. Denker, C. Giannelli, P. Bellavista, and N. Venkatasubramanian "A Software Defined Networking Architecture for the Internet-of-Things," Proc. IEEE/IFIP Network Operations and Management Symposium (NOMS), May 2014, pp. 1-9.

[2] K. Benson and N. Venkatasubramanian, "Improving sensor data delivery during disaster scenarios with resilient overlay networks," Proc. IEEE International conference on Pervasive Computing and Communications Workshops (PERCOM), March 2013, pp. 547-552.

[3] K. Jung, S. Kim, H. Ki, and K. Kim, "Implementation of User Interface for Location-based Social Messaging System," Proc. 2014 KIPS fall conference, November 2014, pp. 1064-1067.

[4] Y. Yoon and K. Kim, "Collection and Analysis of Location Data for Recognizing User Movement Methods," Proc. 2013 KIPS fall conference, November 2013, pp. 509-512.

[5] R. Sokasane and K. Kim, "Towards User Personality Identification by Using Location Clustering," Proc. 2014 KDBS conference, July 2014, pp. 54-59.

[6] S. Hwang and K. Kim, "Message Spreading Model over Online Social Network with Multiple Channels and Multiple Groups," Proc. the Ninth International Conference on Internet and Web Applications and Services (ICIW 2014), July 2014, pp. 124-129.

[7] K. Kim, Y. Zhao, and N. Venkatasubramanian, "GSFord: Towards a Reliable Geo-Social Notification System," Proc. the 31st IEEE International Symposium on Reliable Distributed Systems (SRDS), October 2012, pp. 267-272.

[8] K. Kim and N. Venkatasubramanian, "Assessing the impact of geographically correlated failures on overlay-based data dissemination," Proc. IEEE Globecom, December 2010, pp. 1-5.

 ISBN: 978-1-61208-418-3

APEC: Auto Planner for Efficient Configuration of Indoor Positioning Systems

Ming Jin, Ruoxi Jia, Costas J. Spanos

Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720, USA
Emails: {jinming,ruoxijia,spanos}@berkeley.edu

***Abstract*—Fingerprints-based methods have been prevailing in indoor positioning systems, whereas they have certain drawbacks that fingerprints collection in the offline phase requires considerable manpower and time. Auto Planner for Efficient Configuration (APEC) systematically exploits router setups and fingerprints allocations over space by taking into account user preferences and budget constraints. The task of configuration is formulated as an optimization problem, whose objective is the expected loss based on the Hierarchical Bayesian Signal Model (HBSM) and theoretical results on the misclassification rates. To reduce the computational complexity of large-scale problems, two heuristics are employed, i.e., *the coordinate descent* and the *router-fingerprints decoupling*, which are validated by simulation analysis. Experiments with three mobile devices (Android, iPad, iPhone) in two setups (7 or 9 access points) verify that the expected loss is a reliable predictor of the actual loss of the system (*objective consistency*), and that APEC outperforms the random and uniform approaches (*solution superiority*). Since APEC focuses on the system configuration in the planning stage, it can be combined with other fingerprinting processes in the online phase to improve the utility of the system.**

Keywords–Indoor positioning; Fingerprinting method; System optimization

I. INTRODUCTION

The pervasion of radio-frequency transmitters such as WiFi access points, iBeacons and GSM towers has gathered momentum for indoor positioning without the need for specialized infrastructure. One popular approach, pioneered by RADAR [1] and further developed by [2]–[7], is to employ received signal strength (RSS)-based fingerprinting of locations in the space of interest, where typically multiple access points can be heard at each location. A mobile device is then localized by matching the observed RSS readings against the database by deterministic, e.g., K-Nearest Neighbors (KNN), or stochastic methods, e.g., maximum likelihood criterion. While the fingerprinting approach requires a site survey involving detailed RSS measurements which entail considerable effort, an alternative method is to use RF propagation model, such as the prevalent log-distance pass-loss model, which leads to light-weighted localization schemes [8]–[11]. Model-based localization method itself suffers from reduced accuracy since the model can hardly capture signal variance resulting from complexities of indoor environments. The combination of fingerprinting and model-based methods has been proposed as a trade-off between accuracy and RSS measurement effort, such as [12], [13].

Previous work has been devoting effort to maneuvering fingerprinting matching process in the online phase to achieve better localization accuracy, while it remains untouched that potential performance improvement can be obtained by exploiting an optimized way to collect fingerprints in the offline phase. The focus of this study is on the configuration of fingerprints-based positioning system as motivated by the following problems:

- How to take user preferences, i.e., location priority and visiting frequencies, into account?
- How to place routers and allocate fingerprints to be collected over the space under budget constraints?

The first problem arises, for instance, in the scenarios such as: (i) customer behavior analysis in a supermarket, where merchandise area has higher priority than check-out stations, or (ii) region-based indoor environment control, where climate zones are more important than open spaces. The second problem emerges since router setup involves capital costs and fingerprints collection is time-consuming.

It is, therefore, the objective of APEC *to design the fingerprints-based localization system that takes into account user preferences and budget constraints.* The key contributions of the study are as follow:

- Proposal of Hierarchical Bayesian Signal Model (HBSM), a learning-to-learn framework to improve RSSI estimation over space.
- Formulation of the optimization problem, where the objective as a theoretical solution has strong correlation with the actual loss of the system.
- Design and implementation of APEC as a guidance for field deployments.

The rest of the paper is organized as follow. The HBSM is detailed in Section II. Section III formulates the optimization problem and derives the expected loss based on linear/quadratic discriminant analysis, followed by the illustration of the APEC algorithm. Results of field experiments are reported in Section IV, in addition to a toy example to examine the heuristics and algorithmic performances. Section V draws conclusion and discusses future works. Notations and shorthands in the paper are listed in Table I, as a reference.

II. HIERARCHICAL BAYESIAN SIGNAL MODEL

A. Background

1) Log-Distance Pass-Loss Model: The path loss of signal strength inside a building over distance is modeled as

$$X^{\text{RSSI}}(d) = X^{\text{RSSI}}(d_0) + 10\gamma \log \tfrac{d}{d_0} + \epsilon_\sigma, \qquad (1)$$

 ISBN: 978-1-61208-418-3

TABLE I. NOTATIONS AND SHORTHANDS REFERENCE.

Random variables	
$X^{RSSI}(d)$	RSSI at distance d by log-distance pass-loss model
ϵ_σ	Random variable with $\epsilon_\sigma \sim \mathcal{N}(0, \sigma^2)$
$Z \sim P_Z$	Location Z follows the visiting frequency probability P_Z
$X \sim P_{X\|Z}$	RSSI measurement X at location Z distributed as $P_{X\|Z}$
$\boldsymbol{x}_i^{(t)}$	RSSI measurement at location i and time t, which is the realization of $X \sim \mathcal{N}(\Lambda_{i,\cdot}^\top, \tilde{\Sigma}_i)$
Parameters	
K	Total number of routers
N	Number of subregions
M	Number of places available for routers
N_{tot}	Total number of fingerprints to be collected
$\boldsymbol{\theta}=\{\boldsymbol{\theta}^{fp},\boldsymbol{\theta}^{rt}\}$	fingerprints allocation parameter $\boldsymbol{\theta}^{fp} \in \mathcal{N}_+^N$ and router location parameter $\boldsymbol{\theta}^{fp} \in \mathbb{R}^M$
S^{rt}	Set of possible location candidates for $\boldsymbol{\theta}^{rt}$
$n_i, \boldsymbol{\theta}_i^{fp}$	Number of fingerprints collected at location $i \in [N]$
$\Lambda \in \mathbb{R}^{N\times K}$	Mean matrix, with $\Lambda_{i,\cdot}$ for mean RSSI at location $i \in [N]$ and $\Lambda_{\cdot,j}$ for all location measurements for router $j \in [K]$
$\tilde{\Sigma}_i$	Covariance matrix for RSSI measurement at location i
$\tilde{\Sigma}, \tilde{\sigma}^2$	$\tilde{\sigma}^2$ is the diagonal entry of RSSI covariance matrix $\tilde{\Sigma}$, which by assumption is the same for all $\tilde{\Sigma}_i$, $i \in [N]$
Σ_{ml}, s_{ml}	$\Sigma_{ml} = \Sigma_m^{-1} - \Sigma_l^{-1}$ whose diagonal entries are identically s_{ml} in Section III-B2 QDA misclassification
$\tilde{\Lambda} \in \mathbb{R}^{N\times K}$ $\tilde{\Gamma} \in \mathbb{R}^{NK\times NK}$	Hyperpriors of Λ, $\Lambda_{\cdot,j} \sim \mathcal{N}(\tilde{\Lambda}_{\cdot,j}, \tilde{\Gamma}_{\beta(j)})$ for router j and $\Lambda_{i,\cdot}^\top \sim \mathcal{N}(\tilde{\Lambda}_{i,\cdot}^\top, \tilde{\Gamma}_{\alpha(i)})$ for location i
$\boldsymbol{c} \in \mathbb{R}^N$	local priority map, $\boldsymbol{c}_i \in$ {HIGH, MED, LOW} for location i
$\boldsymbol{\pi} \in \mathbb{R}^N$	local frequency map, $\boldsymbol{\pi}_i \in$ {OFTEN, SOME, SELDOM}
Shorthand notations and functions	
$[K]$	Shorthand notation for $\{1, ..., K\}$
$R(i)$	Neighborhood of location i
$\alpha(i)$ index func.	$\{(x,y) : x=i+pN, y=i+qN, p,q \in \{0,...,K-1\}\}$
$\beta(j)$ index func.	$\{(x,y) : (j-1)N+1 \le x,y \le jN\}$
$h_{\boldsymbol{\theta}}(\boldsymbol{x})$	Output of IPS parameterized by $\boldsymbol{\theta}$ given RSSI $\boldsymbol{x}$
$L(Z,\hat{Z})$	Cost of misclassification given the target is at Z, as in (3)
$P_m(h(\boldsymbol{x})\neq m)$	Misclassification rate of $h(\cdot)$ given $\boldsymbol{x}$ and true location m
$\hat{L}(\boldsymbol{\theta}^{rt}, \boldsymbol{\theta}^{fp})$	Actual loss of an IPS designed by $(\boldsymbol{\theta}^{rt}, \boldsymbol{\theta}^{fp})$ as in (12)
$\xi_i(j)$	Empirical misclassification rate of location i to j (13)

where $X^{\text{RSSI}}(d)$ in Decibel (dB) is the Received Signal Strength Indicator (RSSI) at distance d, d_0 is the reference distance, γ is the path loss exponent (PLE), and $\epsilon_\sigma \sim \mathcal{N}(0, \sigma^2)$ is a random variable reflecting the attenuation caused by flat fading [14].

2) Gaussian Process (GP): Every point in space has a normal distribution by the Gaussian process. A collection of points follows a multivariate Gaussian, $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \tilde{\Sigma})$ that is characterized by $\boldsymbol{\mu}$ and $\tilde{\Sigma}$ given by the mean and covariance functions, respectively. GP has been employed in spatial smoothing, aka kriging, and prediction. The covariance function can take many forms, such as constant ($K_C(\boldsymbol{z}, \boldsymbol{z}') = C$), Gaussian noise ($K_{GN}(\boldsymbol{z}, \boldsymbol{z}') = \sigma^2 \delta_{\boldsymbol{z},\boldsymbol{z}'}$), and squared exponential $\left(K_{SE} = \exp\{-\frac{\|\boldsymbol{z}-\boldsymbol{z}'\|_2^2}{2}\}\right)$, where $\boldsymbol{z}$ and $\boldsymbol{z}'$ are spatial positions of any two points.

We propose a neighborhood covariance function, $K_{NH} = \rho\sigma^2 \delta_{\boldsymbol{z}\in R(\boldsymbol{z}')}$, where $R(\boldsymbol{z})$ is a set of points that are in the neighborhood of $\boldsymbol{z}$, $\rho \in (0,1)$ is the GP coefficient, and $\delta_{\boldsymbol{z}\in R(\boldsymbol{z}')}$ is an indicator function which evaluates to 1 if $\boldsymbol{z}$ is in the neighborhood of $R(\boldsymbol{z}')$ and 0 otherwise. The neighborhood covariance is symmetric, and by appropriate choice of ρ, positive definite.

B. Hierarchical Bayesian Signal Model (HBSM)

The space is instrumented with K routers, each of which can independently produce RSSI measurements within the area. The HBSM proposed in this study is a two-layered model for the RSSI observations. The top layer imposes hyperpriors on the mean of RSSI at any point, whereas the bottom layer accounts for measurement error.

*1) **Bottom layer (observations)**:* Let $\boldsymbol{x}_i^{(t)} \in \mathbb{R}^K$ denote measurement at location $i \in [N]$ and time t for K routers. $\boldsymbol{x}_i^{(t)} \sim \mathcal{N}(\Lambda_{i,\cdot}^\top, \tilde{\Sigma}_i)$ follows a normal distribution with mean $\Lambda_{i,\cdot}^\top$ and covariance $\tilde{\Sigma}_i$, where $\Lambda \in \mathbb{R}^{N\times K}$ is the **mean matrix** whose i-th row corresponds to the mean RSSI at location $i \in [N]$ and j-th column corresponds to all location measurements for router $j \in [K]$, where we use notations with tilde to represent hyperpriors, which are non random and can be estimated by sample averages. The signal can be considered as a summation of the mean signal $\Lambda \in \mathbb{R}^{N\times K}$ with a multivariate Gaussian random noise $\epsilon_t \sim \mathcal{N}(0, \tilde{\Sigma}_i)$, which results from randomness of measurement and environments. We make the following assumption of the system:

Assumption 1: Routers work independently and identically, which, by the log-distance pass-loss model (1), indicates that $\tilde{\Sigma}_i = \tilde{\Sigma}$, $i \in [N]$, with identical diagonal entries $\tilde{\sigma}^2$.

*2) **Top layer (hyperpriors)**:* The rearranged mean matrix $\begin{bmatrix} \Lambda_{\cdot,1} \\ \vdots \\ \Lambda_{\cdot,K} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \tilde{\Lambda}_{\cdot,1} \\ \vdots \\ \tilde{\Lambda}_{\cdot,K} \end{bmatrix}, \tilde{\Gamma} = \begin{bmatrix} \tilde{\Gamma}_{\beta(1)} & & \\ & \ddots & \\ & & \tilde{\Gamma}_{\beta(K)} \end{bmatrix} \right)$ has multivariate Gaussian distribution, where $\tilde{\Lambda} \in \mathbb{R}^{N\times K}$ is the mean and $\tilde{\Gamma}$ is the covariance matrix. We introduce the indexing functions:

$$\alpha(i) = \{(x,y) : x = i+pN, y = i+qN, p,q \in \{0,...,K-1\}\}$$

$$\beta(j) = \{(x,y) : (j-1)N+1 \le x,y \le jN\}$$

where $\alpha(i)$ and $\beta(j)$ extracts the covariance terms from $\tilde{\Gamma}$ for location i and router j, respectively. By rearranging the terms we have $\Lambda_{i,\cdot}^\top \sim \mathcal{N}(\tilde{\Lambda}_{i,\cdot}^\top, \tilde{\Gamma}_{\alpha(i)})$ as can be verified.

Assumption 2: The mean of the hyperprior, $\tilde{\Lambda}$, is given by the log-distance pass-loss model. The diagonal variance and off-diagonal covariance for all locations corresponding to a single router j, $\tilde{\Gamma}_{\beta(j)}$, are given by the log-distance pass-loss model and Gaussian process model respectively.

Remarks: The HBSM model is inspired by the learning to learn framework [15], where the observation at one point in space can refine our estimation of other points through the top layer of hyperpriors. It lays the theoretical foundation of many empirically proven fingerprinting methods, such as Virtual Fingerprints [16], Modellet [13] and CGSIL [17], where the collected fingerprints are used to train a radio propagation model locally in order to estimate the unknown area. The HBSM model also has implications to radio map reconstruction by introducing methods from empirical Bayes and Gaussian process regressions [18].

III. AUTO PLANNER FOR EFFICIENT CONFIGURATION

We describe the APEC framework in this section. The key idea is that given limited resources (routers and fingerprints), critical locations that are visited frequently should be distinguished with high accuracy. APEC requires users to provide two maps as illustrated in Figures 1 and 7 in the Experiment section (Section IV-B):

- **Local priority map**, where each subarea is associated with a priority level $\boldsymbol{c}_i \in \{\text{HIGH, MED, LOW}\}$ to represent costs incurred in case of location confusion.
- **Local frequency map**, where a visiting frequency level $\boldsymbol{\pi}_i \in \{\text{OFTEN, SOME, SELDOM}\}$ is indicated for each subregion.

where each level is given a nonnegative value to quantify the cost. Typical values are HIGH=3, MED=2, LOW=1, which also applies to the local frequency map.

The practicality of the nonuniform treatment of positioning accuracy is obvious. In an office building, most occupants will spend their time in their cubicles (HIGH cost of confusion, OFTEN frequency) and public areas such as conference rooms (MED, SOME) and kitchen (HIGH cost for energy apportionment, SOME) as compared to corridors (LOW, SELDOM). Another use case is the supermarkets, where the store manager might put HIGH value to food shelves to learn customer behaviors and LOW to open spaces. In the following, we formulate the problem in an optimization framework.

A. Optimization Framework

Our objective is to minimize the expected cost subject to router and fingerprints constraints, i.e., number/locations of routers/fingerprints:

$$\begin{aligned} \underset{\boldsymbol{\theta}}{\text{minimize}} \quad & \underset{Z\sim P_Z, X\sim P_{X|Z}}{\mathbb{E}}[L(Z, h_{\boldsymbol{\theta}}(X))|Z] \\ \text{subject to} \quad & \boldsymbol{\theta} = \{\boldsymbol{\theta}^{fp}, \boldsymbol{\theta}^{rt}\} \in \Theta \end{aligned} \tag{2}$$

where the expectation is with respect to $Z \sim P_Z$, the visiting probability given by the local frequency map, and $X \sim P_{X|Z}$, the fingerprints observation at location Z. To make the problem computationally tractable, we divide the space into N subregions, so the fingerprints decision variable $\boldsymbol{\theta}^{fp} \in \mathbb{N}_+^N$, $\mathbf{1}^\top \boldsymbol{\theta}^{fp} \leq N_{tot}$ where $\boldsymbol{\theta}_i^{fp}$ is the number of fingerprints collected at subregion i, and N_{tot} is the total number to be collected. As for the routers parameter, $\boldsymbol{\theta}^{rt}$, we allow the user to provide M locations to place the K routers (later we describe a heuristic of choosing the valuable router locations to lessen the computational burden).

The loss function $L(Z, h_{\boldsymbol{\theta}}(X))$ represents the cost of misclassification given the target is at location Z, i.e.,

$$L(Z, h_{\boldsymbol{\theta}}(X)) = \boldsymbol{c}_Z P_Z(h_{\boldsymbol{\theta}}(X) \neq Z), \tag{3}$$

where $\boldsymbol{c}_Z$ is indicated by the local priority map, and $h_{\boldsymbol{\theta}}(X)$ is the Bayes optimal classifier which in our case is the linear/quadratic discriminant analysis (LDA/QDA):

Linear/Quadratic Discriminant Analysis: Given two distributions $\mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$, $k \in \{m, l\}$, for an observation $\boldsymbol{x}$, define the discriminant score for distribution k as $d_k(\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{x}_m{}^\top \Sigma_k^{-1}\boldsymbol{x}_m + \boldsymbol{\mu}_k^\top \Sigma_k^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top \Sigma_k^{-1}\boldsymbol{\mu}_k + \ln(\boldsymbol{\pi}_k) - \frac{1}{2}\ln|\Sigma_k|$, the optimal classification rule is given by

$$h(\boldsymbol{x}) = \arg\max_k d_k(\boldsymbol{x}) \tag{4}$$

For distinct covariance matrices, $\Sigma_m \neq \Sigma_l$, the above classification is known as Quadratic Discriminant Analysis (QDA). If the covariances are equal, the discriminant score can be simplified as $d_k(\boldsymbol{x}) = \boldsymbol{\mu}_k^\top \Sigma_k^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top \Sigma_k^{-1}\boldsymbol{\mu}_k + \ln(\boldsymbol{\pi}_k)$, and the corresponding classifier becomes Linear Discriminant Analysis (LDA) [19]. Both QDA and LDA yield maximum posterior distribution, thus are reasonable classifers to employ for fingerprints-based positioning. In the following, we derive the analytic form of the misclassification rate in (3) to evaluate our objective function in (2) explicitly.

B. Theoretical Results for Misclassification Rate

Assume that the true class of $\boldsymbol{x}$ is m, i.e., $\boldsymbol{x}_m \sim \mathcal{N}(\boldsymbol{\mu}_m, \Sigma_m)$, then we will have misclassification if:

$$d_m(\boldsymbol{x}_m) < d_l(\boldsymbol{x}_m) \tag{5}$$

between two classes m and l. In the following, we consider the cases for the LDA and QDA based on the relation of covariance matrices.

*1) **LDA** [20]:* For $\Sigma_m = \Sigma_l$, (5) is equivalent to:

$$\boldsymbol{v}_{ml}^\top \boldsymbol{x}_m + a_{ml} < 0 \tag{6}$$

where $a_{ml} = -\frac{1}{2}\left(\boldsymbol{\mu}_m^\top \Sigma_m^{-1}\boldsymbol{\mu}_m - \boldsymbol{\mu}_l^\top \Sigma_l^{-1}\boldsymbol{\mu}_l\right) + \ln\frac{\boldsymbol{\pi}_m}{\boldsymbol{\pi}_l}$, $\boldsymbol{v}_{ml} = \Sigma_m^{-1}(\boldsymbol{\mu}_m - \boldsymbol{\mu}_l)$. Since $\boldsymbol{x}_m \sim \mathcal{N}(\boldsymbol{\mu}_m, \Sigma_m)$, we have

$$\underbrace{P\left(\boldsymbol{v}_{ml}^\top \boldsymbol{x}_m + a_{ml} < 0\right)}_{P_m(h(\boldsymbol{x})\neq m)} = \Phi\left(-\frac{\boldsymbol{v}_{ml}^\top \boldsymbol{\mu}_m + a_{ml}}{\boldsymbol{v}_{ml}^\top \Sigma_m \boldsymbol{v}_{ml}}\right), \tag{7}$$

where $\Phi(\cdot)$ is the cumulative distribution function for standard Gaussian variable $\mathcal{N}(0, 1)$.

*2) **QDA**:* Inspired by the derivation of misclassification rate for LDA [20], for $\Sigma_m \prec \Sigma_l$ (the case of $\Sigma_m \succ \Sigma_l$ is similar), condition (5) is equivalent to

$$-\tfrac{1}{2}\boldsymbol{x}_m{}^\top \Sigma_{ml}\boldsymbol{x}_m + \boldsymbol{v}_{ml}^\top \boldsymbol{x}_m + a_{ml} < 0 \tag{8}$$

where $\Sigma_{ml} = \Sigma_m^{-1} - \Sigma_l^{-1}$, $\boldsymbol{v}_{ml} = \Sigma_m^{-1}\boldsymbol{\mu}_m - \Sigma_l^{-1}\boldsymbol{\mu}_l$, $a_{ml} = -\frac{1}{2}\left(\boldsymbol{\mu}_m^\top \Sigma_m^{-1}\boldsymbol{\mu}_m - \boldsymbol{\mu}_l^\top \Sigma_l^{-1}\boldsymbol{\mu}_l\right) + \ln\frac{\boldsymbol{\pi}_m |\Sigma_l|^{1/2}}{\boldsymbol{\pi}_l |\Sigma_m|^{1/2}}$. We can rewrite the left hand side term as follows:

$$\begin{aligned} & -\tfrac{1}{2}\boldsymbol{x}_m{}^\top \Sigma_{ml}\boldsymbol{x}_m + \boldsymbol{v}_{ml}^\top \boldsymbol{x}_m + a_{ml} \\ &= -\frac{1}{2}\left\|\Sigma_{ml}^{1/2}\boldsymbol{x}_m - \underbrace{\Sigma_{ml}^{-1/2}\boldsymbol{v}_{ml}}_{\boldsymbol{\omega}_{ml}}\right\|_2^2 + a_{ml} + \tfrac{1}{2}\boldsymbol{v}_{ml}^\top \Sigma_{ml}^{-1}\boldsymbol{v}_{ml} \\ &= -\frac{1}{2}\sum_{i=1}^K \left(s_{ml}^{1/2}[\boldsymbol{x}_m]_i - [\boldsymbol{\omega}_{ml}]_i\right)^2 + a_{ml} + \tfrac{1}{2}\boldsymbol{v}_{ml}^\top \Sigma_{ml}^{-1}\boldsymbol{v}_{ml} \end{aligned}$$

where $[\boldsymbol{x}]_i$ is the i-th component of $\boldsymbol{x}$. Since in our formulation, Σ_{ml} is diagonal matrix with identical diagnoal entries s_{ml} for $i \in [K]$, we can see that $Y^{(m)} = \frac{1}{s_{ml}\sigma_m^2}\sum_{i=1}^K \left(s_{ml}^{1/2}[\boldsymbol{x}_m]_i - [\boldsymbol{\omega}_{ml}]_i\right)^2$ is a noncentral chi-squared distribution with the degrees of freedom K and noncentrality parameter $\lambda^{(ml)} = \sum_{i=1}^K \left(\frac{s_{ml}^{1/2}[\boldsymbol{\mu}_m]_i - [\boldsymbol{\omega}_{ml}]_i}{s_{ml}^{1/2}\sigma_m}\right)^2$. Thus the probability of misclassification given that the true class is m is given by:

$$P_m(h(\boldsymbol{x}) \neq m) = 1 - F\left(\frac{\boldsymbol{v}_{ml}^\top \Sigma_{ml}^{-1}\boldsymbol{v}_{ml} + 2a_{ml}}{s_{ml}\sigma_m^2}; K, \lambda^{(ml)}\right), \tag{9}$$

where $F(\cdot; k, \lambda)$ is the cumulative distribution function of the noncentral chi-squared distribution with degrees K and noncentrality parameter λ.

Applications to APEC: The gist of APEC is to select the number of fingerprints and placement of routers so that

the location misclassification rate weighted by the priority and visiting frequency is minimized. Intuitively, the more fingerprints $\boldsymbol{x}_i^{(t)}$ collected the better we can estimate the mean at that location $\Lambda_{i,\cdot}^\top$ and the less the misclassification rate (refer Section II.B for the relations). By the maximum likelihood method, our estimates of $\Lambda_{i,\cdot}^\top$ is the sample mean of fingerprints $\boldsymbol{x}_i^{(t)}$, which is normally distributed with mean $\Lambda_{i,\cdot}^\top$ and covariance $\frac{1}{n_i}\tilde{\Sigma}_i$, where n_i is the number of fingerprints collected. By *Assumption 1*, $\tilde{\Sigma}_i$ is the same for all locations $i \in [N]$. Consider m and l to be neighboring two regions, then by applying the results in (7) and (9) with $\Sigma_m = \frac{1}{n_m}\tilde{\Sigma}$ and $\Sigma_l = \frac{1}{n_l}\tilde{\Sigma}$ ($\tilde{\Sigma}$ is diagonal with entries $\tilde{\sigma}^2$ by *Assumption 1*),

$$P_m(h(\boldsymbol{x}) \neq m) = \begin{cases} \Phi\left(-\frac{\boldsymbol{v}_{ml}^\top \boldsymbol{\mu}_m + a_{ml}}{\sqrt{\frac{1}{n_m}\boldsymbol{v}_{ml}^\top \tilde{\Sigma} \boldsymbol{v}_{ml}}} \right), & n_m = n_l \\ 1 - F\left(\frac{\boldsymbol{v}_{ml}^\top \Sigma_{ml}^{-1} \boldsymbol{v}_{ml} + 2a_{ml}}{s_{ml}\sigma_m^2}; K, \lambda_1^{(ml)} \right), & n_m > n_l \\ F\left(\frac{\boldsymbol{v}_{ml}^\top \Sigma_{ml}^{-1} \boldsymbol{v}_{ml} - 2a_{ml}}{s_{ml}\sigma_m^2}; K, \lambda_2^{(ml)} \right), & n_m < n_l \end{cases} \tag{10}$$

where $\boldsymbol{v}_{ml} = \tilde{\Sigma}^{-1}(n_m\boldsymbol{\mu}_m - n_l\boldsymbol{\mu}_l)$, $\Sigma_{ml} = |n_m - n_l|\tilde{\Sigma}^{-1}$, $s_{ml} = |n_m - n_l|\,\tilde{\sigma}^{-2}$, $a_{ml} = -\frac{1}{2}\left(n_m\boldsymbol{\mu}_m^\top\tilde{\Sigma}^{-1}\boldsymbol{\mu}_m - n_l\boldsymbol{\mu}_l^\top\tilde{\Sigma}^{-1}\boldsymbol{\mu}_l\right) + \ln\frac{\boldsymbol{\pi}_m n_m^{K/2}}{\boldsymbol{\pi}_l n_l^{K/2}}$, $\boldsymbol{\omega}_{ml} = \Sigma_{ml}^{-1/2}\boldsymbol{v}_{ml}$, σ_m^2 is the diagonal element of $\frac{1}{n_m}\tilde{\Sigma}$, and the noncentrality parameters $\lambda_1^{(ml)} = \sum_{i=1}^K \left(\frac{s_{ml}^{1/2}[\boldsymbol{\mu}_m]_i - [\boldsymbol{\omega}_{ml}]_i}{s_{ml}^{1/2}\sigma_m}\right)^2$, $\lambda_2^{(ml)} = \sum_{i=1}^K \left(\frac{s_{ml}^{1/2}[\boldsymbol{\mu}_m]_i + [\boldsymbol{\omega}_{ml}]_i}{s_{ml}^{1/2}\sigma_m}\right)^2$. The above formula ties the HBSM model and optimization framework to allow us evaluate the objective function efficiently. Now, we introduce the APEC algorithm.

C. APEC Algorithm

We can write out the expected loss in (2) as follow:

$$\mathop{\mathbb{E}}_{\substack{Z\sim P_Z,\\ X\sim P_{X|Z}}} L(Z, h_{\boldsymbol{\theta}}(X)) = \sum_{i\in[N]} \boldsymbol{\pi}_i \underbrace{\sum_{j\in R(i)} \boldsymbol{c}_i P_i(h_{\boldsymbol{\theta}}(\boldsymbol{x}) = j)}_{\text{Weighted cost of location confusion}}, \tag{11}$$

where $\boldsymbol{\pi}_i$ is the visiting frequency (normalized from the local frequency map), $\boldsymbol{c}_i$ is the location confusion coefficient given by the local priority map, $R(i)$ is the neighborhood of point i. APEC optimizes over the following parameters:

- $\boldsymbol{\theta}^{rt}$ (router locations): Given M possible locations, choose K to place the routers.
- $\boldsymbol{\theta}^{fp}$ (fingerprints): Plan the number of fingerprints to be collected at each subregion $i \in [N]$ such that $\boldsymbol{\theta}^{fp} = [n_1, \cdots, n_N]^\top \in \mathbb{N}_+^N$, $\mathbf{1}^\top\boldsymbol{\theta}^{fp} \leq N_{tot}$.

It can be seen that the problem is combinatorial in nature, which requires integer programming. The computation is formidable for large scale problems. For instance, there are $\binom{N + N_{tot} - 1}{N - 1}$ possible solutions to distribute N_{tot} fingerprints to N subregions. APEC Greedy, therefore, is proposed to solve the problem efficiently, as shown in Algorithm 1, which is based on the HBSM (*line 5*, see Section II) and weighted misclassification cost (*line 8*, see Section III).

Algorithm 1: Pseudo-code of APEC Greedy

```
APEC_Greedy(Maps, N_tot, K)
Input: Maps: possible M locations to place K routers (S^rt),
       centers of N subregions to collect N_tot fingerprints
       (S^fp), local priority/frequency maps
Initialization:
1  S^rt ← Comb(Maps, K)        // Set of router locations
2  b ← Number of fingerprints (fp) increment
3  B ← Batch size
   θ ← { }   // Cell to store the history of {θ^fp, θ^rt}
   V ← [ ]   // Vector to store the history of costs
   bookInd ← 1
Main program:
4  for θ^rt ∈ S^rt do // Scan possible router locations
5      Λ̃, Γ̃, Σ̃ ← HBSM(θ^rt, S^fp)                    // Sec.II
       θ^fp ← 1(N, 1)  // Start with 1 fp per location
       k ← N                          // Current number of fps
       while k < N_tot do
           θ_t ← { }, V_t ← [ ]
           for i ∈ {1, ..., B} do
               /* Randomly choose b indices out of N
                  with replacement, then increment
                  the corresponding entries in θ̂^fp  */
6              u ← RandInd(N, b)
7              θ̂^fp ← θ^fp, θ̂^fp_u ← θ̂^fp_u + 1
8              v ← E_{Z,X} L(Z, h_θ(X))              // Equ. (11)
               θ_t(i) ← θ̂^fp, V_t(i) ← v
           i_bst ← arg min_i V_t(i)
           θ^fp ← θ_t(i_bst), v_bst ← V_t(i_bst)
           k ← k + b          // Increment fps by stepsize
       θ(bookInd) ← {θ^fp, θ^rt}, V(bookInd) ← v_bst
       bookInd ← bookInd + 1
9  i_bst ← arg min_i V(i)
   {θ^rt_bst, θ^fp_bst} ← θ(i_bst)
   Output: {θ^rt_bst, θ^fp_bst}          // APEC Greedy solution
```

The APEC Greedy algorithm exhaustively searches for router locations ($\boldsymbol{\theta}^{rt} \in S^{rt}$, where S^{rt} is all possible combinations of M choose K locations, *line 1*), and stochastically optimizes for fingerprints vectors $\boldsymbol{\theta}^{fp}$. The asymptotic case of choosing $b = N_{tot}$ and batch size $B \to \infty$ (*lines 2,3*) produces the same result of exhaustive search at the cost of infeasible computation time. The simulation experiments in the following verify that APEC Greedy performs almost as well as exhaustive search at computational advantage. We also propose heuristics in selecting the most useful locations to place routers through router-fingerprints decoupling to resolve scalability issue in field deployment.

IV. Experimental Results

A. Toy Case Study

Through the simulation, we compare the performance of APEC Greedy with APEC Exhaustive, and understand the router and fingerprints placement in relation to the local priority and frequency maps. Specifically, we will examine the following heuristics that APEC Greedy (Algorithm 1) employs to reduce computational complexity:

- **Heuristic 1 (Coordinate Descent)**: Fingerprints budget is allocated in $\boldsymbol{\theta}^{fp}$ by random selection of location candidates (*lines 6,7*) and choosing the set that makes the most cost reduction (*line 9*).
- **Heuristic 2 (Router-Fingerprints Decoupling)**: The set of optimal router locations S^{rt} that minimizes the loss (11) is chosen by assuming a uniform allocation of fingerprints budget, i.e., each spot has fingerprints $\frac{N_{tot}}{N}$. Then, the fingerprints $\boldsymbol{\theta}^{fp}$ are optimized within the reduced set (*line 4*) instead of the full set with combinatorial number of router candidates.

We design a simple problem, as shown in Figure 1, with $M = 9$ possible router locations and $N = 5$ subregions, where the location priority is color coded, and the frequency is identical for all 5 subregions. The total number of fingerprints N_{tot} is restricted so that APEC Exhaust is tractable. A random approach is also implemented where the fingerprints are distributed randomly.

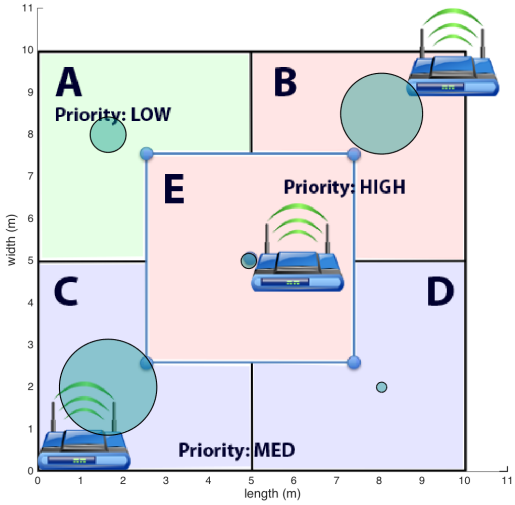


Figure 1. Map of the toy example showing the access points locations and fingerprints allocations (radius of the green circle) determined by APEC Exhaust (also APEC Greedy since they agree).

1) Examination of Heuristic 1: The losses (8) of fingerprints allocation $\boldsymbol{\theta}^{fp}$ given by APEC Exhaust, Greedy, and random approach corresponding to each router configuration is shown in Figure 2. The router setup is indexed by the expected loss of APEC Exhaust, so ideally other methods should incur similar loss and exhibit descending trend to match the optimal solution.

2) Examination of Heuristic 2: The router-fingerprints decoupling heuristic makes a trade-off between cost and computation by sequentially optimizing over $\boldsymbol{\theta}^{rt}$ and $\boldsymbol{\theta}^{fp}$ in problem (2), which brings computational advantage for large-scale problems. Even though the heuristic relies on estimation of the optimal cost that is not the most accurate, the performance is guaranteed if the estimation preserves rankings among the candidates, so that the best router setups are revealed nevertheless, as shown in Figure 3. As it can be seen, though the estimation is noisy, the decoupling heuristic can recover

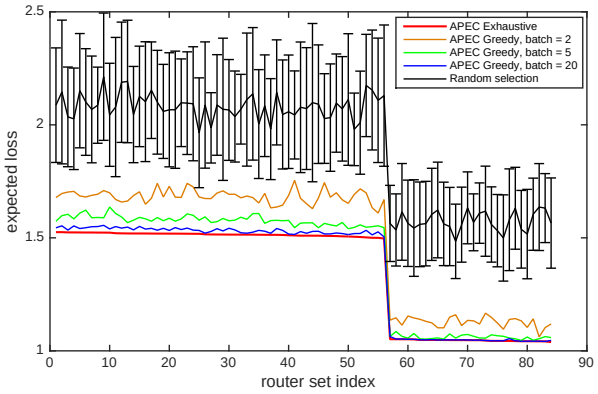


Figure 2. Expected loss of fingerprints allocation $\boldsymbol{\theta}^{fp}$ for each router setup (out of 84 candidates) with $N_{tot} = 15$ total budget, $N = 5$ subregions, $K = 3$ routers, as selected by APEC Exhaust, APEC Greedy (with different batch sizes B), and random approach (black error bar showing mean$\pm$1 standard deviation over 20 test runs).

the best 20 router setups with high probability, which is not likely for the random selection approach.

Another possible heuristic, *random selection*, is to treat each setup uniformly, where they have identical distribution of expected loss, as shown in Figure 3. We might end up in a region, i.e., the right half of the graph, where even with the optimized fingerprints $\boldsymbol{\theta}^{fp}_{*}$ the incurred loss is significantly higher then the optimal, since unlike the decoupling heuristic, the ranking information is lost.

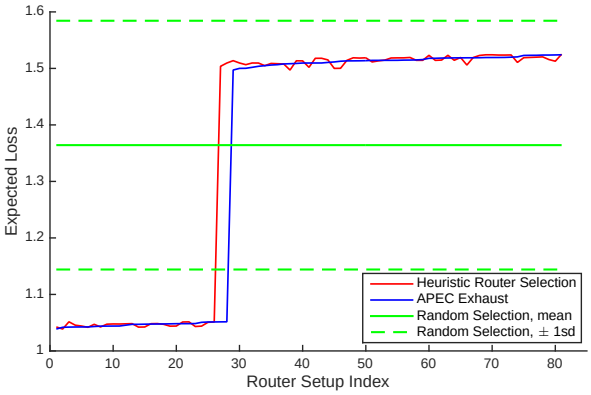


Figure 3. Expected loss vs. router setup indexed by the optimal costs given by solving the fully coupled problem (APEC Exhaust). The plot also shows the loss estimated by random selection (green, mean $\pm$ 1 standard deviation) and heuristic router selection (red).

B. Field Deployment

The theoretical foundation of APEC Greedy, a method to predetermine the optimal router locations and fingerprints allocation for indoor positioning system (IPS), is the formulation of the optimization problem (2), which takes into account the HBSM (see Section II), location priority, and visiting frequencies. As most fingerprinting system employs methods such as K-Nearest Neighbor (KNN) for positioning [1], the assessment of APEC requires the examination of the following:

 ISBN: 978-1-61208-418-3

TABLE II. SUMMARY OF DATA COLLECTED IN EXPERIMENT A (WITH 5 ROUTERS) AND B (7 ROUTERS) FOR THREE DEVICES (IPHONE, IPAD, ANDROID).

	iPhone	iPad	Android
Exp A: 7 RTs	Avg: 72±18 pts Tot: 144 minutes	Avg: 72±15 pts Tot: 144 minutes	Avg: 96±25 pts Tot: 193 minutes
Exp B: 9 RTs	Avg: 84±17 pts Tot: 167 minutes	Avg: 80±19 pts Tot: 160 minutes	Avg: 98±12 pts Tot: 196 minutes

- **Hypothesis 1 (Objective Consistency)**: Given $(\boldsymbol{\theta}^{rt}, \boldsymbol{\theta}^{fp})$, the expected loss given in (11) is a good indicator of the actual loss of the system. *Objective consistency* ensures that the solution of (2) is (near-) optimal in practice (see Figure 5).
- **Hypothesis 2 (Solution Superiority)**: Though the objective consistency if met can guarantee optimal solution, we still want to verify that APEC, with the application of heuristics proposed in Section IV-A, performs well with respect to the actual cost (Figure 6).

Data collection takes place in the Center for Research in Energy Systems Transformation (CREST) on the UC Berkeley campus, where Figure 7 shows the floor plan and location priority. The priority is set HIGH to facilitate region-based thermal and lighting control [21], MEDIUM for shared spaces such as kitchen and conference room for energy apportionment, and LOW for corridors. Fingerprints are collected by an Android phone (Nexus 5), iPad, and iPhone, in two independent experiments, where 7 or 9 access points (D-Link DIR-605LWiFi Cloud Router) are installed as summarized in Table II. The average number of fingerprints $\pm$ 1 standard deviation for all 40 subregions ("Avg"), and the total time for the experiment, which accounts for roughly 3 seconds required to collect one fingerprint ("Tot") are indicated in the table.

Figure 4 shows the RSSI with respect to the log distance from the fingerprint to the access point. Generally, a linear relation is observed (with correlation score .77), though the variance can be reduced by accounting for walls as suggested by Bahl and Padmanabhan, [1]. It can be seen that the dependency is stable for all devices (Android, iPad, iPhone), which ensures cross-platform performance of fingerprints configuration.

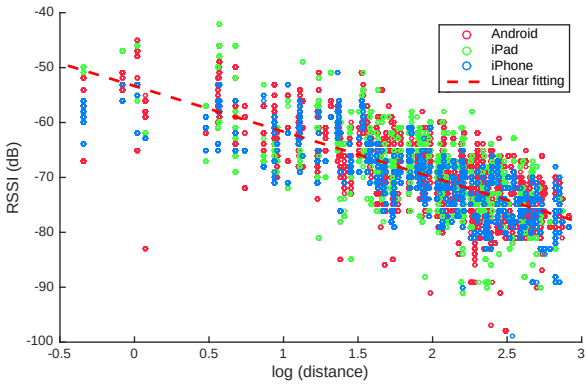


Figure 4. Plots of RSSI vs. log distance color coded by devices. A linear regression line is fitted on the data: RSSI $= -8.31 \cdot \log(d) - 53.4$.

1) Examination of Objective Consistency: Given location priority map $\boldsymbol{c}$ and frequency map $\boldsymbol{\pi}$, the actual loss of an IPS system parameterized by $(\boldsymbol{\theta}^{rt}, \boldsymbol{\theta}^{fp})$ is given by

$$\hat{L}(\boldsymbol{\theta}^{rt}, \boldsymbol{\theta}^{fp}) = \sum_{i \in [N]} \boldsymbol{\pi}_i \underbrace{\sum_{j \in R(i)} \boldsymbol{c}_i \xi_i(j)}_{\text{Weighted cost of location confusion}}, \quad (12)$$

where $\xi_i(j)$ is the empirical evaluation of the misclassification rate of location i to its neighboring spots j,

$$\xi_i(j) = \frac{1}{|\mathcal{X}_i|} \sum_{\boldsymbol{x} \in \mathcal{X}_i} \mathbb{1}(h_{\boldsymbol{\theta}}(\boldsymbol{x}) = j, j \neq i). \quad (13)$$

$\mathcal{X}_i$ is the set of test points at location i, and $\mathbb{1}(h_{\boldsymbol{\theta}}(\boldsymbol{x}) = j, j \neq i)$ is the indicator function valued 1 if the IPS function $h_{\boldsymbol{\theta}}(\boldsymbol{x})$, e.g., 1-nearest neighbor, outputs $j \neq i$, and 0 otherwise. Given a particular setup, the actual cost and the expected cost given in (11) is shown in Figure 5.

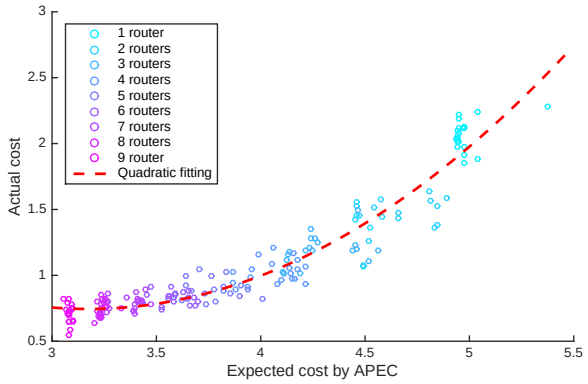


Figure 5. Plots of the actual cost (12) vs. the expected cost (11), color coded by the number of routers deployed. A quadratic curve is fitted to the data: $\hat{L} = 0.37 \cdot L_e^2 - 2.34 \cdot L_e + 4.46$, where $\hat{L}$ and L_e are actual and expected losses given in (12) and (11) respectively.

As it can be seen, the expected cost by APEC is a strong indicator of the actual cost of IPS. Compared to $\xi_i(j)$ which is hard to determine *a priori*, the expected cost is easy to calculate as a function of $(\boldsymbol{\theta}^{rt}, \boldsymbol{\theta}^{fp})$, as a closed form solution of the misclassification rate $P_i(h_{\boldsymbol{\theta}}(\boldsymbol{x}) = j)$ is derived in (10). In other words, *we can predict the fingerprints-based IPS performance based on the router-fingerprints configuration*, which can be useful for other problems as well, such as the optimization of the number of routers to be deployed under budget constraints.

2) Examination of Solution Superiority: Though it is difficult to check strict optimality due to the non-convexity and intractable state space of the problem, solution superiority can be demonstrated by comparing to the two common practices, i.e., *random* and *uniform* selections, where the former randomly allocates fingerprints, the latter maintains a balanced profile over the space. Both methods, nevertheless, ignores user preferences encoded in the location priority and frequency maps. As the expected cost is shown to be a strong predictor of the actual cost, APEC, theoretically, can reach a *preferred* solution as guided by the optimization of (2).

As it is shown in Figure 6, which illustrates the distribution of actual costs (12) in Experiment A and B (see Table II),

 ISBN: 978-1-61208-418-3

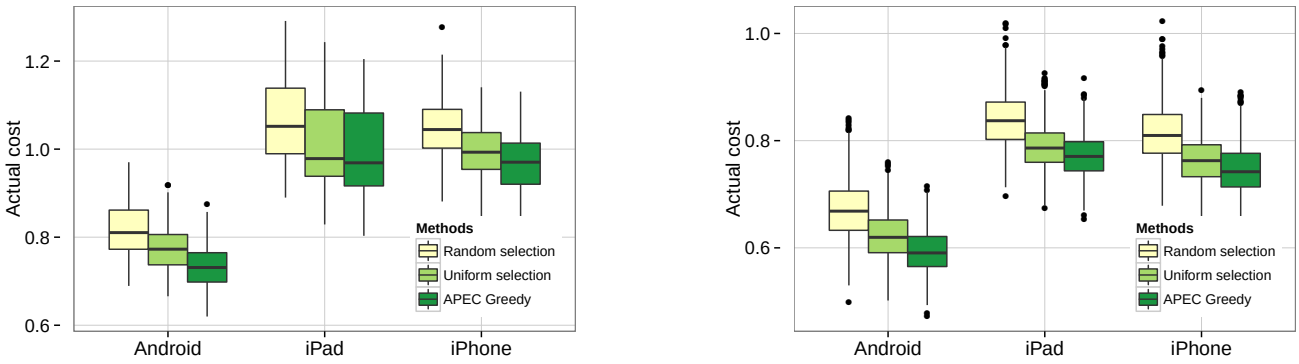


Figure 6. Actual cost distribution in Experiment A (left) and B (right) for all devices color coded by the fingerprinting methods, i.e., random selection, uniform selection, and APEC Greedy. The box goes from the 1^{st} quantile to the 3^{rd} quantile, with the black line indicating the mean. As a side note, Android device incurs less cost in both experiments as a result of its stability of the RSSI signal.

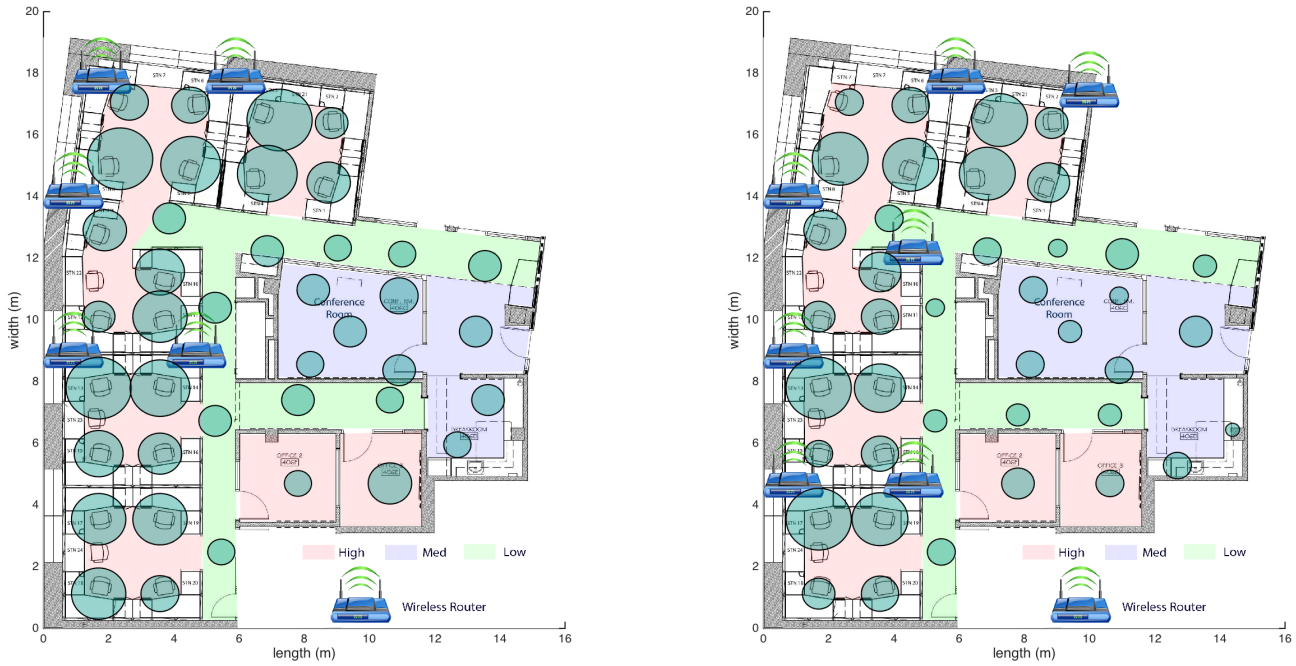


Figure 7. Visualization of IPS configuration in CREST for experiment A (left) and B (right), where 5 or 7 access points are deployed according to APEC Greedy. Location priority map is color coded as illustrated in the legend. Location frequency map is not shown, which is HIGH for cubicles, MED for shared spaces, and LOW for corridors. The radius of the green circle indicates the number of fingerprints at the spot.

the strategy suggested by APEC performs better than the others for all devices. Since a lower actual cost is a result of accurate classification in high priority areas, APEC is effective in catering to user needs.

3) Visualization: The best fingerprints configuration, $(\boldsymbol{\theta}^{rt}, \boldsymbol{\theta}^{fp})$, as chosen by APEC Greedy for Experiment A and B are shown in Figure 7. Some simple guiding rules can be learned from observations:

- Routers are placed in regions with high priority to ensure fingerprints distinction.
- More fingerprints are needed for high priority areas.
- Region close to the routers requires more fingerprints.

The third point, though less intuitive, can be explained by inspecting Figure 4, where the RSSI changes slowly in near-router regions, and the difference of distances are not sufficient for fingerprints separations.

As future work, we would like to implement the visualization in mobile platforms to further assist planning when fingerprints IPS is in demand.

V. Conclusion and Future Work

APEC systematically optimizes the locations of access points and allocations of fingerprints over space by taking into consideration user preferences through local priority/frequency maps and budget constraints, which can be combined with

 ISBN: 978-1-61208-418-3

existing fingerprinting-based methods to improve utility of the indoor positioning system.

The core of APEC is the optimization problem (2), where the objective is the expected loss (11) based on the proposed HBSM and theoretical results on the misclassification rates. As verified by *objective consistency* (Figure 5), the expected loss is a strong predictor for the actual loss incurred by the IPS system, a useful result to determine the performance of the system in the planning stage.

To make APEC computationally tractable, the coordinate descent and router-fingerprints decoupling heuristics are proposed, which are validated by simulation. Experiments with three devices (Android, iPad, iPhone) in two different setups (7/9 routers) demonstrates *solution superiority* of APEC as compared to the uniform and random approaches. Through visualization, several simple rules are developed, while the map serves as a visual guidance for field deployment. As future work, we want to implement and visualize APEC configuration on mobile platforms to facilitate regular planning.

ACKNOWLEDGMENT

This research is funded by the Republic of Singapore's National Research Foundation through a grant to the Berkeley Education Alliance for Research in Singapore (BEARS) for the Singapore-Berkeley Building Efficiency and Sustainability in the Tropics (SinBerBEST) Program. BEARS has been established by the University of California, Berkeley as a center for intellectual excellence in research and education in Singapore.

REFERENCES

[1] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, vol. 2, 2000, pp. 775–784.

[2] T. Roos, P. Myllymki, H. Tirri, P. Misikangas, and J. Sievnen, "A probabilistic approach to wlan user location estimation," International Journal of Wireless Information Networks, vol. 9, 2002, pp. 155–164.

[3] M. Youssef and A. Agrawala, "The horus wlan location determination system," in Proceedings of the 3rd international conference on Mobile systems, applications, and services. ACM, 2005, pp. 205–218.

[4] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: wireless indoor localization with little human intervention," in Proceedings of the 18th annual international conference on Mobile computing and networking. ACM, 2012, pp. 269–280.

[5] A. Rai, K. K. Chintalapudi, V. N. Padmanabhan, and R. Sen, "Zee: zero-effort crowdsourcing for indoor localization," in Proceedings of the 18th annual international conference on Mobile computing and networking. ACM, 2012, pp. 293–304.

[6] H. Liu, Y. Gan, J. Yang, S. Sidhom, Y. Wang, Y. Chen, and F. Ye, "Push the limit of wifi based localization for smartphones," in Proceedings of the 18th annual international conference on Mobile computing and networking. ACM, 2012, pp. 305–316.

[7] J.-g. Park, B. Charrow, D. Curtis, J. Battat, E. Minkov, J. Hicks, S. Teller, and J. Ledlie, "Growing an organic indoor location system," in Proceedings of the 8th international conference on Mobile systems, applications, and services. ACM, 2010, pp. 271–284.

[8] K. Chintalapudi, A. Padmanabha Iyer, and V. N. Padmanabhan, "Indoor localization without the pain," in Proceedings of the sixteenth annual international conference on Mobile computing and networking. ACM, 2010, pp. 173–184.

[9] H. Lim, L.-C. Kung, J. C. Hou, and H. Luo, "Zero-configuration indoor localization over ieee 802.11 wireless infrastructure," Wireless Networks, vol. 16, 2010, pp. 405–420.

[10] Y. Ji, S. Biaz, S. Pandey, and P. Agrawal, "Ariadne: a dynamic indoor signal map construction and localization system," in Proceedings of the 4th international conference on Mobile systems, applications and services. ACM, 2006, pp. 151–164.

[11] Y. Gwon and R. Jain, "Error characteristics and calibration-free techniques for wireless lan-based location estimation," in Proceedings of the second international workshop on Mobility management & wireless access protocols. ACM, 2004, pp. 2–9.

[12] D. Roland, E. Martin, and B. Robert, "Indoor navigation by wlan location fingerprinting," in UBICOMM 2014, The Seventh International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Porto, Portugal, 2013, pp. 1–6.

[13] L. Li, G. Shen, C. Zhao, T. Moscibroda, J.-H. Lin, and F. Zhao, "Experiencing and handling the diversity in data density and environmental locality in an indoor positioning service," in Proceedings of the 20th annual international conference on Mobile computing and networking. ACM, 2014, pp. 459–470.

[14] J. Goldhirsh and W. J. Vogel, "Handbook of propagation effects for vehicular and personal mobile satellite systems," NASA Reference Publication, vol. 1274, 1998, pp. 40–67.

[15] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," Machine Learning, vol. 28, no. 1, 1997, pp. 7–39.

[16] S.-T. Sheu, Y.-M. Hsu, and H.-Y. Chen, "Indoor location estimation using smart antenna system with virtual fingerprint construction scheme," in UBICOMM 2014, The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Rome, Italy, 2014, pp. 281–286.

[17] H. D. Nguyen, T. M. Doan, and N. T. Nguyen, "Cgsil: A viable training-free wi-fi localization," in UBICOMM 2014, The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Rome, Italy, 2014, pp. 268–274.

[18] C. E. Rasmussen, Gaussian processes for machine learning. Citeseer, 2006.

[19] G. McLachlan, Discriminant analysis and statistical pattern recognition. John Wiley & Sons, 2004, vol. 544.

[20] B. Klaus, "Effect size estimation and misclassification rate based variable selection in linear discriminant analysis," Journal of Data Science, vol. 11, 2013, p. 537.

[21] L. J. Ratliff, M. Jin, I. C. Konstantakopoulos, C. Spanos, and S. S. Sastry, "Social game for building energy efficiency: Incentive design," in 52nd Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, 2014, pp. 1011 – 1018.

 ISBN: 978-1-61208-418-3

Intelligent Manufacturing based on Self-Monitoring Cyber-Physical Systems

Simon Bergweiler

German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
Email: simon.bergweiler@dfki.de

***Abstract*—This paper describes the approach and implementation of a system that combines real industrial environments with a virtual copy of these components. The coupling elements for communication and data management are cyber-physical systems and active digital object memories. The idea of this approach is to create an assistance system that relies on these virtual digital object memories to ensure quality characteristics and to describe any information in a unified structured format. Up to a certain level of complexity, state changes and feature checks are done decentralized by each object memory, in the way of autonomous control.**

Keywords–active digital object memory; cyber-physical systems; cyber-physical production system.

I. INTRODUCTION

The current order of the market is shifting more and more towards the idea of an individual production, different product variants need to be made in almost no time. However, this flexible approach also requires that future factories must be easily adapted and converted to the order situation, but this is time-consuming and costly. To make such a complex task more manageable, parts of the plant, e.g., sensors, machinery and products need to be developed that will make a flexible and modular engineering possible. Nowadays, as a general trend, the focus shifts from pure engineering, which is based on mechanical processes, to software-controlled processes [1]. Future-oriented technologies will increase efficiency in production, for instance through the application of self-monitoring for manufactured products and field devices [2].

The evolution of the Internet to the Internet of Things (IoT) corresponds to the fusion of the real and the virtual world. When considering this trend, Cyber-Physical Systems (CPS) play a main role by coupling the different scientific worlds - mechanical engineering, electrical engineering and computer science. This trend reveals the German industry that it stands on the threshold of the fourth industrial revolution (Industrie 4.0) [2]. Future production processes are characterized by specific requirements to the individual manufacturing of products. This opens up new requirements for highly flexible production systems, and increasing efficiency in industrial production processes will become a significant competitive factor. CPS form a solid basis for Industrie 4.0 [3], and this approach shows the integration of these systems in a real production environment.

The development of component-based machine-to-machine (M2M) communication technologies enable field devices to exchange information with each other in an autonomous way without human intervention. The concept of IoT extends this M2M concept by the possibility to communicate and interact with physical objects, which are represented by CPS. These CPS provide the necessary computing power, storage, sensors and ubiquitous access to the functionality of the instrumented machines and field devices. In this approach, all major field devices are equipped with CPS and installed in spatially separated production lines. The idea goes here towards the concept of "retrofitting". Retrofitting means the advanced equipment of existing facilities through additional hardware: function-enhancing modules for communication and distributed processing. With this instrumentation, it is possible that individual field devices and the manufactured products communicate with each other, until the industrial plant meets the standards and directives of future factories and principles of Industrie 4.0.

In Section II, this paper gives an overview of used technologies and introduces the terms field devices, IoT, CPS, active digital object memories, smart factories and smart products. The Section III describes the concept of distributed decentralized CPS and corresponding locally and globally stored data structures. Section IV describes the scenario and application domain and shows how the approach and the developed framework can be used in this industrial environment. In the following Section V the technical creation of an infrastructure for distributed CPS-based product memories is shown in detail, and Section VI gives a conclusion and an outlook on future work.

II. BACKGROUND

A. Field Devices

Field devices are electronic devices that are located at the field level, the lowest level in the hierarchical level model for automation. They are associated with sensors that, on one hand, detect the data of the measuring points and on the other pass the control data to the actuators. At certain time intervals, field devices continuously supply measured data for process control and receive control data for the actuators.

B. Internet of Things

The inexorable growth and innovation diversity of information and communication technologies leads to a fundamental change in daily life. Computers are becoming smaller and can be used almost anywhere. They are built almost inside of all of our technical equipment, e.g., smart watches that track bio-physical data. These devices provide a wide range of technical capabilities that can be used quite comfortable and allow individual components to communicate and cooperate by constantly exchanging sensor information. Following this future trend it can be expected that all utensils of our daily life are turning into smart nodes within a global communication network: this is called the IoT [4], a trend that will also find its way into domains such as consumer electronics and also industrial production.

The term *Internet of Things* was coined and popularized by the work of the Auto-ID Center at the Massachusetts Institute of Technology (MIT), which in 1999 started to design and propagate a cross-company RFID infrastructure. In 2002, its

 ISBN: 978-1-61208-418-3

co-founder and former head Kevin Ashton was quoted in Forbes Magazine as saying, "We need an internet for things, a standardized way for computers to understand the real world" [5]. This article was entitled "The internet of things", and was the first documented use of the term in a literal sense [6].

C. Cyber-Physical Systems

In the fields of agriculture, health, transport, energy supply and industry, we are facing a revolution, hat will open up new ways and possibilities in the upcoming years. Modern information technologies connect data out of different areas and bring them together. This works, if there is a virtual counterpart for every physical product, that can reproduce, by means of sensors and cameras, the environment and the context to combine simulation models and predictive models.

Therefore, the paradigm of the IoT describes distributed networks, which in turn are composed of networks of smart objects. As a technical term for such smart objects, the term Cyber-Physical Systems (CPS) was coined [7]. The main feature of a CPS is that the information and communication technologies were developed and finely tuned to create virtual counterparts to physical components. CPS link data of the real world and this increases the effectiveness and does not encapsulate computing power in an embedded system. Over the communication channel available distributed computing power can be used to solve problems within a network. The IoT and CPS are not fundamentally new concepts. Indeed, Simon [8] already identified the importance and benefits of combining both, physical and virtual domains. His approach was presented many years ago, when not all embedded platforms and manufacturing techniques were developed as today. In fact, the possibility to develop and use a mature platform and techniques are nowadays widely accepted by the industry. Production processes in the context of the initiative "Industrie 4.0" of the federal German government can be fine-grained equipped with sensors and deliver real-time internal and external production parameters in an very high level of detail [2][9].

These following four features typically characterize CPS [10]:

- A physical part, e.g., sensors and actuators capture physical data directly. This allows a direct influence on physical processes.
- A communication part, e.g., connected to digital networks: wireless, bound, local, global. This allows the use of globally available data and services.
- A computation part, e.g., save and evaluate data and interact on this basis, active or reactive with physical and digital worlds.
- An interaction-layer for HMI, e.g., feature a range of interfaces for multi-modal human-machine interaction. This provides dedicated facilities for communication and control, like control by speech and gestures.

In this approach, CPS are embedded micro-controllers installed either inside or outside of physical objects, responsible for the connection and communication over a network, e.g., the Internet. The technical aspect of classical embedded systems is extended by the idea of *Real World Awareness* and tight integration in digital networks. In the context of this implementation, CPS act as digital counterpart and couples the real and the virtual worlds [3][11]. Furthermore, the "Real World Awareness" and dynamic integration of CPS is based on three basic principles: self identification (*Who am I?*), service exploration (*What do I offer?*) and active networking (*Where are my buddies?*).

D. Cyber-Physical Production Systems

The application of CPS in production systems leads to the Cyber Physical Production Systems (CPPS), in which products, machines and other resources are represented by CPS sharing information and services across the entire manufacturing and value network. Future factories use CPPS, semantic machine to machine communication (M2M) and semantic product memories to create smart products [12]. These smart products are the basis for smart services that use them as a physical platform.

Overall, a CPPS, which is based on decentralized production logic and networked principles, offers advantages in terms of transparency, adaptivity, resource efficiency and versatility over traditional production systems. In the context of CPPS, CPS are fundamental units that have almost instant access to relevant information and parametrization of machines, production processes and the product itself. On the automation level of a CPPS all these information out of the CPS-network is needed to run the manufacturing process successfully and to make strategic decisions. For decision making and control of the manufacturing processes, consistent and coherent information of the "real" world is needed.

E. Active Digital Object Memories

The development of the IoT makes it possible to assign a digital identity to physical objects [13][14]. Paradigms, such as human-machine interaction and machine-to-machine communications are implemented by the use of clearly identifiable markers, so-called smart labels. However, the identification is not only bound to those labels, it can be also achieved by integrated sensors or by providing identification methods.

These developments pave the way for the concept of Active Digital Object Memories (ADOMe), which extend the usage of smart labels by additional memory and processing capabilities [15]. By the use of the product memory concept all data in the life cycle of a product (manufacturer information, suppliers, dealers and users) can be added, and furthermore, the data exchange can be made over this specific memory model. Also, memory-related operations can be performed by small scripts in a local runtime environment directly on the ADOMe [16]. According to the functionality of these scripts it is possible to closely monitor decentralized production processes and resource consumption, to impove the quality of the products [17].

These innovative technologies and techniques are crucial parts and the further development is highly supported in national research initiatives, such as *Smart Manufacturing Leadership Coalition* in the US [18] and *Industrie 4.0* in Germany [2].

The next step in the development and to establish new technologies is to evaluate, process and merge data from existing enterprise resource planning systems (ERP) [19] and data from different ADOMes. Both sources, considered as a single unit, offer comprehensive access to domain knowledge and contextual information. A more concrete description of the industrial environment and the running manufacturing processes enables a better user assistance to automatically recognize intentions and activities of the worker. Recommendations for improvements of the current activity of the worker can be

 ISBN: 978-1-61208-418-3

presented proactively by the system. The approach of Haupert et al. [20] refers to a system for intention recognition and recommendation that shows an example scenario also based on ADOMes.

Furthermore, the concept of digital product memories still has an active part. This activity is realized in the form of small embedded scripts that can be run in a separate runtime environment on the specific CPS. Thus, according to the computing power and storage capacity autonomously simple tasks can be executed independently in a decentralized way. In a certain interval or linked to events, deployed scripts are executed and perform small tasks such as storage cleaning, threshold value monitoring or target/actual-value comparisons.

The present work uses the idea of the Object Memory Modeling (OMM) [21] and implemented an own Application Programming Interface (API) on this basis. OMM is an XML-based object memory format, which can be used for modelling events and it also defines patterns, so called block structures, to store information about individual physical objects. Moreover, this format is designed to support the storage of additional information of physical artifacts or objects. The present work uses the idea of the Object Memory Modeling (OMM) [21] and implemented an own Application Programming Interface (API) on this basis. OMM is an XML-based object memory format, which can be used for modelling events and it also defines patterns, so called block structures, to store information about individual physical objects. Moreover, this format is designed to support the storage of additional information of physical artifacts or objects.

F. Fields of Application - Smart Factories and Smart Products

Powerful computers are becoming smaller, inexpensive and energy efficient and suitable for the integration in devices, the instrumentation of everyday objects and integration in clothes - *smart products*. Tiny CPS-adapted sensors and actuators are able to perceive and respond to their environment and interact with connected services in the network. These sensor networks are an essential piece of the foundation for future factories - *smart factories*. Software-defined platforms, like CPPS, make sensor data available and processable, enriched with intelligence by integrated analysis methods for monitoring and controlling. CPS-enabled factory modules or factory parts and the produced smart products communicate and interact with each other. In this context, ADOMes provide a way to collect and analyze structured data and gives an answer to the question in which format the obtained data sets of all connected CPS could be stored. A smart service uses a smart product of the smart factory, to use smart data as an asset, linked via semantic technologies, see Figure 1 [22].

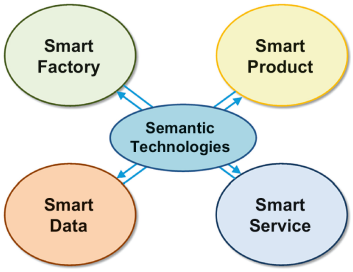


Figure 1: Customization based on semantic technologies [22].

Smart factories and smart products characterize a generation change to new, highly flexible and adaptive manufacturing technologies for the production.

- More computing power in many small devices - extend functionality of existing industrial plants with several CPS.
- Better networked via Cloud-services.
- Gathering and fusion of information - local and global data processing (sensors, actuators).
- Create object memories, and store product/object-specific data.

III. Concept of Distributed Manufacturing Data

In our approach, we consider a production line of a smart factory as a sum of several autonomous CPS. In addition to these aforementioned smart products, there are also intelligent CPS-enabled ADOMes that structure the accruing data of field devices and produced objects and make them accessible. Accordingly, each of these systems is able to act self-regulating and self-monitoring as autonomous factory component, consequently they are able to communicate with each other.

The idea of this approach is to distinguish between locally and globally accessible data structures, respectively represented by an ADOMe. Large amounts of data or storage-intensive data types (e.g., CAD drawings, manufacturer documentation and other internal company documents, videos and examples, electrical wiring diagrams, data history) must be stored in the global version of the ADOMe, because the storage capacity of embedded systems is usually tight. Taking into account these memory restrictions, the local version is an adaptation or filtered version of the global ADOMe, only the necessary information, required for operation and production are stored here. But to accomplish this and to create a special limited local version of an ADOMe, there exist synchronization points and communication structures to ensure the correct synchronization when modifying local or global variables or parameters. Nevertheless, the specific parametrization of field devices should be done first on the unit's local ADOMe and shall be directly accessible. For the fine tuning of dedicated field devices, it is to complex and not practicable to access the central CPPS or global ADOMe. This decentralized parametrization can also be advantageous by setting up a new plant whose infrastructure is also still under construction, or when plant parts are reconstructed and quick compatibility checks must be performed using local data access. Moreover, by the idea that the data is available on the produced object, the ability is given to access these information, just in other factory halls or other companies without access to the central network. Due to the possibility, to keep only certain production data locally in the product's object memory, no sensible production data leaves the factory.

IV. Scenario

In our scenario, depicted in detail in Figure 2, we take the specific case of the production of a gearbox that should be improved or modified during the manufacturing stage. The focus is on the milling of the base plate and the subsequent process of assembling the individual parts. First, the bottom plate is milled and verified by camera, before in a second production step, the product is assembled. These processes take place in

 ISBN: 978-1-61208-418-3

different production lines, which are coupled via a workpiece carrier (WPC). The WPC accompanies the product through the milling, assembling and processing cycle and carries the product physically. The WPC is also equipped with a CPS-enabled ADOMe, that couples the physical product part with its virtual counterpart, which represents all product-specific data. Within this interconnected infrastructure, the WPC has access to all information of the product, to provide relevant and necessary data at the respective part of the industrial plant. The WPC communicates with the ADOMe of the respective object, to provide information for the next production step. Thus, produced objects can be registered early in the process flow.

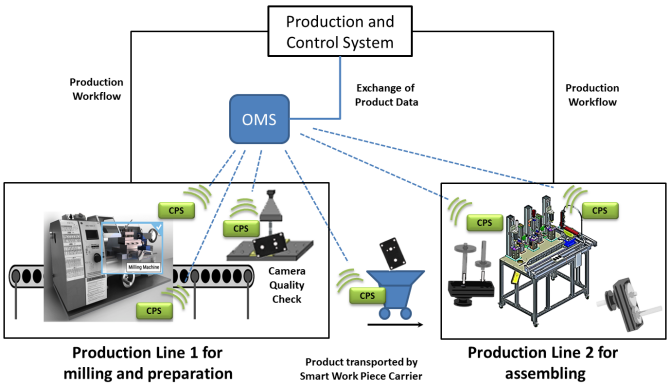


Figure 2: Production scenario.

Beside the idea to structure information in a unified structured format, another goal of this approach is the decentralized autonomous processing of information and immediate derivation of a solution on a CPS-enabled ADOMe. After milling a small script, that has already be embedded to the local ADOMe of the product, checks in a comparison task, whether the actual values match to the specified target values, which are also stored in the same ADOMe. This review will determine, whether the product is fine and meets the quality requirements for the production order, if rework is necessary or it is a faulty product. If reworking is required for that workpiece, a note is stored in the product's memory and the product can be supplied to the production cycle again, when a correctable deviation can be solved directly in the production line. The delayed delivery of produced products, because of reworking, can bring the production process to a standstill. Such bottlenecks can be identified and communicated early enough, so that the overall system is able to reschedule the production workflow.

The smart product knows the sequence and which operations a machine did during the production cycle. Each action is stored by timestamp in an ADOMe. In this assembling scenario of a gearbox, many parts exist that look very similar and have to be prepared and assembled in a certain order. In many cases, it is difficult or not possible to distinguish the material characteristics and the suitability of the gear parts with the naked eye. In this special case, every produced part has its own ADOMe that allows access to the data, which are needed for the next processing or assembling step and for reasons of quality assurance. Furthermore, every single processing step is registered and must be compared with the desired processing steps, defined in the detailed construction phase of the product.

In order to deploy and synchronize a global ADOMe, an own server platform was created, the Object Memory Server (OMS), which provides service functionality in the cloud or the local network. This component is described in detail in Section V-B, cloud-based manufacturing.

V. Technical Components of the Framework

The approach can be subdivided into three processing areas that need to interact with each other. Figure 3 shows the actual products and field device level, represented by each CPS and the associated ADOMe, furthermore, the supply level, where services, snippets and ADOMes are hosted as cloud-based networked solutions, and the assistance level for decision support and knowledge acquisition of the *CPPS*. Decision making is based on the dedicated processing steps and the context-adaptive provision of information of field devices and manufactured products stored in their ADOMes. Each product or field device has both a local and a global ADOMe. The local ADOMe is stored directly on the CPS with limited memory, and the global ADOMe, for storage-intensive data types, is stored by a central server, the OMS.

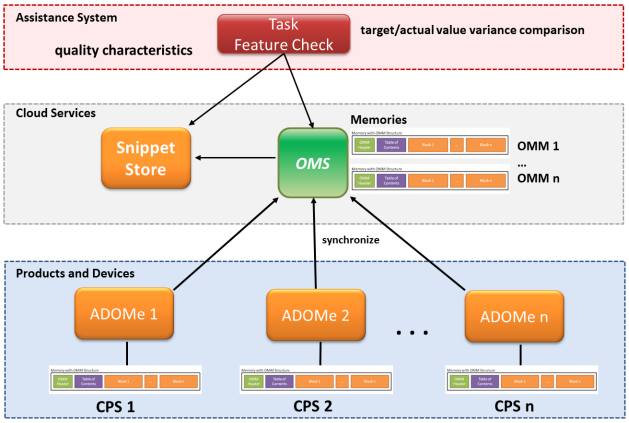


Figure 3: Interaction of the individual components of the framework.

A. Production Assistance System

In an CPPS with many decentralized CPS-enabled modules, condition reports to the overall system are very important. The adopted assistance system for CPPS acts as logical parent unit and is based on managed information out of individual product ADOMes. As presented on Figure 4, the contextual evaluation and context-specific management of processes and procedures is based on facts about the manufactured products (ADOMe), the factory parts (factory model) and the current situation, influenced by the manufacturing process and the skills and role of the user (user model, situational model). The assistance system monitors and supervises the course of production based on process data of the production cycle, and it also monitors and supports the decisions taken by the decentralized scripts. If an intervention in the workflow of the current manufacturing process is needed, based on all converging information here, it generates precise instructions for handling and rescheduling of the production order, or triggers actions, such as maintenance, alteration, or replacement of system components. These reactions of the system are defined in context-dependent rules based on described models, which represents the domain knowledge and the special vocabulary

 ISBN: 978-1-61208-418-3

and terminology. The system decides, whether the manufactured parts are ready for further processing, if they must be revised or if it is rejected goods. These actions are transferred to and processed by the module for output presentation and communicated to the registered clients and subsequent actors.

However, the focus of this approach is not on the consideration and evaluation of complex relationships, for which this assistance system has been designed, but first on simple evaluation purposes, such as self-monitoring and the self-check of quality parameters of a manufactured object or a single field device within its ADOMe and its aligned embedded system. Each distributed ADOMe performs its individual quality checks and returns the data to the assistance system. For example, when a field device is re-parametrized, recommendations are formulated that rely upon the data stored in the history of the memories of this device. Within the system infrastructure, the

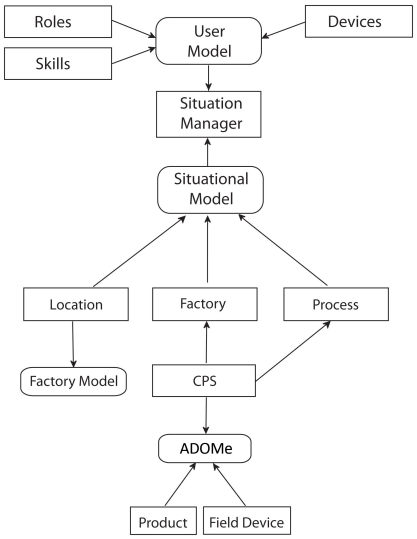


Figure 4: Contextual management based on a situational model.

tiny scripts, we named them snippets, will be hosted in a central cloud-based *Snippet Store*, see Figure 3. Furthermore, based on the task description of the assistance system, e.g., “quality control by target value comparison”, concrete recommendations for scripts are given which adequately provide the required skills. An appropriate matching script is installed in the local ADOMe, when it is compatible with the existing combination of hardware and software of the CPS. The assistance system administrates the runtime of the local ADOMe and sets the execution interval of the script. This scheduling job of the script runs the small tasks, like memory operations or maintenance procedures, based on necessary boundary parameters made dynamically available on the product memory. Moreover, the assistance system must react according to the notification or event mechanism and create a listener functionality for this device configuration. This means that the overall CPPS must check within a time interval, whether the message or event status of an ADOMe has changed. In accordance to these message or event types a recommendation is triggered of the CPPS, which may affect the current production process.

B. Cloud-based Manufacturing

This approach makes use of the potential of a cloud-based networked solution to improve the production process, information sharing, and quality management. Within the cloud, resources, such as processing power, memory or software, in the form of little scripts, are provided dynamically and appropriately over a network. The Object Memory Server (OMS) is the main infrastructure component that stores ADOMes, according to the model of an application server to serve a large number of users. Via a RESTful Web service interface the OMS permits access to process data of each manufacturer and provides the functions to create, store, replace, and modify the data structures in a uniform and consistent manner. Figure 5 shows the interaction of the CPS’ client layer with the OMS. The

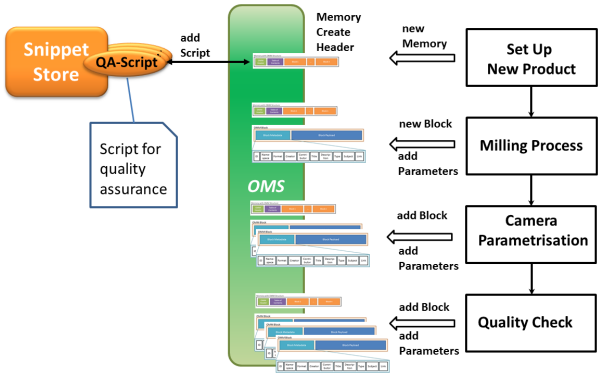


Figure 5: OMS creates ADOMe in production.

OMS uses an own implementation (API) of the Object Memory Model (OMM) to structure and represent the delivered data in an appropriate format. This entails the creation of OMS-records, all communicated data are checked and traceably documented at the time the information was accepted and inserted in the CPS’ ADOMe. But upon closer examination of the data structures from different manufacturers, it becomes evident that no approach is suitable for all requirements, hence the OMS will always be characterized by a certain heterogeneity.

C. Interaction and Output Presentation

A smart factory can never operate without human employees, so one key issue is the human to machine interaction. In a production process, a lot of information passes from monitoring and control, but the problem usually lies in the overview and the appropriate visualization. When people work together with self-learning and self-adapting systems like CPS-based systems, they need to understand each other and which processes are internally occurring. Therefore, the user interface for technical experts or operators is dynamically adapted by a personal assistance system and its module for situational management. This system creates specific UI-layouts or templates for the presentation of contents for diverse mobile devices of the workers (notebooks, smartphones, tablets, smart watches). Currently available monitoring data are presented in adaptable views in form of a curve visualization as depicted in Figure 6. This overview allows the trained experts to draw conclusions about the manufacturing process and possible bottlenecks. First, the situational management component selects the appropriate visualization for a registered device. This selection is based on the situational model, that provides all gathered information about the present situational factors (e.g., user model, parametric influences of the location, factory and production process). According to specific predefined inference rules, which are applied to this model, a visualization pattern is determined

 ISBN: 978-1-61208-418-3

Figure 6: Worker performing a maintenance task with a mobile device.

and prepared for different devices. In this consideration, the special privileges and responsibilities play a major role for an adaptive intelligent visualization, because a technician requires a different view in error or maintenance purposes, as a machine operator who inspects the up and running plant.

VI. Conclusion and Outlook

This article described the conceptualization and implementation of a cyber-physical industrial environment and the use of virtual counterparts of real physical objects, whose data is stored in active digital object memories, hosted on a dedicated Object Memory Server. The described cyber-physical systems enable these memories to communicate over the network and to fulfill small tasks in a decentralized autonomous way, which contribute to the production cycle, like storage cleaning, threshold value monitoring or target/actual-value comparisons. This could even reach a stage, referred to the case of maintenance, in which production systems autonomously order spare parts long before a component fails. With these interconnected cyber-physical systems, it will be possible to implement further product requirements, such as the efficient use of energy and raw materials in production. Furthermore, it will be possible to personalize products and adapt product features in regards to local needs and their individual manufacturing process.

A smart factory can never operate without human employees, so one key issue is the visualization of the stored contents of a dedicated ADOMe. Future work will cover this topic and will further develop strategies that will help to identify and visualize important key values and how these should be presented to the worker (e.g., via tablets or smart watches).

VII. Acknowledgment

This research was funded in part by the German Federal Ministry of Education and Research under grant number 02PJ2477 (project CyProS), 01IA11001 (project RES-COM), and the EIT project CPS for Smart Factories. The responsibility for this publication lies with the authors.

References

[1] M. Loskyll, I. Heck, J. Schlick, and M. Schwarz, “Context-based orchestration for control of resource-efficient manufacturing processes,” Future Internet, vol. 4, no. 3, 2012, pp. 737–761.

[2] H. Kagermann, W. Wahlster, and J. Helbig, Eds., Securing the future of German manufacturing industry - Recommendations for implementing the strategic initiative INDUSTRIE 4.0, Final report of the Industrie 4.0 Working Group. Berlin: acatech National Academy of Science and Engineering, 2013, [retrieved: June 2015]. [Online]. Available: http://www.acatech.de/fileadmin/user_upload/Baumstruktur_nach_Website/Acatech/root/de/Material_fuer_Sonderseiten/Industrie_4.0/Final_report__Industrie_4.0_accessible.pdf

[3] J. Schlick, P. Stephan, M. Loskyll, and D. Lappe, “Industrie 4.0 in der praktischen Anwendung [Industrie 4.0 implemented in practical applications],” in Industrie 4.0 in Produktion, Automatisierung und Logistik, T. Bauernhansl, M. ten Hompel, and B. Vogel-Heuser, Eds. Springer Fachmedien Wiesbaden, 2014, pp. 57–84.

[4] J. Höller, V. Tsiatsis, C. Mulligan, S. Karnouskos, S. Avesand, and D. Boyle, From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence. Elsevier, 2014.

[5] C. R. Schoenberger, “The internet of things,” 2002, [retrieved: June 2015]. [Online]. Available: http://www.forbes.com/global/2002/0318/092.html

[6] F. Mattern and C. Floerkemeier, “From the internet of computers to the internet of things,” in From Active Data Management to Event-Based Systems and More, ser. Lecture Notes in Computer Science, K. Sachs, I. Petrov, and P. Guerrero, Eds. Springer Berlin Heidelberg, 2010, vol. 6462, pp. 242–259.

[7] E. Lee, “Cyber physical systems: Design challenges,” in Object Oriented Real-Time Distributed Computing (ISORC), 2008 11th IEEE International Symposium on, 2008, pp. 363–369.

[8] H. A. Simon, The Sciences of the Artificial (3rd Ed.). Cambridge, MA, USA: MIT Press, 1996.

[9] M. Broy, Cyber-Physical Systems: Innovation durch softwareintensive eingebettete Systeme [Innovation through software-intensive embedded systems], ser. acatech Diskutiert. Springer, 2010.

[10] B. Vogel-Heuser, “Automation als enabler für industrie 4.0 in der produktion auf basis von cyber physical systems [automation as an enabler for industrie 4.0 in production on the basis of cyber physical systems],” Engineering von der Anforderung bis zum Betrieb, vol. 3, 2013, p. 1.

[11] F. Hu, Cyber-Physical Systems: Integrated Computing and Engineering Design. Taylor & Francis, 2013.

[12] W. Wahlster, Ed., SemProM: Foundations of Semantic Product Memories for the Internet of Things. Heidelberg: Springer, 2013.

[13] L. Atzori, A. Iera, and G. Morabito, “The Internet of Things: A survey,” Computer Networks, vol. 54, no. 15, 2010, pp. 2787–2805.

[14] H. Chaouchi, Ed., The Internet of Things: Connecting Objects. Hoboken, NJ, London: Wiley-ISTE, 2010.

[15] J. Haupert, “DOMeMan: A framework for representation, management, and utilization of digital object memories,” in 9th International Conference on Intelligent Environments (IE) 2013, J. C. Augusto et al., Eds. IEEE, 2013, pp. 84–91.

[16] A. Kröner, J. Haupert, C. Hauck, M. Deru, and S. Bergweiler, “Fostering access to data collections in the internet of things,” in UBICOMM 2013, The Seventh International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, W. Narzt and A. Gordon-Ross, Eds. IARIA, 2013, pp. 65–68, best Paper Award.

[17] L. Abele, L. Ollinger, I. Dahmann, and M. Kleinsteuber, “A decentralized resource monitoring system using structural, context and process information,” in Trends in Intelligent Robotics, Automation and Manufacturing. Intelligent Robotics, Automation and Manufacturing (IRAM-2012), vol. 330. Springer Berlin Heidelberg, 2012, pp. 371–378.

[18] “SMLC Forum: Priorities, Infrastructure, and Collaboration for Implementation of Smart Manufacturing: Workshop Summary Report,” Workshop Summary Report, 2015, [retrieved: June 2015]. [Online]. Available: https://smartmanufacturingcoalition.org/sites/default/files/smlc_forum_report_vf_0.pdf

[19] L. Wylie, “A Vision of Next Generation MRP II,” Gartner Group, 1990, pp. 300–339.

[20] J. Haupert, S. Bergweiler, P. Poller, and C. Hauck, “Irar: Smart intention recognition and action recommendation for cyber-physical industry environments,” in Intelligent Environments (IE), 2014 International Conference, 2014, pp. 124–131.

[21] A. Kröner et al., “Object Memory Modeling,” Worldwide Web Consortium, Tech. Rep., 2011, [retrieved: June 2015]. [Online]. Available: http://www.w3.org/2005/Incubator/omm/XGR-omm/

[22] W. Wahlster, “Semantic Technologies for Mass Customization,” in Towards the Internet of Services: The THESEUS Research Program, ser. Cognitive Technologies, W. Wahlster, H.-J. Grallert, S. Wess, H. Friedrich, and T. Widenka, Eds. Springer International Publishing, 2014, pp. 3–13.

 ISBN: 978-1-61208-418-3

Collaborative Detection with Uncertain Signal Distributions in Wireless Sensor Networks

Tai-Lin Chin, Jiun-Hao Chen and Cheng-Chia Huang
Department of Computer Science and Information Engineering
National Taiwan University of Science and Technology
Taipei, Taiwan 106
Email: tchin@mail.ntust.edu.tw

***Abstract*—Sensor networks are envisioned to have the capability to detect the presence of an event or target in a monitored region. Sensors can collect measurements about the target and make local decisions about the presence or absence of the target. To reduce probability of false alarms, collaborative detection is usually exploited, where the local decisions are fused to arrive at a consensus about the target presence. In general, the performance of a sensor network can be evaluated in terms of detection probability and false alarm probability. This paper adopts the Constant False Alarm Rate (CFAR) detector for sensors to make local decisions. The distributions of the target signal and noise are assumed unknown a priori. Simple and effective methods are proposed to estimate the distributions of sensor measurements. The AND and OR fusion methods are exploited in making the final decisions. Simulations are conducted to verify the analytic results to the simulated results. The best selection of sensors to participate the fusion in order to protect a particular location in the monitored region is also shown by experiments. Essentially, the paper analyzes the approximated detection probability and false alarm probability based on the estimated distributions of the unknown target signal and noise. Through simulations, it is shown that those approximated results could be close to the true values.**

Keywords–Sensor networks; Target Detection; Data Fusion; Constant False Alarm Rate.

I. INTRODUCTION

The advances of technologies in sensor networks have made it possible to improve the capability of human to monitor a region of interest. One potential application of sensor networks is to detect the presence of abnormal events or targets in the monitored region. For instance, sensor networks have been used in battlefield monitoring in order to detect unauthorized intrusions[1], [2], or in wildfire detection to protect forests[3]. In such networks, sensors deployed in the monitored region can sample the environment, exchange information with other sensors, and make decisions about the presence or absence of the events or targets. In many cases, the targets may be dangerous or malicious. Consequently, to design effective and reliable detection methods is an important issue for such applications in sensor networks.

Many studies use data fusion to improve detection performance in sensor networks[4], [5], [6], [7]. In most of the studies, sensors are usually used to detect certain signals for which the probabilistic distributions are assumed to be known. However, in practice, the distributions of the target signal and noise might not be known in advance. Furthermore, the distributions could change from time to time caused by the unpredictable and variant physical environment conditions.

This paper considers the problem of detecting a target with unknown signal distribution in sensor networks. Basically, sensors take samples of the target signal and measure the average signal energy periodically. In each period, a local decision about the presence or absence of the target is made by each sensor using Constant False Alarm Rate (CFAR) detector. If the measurements is greater than a carefully selected threshold, it decides that a target is present. Otherwise, it decides that no target is present. An appropriate threshold depends on the distributions of the measurements in both cases when the target is present and absent. A simple and effective method is proposed to estimate the probabilistic distributions of the target signal and the noise. Without complicated calculations, approximate distributions of the measurements are estimated. Moreover, data fusion is also used to further improve the performance of the network. A consensus decision is arrived at a fusion center periodically based on the local decisions reported from sensors. Two fusion methods, namely AND-fusion and OR-fusion, are investigated for data fusion in the network. In particular, the global detection performance in terms of detection probability subject to a fixed false alarm probability is derived analytically based on the estimated distributions.

Simulations are conducted to verify the correctness of the analytic derivations of the detection performance. The comparisons show that the analytic results are very close to the simulation results. The discrepancy between the analytic results and simulation results is mainly caused by the approximations of the signal distributions. In addition, it can be found that not all sensors need to participate in the fusion. In fact, the detection performance may decrease as the number of sensors increases in the fusion. However, if too few sensors in the fusion, it is not beneficial to target detection, either. The best set of sensors to participate in the fusion for a certain target location is also selected by simulations for both AND and OR fusion. The selection can be the basis to generate an efficient and high-performance surveillance sensor network.

The rest of the paper is organized as follows. Section II reviews the related work. Section III addresses the detection mechanism and the analytic derivations of the detection performance. Section IV shows the simulation results. The paper concludes in Section V.

II. RELATED WORK

Target detection using sensor networks has been extensively studied in the literature. A variety of detection methods have been proposed for target detection in sensor networks[8],

 ISBN: 978-1-61208-418-3

[9]. Some studies in the literature assume that the sensing range of a sensor is a disc[10]. A sensor will report a positive decision if the target is present within its sensing range. However, this model may not conform to real situations since it does not capture the stochastic nature of sensing data. Detection errors like false alarm and missing of target's presence may occur from time to time in real detection operations. Some other studies use different probabilistic models to catch the uncertain characteristics of detection operations in sensor networks[11], [12]. In general, the detection probability is usually assumed to be degraded with the distance from the target to the sensor. Based on the assumption, in [11], the coverage of a location in the monitored region is defined as the probability that at least one sensor detects the target if it is present at the monitored location. A sensor deployment strategy is proposed to achieve that the minimal coverage over the region is greater than a threshold. In [12], the detection probability of a mobile target is analytically evaluated when a group of sensors deployed in the monitored region. A probabilistic detection model where each sensor can have heterogeneous sensing area is developed.

In order to improve detection performance, data fusion is usually used to reduce the probability of false alarm or missing[4], [5], [6], [7]. In [4], the monitored region is divided into a grid and two data fusion methods applying to the grid are investigated. One uses data reported from an individual cell and the other uses data from adjacent cells. The latter is shown to be able to generate better performance for the coverage. In [5], the lower and upper bounds of fusion threshold is analytically derived to ensure that a higher detection probability and lower false alarm probability can be obtained compared to those derived from the weighted averages of individual sensors. In [6], the paper investigates collaborative target detection based on data fusion. The optimal detector, which is proven to be uniformly most powerful, is derived. In [7] and [13], the latency of detecting a target based on data fusion in sensor networks is also analyzed. Detection latency is an important issue for real time detection. Recently, there is also work on the problem of target detection in mobile sensor networks [14], [15]. All of the above studies use data fusion, but derives the detection mechanism based on known signal or noise distributions. Our work is different from those previous studies in that the distributions of the target signal and noise are not necessary to be known in advance, which could be much closer to real situations.

III. DETECTION

A. Sensing Model

Suppose that a target at location r emits a signal S_t at time t. The distribution of S_t is unknown, but the mean and variance of S_t are easy to evaluate. Let μ_s and σ_s^2 denote the mean and variance of S_t. Let S_t^i be the signal sensed by sensor i at location r_i. The signal strength is assumed to be degraded with distance. Thus, S_t^i can be modeled as follows:

$$S_t^i = \frac{S_t}{|r - r_i|^\alpha}, \tag{1}$$

where $|r - r_i|$ is the Euclidean distance from the target to sensor i and α is the decay factor. Note that since $|r - r_i|^\alpha$ is a constant, the mean and variance of S_t^i are given as follows:

$$\mu_{s,i} = \frac{\mu_s}{d_i} \text{ and } \sigma_{s,i}^2 = \frac{\sigma_s^2}{d_i^2}, \tag{2}$$

where $d_i = |r - r_i|^\alpha$. Usually, the signal sensed by a sensor is corrupted by noise. Let X_t^i denote the noise signal at sensor i, and is modeled as a random variable with mean $\mu_{x,i}$ and variance $\sigma_{x,i}^2$. Noise can follow a variety of distributions in different environment conditions. In this paper, the distribution of X_t^i is also assumed to be unknown. The final signal sensed by sensor i is $y_t^i = S_t^i + X_t^i$.

Generally, each sensor in the network measures the average signal energy over a sampling period. The measurement of a sensor can be expressed as

$$M_i = \frac{1}{T}\sum_{t=1}^{T} \left|y_t^i\right|^2 = \frac{1}{T}\sum_{t=1}^{T} \left|S_t^i + X_t^i\right|^2, \tag{3}$$

where T is the number of samples in one period.

B. Approximation of Measurement Distributions

In each sampling period, each sensor would make a local decision based on its measurements. Assume that the sampling result at each time instant is independent. The distribution of the measurement M_i can be approximated by Central Limit Theorem (CLT). When the target is absent, the measurement contains only noise, i.e.,

$$M_i = \frac{1}{T}\sum_{t=1}^{T} \left|X_t^i\right|^2. \tag{4}$$

Suppose X_t^i is an independent identically distributed (i.i.d.) random variable. First, the distribution of ${X_t^i}^2$ is determined. In order to apply CLT, the mean and variance of ${X_t^i}^2$ need to be determined. It is easy to get $E[|X_t^i|^2] = \mu_{x,i}^2 + \sigma_{x,i}^2$. However, there is no closed form expression for the variance of $|X_t^i|^2$. To solve the problem, "Delta Method", which finds the approximation of a function of a random variable is exploited. Delta Method is described as follows.

Proposition 1: Let x be a random variable with mean μ and variance σ^2. The variance of a function $f(x)$ can be approximated by

$$Var(f(x)) \approx [f'(\mu)]^2 \times \sigma^2.$$

Proof: The Taylor series expansion of a function $f(\cdot)$ at value a is given by

$$f(x) = f(a) + f'(a)(x-a) + f"(a)\frac{(x-a)^2}{2!} + \cdots.$$

Take the first two terms as an approximation and let $a = \mu$,

$$f(x) \approx f(\mu) + f'(\mu)(x - \mu).$$

Take the variance of both sides, one can have

$$Var(f(x)) \approx [f'(\mu)]^2 \times Var(x).$$

■

Let $f(X_t^i) = {X_t^i}^2$. From Proposition 1, the variance of ${X_t^i}^2$ can be approximated by $4\mu_{x,i}^2\sigma_{x,i}^2$. Using the approximation as the variance and the mean of ${X_t^i}^2$ obtained previously, by CLT, when T is large, M_i converges in distribution to a Gaussian distribution $N(\mu_{i,0}, \sigma_{i,0}^2)$, where

$$\mu_{i,0} = \mu_{x,i}^2 + \sigma_{x,i}^2 \text{ and } \sigma_{i,0}^2 = 4\mu_{x,i}^2\sigma_{x,i}^2/T. \tag{5}$$

 ISBN: 978-1-61208-418-3

Note that the mean $\mu_{i,0}$ is accurate, but the variance $\sigma^2_{i,0}$ is approximated based on Proposition 1.

The approximated distribution of M_i when the target is present can also be derived in a similar way. If the target is present, sensor measurements are mixed by target signal and noise as in (3), which can be rewritten as follows:

$$M_i = \frac{1}{T}\sum_{t=1}^{T}\left|S_t^i\right|^2 + \frac{1}{T}\sum_{t=1}^{T}\left|X_t^i\right|^2 + \frac{2}{T}\sum_{t=1}^{T}\left|S_t^i X_t^i\right| \quad (6)$$

Similar to the target absence case, using CLT and Proposition 1, the distributions of the first two terms in (6) can be approximated by the following two Gaussian distributions.

$$\frac{1}{T}\sum_{t=1}^{T}\left|S_t^i\right|^2 \sim N\left(\mu_{s,i}^2 + \sigma_{s,i}^2, \frac{4}{T}\mu_{s,i}^2\sigma_{s,i}^2\right) \quad (7)$$

$$\frac{1}{T}\sum_{t=1}^{T}\left|X_t^i\right|^2 \sim N\left(\mu_{x,i}^2 + \sigma_{x,i}^2, \frac{4}{T}\mu_{x,i}^2\sigma_{x,i}^2\right) \quad (8)$$

For the third term, assume that target signal and noise are independent, one can also get the mean and variance of $S_t^i X_t^i$ as follows:

$$E\left[S_t^i X_t^i\right] = E\left[S_t^i\right] E\left[X_t^i\right] = \mu_{s,i}\mu_{x,i}$$

$$\begin{aligned} Var\left(S_t^i X_t^i\right) &= E\left[\left(S_t^i X_t^i - \mu_{s,i}\mu_{x,i}\right)^2\right] && (9)\\ &= E\left[S_t^{i2}\right] E\left[X_t^{i2}\right] - \mu_{s,i}^2\mu_{x,i}^2 && (10)\\ &= \mu_{s,i}^2\sigma_{x,i}^2 + \mu_{x,i}^2\sigma_{s,i}^2 + \sigma_{s,i}^2\sigma_{x,i}^2 && (11)\end{aligned}$$

Again, by CTL, when T is large, the distribution of the third term in (6) can be approximated by a Gaussian distribution as follows.

$$\frac{2}{T}\sum_{t=1}^{T} S_t^i X_t^i \sim N\left(2\mu_{s,i}\mu_{x,i}, \frac{4}{T}\left(\mu_{s,i}^2\sigma_{x,i}^2 + \mu_{x,i}^2\sigma_{s,i}^2 + \sigma_{s,i}^2\sigma_{x,i}^2\right)\right) \quad (12)$$

Obviously, the distribution of measurement M_i when the target is present can be approximated by adding the three Gaussian distribution in (7), (8) and (12). Therefore, when the target is present, M_i can also be approximated by a Gaussian distribution $M_i \sim N(\mu_{i,1}, \sigma^2_{i,1})$, where

$$\mu_{i,1} = \sigma_{s,i}^2 + \sigma_{x,i}^2 + \left(\mu_{s,i} + \mu_{x,i}\right)^2, \quad (13)$$

and

$$\sigma_{i,1}^2 = \frac{4}{T}\left(\mu_{s,i}^2 + \mu_{x,i}^2\right)\left(\sigma_{s,i}^2 + \sigma_{x,i}^2\right) + \sigma_{s,i}^2\sigma_{x,i}^2. \quad (14)$$

Again, the mean $\mu_{i,1}$ is accurate, but the variance $\sigma^2_{i,1}$ is approximated based on Proposition 1.

C. Local Detection

With the distributions of sensor measurements, it is possible to control the performance of the detection operations. By the detection procedure, sensors would collect measurements and make a local decision about the presence or absence of the target. A potential method to make the decision is the CFAR detector as shown in Figure 1. Essentially, the detector compares the signal energy measurement, M_i, to a threshold, η_i. If the measurement is greater than the threshold, the detector reports a positive detection. Otherwise, it reports a negative detection. The false alarm probability is defined as the probability that the detector makes a positive decision (i.e., the target is present) when the target is actually absent.

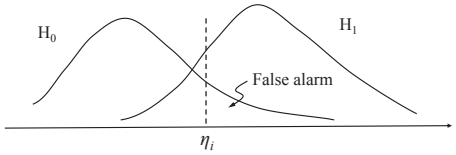


Figure 1. Constant false alarm rate detection

From the previous derivations, the approximate distributions of measurements when the target is present and when the target is absent are known. The detection threshold η_i can be determined subject to a false alarm probability constraint. Specifically, let H_0 be the null hypothesis for the condition that the target is absent and H_1 be the alternative hypothesis for the condition that the target is present. When the target is actually absent, since M_i conforms to a Gaussian $N(\mu_{x,i}^2 + \sigma_{x,i}^2, \frac{4}{T}\mu_{x,i}^2\sigma_{x,i}^2)$, the false alarm probability is given by

$$\begin{aligned} P_{f,i} &= P(M_i > \eta_i \,|H_0\,) && (15)\\ &= Q\left(\frac{\eta_i - \mu_{i,0}}{\sigma_{i,0}}\right), && (16)\end{aligned}$$

where $\mu_{i,0}$ and $\sigma_{i,0}$ are from (5), and $Q(x)$ is the tail probability of a standard normal distribution, i.e.,

$$Q\left(x\right) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left(-\frac{u^2}{2}\right) du.$$

Thus, given a tolerable false alarm probability, one can determine the threshold η_i for the detection operation.

Furthermore, given the target is present, M_i conforms to $N(\mu_{i,1}, \sigma_{i,1})$. Detection probability is defined as the probability that the detector decides a target is present while there is a target present. Therefore, the detection probability can be derived as

$$\begin{aligned} P_{d,i} &= P(M_i > \eta_i \,|H_1\,) && (17)\\ &= Q\left(\frac{\eta_i - \mu_{i,1}}{\sigma_{i,1}}\right), && (18)\end{aligned}$$

where $\mu_{i,1}$ and $\sigma_{i,1}$ are from (13) and (14).

D. Fusion

To further reduce potential false alarms, local decisions of sensors are sent to a fusion center where a consensus decision about the presence or absence of the target is made. Two common used fusion methods are AND fusion and OR fusion. For the AND fusion, the fusion center decides that a target is present if all sensors participating the fusion report positive local decisions. Otherwise, it decides that no target is present. Therefore, the false alarm probability of the final consensus is the probability that all sensors report positive local decisions when there is no target present, i.e.,

$$P_f = \prod_{i=1}^{n} P_{f,i}. \quad (19)$$

 ISBN: 978-1-61208-418-3

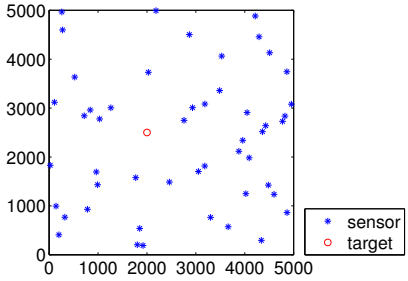


Figure 2. Sensor deployment

In contrast, the detection probability of the final consensus is the probability that all sensors report positive local decisions when a target is present indeed, i.e.,

$$P_d = \prod_{i=1}^{n} P_{d,i}. \tag{20}$$

Similarly, for the OR fusion, the fusion center decides that a target is present if at least one sensor participating the fusion reports a positive local decision. Otherwise, it decides that no target is present. Therefore, the final false alarm probability and detection probability can be derived as in (21) and (22), respectively.

$$P_f = 1 - \prod_{i=1}^{n} (1 - P_{f,i}) \tag{21}$$

$$P_d = 1 - \prod_{i=1}^{n} (1 - P_{d,i}) \tag{22}$$

IV. SIMULATIONS

In the simulations, a sensor network is deployed in a 5000 meter $\times$ 5000 meter field. There are 50 sensors deployed in the field randomly as shown in Figure 2. The target is assumed to be at location (2000, 2500) if it is present. The signal of the target is assumed to be a random variable with mean 50000 and variance 3. The signal decays with distance and the decay factor is $\alpha = 2$. Without loss of generality, a Gaussian random process is used to generate the target signal. It is noted that any other random process can be used without affecting the correctness of the proposed method. The noise process at each individual sensor is also assumed to be Gaussian with mean between 1 to 3 and variance between 0 to 1.

The distribution of the measurement M_i is first evaluated. Sensor measurement M_i is the average of signal energy taken during a sampling period as shown in (3). The mean and variance of M_i when the target is absent are derived in (5) and when the target is present in (13) and (14), respectively. The mean in (5) and (13) are accurate, but the variance in (5) and (14) are estimated by Proposition 1. However, since the target signal and noise signal are assumed to be Gaussian, the true variance of M_i can also be derived as follows. Let x be a Gaussian random variable with mean μ_x and variance σ_x^2. Assume that

$$y = \left(\frac{x - \mu_x}{\sigma_x} \right).$$

TABLE I. Measurement statistic characteristics of a node at (1031.713, 2778.091)

H_0	Mean		Variance		
T	Simulation	True	Simulation	Estimation	True
30	1.313263	1.317921	0.015507	0.014908	0.015462
300	1.318351	1.317921	0.001414	0.001491	0.001546
800	1.317386	1.317921	0.000546	0.000559	0.000580
H_1	Mean		Variance		
T	Simulation	True	Simulation	Estimation	True
30	1.424624	1.429481	0.016872	0.014938	0.016818
300	1.429922	1.429481	0.001537	0.001494	0.001682
800	1.428929	1.429481	0.000595	0.000560	0.000631

Taking the square of both sides of the equality, one can have

$$\sigma_x^2 y^2 = x^2 - 2\mu_x x + \mu_x^2.$$

Therefore,

$$Var(x^2) = Var(\sigma_x^2 y^2) + Var(2\mu_x x).$$

Note that y is a standard Gaussian random variable and y^2 is a Chi-square random variable with one degree of freedom. Thus, y^2 has mean 1 and variance 2. Then,

$$Var(x^2) = 2\sigma_x^4 + 4\mu_x^2\sigma_x^2.$$

Finally, the variance of $\frac{1}{T}\sum_{t=1}^{T} x^2$ is

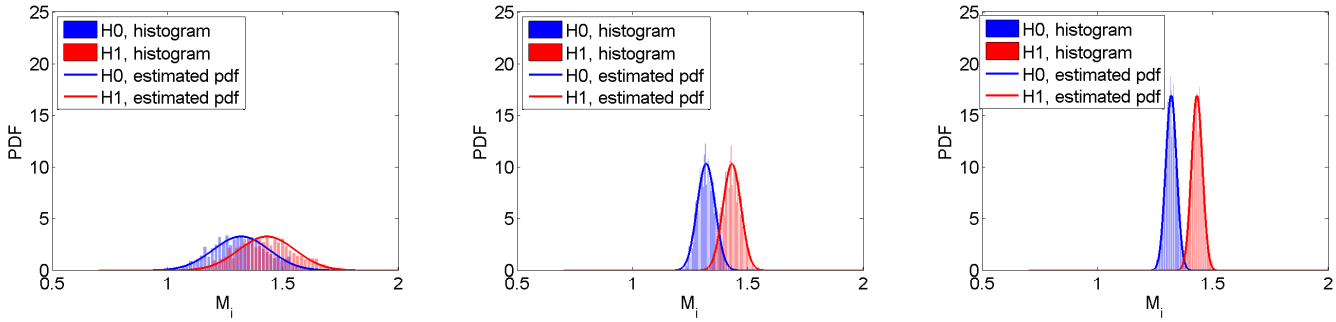
$$\frac{1}{T}\left(2\sigma_x^4 + 4\mu_x^2\sigma_x^2\right).$$

Consequently, let x be the random variable for noise, the variance of measurement M_i when the target is absent can be derived. Analogously, let x be the random variable for target signal plus noise, the variance of measurements when target is present can be obtained.

One sensor located at (1031.713, 2778.091) is chosen to investigate the statistics of its measurements. The target signal at the sensor is a Gaussian random variable with mean 0.05 and standard deviation 0.000003, and the noise is assumed to be Gaussian with mean 1.1 and standard deviation 0.3. The true mean and variance of the measurement are derived and shown in Table I. The approximated variance estimated based on Proposition 1 is also shown in the table. The results of the estimated variance are pretty close to the true variance. The histograms of the measurements are shown in Figure 3. From the figures, when the number of samples T is small, the measurement distributions for target absence and target presence overlap in quite a lot area. It implies that it would be more difficult for the sensor to tell whether the target is present or absent. In contrast, when T is large, the distributions are more concentrated and separated in two groups. Obviously, it is easier for the sensor to tell whether the target is present. Therefore, based on a fixed local false alarm probability, the local detection probability would be higher if T is large.

Figures 4 and 5 show the Receiver Operating Characteristic (ROC) curves for the AND fusion and OR fusion. In each figure, the figure shows global false alarm probability versus global detection probability. In order to get a proper threshold for each sensor such that the fixed global false alarm probability is sustained, the local false alarm probability for AND fusion and OR fusion is choosen as in (23) and (24), respectively.

$$P_{f,i} = \sqrt[n]{P_f} \tag{23}$$



(a) T=30 (b) T=300 (c) T=800

Figure 3. The histograms of sensor measurements

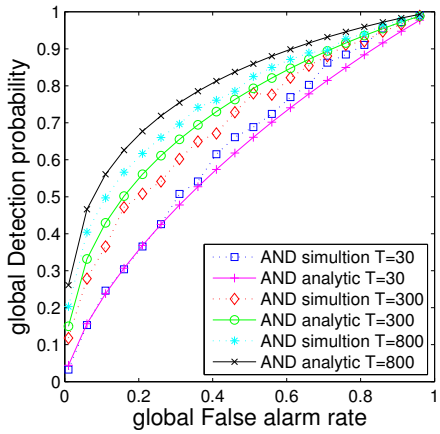


Figure 4. The ROC curves of the AND fusion

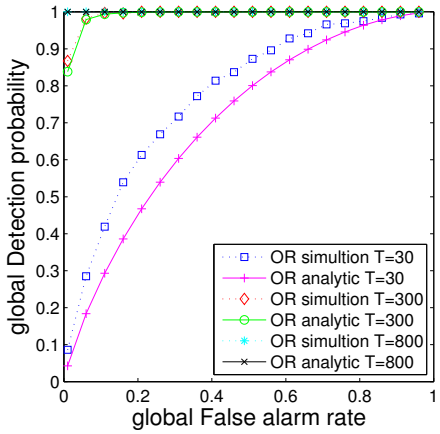


Figure 5. The ROC curves of the OR fusion

$$P_{f,i} = 1 - \sqrt[n]{1 - P_f} \tag{24}$$

Note that local false alarm probability for different sensors can be different for a fixed global false alarm probability. For the sake of simplicity, local false alarm probability of all the sensors are set to be identical. Given the false alarm probability, the detection threshold η_i for each sensor can be determined.

For the AND fusion shown in Figure 4, the analytic results are pretty close to the simulated results. The discrepancy between the simulated and analytic results primarily because the approximations made by CLT and the evaluations of the variance based on Delta method. In general, the system has higher detection probability if it can tolerate higher false alarm probability. In addition, when the number of samples T is large, the detection probability is higher. This is because local detection probability is higher if T is larger, and, thus, from (20), the global detection probability would be higher.

Similar results can be found in the OR fusion shown in Figure 5. Comparing the results of the AND fusion and OR fusion, the detection probability is higher for the OR fusion based on a fixed global false alarm probability. Obviously, the AND fusion requires all the sensors to report a positive detect decision in order to arrive at a positive consensus, while the OR fusion only requires at least one sensor reports a positive decision. The performance of the AND fusion may be degraded by the strict detection rule.

Figure 6 shows the number of sensors participating in the fusion versus the global detection probability using AND fusion. The sensors are ordered by signal to noise ratio, which is defined as the mean of the target signal over the mean of noise signal, i.e., $SNR = \mu_{s,i}/\mu_{x,i}$. The sensors are added to the fusion from the one with the highest SNR. In general, the sensors are added to the fusion roughly in the order of distance from the target location. From the result, it is obvious to see that using all sensors in the fusion is not the best choice. In fact, the detection probability decreases as the number of sensors increases after using three sensors in the fusion. On the other hand, if less than three sensors participate in the fusion, the detection performance also decreases. Consequently, from the simulations, the best choice for data fusion for the specified target location shown in Figure 2 is to choose the three sensors with the highest SNR. Similarly, for the other locations, one can also find the best sets of sensors to monitor the corresponding locations. The results could generate an efficient and high-performance strategy for monitoring the region of interest.

Figure 7 shows the results of similar experiments for

 ISBN: 978-1-61208-418-3

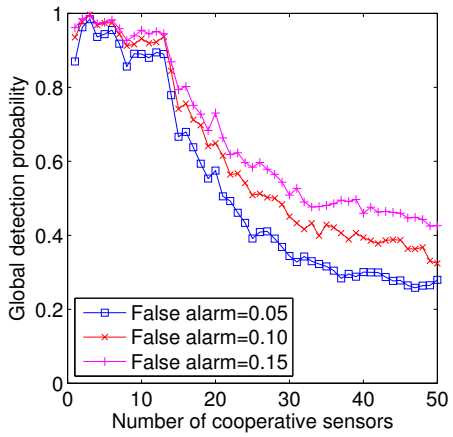


Figure 6. The impact of number of cooperative sensors for AND fusion

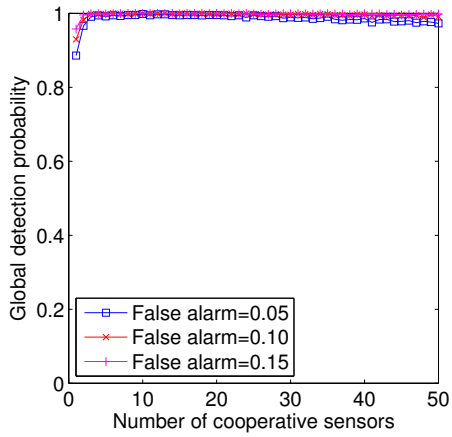


Figure 7. The impact of number of cooperative sensors for OR fusion

the number of participating sensors except using OR fusion. As shown in the figure, the system has highest detection performance when the three sensors with the highest SNR participates in the fusion. Either too many sensors or too few sensors participating in the fusion are not beneficial to the detection performance. It is consistent with the results in AND fusion. Although, in OR fusion, getting more sensors in the fusion would not hurt the performance much, using fewer sensors to monitor the region is always desired.

V. Conclusion

This paper investigates collaborative detection for a target in sensor networks when the distributions of the target signal and noise are unknown. A simple method is proposed to evaluate the approximations of the distributions of the sensor measurements. Using the approximated distributions, local detection decision thresholds can be derived for sensors based on CFAR detection. The global consensus decisions are made by the AND fusion and OR fusion rules. The detection performance in terms of the detection probability subject to a fixed false alarm probability is derived. The performance of both the fusion methods is verified by simulations. From the results, the analytic results are very close to the simulated results. In addition, the best set of sensors to participate in the fusion for monitoring a particular target location is also obtained by simulations. Selecting the best sets of sensors to monitor potential target locations in the region of interest can generate an efficient and high-performance surveillance sensor network.

This paper only investigates the AND and OR fusion rules. From the results, the OR fusion out performs the AND fusion in terms of detection probability subject to a fixed false alarm probability. However, these fusion methods may not be optimal in certain circumstances. For the future work, developing better fusion methods is worth to be explored.

Acknowledgment

This work was supported in part by the Ministry of Science and Technology of Taiwan under grant MOST 103-2221-E-011 -095.

References

[1] D. Li, K. D. Wong, Y. H. Hu, and A. Sayeed, “Detection, classification, and tracking of targets,” IEEE Signal Processing Magazine, vol. 19, no. 2, Mar. 2002, pp. 17–29.

[2] V. Phipatanasuphorn and P. Ramanathan, “Vulnerability of sensor networks to unauthorized traversal and monitoring,” IEEE Transactions on Computers, vol. 53, no. 3, Mar. 2004, pp. 364–369.

[3] M. Hefeeda and M. Bagheri, “Forest fire modeling and early detection using wireless sensor networks,” Ad Hoc & Sensor Wireless Networks, vol. 7, 2009, pp. 169–224.

[4] G. Xing et. al., “Efficient coverage maintenance based on probabilistic distributed detection,” IEEE Transactions on Mobile Computing, vol. 9, no. 9, Sep. 2010, pp. 1346–1360.

[5] M. Zhu, S. Ding, Q. Wu, R. R. Brooks, N. S. V. Rao, and S. S. Iyengar, “Fusion of threshold rules for target detection in wireless sensor networks,” ACM Transactions on Sensor Networks, vol. 6, no. 2, Mar. 2010, pp. 18:1–18:7.

[6] T.-L. Chin and Y. H. Hu, “Optimal detector based on data fusion for wireless sensor networks,” in IEEE Global Telecommunications Conference (GLOBECOM), 2011, pp. 1–5.

[7] R. Tan, G. Xing, B. Liu, J. Wang, and X. Jia, “Exploiting data fusion to improve the coverage of wireless sensor networks,” IEEE/ACM Transactions on Networking, vol. 20, no. 2, Apr 2012, pp. 450–462.

[8] T. Clouqueur, K. K. Saluja, and P. Ramanathan, “Fault-tolerance in collaborative sensor networks for target detection,” IEEE Transactions on Computers, vol. 53, no. 3, Mar. 2004, pp. 320–333.

[9] R. Niu and P. K. Varshney, “Distributed detection and fusion in a large wireless sensor network of random size,” EURASIP Journal on Wireless Commununications and Networking, vol. 2005, no. 4, Sep. 2005, pp. 462–472.

[10] C.-F. Huang and Y.-C. Tseng, “The coverage problem in a wireless sensor network.” MONET, vol. 10, no. 4, 2005, pp. 519–528.

[11] Y. Zou and K. Chakrabarty, “Sensor deployment and target localization based on virtual forces,” in INFOCOM, vol. 2, March 2003, pp. 1293–1303.

[12] L. Lazos, R. Poovendran, and J. Ritcey, “Probabilistic detection of mobile targets in heterogeneous sensor networks,” in IPSN, 2007, pp. 519–528.

[13] T.-L. Chin and W.-C. Chuang, “Latency of collaborative target detection for surveillance sensor networks,” IEEE Transactions on Parallel and Distributed Systems, vol. 26, no. 2, Feb 2015, pp. 467–477.

[14] N. Bisnik, A. Abouzeid, and V. Isler, “Stochastic event capture using mobile sensors subject to a quality metric,” IEEE Transactions on Robotics, vol. 23, no. 4, 2007, pp. 676–692.

[15] T.-L. Chin, P. Ramanathan, and K. Saluja, “Modeling detection latency with collaborative mobile sensing architecture,” IEEE Transactions on Computers, vol. 58, no. 5, May 2009, pp. 692–705.

 ISBN: 978-1-61208-418-3

Delay Prediction Approach for Cyclic Mobility Models in Ad hoc Networks

Jihen Bokri, Sofiane Ouni and Leila Saidane
CRISTAL laboratory
Ecole Nationale des Sciences de l'Informatique (ENSI)
Tunis, Tunisia
e-mails : jihen.bokri@ensi-uma.tn, Sofiane.Ouni@insat.rnu.tn, leila.saidane@ensi.rnu.tn

***Abstract* – Guaranteeing the respect of the deadline on the end-to-end transmission delay for a hard real time traffic is a big challenge, especially in mobile Ad hoc networks. In fact, in such networks, the movement of the nodes can lead to the breakage of the current path between the source and the destination and therefore the establishment of a new path which may provide a transmission delay that exceeds the deadline. In this paper, we propose a delay prediction approach which allows guaranteeing the respect of the deadline regardless of the used path for cyclic mobility models. Indeed, we aim to predict all the possible paths between two nodes and calculate the end-to-end transmission delay in the worst case. Thus, we can guarantee the respect of the deadline if this latter is superior to the calculated delay. By comparing the worst end-to-end delay values calculated by our prediction approach with the delay values obtained by the network simulator NS2 (Network Simulator 2) for the same network, we perceive that the NS2 delays are always lower than the end-to-end transmission delay of the worst case calculated by our approach. This proves that our prediction approach allows guaranteeing the respect of the deadline on transmission delays for hard real time applications.**

Keywords-hard real-time; deadline guarantee; communication delay; prediction; mobility

I. INTRODUCTION

The hard real-time applications, which are used in so many domains such as industrial environments, have strict requirements regarding reliability, latency and end-to-end transmission delay. However, respecting the required transmission delay is the most challenging issue since the transmitted data can be easily delayed because of the nodes mobility. In fact, the mobile nodes movements can break the already established path between the source and the destination which requires a new path finding procedure. This can prevent the real-time traffic from satisfying the deadline on the end-to-end transmission delay. For that, many researches are interested in the mobility prediction for Ad hoc networks. However, most of the researches are based on specific equipment, such as GPS (Global Positioning System) [3][6][9][10] or based on non-accurate methods such as probability approaches [4] or approaches which are based on transmission strength [11].

In this paper, we propose a deterministic prediction approach which can guarantee the respect of the deadline on the end-to-end transmission delay for a real-time traffic even in a mobile environment. In fact, our approach predicts all the possible paths between the source and the destination and calculates the end-to-end transmission delay in the worst case for Ad hoc networks having nodes with cyclic mobility models. Thus, we can know in advance if the deadline on the transmission delay of the real traffic will be respected by comparing it with the calculated worst case transmission delay.

This paper is organized as follows. Section 2 presents the mobility prediction approaches in the literature. Section 3 describes our delay prediction approach. In Section 4, we explain the response time analysis of our approach. Sections 5 and 6 are dedicated to the evaluation of our approach and the conclusion, respectively.

II. MOBILITY PREDICTION APPROACHES IN THE LITERATURE

So many researches were interested in the mobility prediction for Ad hoc networks. For example, in [1], J. Wang proposes a mobility prediction method for Wireless Ad hoc networks having a reliable service composition. This latter consists in integrating many services which are provided by different service providers in the network. In this research, the only required information for the mobility prediction is the estimated time that a service provider will be present in the current environment. However, some service providers can overestimate this time of availability which leads to the uncertainty in the mobility prediction. Hence, the author aims to characterize this uncertainty using two different models which are the probabilistic-free model and the probabilistic model. For that, J. Wang presents heuristic algorithms which are based on the fact that each service provider has a predicted future location described as a function of time.

While the probabilistic model cannot be used for Hard real time communications since it is not deterministic, the probability-free model is also not suitable for such kind of transfer. In fact, it aims to minimize the risk of uncertainty about the links between nodes but it does not provide a sophisticated solution.

In [3], a mobility prediction solution is proposed, which exploits a mobile user's non-random traveling pattern. In fact, the author presents an enhancement to unicast and multicast routing protocols which predict the network topology changes. To this end, he utilizes GPS location information to predict disconnections by estimating the expiration time of the link between two adjacent nodes. For that, the researcher assumes that the velocities of the nodes are constant and he presents an

 ISBN: 978-1-61208-418-3

algorithm which decides whether a link between two given nodes has expired.

The decision is based on the probability **p,** which is the result of the current time of separation (T_{jk}) for the pair of nodes j and k divided by their maximum time of separation ($p= T_{jk}/Max(T_{jk})$). The time of separation is given by the diameter of the node's coverage area (D) divided by the relative velocity of these two nodes (V_{jk}). Hence, if $q(=1-p)<\alpha p$, then the link is expired (while α is a constant factor of persistence).

Although this research presents a mobility prediction solution which allows to identify in advance the links which will expire, this approach is not suitable for Hard real time transmissions since it is based on probability to decide about the link expiration.

Some researches concentrate on the link availability estimation to deal with the mobility issue. In fact, Huang and Bai [4] presented an approach of link availability estimation within a random mobility Ad hoc network. According to the researchers, the link availability is defined as the probability that a link remains available for a period of time t. Hence, they present an analytical expression of link availability based on the estimation of the initial distance between two mobile nodes.

Obviously, this approach cannot be used in the case of Hard real time transmissions since it is based on estimations and probabilities.

Other researches, such as Chegin et al. [5], focus on some specific areas to predict the movements of the mobile nodes and consequently the links duration. In fact, Chegin et al. [5] are interested in predicting the links expiration time in an urban area. For this purpose, they use a map file containing the locations of all the vertical and horizontal streets, as well as the cross points of the roads. Then, they run a prediction algorithm which brings out a prediction table including the links expiration times. This prediction table will be used later by the routing algorithm in order to select the optimal path.

However, the presented solution in [5] is specifically used, as the authors said, with the proactive routing protocols. Also, they assume that the mobile nodes are localized at any time using the GPS which is not always obvious.

In [6][9], the authors present a mobility prediction method which predicts the future distance between two neighboring nodes using learning automaton. In fact, Mousavi et al. [6][9] base their work on the proposed prediction scheme in [7] and they enhance it to have more accurate results.

In [7], a node predicts its future position according to its current position, speed and direction (which are given by a GPS) using the following equations ((1) and (2)):

$$x(t_{0+a}) = x(t_0) \pm s * (t_{0+a} - t_0) * \cos(\theta) \quad (1)$$

$$y(t_{0+a}) = y(t_0) \pm s * (t_{0+a} - t_0) * \sin(\theta) \quad (2)$$

where $\mathbf{t_0}$ is the current time, **α** is the time increment in seconds so that $\mathbf{(t_0+\alpha)}$ is the next sampling time, $\mathbf{(x(t_0+\alpha),y(t_0+\alpha))}$ is the position of a node at $\mathbf{(t_0+\alpha)}$, **s** is the current speed and **θ** is the direction angle of node motion.

In [6], Mousavi et al. use a similar predictor to the one proposed by Mir et al. [7], but with an additional term **(1/alpha)** called the scaling coefficient of the estimator. Hence, the calculation of the future position in works [6] and [9] will be given by the following coordinates ((3) and (4)):

$$x(t_{0+a}) = x(t_0) \pm 1/alpha * (s * (t_{0+a} - t_0) * cos(\theta)) \quad (3)$$

$$y(t_{0+a}) = y(t_0) \pm 1/alpha * (s * (t_{0+a} - t_0) * sin\,\theta) \quad (4)$$

The added coefficient (1/alpha) varies according to the mobility models, speeds and sampling rates in order to have a more accurate prediction. The added coefficient is estimated using a learning automaton [8].

Although this work presents a prediction mobility method providing future positions which are close to the reality (as their results showed), it cannot be used for Hard real time transmissions since such transfers require a strict accuracy.

Other mobility prediction methods are interested in cluster based Ad hoc networks. For example, the proposed research in [10] aims to predict the cluster changes in order to use the provided information in the route maintenance. Sathyaraj et al. [10] integrate the proposed scheme into the reactive routing protocol DSR (Dynamic Source Routing) and show that it offers better packet delivery ratio.

However, the solutions provided by these studies are valid only with clustering mechanisms which can be wasteful in time and bandwidth (Control overhead).

In [11], the authors propose a novel routing protocol (MAODV: Multicast Ad-Hoc On-Demand Distance Vector), which is an enhancement for the Ad hoc On-demand Vector routing protocol (AODV: Ad-Hoc On-Demand Distance Vector). In fact, in MAODV, a mobility prediction algorithm is added to AODV in order to control the detected routes and predict the neighbor node's mobility, so that it can deal with the network topology changes. The prediction is based on estimating the neighboring nodes distance each period of time according to the transmission strength. Thus, if the neighboring node is moving away, a new route should be established before the breakage of the current path.

The main idea of the proposed solution by Meng et al. [11] is the estimation of the distance between the current node and its neighbor. Since this estimation is only based on the transmission power, it cannot provide an accurate value of the distance.

Most of these researches are based on the GPS which provides information about the localization of the nodes. In fact, Mehdi [3] uses the GPS to predict the topology changes and predict thereafter the disconnections between the adjacent nodes, while in [6][9][10], the GPS is used to predict the future position of the mobile node.

 ISBN: 978-1-61208-418-3

Furthermore, the mobility solution presented by Chegin and Fathy [5] uses the GPS to build a prediction table of the links which will be used by the proactive routing algorithms. However, these methods require an additional equipment to provide the basic information which is the GPS, which is not always obvious.

Other researches are based on probabilistic methods as a research presented by Huang and Bai [4] which proposes an estimation approach of the links availability. This approach is defined as a probability that the link remains available during a period of time t. Also, the heuristic algorithm presented by Wang [1] provides two models: A probabilistic model and a non-probabilistic model which aims only to minimize the risk of uncertainty about the links. Some other mobility solutions are based on inaccurate information. In fact, in [11], the mobility prediction algorithm which was added to the routing protocol AODV is based on the transmission strength. This latter is used to estimate the distance between two neighboring nodes which cannot provide an accurate value.

Other mobility proposals are dedicated to the cluster-based Ad hoc networks like in [10] where the authors are interested in predicting the clusters change in order to maintain the constructed paths between the nodes.

In our mobility approach, we developed a deterministic delay prediction method for mobile Ad hoc networks. This method is based on accurate and certain information and it does not require additional equipment like the GPS.

III. OUR DELAY PREDICTION APPROACH

In order to check if a specific deadline for a hard real-time traffic can be satisfied, we should be sure that each used path between the source and the destination provides an end-to-end transmission delay which is lower than the deadline (We assume that there is always an available path between the source and the destination). For that, our approach allows to predict all the possible paths between two given nodes, so that we can calculate the worst end-to-end delay for each path and check if the deadline on the transmission delay will be satisfied. Furthermore, we should predict the instants in which the topology changes and consequently we should re-verify the available paths and the deadline satisfaction.

In this approach, we assume that the coordinates of the node n_i, in its trajectory at the instant $\boldsymbol{t}$ are described by the time functions $\boldsymbol{x_i(t)}$ and $\boldsymbol{y_i(t)}$. So, we note the position of the node n_i at the instant $\boldsymbol{t}$: $\boldsymbol{M_i(t)(x_i(t), y_i(t))}$. We suppose also that each mobile node n_i has a closed predefined trajectory. Actually, each node has a cyclic movement within a defined route (Like the industrial robots which move in a factory). With this hypothesis, we have two different cases. The first case is the periodic movement which occurs when the velocity is constant. The second case is the non-periodic movement which occurs when the velocity is variable. However, in the first case, in which all the nodes have periodic movements; the prediction of the paths change is easier since we can find a global period in which the same network behavior is repeated. So, we will start with the first case: periodic movements.

A. Paths prediction in the case of periodic movements

The periodic movement is the repetition of the same behavior each time period *T*. If we suppose that each node n_i has a periodic movement with its own time period T_i, we can find a time period *T* in which the same network behavior is repeated. Hence, if we verify the deadline satisfaction in the time period *T*, we will have a verification result for all time. Thus, we can predict the possible paths and calculate their transmission delays only in the instants of topology change belonging to the time period *T*.

In this section, we assume that all the nodes of the Ad hoc network have a periodic movement and that they start their movement at the same time, which is the case of mobile robots in most industrial environments. Based on this assumption, we will define the set of available paths between the source and the destination in each topology change instant belonging to the period *[0, T]*. Thus, we have to determine the value of this period *T*. For that, we use property 1.

Property 1: The same network behavior is repeated each period *T* if and only if *T* is the Smallest Common Multiple (SCM) of the time periods T_i which are the time periods of the mobile nodes n_i according to their trajectories, as shown in (5):

$$T = SCM_i\{T_i\} \quad (5)$$

where $\boldsymbol{T_i}$ is the period in which the movement of the mobile node n_i is repeated according to its trajectory.

The set of possible paths between the source and the destination (noted E_{ch}) is, therefore, defined as the set of possible paths between *0* and *T*, since the same network behavior is repeated each time period *T*.

To collect the set of available paths, we start with defining the set of neighbors of a given node n_i at the instant t (noted $NE_i(t)$). $NE_i(t)$ is the set of nodes n_j such as n_j is the neighbor of the node n_i at the instant t. Hence:

$$\boldsymbol{NE_i(t)} = \{\boldsymbol{n_j / n_j}\ \textit{is the neighbor of}\ \boldsymbol{n_i}\ \textit{at the instant}\ \boldsymbol{t}\} \quad (6)$$

When one of the sets of neighbors changes, the available paths between the source and the destination may also change. In fact, a link breakage can occur on one of the paths because of the movement of the node which maintains this link. For that reason, we consider that the instants of neighboring change are the instants of paths change (E_{inst}). So, we should define at these instants, the new possible paths between the source and the destination.

 ISBN: 978-1-61208-418-3

We define E_{inst} as the set of instants of paths change between *0* and *T*. Therefore, the set of all the possible paths between the source and the destination (E_{ch}) is defined as the union of the sets of possible paths at the instants belonging to the set E_{inst}. Hence:

$$E_{ch} = \cup_{t \, \epsilon \, E_{inst}} E_{ch}(t) \quad (7)$$

where $E_{ch}(t)$ is the set of possible paths at the instant of paths change *t*. Thus, $E_{ch}(t)$ is defined as the union of all the possible paths at the instant *t*.

$$E_{ch}(t) = \cup_k P_k(t) \quad (8)$$

where $P_k(t)$ is the k[th] path linking the source and the destination at the instant *t*. Therefore, $P_k(t)$ is represented as the set of nodes which make a link between the source and the destination at the instant *t*:

$$P_k(t) = \{S, n_j, n_k, \dots, n_n, D\} \quad (9)$$

To illustrate this principle, we take the example of the network represented in Figure 1. This network consists of 12 nodes numbered from 0 to 11. At the instant *t*, these nodes are arranged as described in Fig. 1. If we suppose that node 0 is the source and node 11 is the destination, the set of possible paths between the source and the destination at the instant *t* are described in Figure 1. Indeed, the possible paths are the following:

$P_0(t) = \{0, 1, 4, 8, 11\}$
$P_1(t) = \{0, 1, 3, 5, 9, 11\}$
$P_2(t) = \{0, 2, 3, 5, 9, 11\}$
$P_3(t) = \{0, 2, 6, 7, 10, 11\}$

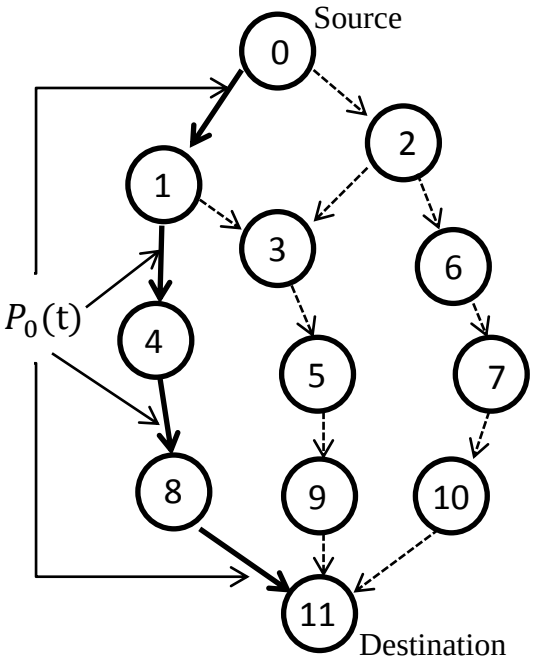


Figure 1. Set of possible paths between the source and the destination

So, the set of all the available paths at the instant t ($E_{ch}(t)$) is equal to:

$$E_{ch}(t) = \bigcup_k P_k(t) = P_0(t) \cup P_1(t) \cup P_2(t) \cup P_3(t)$$

Thus, we propose an algorithm which allows us to find out the set of possible paths between the source and the destination at any instant t belonging to the set of instants of topology change E_{inst}.

1) Presentation of the proposed algorithm

The goal of the proposed algorithm, which we call $ConsPaths$, is to collect and save all the possible paths between any two nodes from the network. For that, the algorithm starts from the source and browses the links between neighbors until reaching the destination. Once arriving to the destination, each traversed path is added to the set of possible paths. To do so, we associate four parameters to the algorithm *ConsPaths*. The first parameter is the global variable $\boldsymbol{G_{SD}}$ which will contain all the available paths. The parameters $\boldsymbol{n_i}$ and ***D*** will represent respectively the source and the destination. The fourth parameter ***P(t)*** is a temporary variable containing the current path which we are discovering. The variable ***P(t)*** is reset each time we start discovering a new path (see algorithm).

The principle of the algorithm $ConsPaths$ is to record all the possible paths between the source and the destination in the global variable $\boldsymbol{G_{SD}}$. For that, we use the temporary variable ***P(t),*** which allows to keep each discovered path so that we can add it to the global variable $\boldsymbol{G_{SD}}$. So, ***P(t)*** contains temporarily the path which we are discovering. Indeed, it is initiated with the node source. Then, each node n_i having a direct link (neighbor) with the source node is added to a separate initiated variable ***P(t)*** (We will have an independent temporary variable ***P(t)*** for each neighbor of the source node). After that, each node n_j having a link with the node n_i is added to a separate *P(t)* and so forth until reaching the destination or reaching a node which doesn't have neighbors. In the first case, when we reach the destination, we add the recorded path *P(t)* to the set $\boldsymbol{G_{SD}}$. In the second case, when we reach a node which hasn't neighbors, we discard the current content of *P(t)*, since it does not have a path leading to the destination.

Variables of the algorithm ConsPaths :

$\boldsymbol{G_{SD}}$**:** Global variable which contains the set of all the possible paths.
S : Source node.
D : Destination node.
$\boldsymbol{n_i}$***:*** A node **i** belonging to the network.
P(t) : Temporary variable containing the path which we are discovering.
CP(t) : Intermediate variable containing the current content of ***P(t)***.

 ISBN: 978-1-61208-418-3

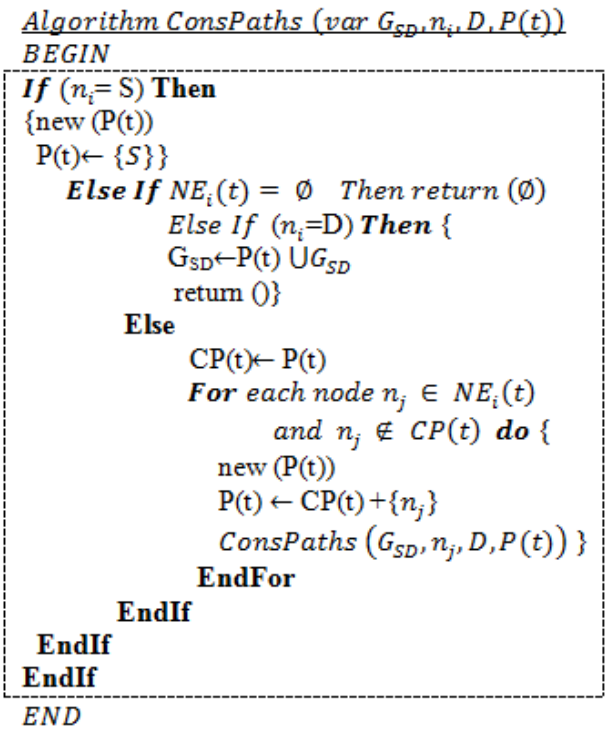


Figure 2. Algorithm of paths construction in the case of periodic movements

This algorithm is executed at each instant of topology change belonging to the time interval *[0, T]*. So, we should determine these instants.

2) The instants of topology change:

To identify the instants of topology change in the time period between 0 and T, we will use property 2 which allows checking if the neighborhood (and consequently the topology) has been changed.

Property 2: We identify t_1 as the instant of neighborhood change, if the neighborhood of at least one node n_i from the network changes compared with the last instant of neighborhood change t_2. Thus, if t_2 is an instant of neighborhood change, then the next instant of neighborhood change will be t_1 if and only if:

$$For\ t_1 > t_2, \exists i / NE_i(t_1) \neq NE_i(t_2) \qquad (10)$$

where $NE_i(t)$ is the set of the neighbors of the node n_i at the instant *t.*

Based on property 2, we can check if a given instant is an instant of topology change. So, it remains to determine the instants we should verify if there are instants of topology (neighborhood) change.

3) The instants of verification of neighborhood change:

The neighborhood of a node may change in three cases. The first case: If the node moves and its position changes. The second case: If one of its neighbors moves and comes out from its range. The third case: if a new neighbor comes into its range. All those cases depend on nodes movement. Thus, the instants of verification may be identified according to the nodes movements which are characterized by the speed and the running distance.

Since we are interested in the worst case, we will consider the minimum running distance which can change the neighborhood of a given node regarding the predefined trajectories ($Distance_{min}$) and the maximum speed on the network ($Speed_{max}$). Therefore, the minimum time after which the network topology may change (t_{min}) is the quotient of the division of the minimum running distance by the maximum speed ($Speed_{max}$) (as in (11)).

$$t_{min} = \frac{Distance_{min}}{Speed_{max}} \qquad (11)$$

Therefore, if t_{i-1} is the precedent instant of verification, the next instant of verification t_i will be (as in (12)):

$$t_i = t_{i-1} + t_{min} \qquad (12)$$

B. Paths prediction in the case of non-periodic movements

In the case of non-periodic movements, we cannot find a time period in which the same network behavior is repeated. Thus, we cannot predict all the instants of topology change. In this case, we can only define the set of possible paths between the source and the destination regardless of the time. For that, we use the property of the possible neighbors (property 3).

Property 3: We define $M_i(t)$ as the position of the node n_i at the instant *t*, NP_i as the set of possible neighbors for the node n_i and *R* the transmission range of the nodes. A node n_j can belong to the set of possible neighbors of the node n_i (NP_i) if and only if there are two instants t_1 and t_2 such as the distance between the position of the node n_i at the instant t_1 ($M_i(t_1)$) and the position of the node n_j at the instant t_2 ($M_j(t_2)$) is inferior to the range of the nodes (*R*) (as in (13)):

$$n_j \epsilon\ NP_i\ if\ and\ only\ if$$
$$\exists\ t_1, t_2 / \left| M_i(t_1) - M_j(t_2) \right| < R \qquad (13)$$

Based on property 3, we can find out the possible paths between a source and a destination. Thus, the algorithm which allows getting the possible paths in the case of non-periodic movements ($ConsPathsNP$) has the same principle as that of the periodic movements except that it does not depend on the time. It allows only getting the set of possible paths according to the possible neighbors from the source to the destination.

Variables of the algorithm ConsPathsNP :

G_{SD}: Global variable which contains the set of all the possible paths.
S: Source node.
D: Destination node.
n_i: A node **i** belonging to the network.

 ISBN: 978-1-61208-418-3

P: Temporary variable containing the path which we are discovering.
CP: Intermediate variable containing the current content of ***P***.

```
Algorithm ConsPathsNP (var G_SD, n_i, D, P)
BEGIN
If (n_i = S) Then
{new (P)
 P← {S}}
    Else If NP_i = ∅  Then return (∅)
         Else If (n_i=D) Then {
         G_SD←P ∪G_SD
         return ()}
       Else
            CP← P
            For each node n_j ∈ NP_i
                  and n_j ∉ CP do {
              new (P)
              P ← CP+{n_j}
              ConsPathsNP (G_SD, n_j, D, P)}
             EndFor
       EndIf
 EndIf
EndIf
END
```

Figure 3. Algorithm of paths construction in the case of non-periodic movements

After finding out the possible paths between the source and the destination, we should analyze the response times of these paths so that we can get the response time (end-to-end transmission delay) in the worst case.

IV. RESPONSE TIME ANALYSIS

If we assume that the longest path has the largest response time, the transmission delay of the longest possible path between two nodes will be the maximum delay and consequently the transmission delay of the worst case [12].

In the case of periodic movements, the set of possible paths is the union of the sets of possible paths at the instants of topology change. So, if we have the maximum transmission delay of each instant of topology change, the worst case delay between the source and the destination will be the maximum of these delays. Thus, $TR_{E_{ch}}$, which is the response time between two nodes of the network in the worst case, is given by the following formula:

$$TR_{E_{ch}} = \max_{t\epsilon[0,T]} TR_{E_{ch}(t)} \quad (14)$$

where $TR_{E_{ch}(t)}$ is the response time in the worst case at the instant of topology change *t*. Hence, its value is the maximum of the response times of the discovered paths at the instant *t*. $TR_{E_{ch}(t)}$ is given by the following formula:

$$TR_{E_{ch}(t)} = \max_k TR_{P_k(t)} \quad (15)$$

where $TR_{P_k(t)}$ is the response time of the path $P_k(t)$ discovered at the instant of topology change *t*.

The response time $TR_{P_k(t)}$ for the path $P_k(t)$ is the sum of the elementary transfer times between two each neighboring nodes which belong to the path, as shown in the following formula:

$$TR_{P_k(t)} = \sum_{(N_i,N_{i+1})\in P_k(t)} ETT\ (N_i, N_{i+1}) \quad (16)$$

where $ETT\ (N_i, N_{i+1})$ is the elementary transfer time between the neighboring nodes N_i and N_{i+1}. It depends on the number of neighbors (According to Opt/TDMA [13]).

In the case of non-periodic movements, the response time between a source and a destination in the worst case is the maximum of the response times of all the possible paths (as in (17)) :

$$TR_{E_{ch}} = \max_k \mathrm{TR}_{P_k} \quad (17)$$

The response time TR_{P_k} of the path P_k is the sum of the elementary transfer times between two neighboring nodes which belong to the path as in (16). However, it doesn't depend on time.

Thus, we have determined the response time in the worst case between two given nodes from the Ad hoc network. In the following section, we will evaluate the results of this analysis by comparing them to the simulation results.

V. EVALUATION

To evaluate our delay prediction, we have performed simulations on NS2 simulator [14] for an Ad hoc network having mobile nodes with closed trajectories. Then, we compared the end-to-end delays obtained from simulations with the worst case delays obtained from our analysis approach (see previous section).

A. Simulation scenarios

The simulated Ad hoc network is composed of twenty nodes randomly positioned in a 1000×1000 area and moving with cyclic movements. In this network, we use Time Division Multiple Access (TDMA) [15] as access method and RT-DSR [2] as routing protocol.

To perform the simulations, we created three traffics which are transmitting in the network and we are interested to the worst case end-to-end delay of the first real-time traffic.

B. Variation of the transmission delay over time

First, we observed the variation of the end-to-end transmission delay when the mobile nodes of the simulation are moving with a radius which is equal to 100 meters. Then, we calculated the worst case end-to-end delay with our analytic approach and we compared the obtained value with the simulation delays. The results are represented in Figure 4.

 ISBN: 978-1-61208-418-3

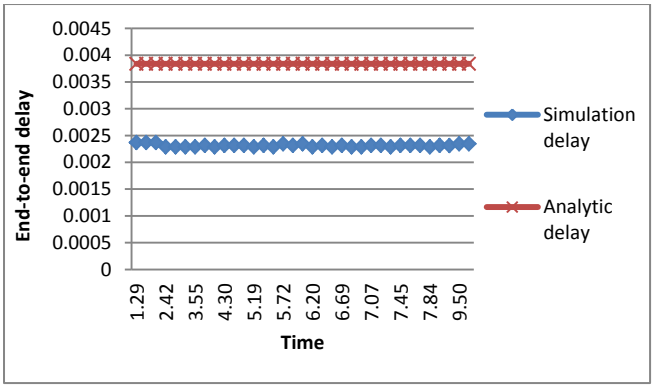


Figure 4. Variation of the transmission delays over time

We note, according to Figure 4, that all the simulation delays are inferior to the delays calculated by our analytic approach. This is due to the fact that our analytic approach gives us the end-to-end delay in the worst case which is a specific case and it may not occur in the simulations. So, we conclude that our analytic approach allows getting an upper bound of end-to-end transmission delays. Thus, if a delay deadline is superior or equal to this boundary, we can be sure that it will be respected.

C. *Variation of the worst case transmission delays according to the moving radius*

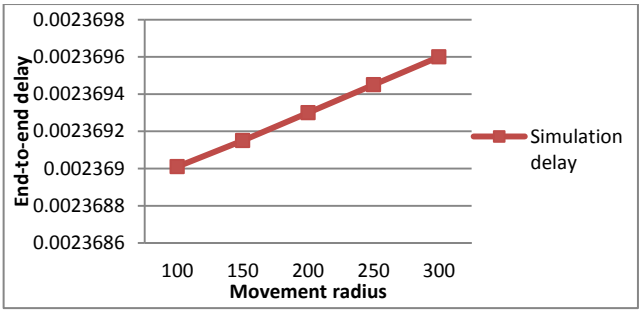
Then, we were interested in the variation of the end-to-end delays according to the moving radius of the mobile nodes. For that, we present in Figure 5 the maximum transmission delays obtained with the simulator NS2 (a) and the comparison between those simulation delays and the transmission delays calculated with our analytic approach (b).

Figure 5 shows that the worst case end-to-end delay increases when the moving radius of the mobile nodes increases in the simulation results as well as in the analytic results. However, the slope of the results calculated by our analytic approach is higher.

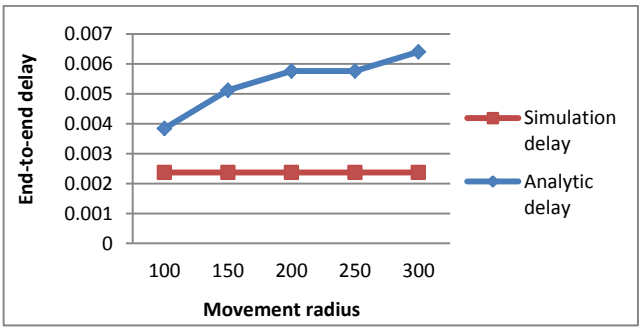
These results show also that the worst case analytic delays are always superior to the simulation delays with a small difference which proves the effectiveness of our approach in estimating the transmission delays.

VI. CONCLUSION

In this paper, we proposed a delay prediction approach for Ad hoc networks. In fact, we presented an algorithm which allows getting the worst case end-to-end transmission delay. By comparing the delay values calculated by our approach with the simulation results obtained from NS2 simulator for the same network, we conclude that the simulation delays are always inferior to the calculated worst case delays with a small difference. This proves the effectiveness of our approach which allows deciding if we can satisfy a required deadline for real time traffic.



-a- Variation of the maximum transmission delay in the simulations according to the moving radius of the nodes



-b- Comparison between the simulation transmission delays and the analytic transmission delays according to the moving radius of the nodes

Figure 5. Variation of the worst case transmission delay according to the moving radius of the nodes

REFERENCES

[1] J. Wang, "Exploiting Mobility Prediction for Dependable Service Composition in Wireless Mobile Ad Hoc Networks", IEEE Transactions on services computing, vol. 4, no. 1, January-March 2011.

[2] S. Ouni, J. Bokri, and F. Kamoun, "DSR based Routing Algorithm with Delay Guarantee for Ad Hoc Networks", Journal of Networks, Academy Publisher, vol. 4, iss. 5, July 2009, pp. 359-369.

[3] H. Mehdi, "Mobility Prediction With LLT Algorithm In Wireless Networks", International Conference on Information, Networking and Automation (ICINA), 2010.

[4] J. Huang and Y. Bai, "A Novel Approach of Link Availability Estimation for Mobile Ad Hoc Networks", Vehicular Technology Conference, 2008. VTC Spring 2008. IEEE, May 11-14, 2008.

[5] M. Chegin and M. Fathy, "Optimized Routing Based on Mobility Prediction in Wireless Mobile Adhoc Networks for Urban Area", Fifth International Conference on Information Technology: New Generations , Las Vegas, Nevada, April 2008.

[6] S. M. Mousavi, H. R. Rabiee, M. Moshref, and A. Dabirmoghaddam, "Model Based Adaptive Mobility Prediction in Mobile Ad-Hoc Networks", Wireless Communications, Networking and Mobile Computing, 2007 (WiCom 2007), September 2007, pp. 1713 – 1716.

[7] Z. H. Mir, D. M. Shrestha, G. Cho, and Y. Ko, "Mobility Aware Distributed Topology Control for Mobile Multi-hop Wireless Networks", ICOINS 2006, LNCS 3961, 2006, pp. 257-266.

[8] A.S. Poznyak and K. Najim, "Learning Automata and Stochastic optimization", New York: Springer, 1997.

[9] S. M. Ousavi, H. R. Rabiee, M. Moshref, and A. Dabirmoghaddam , "Mobility Aware Distributed Topology Control in Mobile Ad-Hoc Networks with Model Based Adaptive Mobility Prediction", Wireless and Mobile Computing, Networking and Communications, 2007. WiMOB 2007, October 2007, pp. 86-86.

 ISBN: 978-1-61208-418-3

[10] B. H. Sathyaraj and R. C. Doss, "Route Maintenance using Mobility Prediction for Mobile Ad hoc Networks", Mobile Adhoc and Sensor Systems Conference, November 2005.

[11] L. M. Meng, J. X. Zang, W. H. Fu, and Z. J. Xu, "A Novel Ad hoc Routing Protocol Research Based on Mobility Prediction Algorithm", Wireless Communications, Networking and Mobile Computing, 2005, 23-26 September 2005.

[12] J. Bokri, S. Ouni, and F. Kamoun, "The Worst Path Estimation for Real Time Communications over Ad hoc Networks", The International Conference on Wireless Networks (ICWN'12), Las Vegas, Nevada, 16-19 July 2012.

[13] S. Ouni, J. Bokri, and F. Kamoun, "Opt-TDMA/DCR: Optimized TDMA Deterministic Collision Resolution Approach for Hard Real-Time Mobile Ad hoc Networks", WSEAS Transactions on Communications 11/2013, 2013, pp. 12(11):570-583.

[14] Z. Wu, "Network simulator 2 for wireless : My experience", Technical report, Rutgers University, 2007.

[15] I. Jawhar, and J. Wu, "Qos support in Tdma-based mobile ad hoc networks", Journal of Computer Science and Technology, pp. 20(6):797–810, 2005.

 ISBN: 978-1-61208-418-3

Advertising Method via Smart Device Based on High Frequency

Myoungbeom Chung
Division of Computer
Sungkyul University
Anyang City, Korea
e-mail: nzin@sungkyul.ac.kr

Green Bang
Dept. of Digital Media
Soongsil University
Seoul, Korea
e-mail: banglgreen@ssu.ac.kr

Ilju Ko
Dept. of Global Media
Soongsil University
Seoul, Korea
e-mail: andy@ssu.ac.kr

Abstract—**In this paper, we propose a high frequency-based advertising method using a smart device. This method supports the transfer of advertising content to the smart device user with no additional action or TV audio signal required to access that content. Because the proposed method uses the high frequencies of sound signals via the inner speaker of the smart device, its main advantage is that it does not affect the audio signal of TV content. Furthermore, this method allows large numbers of smart device users to see advertising content continuously and automatically. To evaluate the efficacy of the proposed method, we developed an application to implement it and subsequently carried out an advertisement transmission experiment. The success rate of the transmission experiment was approximately 97%. Based on this result, we believe the proposed method will be a useful technique in introducing a customized user advertising service.**

Keywords-Advertising method; high frequency; smart device; wireless communication.

I. INTRODUCTION

In recent years, advertising service technologies based on smart devices have been introduced gradually in line with performance improvements in smart devices. When feature phones first became popular, advertising technologies using Short Message Service(SMS) or Multimedia Messaging System(MMS) were the most prevalent. However, with the advent of smart devices, various technologies such as push services, location-based services [1], advertising using Quick Response code(QR code) or Near Field Communication(NFC) [2]-[4], and cross-media advertising [5] became the new trend. Most significantly, the mainly passive role of advertising services changed to a more active one.

Good examples of advertising technology based on smart devices include Fujitsu's new data transmission technology using video data [6] and KT Media Hub's matching cross-media advertising service. Fujitsu's advertising technology combines visible light communication technology and digital watermarks [7]. This service can send advertising data to smart devices via the device's camera by using lighter and darker parts of the full screen. The technology supports 16 bits of data per second. If smart device users want to access the advertising information, they need to perform a light detection action using the camera of the smart device. However, in the case of TVs, this method requires extra technology to insert the advertising data into TV content. Also, it is inconvenient for the user, who has to perform an advertisement detection action while watching TV.

KT Media Hub's matching cross-media advertising service can support event coupons or the transmission of additional information to users through the audio signal of TV content. With this method, when users want to access advertisements while watching TV, they only have to shake their smart device. The advertisements are then transmitted via the TV audio, and the smart device first analyzes the advertising information that has been sent from the server and then shows the advertisements to the user. The advantage of this method is that it does not need any process enabling it to utilize the audio of TV content for advertisement transmission, because it already uses inaudible audio information from TV content. However, even if there was no inaudible audio, this method could not support sending of advertising information to the user. In addition, it is prone to detection errors due to the shaking duration or shaking intensity of the smart device.

Therefore, in this paper, we propose a new advertisement transmission method using the high frequencies of inaudible sound signals, which is aimed at addressing the disadvantages of existing methods. The proposed method uses the 18kHz ~ 22kHz frequency of the audible sound range (20Hz ~ 22kHz). We used the 18kHz ~ 22kHz frequency because it is the defined audible sound range, despite the fact that most people could not listen to sound in this range. Furthermore, to prevent signal detection errors from surrounding noise, we used two sequential high frequencies as the audio signal of the TV content, and smart devices that detect the high frequencies receive the related advertising content from the advertising content server. Similar to Fujitsu's technology, the two high frequencies of the TV content are generated consistently during the time needed for advertisement transmission. High frequencies can then be added easily to the sound component of the TV content and can reach any smart device in an indoor space where people are watching TV. Also, because the high frequencies used in the proposed method do not influence the original sound of the TV content or vice versa, the proposed method can transmit advertising data to smart devices very accurately. To evaluate the performance of the proposed method, we devised TV content that included the high frequencies needed to supply advertisements, and we then developed the advertising service application based on a smart device using the proposed method. During this process,

 ISBN: 978-1-61208-418-3

we carried out performance evaluation tests according to frequency changes at high frequencies, and we then tested the advertisement transmission capability via distance. The success rate of the results in terms of frequency change was 97%, while the success rate in terms of distance was 98.5% within 5m. These results indicate that the proposed method will prove useful with regard to the effective transmission of advertising information based on smart devices used in an indoor space.

The present paper is organized as follows: In Section 2, we explain existing technologies that use high frequency for information transmission. In Section 3, we describe the general architecture of advertising information based on smart devices using high frequencies, and what methods smart devices use to handle high frequencies. In Section 4, we evaluate the performance of the proposed method, followed by a conclusion in Section 5 and by a future research in Section 6.

II. RELATED WORK

This section explains existing technology that uses high frequency to transmit information and data. Early researchers in this area used high frequency to trace the position of mobile phone users indoors [8][9]. In 2011, Bihler proposed a method of transmitting information to smart devices using high frequencies and Wireless Fidelity(WiFi). Bihler's method availed of two high frequencies, 20kHz and 22kHz [10]. According to this method, eight bits of data were sent within 208 ms and Hamming code schemas were used for error correction. However, because the method involved fast changes between those two high frequencies, the result was a noise familiar to many, and the attainable distance of the high frequencies was very short.

Lee then proposed smart phone user authentication using audio channels [11]. This method used high frequencies from 15,800Hz to 20,000Hz as authentication signals and implemented user authentication between smart devices and personal computers (PC). Because the method generated two high frequencies simultaneously, the spacing of each frequency was 600Hz too much. Table 1 below shows the assigned frequencies according to bit; using this method, two bytes of data can be sent in eight seconds. In Table 1, Lee's method generates a beep four times in eight seconds using assigned frequencies at one-second intervals, and the smart device, on receiving the signal, sends the received data to the authentication server.

While this method is more stable than Bihler's method in terms of data transmission accuracy, data transmission takes too long.

TABLE I. ASSIGNED FREQUENCIES FOR EACH HEXADECIMAL DIGIT

Digit	Frequencies (kHz)	Digit	Frequencies (kHz)
0 x 0	15.8, 18.0	0 x 8	17.0, 18.8
0 x 1	16.4, 17.6	0 x 9	17.6, 19.4
0 x 2	17.0, 18.2	0 x A	18.2, 20.0
0 x 3	17.6, 18.8	0 x B	15.8, 18.2
0 x 4	18.2, 19.4	0 x C	16.4, 18.8
0 x 5	18.8, 20.0	0 x D	17.0, 19.4
0 x 6	15.8, 17.6	0 x E	17.6, 20.0
0 x 7	16.4, 18.2	0 x F	15.8, 18.8

Another researcher, Chung, proposed a near wireless control technology using high frequencies [12][13]. In this method, he defined high frequencies as base signals and low latency as a control signal. Base signals comprise more than two high frequencies and note the existence of the control signal to the smart device. Then, because low latency accesses various control data by frequency value, low latency is generated by the Central Processing Unit(CPU) of the smart device. The distance of the control signal in this method is greater than in Bihler's method and the data transmission time is faster than in Lee's method. Thus, Chung's method can be used as a trigger signal for the effective transmission of advertising information.

III. ADVERTISING METHOD VIA SMART DEVICE BASED ON HIGH FREQUENCY

This section explains the general architecture relating to advertisement transmission to smart devices using high frequencies, in addition to the methods used by smart devices to process high frequencies. Figure 1 below represents the general architecture of the proposed method.

First, the smart TV pictured on the left in Figure 1 starts sending the first high frequency signal to the smart device (①), and the smart device verifies the presence of the first high frequency using Fast Fourier Transform (FFT) (②). Then, if the smart device detects the first high frequency consistently, it checks whether the number of first high frequency signals is ***k*** times over (③). Once this has been

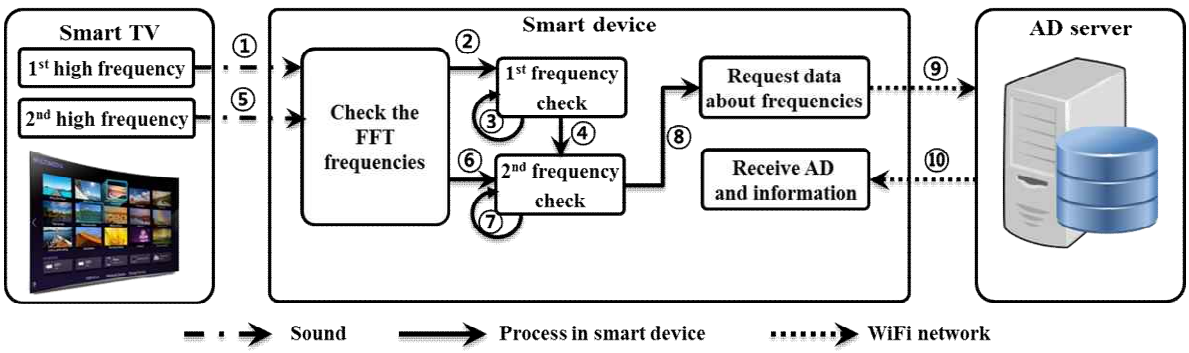


Figure 1. Work flow of advertising method via smart device using high frequencies.

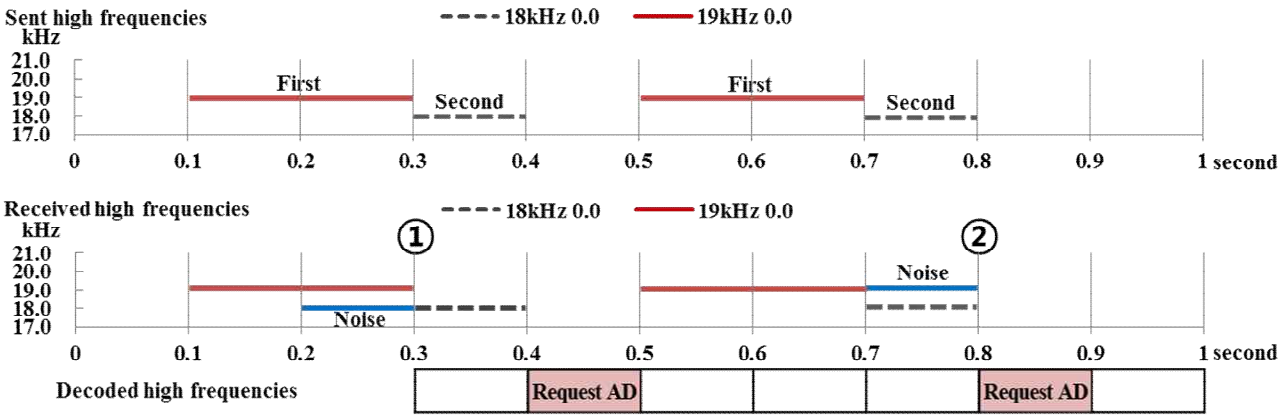


Figure 2. Example of high frequency use in advertising information requests

confirmed, the smart device waits to verify the presence of the second high frequency (④). When the smart TV sends the second high frequency signal to the smart device (⑤), the smart device verifies the signal's presence (⑥) and checks whether the number of second high frequency signals is ***i*** times over (⑦). If the smart device detects that the first high frequency signal is ***k*** times over and the second is ***i*** times over, it then requests and receives advertising information in the advertisement server via WiFi (⑧～⑩). At this point, the proposed method uses a high frequency at an inaudible frequency range, namely 19kHz～22kHz, as the first high frequency and 18kHz as the second high frequency. Figure 2 below shows how high frequencies are used in the proposed method. And then, the proposed high frequencies are contained in audio of TV contents by the broadcasting company.

Figure 2 shows how high frequencies are sent as advertisement transmission signals, with the first high frequency at 19kHz (0.2s) and 0.1s, and the second high frequency at 18kHz (0.1s) and 0.3s. The same high frequencies are sent again at 0.5s. In the case of the high frequencies being received by the smart device, the first high frequency is received at 19kHz (0.2s) and 0.1s, while the second high frequency is received at 18kHz (0.2s) and 0.2s. In fact, the 18kHz second high frequency (0.2s～0.3s) only amounts to noise. Nevertheless, the smart device can receive the advertisement signal accurately because it is still undergoing the process outlined in ③ in Figure 1. Furthermore, in step ② of Figure 2, the receiving high frequency results in noise at 19kHz from 0.7s to 0.8s. However, because the smart device has already completed process ③ in Figure 1 at 0.7s and is now undergoing process ④, it can receive the advertisement signal. The first high frequency can then use 31 signals from 19kHz to 22kHz by 100Hz each time. For example, the advertisement signal that uses 19kHz is different from the advertisement signal that uses 19.1kHz. The smart device, having detected the advertisement signal, sends the values for the advertisement signal to the advertising information server and the advertising information server then sends the related advertising content to the smart device.

IV. EXPERIMENTS AND EVALUATION

This section introduces the advertising application capable of supplying the advertising information to the smart device user by utilizing high frequencies as the proposed method. In addition, we explain the experiments and results used for the performance evaluation of the proposed method. We developed the advertising application based on iOS 6 and created it using Xcode 5. Figure 3 is a screenshot of the main advertising application.

Figure 3. Screenshot of the advertising application: (a) Main screen featuring a graph of the receiving high frequencies; (b) Real advertising information screen by advertisement signal.

 ISBN: 978-1-61208-418-3

In Figure 3 (a), the graph displays the FFT bin number relating to the first high frequency (red line) and the second high frequency (blue line) when the smart device receives the advertisement signal. The x axis of the graph indicates the time sequence, and the flow of time is from right to left. The y axis of the graph is the FFT bin number of high frequencies. Thus, we can see from the graph that the smart device received the advertisement signal four times and the order of receipt was ①, ②, ③, and ④.

Next, we evaluated the performance of the proposed method. We added an advertisement signal to the sound component of the TV content, which was a K-pop music video. We used 19kHz, 20kHz, and 21kHz as the first high frequency and 18kHz as the second high frequency. Each advertisement signal was generated twice at two-minute intervals. Thus, we did the receiving test 50 times for each advertisement signal. Because the advertising application checks high frequencies of FFT per 10ms, we set each threshold value in such a way that ***k*** was 12 times (60% of the first high frequency length) and ***i*** was six times (60% of the second frequency length). The distance between the smart TV and the smart device was 3m and the volume level of the smart TV was 70dB. The test space measured 5m × 4m, and 40dB was maintained as the silence space. Figure 4 below shows the result of the test using each advertisement signal.

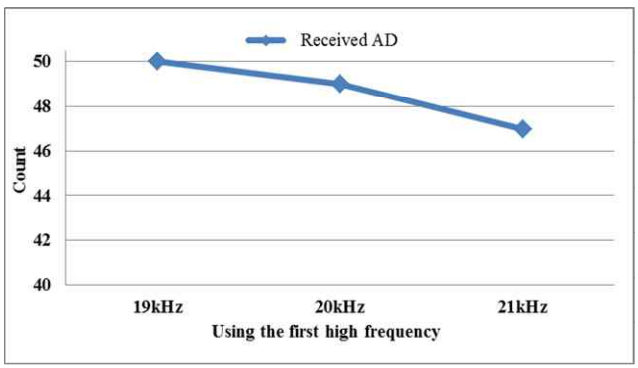


Figure 4. Results of the test using each advertisement signal.

In Figure 4, when the first high frequency was 19kHz, the smart device received the advertisement signal 50 times. Using 20kHz as the first high frequency caused the signal to be received 49 times, 21kHz caused it to be received 47 times, and so on. Thus, the total success rate averaged 97%. This shows that transmission errors increase in line with the rising value of the first high frequency. We believe the reason for this is that the bin number of the higher frequency detected was lower than the bin number of the lower frequency over the same distance and with the same number of decibels. The next task was to test performance according to distance. The distance between the smart TV and the smart device was from 1m to 5m in 1m intervals, and we did the receiving test 50 times with each advertisement signal (19kHz, 20kHz, and 21kHz) and at each distance. The test space and every threshold value were the same. Figure 5 below shows the results of the advertisement transmission test according to distance.

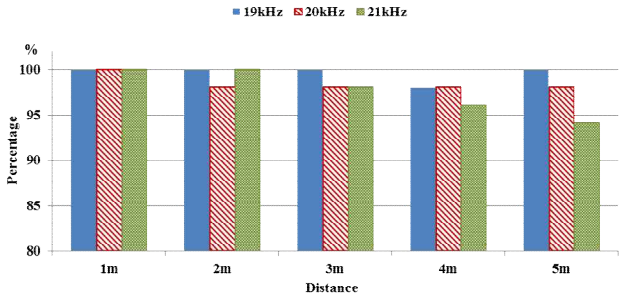


Figure 5. Result of advertisement transmission test according to distance.

Figure 5 tells us that the smart device received all of the advertisement signals at a distance of 1m. At a distance of 2m, the first high frequency of 20kHz failed to transmit the advertisement once. At a distance of 3m, the first high frequency of 20 kHz and 21kHz each failed once. At a distance of 4m, the first high frequency of 19kHz and 20kHz each failed once, and twice in the case of 21 kHz. At a distance of 5m, the first high frequency of 20kHz failed once and the first high frequency of 21kHz failed three times. Thus, transmission success rates were 99.3% within 3m and 98.5% within 5m. We think that the first high frequency transmission of 20kHz failed at a 2m distance because the ***k*** and ***i*** thresholds had not yet been optimized. Therefore, the solution to this error is to optimize each threshold. In addition, the 20kHz and 21kHz first high frequencies failed twice or three times at a 3～5 m distance. This was because the bin number of high frequencies decreased according to increasing frequency values.

V. CONCLUSION

In this paper, we proposed an advertisement method via smart device using high frequencies. This is a useful method that can support the transmission of advertising information to smart devices from a smart TV naturally and with no detection action required on the part of the user. In addition, because the method uses inaudible high frequencies, it does not influence the sound of the TV content, making it easy for the high frequencies needed for advertisement transmission to add to the sound of TV content. Therefore, the proposed method can support the transmission of advertising information better than existing methods and would work as an effective advertisement transmission technology for use indoors.

VI. FUTURE RESEARCH

In future research, we will study data and information transmission technology between smart devices using only high frequencies, as well as simultaneous data sharing technology among multiple smart devices indoors. We also aim to research performance improvement at high

 ISBN: 978-1-61208-418-3

frequencies, even in the case of greater distances from each smart device.

ACKNOWLEDGMENT

This research project was supported in part by the Ministry of Education under Basic Science Research Program (NRF-2013R1A1A2061478) and the Sports Promotion Fund of Seoul Olympic Sports Promotion Foundation from Ministry of Culture, Sports and Tourism, respectively.

REFERENCES

[1] S. Dhar and U. Varshney, "Challenges and business models for mobile location-based services and advertising," Communications of the ACM, vol. 54, no. 5, pp. 121-128, May 2011, doi:10.1145/1941487.1941515.

[2] Y. Liu, J. Yang, and M. Liu, "Recognition of QR Code with mobile phones," In Control and Decision Conference IEEE, July 2008, pp. 203-206, doi:10.1109/CCDC.2008.4597299.

[3] S. Narang, V. Jain, and S. Roy, "Effect of QR codes on consumer attitudes," International Journal of Mobile Marketing, vol. 7, no. 2, pp. 52-64, 2012.

[4] J. J. Sánchez-Silos, F. J. Velasco-Arjona, I. L. Ruiz, and M. A. Gomez-Nieto, "An NFC-Based solution for discount and loyalty mobile coupons," In Near Field Communication (NFC), 2012 4th International Workshop on, March 2012, pp. 45-50, doi:10.1109/NFC.2012.12.

[5] Samsung AdHub, Swingo. [Online]. Available from: http://www.samsungadhub.com/pr/ourCapability/swingo.do 2015.06.10

[6] Fujitsu Laboratories, Fujitsu develops new data transmission technology using video data, Diginfo TV. [Online] Available from: http://www.diginfo.tv/v/12-0183-r-en.php 2015.06.10

[7] Outstanding Technology, Visible light communication devices ready for commercialization, Wireless Technology Park 2012. [Online] Available from: http://www.diginfo.tv/v/12-0132-r-en.php 2015.06.10

[8] V. Filonenko, C. Cullen, and J. Carswell, "Investigating ultrasonic positioning on mobile phones," International Conference on Indoor Positioning and Indoor Navigation (IPIN), September 2010, pp. 1-8, doi:10.1109/IPIN.2010.5648235.

[9] V. Filonenko, C. Cullen, and J. D. Carswell, "Asynchronous ultrasonic trilateration for indoor positioning of mobile phones," In Web and Wireless Geographical Information Systems, April 2012, pp. 33-46. doi:10.1007/978-3-642-29247-7_4.

[10] P. Bihler, P. Imhoff, and A. B. Cremers, "SmartGuide–A smartphone museum guide with ultrasound control," Procedia Computer Science, vol. 5, pp. 586-592, 2011, doi:10.1016/j.procs.2011.07.076.

[11] M. K. Lee, J. B. Kim, and J. E. Song, "Smart phone user authentication using audio channels," In Consumer Electronics (ICCE), 2012 IEEE International Conference on, January 2012, pp. 735-736, doi:10.1109/ICCE.2012.6162060.

[12] M. B. Chung and H. S. Choo, "Near wireless-control technology between smart devices using inaudible high-frequencies," Multimedia Tools and Applications, pp. 1-17, 2014, doi: 10.1007/s11042-014-1901-x.

[13] M. B. Chung and H. S. Choo, "New control Method between smart devices using high frequencies," International Conference on Electronics, Information and Communication, January 2015.

 ISBN: 978-1-61208-418-3

MAP-Based Error Correction Mechanism for Five-Key Chording Keyboards

Adrian Tarniceriu, Bixio Rimoldi, Pierre Dillenbourg
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
e-mail: adrian.tarniceriu@epfl.ch, bixio.rimoldi@epfl.ch, pierre.dillenbourg@epfl.ch

***Abstract*—Because of different designs, different text input devices have different error patterns. If we consider these aspects when designing an error correction mechanism, we can obtain significantly lower error rates. In this paper, we propose and evaluate a spelling algorithm specifically designed for a five-key chording keyboard. It is based on the maximum a posteriori probability (MAP) criterion, taking into account a dictionary model and the probability that one character is typed for another. These probabilities are determined experimentally. For the considered evaluation text, the proposed method reduced the error rate from 10.11% to 2.17%. As comparison, MsWord and iSpell reduced the error rate to 5.15% and 6.69%, respectively.**

***Keywords*-error correction; chording keyboard; maximum a posteriori probability; confusion matrix**

I. INTRODUCTION

Most of us use mobile computing devices, such as smartphones or tablets, and would like to use them even more, but there are situations when we cannot easily access their services. For example, when walking in a crowded place, we should focus on what happens around us rather than on the mobile device. This limitation in usability is mainly due to the input interface, which requires visual commitment.

A way to reduce visual constraints while typing is given by chording keyboards. This type of keyboard enables users to generate a character by simultaneously pressing a combination of keys, similarly to playing a note or a chord on a musical instrument. For a keyboard with five keys, there are 31 combinations in which at least one key is pressed. This is enough for the 26 letters of the English alphabet and five other characters. If the keys are in a position that is naturally under the fingertips, a person can type using the fingers of one hand, without committing the eyes to the keyboard. Therefore, we will be able to use a mobile device even during activities for which vision is partially or entirely committed, such as walking in crowded spaces, jogging, or riding a bike (for example, we can place the keys around the phone, or on a bike handlebar).

The reason why chording keyboards are not popular is that they require training. Compared to a QWERTY keyboard, where users can "hunt and peck" from the beginning, for chording keyboards the mapping between keys and characters has to be learned before being able to type. The effort needed to do so depends on the keyboard type and mapping and can vary by several hours. In previous studies [1], we described a five-key chording keyboard and a key-to-character mapping that can be learned in less than 45 minutes. After 350 minutes of practice, the average typing rate was around 20 words per minute (wpm) with a maximum of 31.7 wpm, comparable to iPhone, Twiddler [2], or handwriting [3]. Without correcting the mistakes, the average character error rate at the end of the studies was 2.69%. Automatically correcting these mistakes will probably increase the keyboard's ease of-use and typing speed, because users will not have to stop typing in order to correct errors. In addition, typing in a dynamic environment will probably lead to more errors, so efficient error correction becomes even more important in these situations.

In this paper, we continue our work on error correction mechanisms for five-key chording keyboards [4][5]. Whereas our previous work only considered substitution errors (when a character is replaced by another character), now we will also consider errors such as deletions (when a character is omitted), insertions (when an additional character is inserted), or split or merged words. The correction mechanism is based on the maximum a posteriori probability (MAP) principle [6] and for each typed word, it provides a list of possible candidates and chooses the one that is the most likely. Moreover, it takes into consideration the particularities of the text input device. This is motivated by the fact that different devices lead to different error patterns, and knowledge about these patterns can be used to improve the error correction methods.

The paper is organized as follows. Section II presents a brief overview of existing text error correction mechanisms. In Sections III and IV, we describe the proposed error correction algorithm and the data set used for evaluation. Section V presents the error correction results. In Section VI, we conclude the paper.

II. RELATED WORK

Work on automatically correcting misspelled words in computer-typed text began in the 1960s [7] and the algorithms' efficiency has steadily increased since then. Nevertheless, the correction rates are still far below 100% and

 ISBN: 978-1-61208-418-3

improving them remains a challenge.

Traditionally, error detection and correction mechanisms functioned at word level. Non-words are identified in a typed text and the most likely corresponding words are suggested from a dictionary. The appearing errors are defined at character level and can be classified into three categories: deletions, when a character is omitted; insertions, when an additional character is inserted; substitutions, when a character is substituted by another character. Other, more complex, approaches take into account the context, grammatical and semantical rules, and also detect errors such as missing words, wrong phrase structure, misused inflections, or others.

A detailed overview of the commonly used correction techniques is presented by Kukich in [8]. Research in spelling error detection and correction is grouped in three main categories:

1) Non-word error detection:
   Groups of *n* letters (*n*-grams) are examined and looked up in a table of statistics. The strings that contain non-existing or highly infrequent *n*-grams are considered errors.
2) Isolated-word error correction:
   Each word is treated individually and considered either correct or incorrect. In the latter case, the incorrectly spelled word is compared to entries from a dictionary. Based on similarities between the typed word and dictionary words, a list of possible candidates is proposed. These candidates can be provided using several techniques:
   - minimum edit distance techniques consider the minimum number of editing operations required to transform a string into another. A basic example is to consider the dictionary word that can be obtained from the typed word with a minimum number of insertions, deletions, and substitutions;
   - similarity key techniques map each string to a key which is similar or identical for similarly spelled strings. In this way, the key for a misspelled string can point to similarly spelled candidates from the dictionary. The advantage of this approach is that the misspelled string is not compared to all entries in the dictionary;
   - rule-based techniques propose candidate words by using knowledge of the most common errors;
   - probabilistic techniques, which consider transition and confusion probabilities. The first ones provide the probability that a letter is followed by another given letter (the values are language dependent). Confusion probabilities estimate how often a letter is typed instead of another letter (the values are text-input device dependent);
   - among other possible methods, *n*-gram techniques and neural net techniques can also be efficiently used.

   Most isolated word error correction methods do not correct errors when the erroneously typed word is in the dictionary. For example, if *farm* is typed instead of *form*, no error will be detected. Moreover, these methods cannot detect the use of wrongly inflected words (for example, *they is* instead of *they are*).
3) Context-dependent error correction:
   These methods try to overcome the drawbacks of analyzing each word individually by also considering the context. Errors can be detected by parsing the text and identifying incorrect part-of-speech or part-of-sentence *n*-grams. Or, if enough memory and processing power are available, tables of word *n*-grams can be used. Other approaches consider grammatical and inflectional rules, semantical context, and can also identify stylistic errors.

Most of the methods presented above can be applied to any typed text, regardless of the input device. As various input techniques become more and more popular, the classic correction techniques have been improved to consider both the text and the device particularities. Goodman et al. [9] presented an algorithm for soft keyboards that combines a language model and the probabilities that the user hits a key outside the boundaries of the desired key. Kristensson and Zhai [10] proposed an error correction technique for stylus typing using geometric pattern matching. The T9 text input method for mobile phones can also be included here, as it considers the correspondence between keys and characters to predict words.

An error correction algorithm for chording keyboards is presented by Sandnes and Huang [11]. Firstly, they classify chording errors in three categories similar to the character errors: deletions, when the user does not press one of the required keys, insertions, when the user presses an extra key, and substitutions, when the user makes a mistake between adjacent fingers. Then, starting from the assumption that most words have very few errors, they describe an algorithm that can correct words that contain one deletion, insertion, or substitution.

III. ALGORITHM

The algorithm that we propose focuses on individual errors, without considering any contextual information, and is based on the maximum a posteriori probability principle. As we designed it to correct errors from text typed with a five-key chording keyboard, we will name it 5keys-MAP. The first part of the algorithm was presented in our previous work, but, to make this paper self-contained, we will also present it in the following.

 ISBN: 978-1-61208-418-3

A. MAP Algorithm

The starting point is the noisy channel approach [12], where the typing process is seen as sending information over a communication channel. The symbol at the channel input, x, is the word to be typed and the channel output, y, is what has actually been typed (Figure 1).

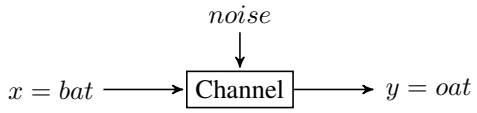


Figure 1. Typing seen as a sending information over a noisy channel

The MAP algorithm will find the string $\hat{x}$, which is the most likely in the sense of maximizing the posterior probability $p(x|y)$ over all $x \in \mathcal{S}$. The set $\mathcal{S}$ contains all the possible candidate strings. If we denote by $p(x)$ and $p(y)$ the distributions for the channel input and output, respectively, then

$$\hat{x} = \underset{x \in \mathcal{S}}{\operatorname{argmax}}\, p(x|y) \tag{1}$$

$$= \underset{x \in \mathcal{S}}{\operatorname{argmax}}\, \frac{p(y|x)p(x)}{p(y)} \tag{2}$$

$$= \underset{x \in \mathcal{S}}{\operatorname{argmax}}\, p(y|x)p(x), \tag{3}$$

where (2) follows from Bayes' rule.

Because our goal is to design a spelling algorithm, we can reduce the set of candidates from all possible strings to dictionary words. Then, assuming that the typing of each letter depends only on the intended letter and not on previous or successive letters, we can write

$$p(y|x) = \prod_{i=1}^{i=N} p(y_i|x_i)\,, \tag{4}$$

where y_i it the ith letter of the typed word, x_i is the intended letter, and N is the word length. The conditional probability $p(y_i|x_i)$ is the probability that the character y_i is typed in lieu of x_i. The prior probability, $p(x)$, is given by the frequencies of the dictionary entries in English language.

For example, given the typed word $y = oat$ and the candidate $x = bat$, we need to compute the posterior probability $p(bat|oat)$. This is proportional the product between the likelihood $p(oat|bat)$ and the prior probability $p(bat)$. We will denote this product by $F(oat|bat)$:

$$F(oat|bat) = p(oat|bat)p(bat) \tag{5}$$

$$= p(o|b)p(a|a)p(t|t)p(bat). \tag{6}$$

The prior probabilities, $p(x)$, were obtained from the British National Corpus, containing approximately 100 million words [13]. From this set, we only considered the items that appeared at least five times, obtaining a dictionary with $100\,944$ entries (including inflected forms, such as declensions and conjugations). The confusion probabilities, $p(y_i|x_i)$, were estimated experimentally.

The method described so far is the same as in our previous work on error correction, and can be applied to substitution errors, when the intended and the typed words have the same length. In the following, we will extend it to also consider error types such as missing or extra characters, or concatenated or split words. For this, besides the prior and confusion probabilities, we will use the probabilities that a letter is added or deleted from a word, the probability that a space character is added, and the probability that a space character between words is deleted. These values will be estimated experimentally from the same data set as the confusion probabilities.

B. Error Types

Before explaining the algorithm, it is useful to enumerate the errors that we will consider.

- Substitutions, when one letter is replaced by another (e.g., *hous**a*** instead of *hous**e***).
- Additions, when an extra letter is added to a word (e.g., *hous**a**e* instead of *house*). The probability to add a letter to a word is denoted as p_{Add}.
- Deletions, when a letter is missing from a word (e.g., *hous* instead of *hous**e***). The probability to delete a letter is denoted as p_{Del}.
- Extra space, when a word is split by an added space character (e.g., *hou se* instead of *house*). The corresponding probability is denoted as p_{SpAdd}.
- Missing space, when the space between consecutive words is missing (e.g., *thehouse* instead of *the house*). The corresponding probability is denoted as p_{SpDel}.
- Replacing a letter by a space (e.g., *ho se* instead of *house*).
- Replacing a space by a letter (e.g., *theohouse* instead of *the house*).
- Any combination of two of the above-mentioned errors.

C. Error Correction Algorithm

For every typed word, where by typed word we mean a set of letters separated by space characters, we will consider as candidates the words or bigrams obtained through the above mentioned operations. Then, we will determine the posterior probabilities using the MAP rule, and choose the most likely candidate. In case of substitutions, the algorithm is the same as in the previous subsection. For other error types, we also use the probabilities of the corresponding operations. In (8) and (10), we provide two examples for computing the posterior probabilities. As now we also analyze sets of two words, we need to know word bigram frequencies. These will also be estimated from the British National Corpus.

 ISBN: 978-1-61208-418-3

If we want to consider split words too, we have to go one step further and analyze groups of two consecutively typed words. For example, if the two words are *dictio* and *nary*, a candidate will be *dictionary*, and the posterior probability is given in (12).

The algorithm can be summarized in three steps, enumerated below. To make things clearer, we also provide an example, when the typed text is *"the dict ionary istoo heavty"*, and the intended text is *"the dictionary is too heavy"*.

1) Analyze each individual word:
   For each of *the*, *dict*, *ionary*, *istoo*, and *heavty*, we find the most likely candidates, named individual candidates, and the posterior probabilities. The results are given in lines 2 and 3 of Table I. We mention that the posterior probabilities from the table have been scaled to avoid working with very small numbers.

TABLE I. THE MOST LIKELY CANDIDATES AND THE POSTERIOR PROBABILITIES FOR EACH TYPED WORD AND FOR GROUPS OF TWO TYPED WORDS

Typed word	*the*	*dict*	*ionary*	*istoo*	*heavty*
Individual candidate	***the***	*diet*	*i nary*	***is too***	***heavy***
Posterior probability	99.99	0.66	0.01	0.86	0.45
Split candidate	*theodicity*		-		
Posterior probability	0.001		-		
Split candidate		***dictionary***		-	
Posterior probability		0.04		-	

2) Analyze groups of two consecutive words:
   At this step, we take groups of two consecutive words and check if they can be part of a split original word. The possible candidates (named split candidates) and the corresponding probabilities are provided in lines 4 - 7 of Table I. A "-" sign means that there were no candidates.
3) Decide if a word was split:
   The last step is to decide between the candidates from the previous steps, by comparing the individual and split probabilities from Table I. If the probability of the split candidate is higher than at least one of the individual probabilities, we decide that a word was split. The probability for *theodicity* is smaller than both the probabilities for *the* and *diet*, so we decide that there was no split. However, the probability for *dictionary* is higher then the probability for *i nary*, so we decide that for the typed bigram *dict ionary*, the intended word was *dictionary*. The candidates in bold font from the table are the most likely.
   Assume now that the probability of the individual candidate *diet* is 0.0001. This is lower than the probability for *theodicity*, but the typed word *dict* cannot be part of two split words. To solve this, we will assign it to the split candidate with higher probability, in our case this being *dictionary*.

IV. EVALUATION DATA

In order to gather enough data to evaluate the proposed algorithm, we asked 10 students from our university to type using a chording keyboard prototype. The prototype has the keys placed around a computer mouse and is presented in Figure 2. We designed the prototype in this way because we wanted the subjects to see a practical application of a chording device: allowing typing and screen navigation at the same time, with only one hand. The buttons are placed so that they can be easily operated while holding the mouse with the palm. The keyboard is designed using an Arduino Pro Mini microcontroller board and communicates with the computer by Bluetooth.

The participants were asked to type for 10 sessions of 30 minutes each, while sitting at a desk. Each session consisted of three rounds of 10 minutes, separated by breaks of two minutes. In the beginning of each round, the participants warmed up by typing each letter of the alphabet. During the warm-up, a help image showing the key combination for the letter to be typed was displayed. Afterwards, the help image was no longer available and the participants typed sentences chosen from a set considered representative for the English language [14]. These sentences were pre-prepared before the experiment to contain only small letters and no punctuation

$$F(housse|house) = p(housse|house)p(house) \tag{7}$$
$$= p(h|h)p(o|o)p(u|u)p(s|s)p_{Add}p(e|e)p(house) \tag{8}$$

$$F(thedob|the\ dog) = p(thedob|the\ dog)p(the\ dog) \tag{9}$$
$$= p(t|t)p(h|h)p(e|e)p_{SpDel}p(d|d)p(o|o)p(b|g)p(the\ dog) \tag{10}$$

$$F(dictio\ nary|dictionary) = p(dictio\ nary|dictionary)p(dictionary) \tag{11}$$
$$= p(dictio|dictio)p_{SpAdd}p(nary|nary)p(dictionary) \tag{12}$$

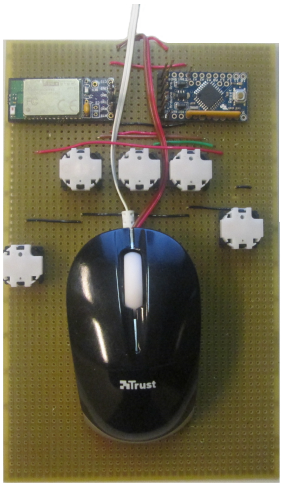


Figure 2. Chording keyboard prototype used during the typing study

signs (for example, *"February is the shortest month."* was changed to *"february is the shortest month"*).

A Java application was designed to display the text to be typed and to monitor the pressed keys. A screenshot of the application is shown in Figure 3. The top-left window contains the text to be typed and the bottom-left window represents the typing area. The help image is displayed on the right.

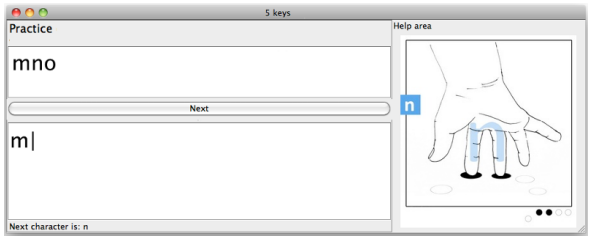


Figure 3. The interface of the Java application used during the study

Because we wanted to evaluate an error correction mechanism, we instructed the participants not to correct their mistakes (however, this was not enforced and they could delete typed text). As a reward for the time commitment during the experiment, they received a fixed monetary compensation for the first nine typing sessions. To provide additional motivation, for the last session, the reward was proportional to the number of typed words and to the typing accuracy.

The total amount of data gathered during the experiment consists of 40 640 words, of which 4109 (10.11%) contain errors. Of these, 3120 (75.93%) are substitution errors. The remaining 989 errors occurred when people did not type a letter (e.g., *hous* instead of *house*), typed an extra letter (*housee* instead of *house*), the space between words was missing (*thehouse* instead of *the house*), or when whole words were missing, added, or the topic of the sentence changed.

The total number of typed characters is 220 910, from which 6428 (2.91%) are errors. We used these characters to determine the confusion matrix, which is a square matrix with rows and columns labeled with all the characters that can be typed. The value at position *ij* shows the frequency of character *j* being typed when *i* was intended. The values are given as percentages from the total number of occurrences for character *i* and represent the confusion probabilities used by the algorithm.

V. Correction Results

The error correction mechanism was implemented in MATLAB and Python. To avoid overfitting, we used 10-fold cross-validation when determining the confusion matrix and evaluating the algorithm. As references, we used MsWord and iSpell. For each typed word, these algorithms return an ordered list of candidates. We considered the first one, which is the most likely, as the correction result. In addition, we compared the results with those from our previous work, which only focuses on substitution errors.

The results for both all-error-types and substitution-only scenarios are shown in Figure 4. The error rate after applying the algorithm is 2.17%, and, as expected, is lower than when considering only substitution errors (3.69%). For MsWord, considering all error types does not bring such a big improvement, the error rates being 5.15% and 5.57%, respectively. It is surprising that when we consider more error types, the error rates increase for iSpell. The most probable explanation is that for each typed word there are more candidates now, some of different lengths, and it is more difficult to choose the correct one. Or that more correctly typed words are changed by this algorithm.

In Table II, we show the initial error distribution and how many errors of each type were not corrected by the algorithm. It can be noticed that the correction method is less efficient for deletions, when a word has been split, and for combined error types. In the case of deletions, the typed

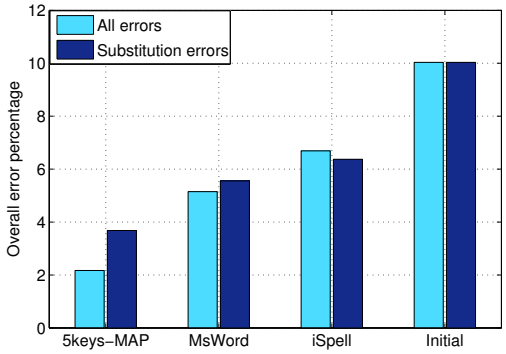


Figure 4. Overall error rates for the 5keys-MAP, MsWord, and iSpell algorithms, when considering all error types and only substitution errors, respectively

TABLE II. ERROR DISTRIBUTION FOR THE EVALUATION TEXT, BEFORE AND AFTER APPLYING THE CORRECTION ALGORITHM

Error type	Substitutions	Additions	Deletions	Extra space	Missing space
Before	3120	248	187	12	93
After	522	17	77	8	14
Error type	Letter → space	Space → letter	Combined	Other	Added errors
Before	134	156	130	29	
After	66	6	69	29	68

word length becomes smaller, and shorter words are usually more difficult to correct [5]. The same reasoning can be applied in the case of split words: the two components are considered as individual words, and their length is obviously smaller than the length of the correct word.

The results of the 5keys-MAP algorithm are clearly better than for MsWord and for iSpell. However, one should not forget that the dictionaries used by the three methods are not the same, and this can affect the results. Moreover, our algorithm is specifically designed for a five key chording keyboard, while MsWord and iSpell can be applied to any text input device with the same results.

VI. CONCLUSION

In this paper, we presented a MAP-based error correction mechanism for five-key chording keyboards. It is an isolated-word correction method, focusing on individual words without considering the context.

For the evaluation text, the algorithm reduced the error rate from 10.11% to 2.17%. This is more than two times lower than for MsWord (5.15%) and more than three times lower than for iSpell (6.69%). This advantage is due to the MAP algorithm, which takes into account the prior distribution of words and the device-dependent confusion probabilities.

The comparison between our algorithm, MsWord, and iSpell was done by only analyzing the first proposed candidate. We chose this approach because one possible use of chording keyboards is in dynamic environments, such as walking in crowded places or riding a bike, when users cannot continuously look at the typed text. Therefore, the error correction mechanism should run automatically, without requiring user supervision. In more static situations (for example when the keys are placed around a computer mouse), the most likely candidates can be displayed and the user will choose the desired one.

The correction algorithm was designed for a specific keyboard and mapping, but can be easily adapted to other input devices by updating the confusion matrix. One possibility to improve the algorithm is to implement an adaptive approach, starting with a general confusion matrix and update it based on what one types. Words that are typed more often can have their prior probability increased, becoming more likely than other candidates. Another natural improvement is to consider the typing context.

REFERENCES

[1] A. Tarniceriu, P. Dillenbourg, and B. Rimoldi, "Single-handed eyes-free chord typing: A text-entry study," International Journal On Advances in Intelligent Systems, vol. 7, no. 1,2, pp. 145–155, 2014.

[2] K. Lyons et al., "Twiddler typing: one-handed chording text entry for mobile phones," Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04), Vienna, Austria: ACM, 2004, pp. 671–678.

[3] I. S. MacKenzie and R. W. Soukoreff, "Text Entry for Mobile Computing: Models and Methods, Theory and Practice," Human-Computer Interaction, vol. 17, pp. 147–198, 2002.

[4] A. Tarniceriu, B. Rimoldi, and P. Dillenbourg, "Error correction mechanism for five-key chording keyboards," The 7th International Conference on Speech Technology and Human-Computer Dialogue (SpeD '13), Cluj-Napoca, Romania, October 2013.

[5] A. Tarniceriu, B. Rimoldi, and P. Dillenbourg, "Fine-tunning a map error correction algorithm for five-key chording keyboards," ECUMICT 2014, Lecture Notes in Electrical Engineering, L. De Strycker, Ed. Springer International Publishing, vol. 302, pp. 153–165, 2014.

[6] S. Kay, Fundamentals of statistical signal processing: estimation theory. Prentice-Hall, 1993.

[7] C. R. Blair, "A program for correcting spelling errors," Information and Control, vol. 3, no. 1, pp. 60 – 67, 1960.

[8] K. Kukich, "Techniques for automatically correcting words in text," ACM Comput. Surv., vol. 24, pp. 377–439, 1992.

[9] J. Goodman, G. Venolia, K. Steury, and C. Parker, "Language modeling for soft keyboards," Proc. of the 7th International Conference on Intelligent User Interfaces (IUI '02), New York, NY, USA: ACM, 2002, pp. 194–195.

[10] P.-O. Kristensson and S. Zhai, "Relaxing stylus typing precision by geometric pattern matching," Proc. of the 10th International Conference on Intelligent User Interfaces (IUI '05), New York, NY, USA: ACM, 2005, pp. 151–158.

[11] F. Sandnes and Y.-P. Huang, "Non-intrusive error-correction of text input chords: a language model approach," Annual Meeting of the North American Fuzzy Information Processing Society, 2005 (NAFIPS 2005), June 2005, pp. 373 – 378.

[12] M. D. Kernighan, K. W. Church, and W. A. Gale, "A spelling correction program based on a noisy channel model," Proc. of the 13th Conference on Computational Linguistics - Volume 2 (COLING '90), Stroudsburg, PA, USA: Association for Computational Linguistics, 1990, pp. 205–210.

[13] URL: http://www.kilgarriff.co.uk/bnc-readme.html [accessed: 2015-03-10].

[14] I. S. Mackenzie and R. W. Soukoreff, "Phrase sets for evaluating text entry techniques," Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI '03), Fort Lauderdale, Florida, United States: ACM, 2003, pp. 766–767.

 ISBN: 978-1-61208-418-3

DDA_{AV} - Student Performance Detector

Andreia Rosangela Kessler Mühlbeier
UFSM – Universidade Federal de Santa Maria
Santa Maria - Brasil
andreiamuhlbeier@yahoo.com.br

Aderson de Carvalho
UFSM – Universidade Federal de Santa Maria
Santa Maria - Brasil
acarvalho@inf.ufsm.br

Fabiana Santiago Sgobbi
UFRGS – Universidade Federal do Rio Grande do Sul
Porto Alegre - Brasil
fabianasgobbi@gmail.com

Roseclea Duarte Medina
UFSM – Universidade Federal de Santa Maria
Santa Maria - Brasil
roseclea.medina@gmail.com

Liane Margarida Rockenbach Tarouco
UFRGS – Universidade Federal do Rio Grande do Sul
Porto Alegre - Brasil
liane@penta.ufrgs.br

Abstract - The accelerated development and increasing use of virtual learning environments (VLE) motivates a transformation in education. This article presents a study of techniques and data mining tools (MD), which aims to research and analyze the student's behavior in the virtual environment, real-time execution of a course, providing feedback to the teacher about the students' academic performance encouraging its participation to improve the performance and preventing the circumvention of the course. The results obtained from the research of tools and techniques demonstrate that it is possible to obtain these inferences during periods.

Keywords – data mining; student achievement; weka; knowledge discovery in databases.

I. INTRODUCTION

The progress and the widespread use of technologies open up new perspectives in terms of classroom teaching, semi-classroom and distance learning. With support tools and access through mobile devices, the Virtual Learning Environments (VLE) are highlighted with great expansion in the educational process. However, virtual learning environments bring a transformation in education, allowing greater interaction in the environment among students, teachers, tutors, content and interfaces; this fact takes interactions as an effective part of the learning processes [1].

The data store large volumes of information in environments that are very rich sources of knowledge which end up being left out, sometimes by lack of knowledge in how to interpret them. [2]

In this context, the process of evaluating the student's performance in virtual learning environments is held at the end of disciplines, scoring the cognitive performance of an often static way, practically without time for actions of retroactive recovery of this student. However, the need for these evaluation and monitoring actions during the course it is evident, to propose alternatives for its best achievement, in order to generate subsidies for early identification, in time to successfully complete the student's learning process.

This research aims to research and analyze the student performance in the virtual learning environment, using data mining (MD) methods, through efficient techniques for Knowledge Discovery in Databases (KDD) in the information stored in the database, providing perform the mapping of student performance, in real-time execution of the course. With this mapping, the teacher will have a feedback that may assist in stimulating the participation and improving the performance of the student in the course.

This paper is organized this way: Section II presents the virtual learning environments. Section III describes the theoretical basis for knowledge discovery in databases. Section IV describes the WEKA tool and its relevant characteristics. Section V presents the related work. Section VI presents the development of the research methodology. Section VII presents the Student Performance Detector (DDA_{AV}), considering its pedagogical and technological aspects, and results. Section VIII presents the conclusions of this paper.

II. VIRTUAL LEARNING ENVIRONMENTS

The Virtual Learning Environments (VLE) are softwares applications on web servers; they have a set of tools which allow the creation of courses and the development of learning. These environments often classify their users in three pre-set profiles: Administrator, Teacher and Student. However, it is valid to mention that there is another profile

 ISBN: 978-1-61208-418-3

as the Tutor, who works with the teacher, being responsible for pedagogical mediation [3].

One environment widely used in educational spaces is Modular Object-Oriented Dynamic Learning Environment (MOODLE); it is an Open Source software, which had its development started in the 1990s by Martin Dougiamas based on learning philosophies of constructivism and social constructivism, supporting the creation and management of courses with a focus on collaborative work and in a simple and intuitive environment to use [4].

Amid other existing free softwares applications which allow the teaching/learning process, it was opted for MOODLE to conduct this research. The choice is given by virtue of being an environment with lots of tools constantly updated where there is a broad group of users who collaborated with their evolution, in addition to integrating other techniques in their repositories.

III. KNOWLEDGE DISCOVERY DATABASES

With the advancement of computer technologies which allows the storage and processing of a large volume of data, new technologies have been developed to assist in the extraction of information from these databases, through techniques such as Knowledge Discovery in Databases (KDD) and Data Mining (MD) [5].

The Knowledge Discovery process in Databases (KDD) presented by Fayyad [6], is "a non-trivial process of identifying valid standards, unknown, potentially useful and interpretable." It is to discover useful knowledge to the stored data from the application of modern techniques of data mining, evaluation of achieved standards and the interpretation of results.

The KDD process involves complex steps and each one must be performed carefully as it is very important that the established objectives and the overall success of the application are achieved. The steps are divided into: Pre-processing, data mining and post-processing [6].

A. Pre-processing

This stage is the identification and understanding of the problem, considering aspects such as goals and the data sources of which you want to extract knowledge. The next step is the selection of data from the sources, according to the objectives of the process, and the processing of data in order to be subjected to the methods and tools, the standards extraction phase.

B. Post-processing

At this stage, the extracted knowledge is evaluated about its quality and/or utility so it can be used to support a process of decision-making, whether by a human expert or an expert system.

C. Data Mining

The data mining phase is the central step that runs the knowledge discovery itself; its algorithms are the responsible to produce, semi-automatically, knowledge from existing data.

The MD process covers data selection, preparation, implementation tasks and/or techniques with their algorithms to make analyzes of the results in order to detect the extracted knowledge. The MD is divided into:

a) Association.
b) Classification.
c) Estimate.
d) Segmentation.
e) Summarization.

These tasks are performed by implementing algorithms with machine learning techniques such as:

a) Genetic algorithms.
b) Decision trees.
c) Association rule discovery.
d) Based reasoning cases.
e) Neural networks.

For the present study, the classification task that performs the decision tree technique, through the J48 algorithm [7] was chosen. The choice reflects the objective of searching, through the conditions offered to the 2 techniques, and identifying the student's performance with the result of some parameters. The decision tree respects a hierarchical test sequence, constructed along a tree structure with leaf nodes representing classes, where the algorithm expresses the rule through the path of the tree, from the root until a leaf node. The J48 algorithm is the ranking algorithm, implemented in Java, from the algorithm C 4.5 release 8; it builds a decision tree model based on a set of attributes, and it uses this model to classify instances in a cluster [7].

IV. WEKA TOOL

With the growing number of digital information, the interest in implicit knowledge discovery of informations grows. According to [8][9], there are some features that should be considered to choose a knowledge discovery tool.

a) Ability to access a variety of data sources, being online and offline.

b) Ability to include data models, object-oriented or non-standard models.

c) Processing capacity related to the maximum number of tables, records, or attributes.

d) Variety of attributes' types which can manipulate the tool.

e) Type of query language.

The Weka tool was developed at the University of New Zealand, in the Department of Computing. This tool uses

 ISBN: 978-1-61208-418-3

techniques to perform the following data mining tasks [10]: association, classification and clustering.

Mining begins by reading data from a file formatted especially for the tool, the ARFF (Attribute-Relation File Format). The ARFF is a text file that describes a list of instances which share a set of attributes [10].

In Weka, there is a variety of techniques for the listed tasks, such: ID3, PRISM, OneR and Naive Bayes [7].

The choice of the tool WEKA for this work is justified because it makes the system portable and it presents a cross-platform object-oriented language. The portability of language allows the tool to run on different platforms, and its object orientation produces advantages such as modularity, polymorphism, encapsulation, code reuse among others [11].

V. RELATED WORK

The work of Maia et al. [12] focuses on the future performance of students in disciplines of an undergraduate degree, are made from the grades achieved in subjects taken already. In this model, students and course subjects were modeled as nodes and their representation as the edges that make up a graph. The authors reported that, among the subjects there is a large variation in the values of the average errors analyzed, ranging from 3.6% to 100%. However, the authors conclude that a significant mean error for a discipline could indicate: that it does not have great relationship with the other subjects in the curriculum, or the assessment has some degree of disconnection with the results obtained in other disciplines.

In [13], to see high rates of dropout students in distance courses, one through an interview fieldwork was performed with a professional distance education, to identify some evidence of evasion courses. Based on the identified attributes, a prototype was designed to identify with the user log records stored in the database, information from these students. The work follows the KDD online database and used the WEKA tool, in particular the J48 algorithm that identifies behavioral prediction by the decision trees show. The author concludes the research, saying it can be identified through access to AVA, use patterns and certain diagnoses with evasion evidence thus propose corrective measures to ensure that a pass student to have a material behavior in the use of a VLE.

Accordingly, VLE [14] used to support classroom courses, are characterized by storing a large volume of data. These environments need tools to filter useful information to detect student performance. The research investigated the data stored in the VLE to extract information related to student performance. To detect this information was necessary to select a set of attributes, considering three dimensions: usage profile of VLE, student-student interaction and two-way interaction student-teacher. The form used RandomForest [7] and MultilayerPerceptron [7] ranking algorithms available in the Weka tool is pointed out that in all the experiments we used the method K-fold Cross-Validation [7] as data layering technique. The results of using the MD techniques on the selected set of attributes demonstrated that it is possible to obtain inferences regarding student performance with overall accuracy rates ranging from 72% to 80%, but leaves specific that the accuracy rate may be insufficient to evaluate the quality of the classification model, since the number of instances of classes is unbalanced in the case study, due to each being in different scenarios.

No analyzes focusing on student performance in the virtual learning environment and real-time course of execution were not identified in the current researches. However, there are indications that this type of analysis is important for the teacher to assist in stimulating participation and improvement in students' learning performance in MOODLE.

VI. RESEARCH METHODOLOGY

The nature of the research is ranked as a field research of qualitative-descriptive type. According to Lakatos and de Marconi [15], a field survey aims to obtain information about an issue, for which it is searched for a response in order to discover the relation between them.

In the first stage of development, to the data mining application in VLE, a literature research was performed in order to have knowledge of how the knowledge discovery in databases was conducted to understand and analyze the operation of data mining steps (tasks and techniques) and the functionality of available data mining tools.

The second stage involved two moments: the assembly of a hardware infrastructure that supported the installation, development and implementation of this work, consisting of a Dell Power Edge T300 server, with Intel Xenon Quadcore X3363 2.83 GHz processor with 4 physical cores and four virtual cores, 8GB of RAM, two 500GB hard drives and Windows Server 2008 operating system 64-bit. On this server, the following programs were installed: WampServer version 2.2, which provides softwares on its package that is necessary for the operation of MOODLE, as Apache version 2.2.22 server; the database MySQL version 5.5.24; PHP version 5.2.13 and phpMyAdmin version 3.4.10.1. After, the installation of MOODLE VLE version 2.5.2 was done. For the development, editing, and environmental manipulation, a Philco laptop with Intel Pentium Dual-Core processor was used, SU 4100 1.3GHz, 2GB of RAM, a 320GB hard drive and the operating system Windows 7 Ultimate 64-bit.

After the installation of MOODLE environment, to compose the research scenario it was worked with the database of the discipline Introduction to the Media Integration in Education. This which composes the curricular base of the course Specialization in Media in

 ISBN: 978-1-61208-418-3

Education, Post-Graduate *lato sensu*, of Universidade Federal de Santa Maria (UFSM), offered in distance education mode, during the second half of 2012. The course integrates in this edition 134 (one hundred and thirty four) students divided into five (5) campuses (Cachoeira do Sul, Cruz Alta, Panambi, Restinga Seca and Santana do Livramento) and it is composed by 10 activities.

In the third stage, the process of modeling the functioning of the block began. The proposed model was executed with the Astah Community tool, which allows the construction of Unified Modeling Language (UML) [16] diagrams, such as use case diagrams, activity diagrams, among others. The Astah Community [16] is afree modeling software for object-oriented systems design, based on the diagrams and in the notations of UML, which can generate code in Java language.

The fourth step involved the installation of WEKA version 3.7.8 tool, developed in the programming language JAVA, which offers several pre-data processing algorithms and results analysis. In the software, files were generated in the extension (* .arff) with their respective rules, to run the J48 algorithm. This algorithm allows the construction of decision trees that classify and display in their branches the most relevant attributes, as name, campus, discipline, notes of the activities and situation.

In the fifth step, it was executed the translation of the rules generated in WEKA software with the extension (* .arff) for the PHP language. The informations were extracted from the MOODLE environment database, in the form of an Excel electronic spreadsheet (note, campus and situation). After, they were processed in WEKA tool, which originated a file in notepad generated in the extension (* .arff). As a result, the generated file has been converted to PHP programming language [17] through the software PHP Editor.

In the sixth stage, the construction of the block was performed, receiving the number of any proposed activities in the discipline to be analyzed, and the integration of it in the MOODLE learning environment. The block which was developed, works by a plug implemented through an application, allowing its application in the environment's interface.

In the seventh step, tests were performed to validate the integration in each development stage, through the white box test (test performed by the developer). According to Sommerville [18], the tests are derived from the knowledge about the structure and implementation of the software, in other words, the developer seeks to test and to know all the system code, by examining the logical path to verify the operation of the tool. For this development, the following was used: basis path testing - aims to check if each statement of the system was performed at least once during the testing activities; condition testing - is based on verify if all logical conditions contained in the system, i.e., common error condition such as parentheses, relational operators and arithmetic expressions [19].

The first test was made after the generation of rules in the format (* .arff), when the consistency of J48 algorithm was verified. The second test was done after translating the rules to the PHP language. In the final test, the block was validated after its integration into the MOODLE environment.

With the active plug-in, the teacher tells the number which corresponds to the activity proposed in the discipline and in response a report is submitted as a web page, informing only the students who obtained low performance. The result is stored in tables in the SQL - based database of MOODLE environment with information about the activities.

Finally, in the eighth stage, monitoring reports were generated about student performance (creation with PHP language) and creation of the decision tree and graphics (WEKA software).

VII. DDA_{AV} – STUDENT PERFORMANCE DETECTOR

The environment aims to detect the student's performance during the execution of the course, using data mining technique in each executed and evaluated activity.

The research scenario consists of the discipline "Introduction to the Media Integration in Education", with 134 students divided into five campuses and subdivided into 10 activities with educational content in correlation to the main subject.

The relevant attributes extracted from the environment are: name, campus, discipline, and the grades of the activities performed by the students and registered by the teacher. In the context of the executed work, performance can mean the evaluation of student interaction in the environment, the learning level in the proposed activities, level of participation and level of difficulty in interpreting the task.

Figure 1 illustrates the course of the environment, displaying the proposed content and activities, the block "Student Performance" and the option of sending the number for any of the activities proposed in the discipline.

As output results, WEKA tool presented the attributes (name, campus, discipline, observation of the activities and situation) after running the J48 algorithm, and some other information relevant for classifying attributes of this work. The selected attributes totaled 134 cases, representing 100% of all the stored in the database; the "Properly Classified Instances" where the attributes were correctly classified contains 123 bodies which means 91.79%, and the "Incorrectly Classified Instances" shows the misclassification of 11 cases which is only 8.21%.

Figure 2 shown the performance of students in a report generated in the MOODLE environment. In the report, one can see the individual performance (student name) and general (discipline), with all the activities.

 ISBN: 978-1-61208-418-3

The proposed work is justified by the importance of the teacher to follow, throughout the course, the implementation process, avoiding a posteriori analysis, i.e., with the results, the teacher is able to give special attention to students with difficulties in constant learning, directly, without requiring more than a tool for analysis.

Figure 1. Environmental progress and block.

Figura 2. Performance report of students.

 ISBN: 978-1-61208-418-3

VIII. CONCLUSION

Currently, teachers and higher education institutions have faced a huge challenge which is to propose quality education and more individualized, to a growing number of students in various courses offered in different modes (presence, semi-presence and distance education). To assist in this process, VLE have been frequently used because they allow a greater control, various types of interaction, and the adoption of different methodologies and strategies. However, all this multiplicity and complexity of information can hamper the task of following and evaluating the student performance.

In this sense, the KDD process that aims to discover new knowledge, assists in the exploration of large volumes of data and detects useful information, through the application of techniques and tasks which implement MD algorithms.

The objective of this research was to apply data mining techniques in a VLE, presenting to the teacher a student's performance report during the execution of a course, which entails the prevention of the student's failure and consequently that occurs the evasion of the course. The report was extracted by means of integration of rules J48 classification algorithm, with the relevant attributes of the environment database.

The DDA_{AV} had as the research scenario, data from VLE MOODLE, which involves pedagogical processes between teachers and students in courses offered in virtual environments. The survey endorsed the difficulty of analyzing a large amount of data, which are available in the VLEs database and then highlighted the importance of using tools that help the teacher to follow the trajectory of the student, and monitor their performance within the course.

In the survey, the difference between the virtual environments of learning existing is that the DDA_{Av} has the advantage of the unification into a single report, information to the teacher about the trajectory of the student, consisting in a set of relevant data so that the teacher can prepare teaching strategies that meet the individual needs of students. Additionally, DDA_{AV} obtains semi-automatically, the variables "Enough" and "Insufficient" which characterize the performance of the student, for the inference of measures by the teacher.

REFERENCES

[1] R. Donnelly, “Interaction analysis in a 'learning by doing' problem-based professional development context”. Computers & Education, vol. 55 no. 3, p. 1357-1366, 2010.

[2] C. Romero, S. Ventura, M. Pechenizkiy, R. Backer, S. J. D., “Handbook of Educational Data Mining”, Ed. C R C, p. 535, 2012.

[3] A. P. Rodrigues, "Virtual Environment Integration with Digital Learning Repository" 2012. Thesis (Ph.D. in Education.) - Federal University of Rio Grande do Sul - UFRGS, Porto Alegre, p.188.

[4] MOODLE. “Statistics Documentation Moodle”. 2011. Available at: <http://docs.MOODLE.org/22/en/Statistics>. Accessed: Mar 2015.

[5] R. Goldschmidt, E. L. Passos, “Data Mining: um guia prático”. Rio de Janeiro: Elsevier, 2005. 2ª. Reimpressão.

[6] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, “The KDD process for extracting useful knowledge from volumes of data”. Communications of the ACM, New York, vol. 39, no. 11, p.27-34, 1996.

[7] WEKA. Waikato environment for knowledge analysis. 2013. Available at: <http://www.cs.waikato.ac.nz/ml/weka/>. Accessed: Abr 2015.

[8] M. Goebel, L. Gruenwald, “A survey of data mining and knowledge discovery software tools”. In: *SIGKDD Explorations*, June, 1999.

[9] L. A. Vieira, “Tools to estimate missing values in a database in the Pre-Processing Step of a KDD”. Work and Conclusion Course (Computer Science), University of Vale do Itajaí, 2008.

[10] I. H. Witten, E. Frank, M. A. Hall, “Data mining: Practical machine learning tools and techniques”. San Francisco: Morgan Kaufmann, 3 ed., 2011.

[11] D. Jacomini, “Entrants of Base Analysis in UNIDAVI”. Work Completion Course in Information Systems. New South Wales, in 2008.

[12] R. F. Maia, E. M. Spina, S. S. Shimizu, “System Student Performance Forecast for Assisted Learning and Course Rating”. Proceedings of the XXI SBIE -XVI WIE, 2010.

[13] C. S. de Afiune, “Educational Data Mining: Prediction Behavior in Distance Education Environments (DE)”. Term paper. State University of Goias, Anapolis, 2012, p. 108

[14] E. Gottardo, “Academic Performance estimation Students in A AVA using Data Mining Techniques”. Dissertation (Master of Applied Computing, Federal Technological University of Paraná (UTFPR), 2012, p.85.

[15] E. M. Lakatos, M. A. de Marconi, “Scientific Methodology fundamentals”. 5th . Ed . Editora Atlas . Faculty of Arts, 2003.

[16] ASTAH, “Astah Community”. 2010. Available at: <http://astah.change-vision.com/en/product/stah-community.html>. Accessed: Mar. 2014.

[17] C.A. J. Oliviero. "Make a site with PHP 5.2 MySQL 5.0, E-Commerce Driven project. "1st Edition. Ed. Erica Ltda. São Paulo, p.412, 2013.

[18] I. Sommerville, “Software Engineering”. Edição 6: Addison-Wesley, 2003.

[19] R. Pressman, “Software Engineering - A Professional Approach ”. 7th Edition, 2011.

 ISBN: 978-1-61208-418-3

Efficient Algorithms for Accuracy Improvement in Mobile Crowdsensing Vehicular Applications

Saverio Delpriori, Valerio Freschi, Emanuele Lattanzi and Alessandro Bogliolo

Department of Basic Sciences and Foundations
University of Urbino, Urbino, Italy
Email: saverio.delpriori@uniurb.it, valerio.freschi@uniurb.it, emanuele.lattanzi@uniurb.it, alessandro.bogliolo@uniurb.it

***Abstract*—Mobile crowdsensing is emerging as a cost-effective solution to conduct extensive monitoring campaigns by exploiting the potential of mobile terminals with unprecedented communication, processing, and sensing capabilities. On the other hand, data provided by a large number of end-users equipped with heterogeneous devices pose trust and accuracy issues that might impair the overall reliability and usability of the system. Trust management techniques developed in the field of online social networks can be effectively used to detect and isolate cheating users, but they cannot avoid the risk of inaccurate data provided by trustworthy agents because of the inherent limitations of their devices or of the adverse conditions in which they operate. This work presents efficient algorithms for compensating the inaccuracy of crowdsensing geospatial data to be reported on a road map. The paper illustrates on a representative case study the main issues of map matching and the effectiveness of the proposed solutions, which belong to the category of incremental online methods targeting dense sampling points to be mapped on road lines without topological annotations.**

Keywords–Map Matching; Crowdsensing; Vehicular Application.

I. INTRODUCTION

The widespread diffusion of smart mobile devices equipped with sensors and GPS receivers has enabled the development of crowdsensing applications engaging end-users in environmental/urban monitoring tasks. If, on one hand, crowdsensing has the potential of providing huge amount of data at negligible cost, it has some inherent limitations that need to be carefully addressed. First of all, end-users need to be motivated to take part in monitoring campaigns in order to make sure that a sufficient amount of data is provided. Second, misbehaviours (due either to unskillfulness or to cheating) need to be recognized and isolated in order not to affect the reliability. Third, data accuracy is not guaranteed because of the lack of control on the inherent accuracy of the devices and on their operating conditions.

The fast growing interest in crowdsensing has prompted for the development of a large number of approaches to address each one of the above mentioned issues, as documented in recent works surveying cooperation incentives [1], trust management [2], and crowdsensing data accuracy [3].

When dealing with geo-spatial data, accuracy is a two-fold issue, in that it concerns both the value of the sensed quantity and the GPS coordinates of the point in which it was measured. When the point needs to be associated with an object on a map, GPS accuracy impacts map matching [4], [5].

This paper focuses on efficient map matching algorithms for crowdsensing road applications. In particular, it provides a classification of map matching issues and proposes incremental real-time algorithms to tackle them and improve the overall mapping accuracy.

A mobile crowd-sensing application for road surface monitoring, called *Smart Road Sense* (SRS), is taken as a case study [6]. Among the wide range of vehicular applications that pose map matching issues, SRS has several peculiarities that give rise to new challenges: i) it requires fine-grained sampling, ii) it targets not only recognized main roads, but also not-yet-tagged road segments, and iii) it collects data from heterogeneous devices installed both on public and on private vehicles. Hence, map matching is addressed without relying on a pre-established trajectory (as in the case of public transportation), nor on the knowledge of roads and links among them. Rather, incoming data are sequences of points that need to be matched on road segments based only on geometrical considerations and on first-order reachability analysis performed on the fly. Moreover, we assume map matching to be applied in real time, so that computational complexity has to be kept under tight control, looking for a solution belonging to the category of incremental online methods.

Experimental results, validated against ground truth datasets collected for one month on known bus lines, show that map matching accuracy can be significantly improved by means of incremental linear algorithms applied on the fly without leveraging any topological information.

The rest of the paper is organized as follows: Section II provides a summary of related work on vehicular crowdsensing systems and map-matching algorithms; Section III introduces concepts and definitions related to the map-matching problem; Section IV describes the proposed approach; Section V describes the experimental setup and presents the results; Section VI concludes the work.

II. PREVIOUS WORK

In this section we report some recent scientific literature related to techniques, methods and systems used in the paper, namely: crowdsensing, system architectures for mobile sensing (with a particular focus on vehicular applications) and algorithmic approaches for map matching problems.

Crowdsensing denotes a whole class of applications executed by many individuals equipped with mobile devices featuring sensing, computational and communication capabilities. Modern smartphones are today commonly equipped with a wide range of sensors (e.g., GPS, accelerometers, microphones, cameras, etc.). This fact, combined to the growing diffusion of these devices and to the possibility of geo-

 ISBN: 978-1-61208-418-3

referencing collected data, makes them particularly suitable as an enabling technology for supporting mobile crowdsensing systems [7], [8].

A. Vehicular crowdsensing systems

One of the first applications of vehicle-based crowdsensing is represented by *CarTel*, a system developed with the aim of detecting road potholes by means of GPS and accelerometers mounted in cars equipped with embedded microprocessors [9]. Mobile sensing for the detection of traffic conditions, bumps and acoustic events has been investigated through the integration of data from accelerometers and microphones into a system called *Nericell* [10]. Thiagarajan *et al.* proposed in 2009 a system (named *VTrack*) targeting the goal of road traffic delays estimation by means of mobile phones [4]. A particular focus of this work is represented by the reduction of smartphone energy consumption during sensing activities and by the compensation of noise associated to sensors sampling. *CTrack* is a system developed by some of the authors of VTrack to achieve accurate trajectory mapping from GSM fingerprints instead of using WiFi signals or GPS traces [11].

Large-scale mobile sensing has been proposed for air pollution monitoring in *OpenSense* [12]. Data gathering in OpenSense is achieved by means of participative sensing of citizens equipped with enhanced modified smartphones or ad hoc pocket sensors, and by means of special sensor stations placed on public transport vehicles.

Recently, a system architecture for road surface collaborative monitoring (termed *SmartRoadSense*, SRS) has been proposed [6], [13]. SRS is designed to integrate mobile sensing and cloud systems to support continuous monitoring of road surface quality. A roughness index is computed on board of end-users' smartphones and then transmitted to be stored, processed, aggregated and visualized in cloud.

B. Map matching

Map matching is an inference process that reports a sequence of location data onto a map. In mobile sensing applications, data is typically a GPS trace (i.e., a sequence of time-stamped *(latitude, longitude)* values) and the map refers to annotated road networks in a digital georeferenced database. Sequence localization data can derive from different sources of information (e.g., GPS devices, GSM fingerprints, WiFi access points position) while target maps could differ in the information content and available details (e.g., topological information, one-way roads annotations, etc.).

Map-matching algorithms are classified into *global* algorithms and *incremental* algorithms. Global approaches [5], [11] process whole input traces in order to achieve a solution, while incremental algorithms [14], [15] work on small segments to be processed in sequential order. On one hand, global algorithms usually result into more accurate solutions but must be inherently run only offline; on the other hand, incremental algorithms make local choices that could possibly impact the accuracy of mapping but are suitable for online execution.

In this framework, many different techniques have been proposed to tackle the map-matching problem. Some authors investigated the use of computational geometry techniques posing the problem as a pattern matching between curves under Frechet distance [16], [17]. Others proposed the use of signal processing methods (e.g., Extended Kalman Filters) [18] or Bayesian estimators [15]. According to recent scientific literature [19], the best performances in terms of accuracy are obtained by algorithms based on *Hidden Markov Models* (HMMs). This statistical approach grants robustness to the matching algorithm, resulting into enhanced accuracy when faced with noisy inputs [5], [11]. A local, incremental variant amenable for an online implementation has been recently presented by Goh *et al.* [19].

Despite significant advancements obtained, the above mentioned algorithmic approaches present some issues with respect to the mobile crowdsensing scenario targeted by this paper. In fact, global solutions incur high computational overhead that make them unsuitable for vehicle-based applications where timeliness is often required. Furthermore, most state-of-the art methods make some assumptions on input data (e.g., the availability of topological information) which are not always guaranteed. Overcoming these issues is one of the main purposes of this work.

III. Problem statement and scope

This section formulates the map-matching problem addressed in this paper referring to known definitions [19] and to the terms adopted in OpenStreetMap [20].

Definition 1: A *trace*, $T = (t_n | n = 1, ..., N)$, is a sequence of N samples collected by a vehicle. Each *trace point* t_n is characterized at least by: time stamp (t_n.t), GPS coordinates (t_n.lat, t_n.lon), and sample (t_n.val). Additional fields can be available, like the GPS accuracy (t_n.acc) or the vehicle speed (t_n.v).

Definition 2: A *line*, $L = (p_m | m = 1, ..., M)$, is a M-point polyline representing a road segment as a series of segments connecting vertices p_1, ..., p_M in order. Each *vertex* p_m is represented by its coordinates (p_m.lat, p_m.lon).

Definition 3: A *map* S is a set of lines representing all the road segments of interest.

According to OpenStreetMap, we consider a *road* as a relation among lines, possibly annotated with viability information (speed limit, directions, permissions, etc.). A road network is a set of roads complemented by a connection matrix which adds topological information to the map. A road network is defined on a map and, in general, it covers just a subset of the lines of the map. The percentage of lines covered by the road network in a given region depends on the maintenance and updating of the underlying data base. Working on a map (rather than on a road network) maximizes the coverage at the cost of giving up topological information, which are usually exploited in map-matching algorithms.

Definition 4: Given a trace T and a map S, the goal of *map matching* is to find the correspondence between each point of T and a line of S.

In general, reducing the sampling rate (i.e., reducing the number and density of the points in the trace) makes it more difficult to determine the actual trajectory of the vehicle on the map. Hence, large research efforts have been devoted to the development of robust map matching solutions able to cope with sparse traces [4], [5]. In the limiting case in which there is less than 1 sample per line, the key problem is to figure out which path the vehicle took between two points. The topology of the road network is essential in this case.

 ISBN: 978-1-61208-418-3

On the other hand, challenging issues can be raised also by the abundance of points provided by high sampling rates. When the rate reaches the order of one sample per second, there are two new problems to face. The first one is accuracy, in that the distance between subsequent points in the trace might fall below the resolution of the GPS, causing many possible artifacts. The second one is performance, in that the sampling rate poses tight constraints on the time taken to process each sample in online real-time applications.

A. Map-matching issues

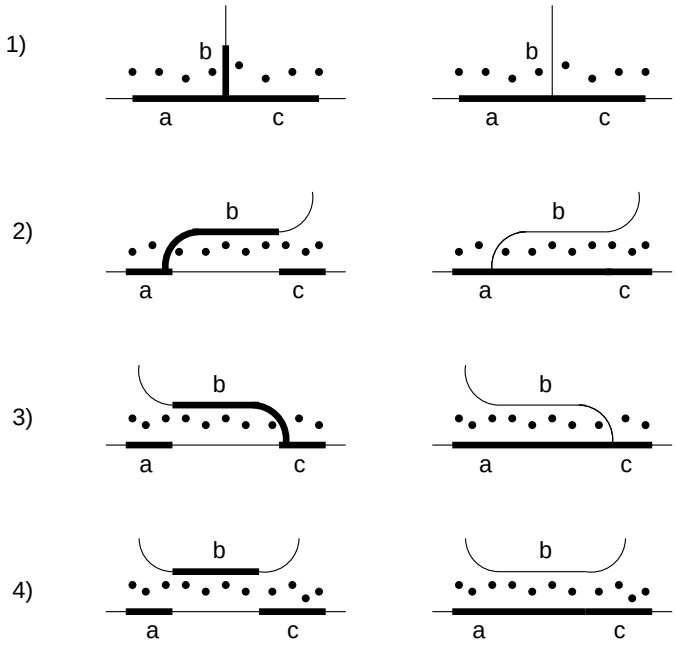


Figure 1. Schematic representation of the 4 main artifacts of map matching applied to dense traces. For each case the artifact is represented to the left and contrasted with the ground truth trajectory, to the right.

Figure 1 provides a schematic representation of the basic artifacts possibly produced by map matching algorithms fed by a dense trace. Lines (i.e., road segments) are denoted by labels a, b and c, while dots represent trace points. Bold lines on the left represent the false trajectory inferred by matching points on the closest lines, while those on the right represent the actual (ground-truth) trajectory. All the traces are assumed to go from left to right (meaning that points to the left have time stamps preceding those to the right) along a ground-truth trajectory that goes from a to c. Label b is used to denote a line which is not in the trajectory but passes so close to some of the trace points to induce mapping artifacts.

In $Case1$ line b is a crossing road orthogonal to a and c. The wrong trajectory looks as if the vehicle had taken the cross and then was immediately turned back on the main road.

In $Case2$ line b is one of the two roads of a fork which runs parallel to c for a non-negligible stretch. If, after the fork, most of the trace points are closer to b than to c, it looks as the fork was taken. But when b and c diverge, the remaining points are mapped back to c without a viable link.

In $Case3$ a fork similar to that of Case 2 is encountered in the opposite direction, so that the fake trajectory jumps from a to b as the two roads get close to each other, and then it rejoin the main road at the fork.

In $Case4$ some of the samples are mapped on line b even if it never encounters lines a and b.

All the artifacts possibly encountered when map-matching a dense trace can be expressed as a combination of the four listed above. Without loss of generality, in cases 1, 2, and 3 we assume the road to switch from line a to line b exactly when it crosses b. All other situations can be easily obtained by considering a=c. For the sake of explanation, the points in Figure 1 are represented as equidistant from each other and laying on an ideal trajectory, even if misplaced with respect to the map. Real traces look much worse, in that the points usually jump on both sides of the ground-truth lines and sometimes they overlap and locally reverse the order. However, persistent errors like the ones schematically represented in Figure 1 are the most difficult to detect and correct. It is also worth mentioning that, in the absence of viability or topological annotations, the artifacts of Figure 1 occur whenever there are lines that cross or get close to each other on the map, even if there are no paths between them in the road network.

The purpose of this work is to find an incremental map-matching algorithm able to detect and correct matching artifacts when dealing with dense traces in the absence of topological information.

IV. Proposed approach

The output of map matching is the association of each trace point to a line. We use t_n.line to denote the property of point t_n that represents its association to a given line. Hence, map matching reduces to setting all the t_n.line properties to the appropriate line id's.

Restricting the range of candidate solutions to the ones that can be executed incrementally in linear time, we start by matching each point to the closest line, as determined by issuing a query to the underlying geospatial data base. To speed up the refinement of this first-cut matching we make use of *run length encoding* (RLE) to represent contiguous subsets of trace points mapped on the same line [21].

Definition 5: A *run* $R = (r_k|k = 1, ..., K)$, is a sequence of K contiguous points taken from a given trace, such that all the points in R are mapped on the same line (denoted by R.line) and the points that precede and follow R in the trace (if any) are mapped on a different line.

Runs are computed on the fly based only on proximity matching, and then incrementally processed to correct artifacts. A sliding window of three runs, denoted by $previous$ (R_p), $current$ (R_c) and $next$ (R_n), is used to this purpose. At each step a decision is taken on the correctness of the matching of R_c, based on the consolidated matching of R_p and on the first-cut matching of R_n.

The decision process is based on the following conditions, which minimize the number of additional queries:

A) Reachability from previous run. R_c.line and R_p.line cross in a point close to the first point of R_c and to the last of R_p (only two points and two lines involved in the query);

 ISBN: 978-1-61208-418-3

B) Reachability of next run. R_n.line and R_c.line cross in a point close to the first point of R_n and to the last of R_c (only two points and two lines involved in the query);

C) Fork. R_p.line, R_c.line and R_n.line cross in the same point, close to the first and last points of R_c (only two points and three lines involved in the query);

D) Short run. The number of points in R_c is lower than a given significance length L (no geospatial queries required).

TABLE I. TRUTH TABLE OF THE ARTIFACT COMPENSATION ALGORITHM.

	A	B	C	D	Action
Case 1	T	T	T	T	split R_c between R_p and R_n
Case 2	T	F			merge R_c in R_p
Case 3	F	T			merge R_c in R_n
Case 4	F	F			split R_c between R_p and R_n

Table I shows the decision criteria used to detect the 4 cases of Figure 1 and the actions taken to correct each of them. Rows are associated with possible artifacts, while columns are associated with conditions. Empy entries represent don't care conditions.

In Cases 1 and 4, the current run (R_c) is split between the previous and the last ones. This is done by assigning R_p.line to the line property of the first points of R_c, and R_n.line to the remaining points of R_c. In Cases 2 or 3, the current run is merged enither in the previous one or in the next one.

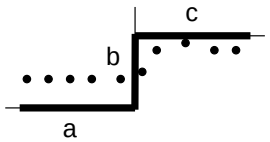


Figure 2. False positive case: actual trajectory that risks to be recognized as a case-1 artifact if the link road *b* is very short.

Looking only at the truth table, conditions C and D seem to be redundant. In fact, the two binary conditions A and B provide the 4 combinations required to encode the 4 cases of interest. However, C and D are needed to distinguish Case 1 from fals positives, like the one in Figure 2, which represents a short link road. Whenever none of the four artifacts is detected, the current run is assumed to be correct and its line labels are left unchanged.

V. IMPLEMENTATION AND EXPERIMENTAL RESULTS

This section presents the experimental results obtained by applying the proposed map matching algorithm to enhance the quality of the SRS road surface monitoring system.

A. Implementation and experimental setup

The architecture of SRS is composed of three components: a mobile application running on Android devices to compute every second a geo tagged estimate of a roughness index, a server that gathers roughness traces from all the end-user devices and map them on OpenStreetMap, and a cloud-based front-end for graphical visualization of geospatial data, based on CartoDB.

The server-side map matcher of SRS, called *Cartesian matcher*, associates each trace point to the closest line, according to the 2-dimensional Cartesian distance. The proposed algorithm is implemented as an additional component, called *match enhancer*, which operates on the runs of trace points annotated by the Cartesian matcher.

Data points are collected by a standard web service endpoint and stored in a PostGIS database system. Both the Cartesian matcher and match enhancer are implemented as PHP scripts, which operate on collected data by means of PostGIS stored procedures written in PL/pgSQL. Maps are taken from the OpenStreetMap project, while CartoDB is used to display data and double-check the correctness of the matches.

Although SRS was conceived as a crowdsensing application engaging end-users in road surface monitoring, the approach was validated on data collected for one month from Android devices installed on public buses operating on a known line in the Province of Pesaro-Urbino, in Italy. All the map elements along the actual bus line were manually sorted to build the ground truth baseline.

B. Results

Figures from 3 to 7 show the results of map matching in the four misleading cases targeted by the proposed approach. Each figure refers to a specific case and reports the trace points projected on the line segments in which they are mapped by the traditional Cartesian matching and by the proposed match enhancer. The effectiveness of the enhancer is apparent, in that the artifacts are completely cancelled. In particular, Figure 3 shows several examples of Case-1 artifacts, Figure 4 shows an example of Case 2, Figure 5 shows an example of Case 3, while both Figure 6 and Figure 7 provide examples of Case 4.

Two metrics can be used to estimate the accuracy of map matching relative to ground truth: the percentage of trace points mapped on lines that belong to the bus line (i.e., the correct segment identification rate [22]), and the ratio between the number of wrong lines and the number of ground-truth lines involved in matching (fake line ratio). Cartesian matching provided a correct segment identification rate of 97.15%, but in spite of the small percentage of wrong matches, the fake lines involved in matching exceeded the number of ground-truth lines (108 against 102), leading to a ratio of 1.06. The difference between the two metrics is explained by the nature of the trace and of the map: the trace includes extra urban roads with a limited occurrence of artifacts, while the map is highly fragmented, containing many unclassified very short segments, which are the most error prone ones.

The application of match enhancement provided sizeable advantages, increasing the correct segment identification rate to 99.10%, and the fake line ratio to 0.47.

VI. CONCLUSIONS

This paper has presented an incremental real-time algorithm to enhance the accuracy of map matching in crowdsensing applications dealing with dense traces to be mapped on maps with unknown topology. The issues raised by the abundance of trace points have been discussed and classified and efficient solutions have been proposed to handle them on the fly. The effectiveness of the match enhancing techniques

 ISBN: 978-1-61208-418-3

(a) (b)

Figure 3. Projected trace points plotted on top of the satellite map showing examples of Case-1: (a) before and (b) after match enhancement.

(a) (b)

Figure 4. Projected trace points plotted on top of the satellite map showing examples of Case-2: (a) before and (b) after match enhancement.

(a) (b)

Figure 5. Projected trace points plotted on top of the satellite map showing examples of Case-3: (a) before and (b) after match enhancement.

presented in this paper has been demonstrated on a real-world case study: a collaborative system for road surface monitoring.

Experimental validation performed on known trajectories shows that all types of artifacts are properly handled by the proposed algorithm with significant improvements of the overall matching accuracy. In particular, the percentage of correct matches increases from 97.15% to 99.10%, while the ratio between fake lines and correct ones reduces from 1.06 to 0.47.

REFERENCES

[1] L. Duan et al., "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in IEEE Proceedings INFOCOM, 2012. IEEE, 2012, pp. 1701–1709.

[2] T. French, E. Asimakopoulou, C. Maple, and N. Bessis, "Trust Issues on Crowd-Sourcing Methods for Urban Environmental Monitoring," Int. J. Distrib. Syst. Technol., vol. 3, no. 1, 2012, pp. 35–47.

[3] S. Sarma, N. Venkatasubramanian, and N. Dutt, "Sense-making from Distributed and Mobile Sensing Data: A Middleware Perspective," in Proc. of the 51st Annual Design Automation Conference, ser. DAC '14. ACM, 2014, pp. 68:1–68:6.

[4] A. Thiagarajan et al., "VTrack: Accurate, Energy-aware Road Traffic Delay Estimation Using Mobile Phones," in Proc. of the 7th ACM Conference on Embedded Networked Sensor Systems. ACM, 2009, pp. 85–98.

 ISBN: 978-1-61208-418-3

(a) (b)

Figure 6. Projected trace points plotted on top of the satellite map showing examples of Case-4: (a) before and (b) after match enhancement.

(a) (b)

Figure 7. Projected trace points plotted on top of the satellite map showing examples of Case-4: (a) before and (b) after match enhancement.

[5] P. Newson and J. Krumm, "Hidden Markov Map Matching Through Noise and Sparseness," in Proc. of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ser. GIS '09. ACM, 2009, pp. 336–343.

[6] G. Alessandroni et al., "SmartRoadSense: Collaborative Road Surface Condition Monitoring," in Proc. of 8th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, Ubicomm 2014. IARIA, 2014, pp. 210–215.

[7] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," Communications Magazine, IEEE, vol. 49, no. 11, 2011, pp. 32–39.

[8] N. D. Lane et al., "A survey of mobile phone sensing," Communications Magazine, IEEE, vol. 48, no. 9, 2010, pp. 140–150.

[9] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The Pothole Patrol: Using a Mobile Sensor Network for Road Surface Monitoring," in Proc. of the 6th international conference on Mobile systems, applications, and services. ACM, 2008, pp. 29–39.

[10] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: Rich Monitoring of Road and Traffic Conditions using Mobile Smartphones," in Proc. of the 6th ACM conference on Embedded network sensor systems. ACM, 2008, pp. 323–336.

[11] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden, and L. Girod, "Accurate, Low-energy Trajectory Mapping for Mobile Devices," in Proc. of the 8th USENIX Conference on Networked Systems Design and Implementation, ser. NSDI'11. USENIX Association, 2011, pp. 267–280.

[12] K. Aberer et al., "Opensense: open community driven sensing of environment," in Proc. of the 1st International Workshop on GeoStreaming (IWGS '10), 2010, pp. 39–42.

[13] V. Freschi et al., "Geospatial Data Aggregation and Reduction in Vehicular Sensing Applications: the Case of Road Surface Monitoring," in Proc. of 9th IEEE International Conference on Connected Vehicles, ICCVE 2014. IEEE, 2014.

[14] N. R. Velaga, M. A. Quddus, and A. L. Bristow, "Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems," Transportation Research Part C: Emerging Technologies, vol. 17, no. 6, 2009, pp. 672–683.

[15] O. Mazhelis, "Using recursive Bayesian estimation for matching GPS measurements to imperfect road network data," in 13th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2010, Sept 2010, pp. 1492–1497.

[16] S. Brakatsoulas, D. Pfoser, R. Salas, and C. Wenk, "On map-matching vehicle tracking data," in Proc. of the 31st international conference on Very large data bases. VLDB Endowment, 2005, pp. 853–864.

[17] D. Chen, A. Driemel, L. Guibas, A. Nguyen, and C. Wenk, "Approximate Map Matching with respect to the Frechet Distance," in Proc. of the 13th Workshop on Algorithm Engineering and Experiments, ALENEX'11. SIAM, 2011, pp. 75–83.

[18] D. Obradovic, H. Lenz, and M. Schupfner, "Fusion of Map and Sensor Data in a Modern Car Navigation System," Journal of VLSI signal processing systems for signal, image and video technology, vol. 45, no. 1-2, 2006, pp. 111–122.

[19] C. Goh et al., "Online map-matching based on Hidden Markov model for real-time traffic sensing applications," in 15th International IEEE Conference on Intelligent Transportation Systems (ITSC), 2012, Sept 2012, pp. 776–781.

[20] "OpenStreetMap," 2014, URL: http://www.openstreetmap.org [accessed: 2015-01-02].

[21] D. Salomon, Data compression: the complete reference. Springer, 2004.

[22] M. Hashemi and H. A. Karimi, "A critical review of real-time map-matching algorithms: Current issues and future directions," Computers, Environment and Urban Systems, vol. 48, 2014, pp. 153–165.

 ISBN: 978-1-61208-418-3

How Internet of Thing Makes the Energy Grid Smart

The rise of the Energy of Things

Giampaolo Fiorentino
Research & Development Lab
Engineering- Ingegneria Informatica
Rome, Italy
e-mail: giampaolo.fiorentino@eng.it

Antonello Corsi
Research & Development Lab
Engineering- Ingegneria Informatica
Rome, Italy
e-mail: antonello.corsi@eng.it

Pietro Fragnito
Research & Development Lab
Engineering- Ingegneria Informatica
Rome, Italy
e-mail: pietro.fragnito@eng.it

***Abstract*—The high penetration of intelligent appliances has turned whole buildings into effective and efficient prosumers. These distributed and autonomous intelligent Commercial Prosumer Hubs, constituted of Distributed Energy Resources (DER) clusters raising an actual decentralized Demand Side Management (DSM), behave like Smart Virtual Power Plants. An Aggregator manages the new Smart Virtual Power Plant enabling the electricity production and consumption to be measured, reported and controlled in real time. This new infrastructure maximizes the response capacity of the vast, small-commercial prosumer base (e.g., tertiary buildings, offices, etc.), presenting incentives and delivering benefits through their automated active participation in the energy market, aligning consumption by asking consumers to reduce their power usage rather than increasing the power generation facilities. Under this approach, prosumers that cooperate might receive incentive payments from the power company. In comparison, the Internet of Things (IoT) paradigm claims to solve this issue expanding Demand Response services based on the analysis of occupants historical interactions with the lighting, ventilation and air conditioning controls. In this respect, an overlay smart network for efficient grid control, running on top of the existing energy grid and incorporating high levels of distributed intelligence within autonomous and semantically enhanced Prosumer Hubs (local hub) will bring to the new concept of Internet Of Energy. This new smart network addresses the present structural inertia of the Distribution Grid by introducing more active elements combined with the necessary control and distributed coordination mechanisms, as well as Demand Side Management Operator.**

Keywords-Internet Of Things; Energy of Things; Smart Grid; Demand Side Management; Energy Flexibility;

I. INTRODUCTION

The great amount of intermittent renewable energy resources injected into electric power systems can significantly modify the net demand profile [1]. This recent phenomenon with the current rather inelastic nature of the demand curve can generate big issues on the electric grid, most important of which are frequency fluctuations and voltage imbalances [2].

These problems are found in the majority of current electric grids due to the large number of passive elements that constitutes today's electrical network.

Different approaches were proposed to overcome these problems [3].

The most common solutions are those that use the devices flexibility to balance the grid. Flexibility is the capability of shifting production or consumption of energy in time following an external signal, in order to provide a service within the energy system.

The problem with this approach is that flexibility is obtained compromising the final user comfort [4].

With this regard, the INERTIA project [5] has been thought to extend Demand Side Management strategies by incorporating a new entity: an enhanced Distributed Energy Resources. This new entity includes local generation and consumption capacity and will provide flexibility without impacting users comfort. Therefore, it addresses the present "structural inertia" of the grid by introducing more active elements combined with the necessary control and distributed coordination mechanisms.

This new kind of DER, semantically enhanced (generation, consumption and flexibility), is the core of our solution and will constitute an active and flexible knot equipped with local information based on environmental, occupancy, and historical data.

The adopted solution enables a mechanism that allows consumers to actively participate in the Demand Side Management without affecting the customer's comfort level, as well as turning the customer into an active and proactive *prosumer*. The entrance in the landscape of this new actor will lead to a most effective participation of all elements in the grid.

Of course, the best natural way to do this is following the *Internet of Thing Paradigm*.

According to this approach, every DER has a computational model and a communication part in embedded systems, that paves the way towards highly sophisticated networked devices that carry out a variety of tasks not in a standalone mode, as usually today, but taking into fully account dynamic and context specific information.

These DER "objects" are able to cooperate, share information, act as part of communities and generally be active elements of a more complex system. Our solution is a real instance of the so called Ubiquitous Computing: different systems and subsystems (each having its

 ISBN: 978-1-61208-418-3

computational capacity) that are driven simultaneously to cooperate with each other as in Figure 1.

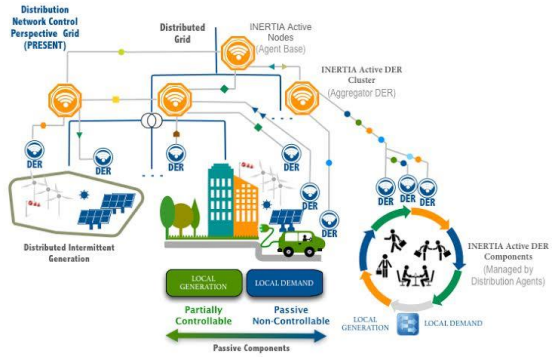


Figure 1. INERTIA framework

This paper learns from the progress of the existing efforts in demand side management that allows for intelligent demand aggregation and dynamic demand on much higher levels than those of individual appliances and is organized as follows. Section II describes the modeling devices that are the focus of INERTIA together with systems and subsystems, that are provided with different communication and integration protocols: DERs, Key Performance indicators, sensors, mobile devices with a specific user interface. Section III describes the approach followed to aggregate these device models in the whole system. Section IV describes the pilot DEMO implemented in Europe. At the end conclusions, close the article.

II. INTERNET OF THINGS REVOLUTION FOR ENERGY EFFICENCY IN SMART GRID

The Internet of Energy (IoE) is a new conception of the power grid that allows promoting a transition from the current energy system to a new modulated one. IoE provides an architecture with distributed embedded systems to implement a real-time interface between the smart grid (which depends of electrical generating energy sources but also of flexibility concept before explained) and a cloud of devices (electric vehicles, commercial and residential buildings, offices, electrical devices, appliances, etc.).

This provides the capacity to produce, store and use energy efficiently, balancing the supply and the demand using a cognitive Internet of Energy, which will harmonize the grid by processing data, information and knowledge through the Internet.

This innovation causes all network elements of a physical model to be instantly connected and able to contribute actively to the network to the purpose of DSM.

A. *DER flexibility modeling*

The DER models are aimed at simulating the individual and aggregated behaviour of the different devices/systems taking into account the complete operational context (environmental conditions, occupants, time, device operational characteristics, etc.). These will allow creation of multi-dimensional DER flexibility profiles reflecting the real-time load demand elasticity as a function of multiple parameters such as price and occupancy prediction.

To this purpose, DER modeling consists of two main parts:

- DER models: DER models contain the mathematical formulation defining the electric demand (consumption, generation and storage) of the DER in function of dynamic input parameters and static parameters (configuration) affecting DER's demand. For example: the DER model for a HVAC (Heating, Ventilating, and Air Conditioning) system contains the mathematical model that calculates the power consumption of the HVAC given the HVAC characteristics (rated power, efficiency, thermal characteristics of the building) and several inputs that change dynamically (temperature set point, outdoor temperature, occupancy, etc.).
- Control models: control models presented in this document refer to the local controllers associated to each controlled DER. The models described here represent the components that contain some local intelligence including the capability to simulate set point schedules provided by the centralized optimizer and by which of means it is possible to extract flexibility.

These models have the objective of simulating the final state of the device and are able to provide the forecasted flexibility. For example an HVAC DER model would have also as input the expected occupancy of a thermal zone and in case that this value is 0 it will tell us that all the energy forecasted will be available under the form of flexibility to the system. The main output of each DER model is the power consumption during the simulation time step.

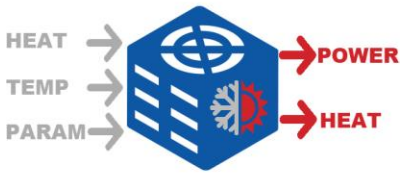


Figure 2. HVAC model

Other auxiliary outputs are also provided. These auxiliary outputs are needed in order to feed the model with input data for the next simulation step. Other key output of the models is the heat gains generated by the DER and the occupants. This heat gains are used together with the thermal models of building zone to calculate the power consumption of HVAC systems which are one of the main power consuming devices in buildings [6]. Inputs needed by each DER model are separated into configuration and dynamic type.

For the correct creation of a profile, we need to divide the DER models in four categories: local demand, generation, thermal zone and thermostatically controlled appliance [7]. While local demand and generation are models that simulate consumption and electricity production, a

thermal zone model represents the thermal losses and gains of a building area, which is controlled by a thermostat.

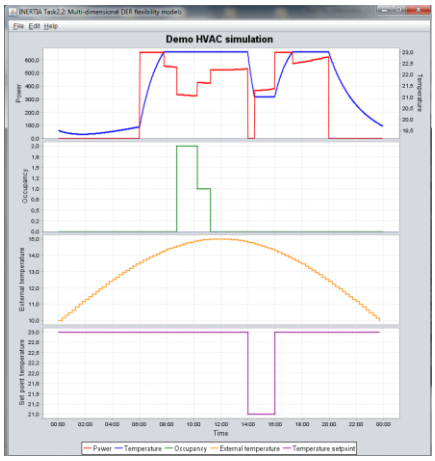


Figure 3. HVAC power consumption as parameter function

A thermal zone contains a set of construction elements describing how the heat is transferred from one area to another, also the loads and the occupants of the thermal zone need to be considered since they act as heat producing elements. All these elements together with the desired temperature set point in the thermal zone define the required parameters to obtain the heat demand that is used as input for the HVAC DER as in Figure 3.

At last, we have the thermostatically controlled appliances whose operations are driven by a temperature thermostat.

Thermostatically controlled appliances can be used to cool or heat a certain space or element. For a cooling appliance, when the temperature measured by the thermostat reaches the high temperature limit the appliance starts cooling and therefore consuming power and when the temperature reaches the lower limit the appliance stops cooling and stops consuming power. For a heating appliance,

the operation procedure is exactly the same but the consuming period is started at the low temperature limit and stops at the high temperature limit.

There are two main appliances operating as thermostatically controlled appliances

- Refrigerators/Freezers: their objective is to maintain the temperature between a certain temperature range where this temperature range is usually below the ambient temperature.

- Water Heaters: They are in charge of heating the water inside a tank maintaining it within a predefined temperature range.

To simulate a thermostatic device one important parameter to keep in mind is the percentage of consumption time (on time) with respect to the operating cycle (on + off times) that is called duty cycle and defines the power consumption profile of the appliance.

The shorter duty cycle time periods reflect low activity indicating that the appliance is only supplying the thermal losses through its shell. While the larger duty cycle time periods reflect increased heating/cooling demand indicating that the appliance is being used (hot water is being drawn from the water heater or food is being filled or removed from the fridge).

Depending on the usage of the appliance, the duty cycle is larger when the appliance is being used and shorter when the appliance is not being used.

This consideration leads to the approach that calculates energy consumption of the appliance, according to the usage patterns that may be inferred from occupancy data. This means that during time periods with higher usage of the appliance the on time periods will be larger than at times where the appliance is not used. Given therefore the energy consumption during the operating cycle, the duty cycle linked to different usage levels and the occupancy data is possible to model the appliance.

B. Sensor and Occupancy

The behavior of occupant has been shown to have large impact on building control and appliances consumption. Consequently, user activity and presence is considered as a key element and has been used for control of various devices. Innovative approach of INERTIA will consist on bring occupancy related information in the DER models. The most natural way to do this is through a sensor cloud that will provide the model for the real time occupancy extraction, along with the monitoring part related to the energy consumption and production. This kind of behavioral modelling approach is going to take occupancy related data, control actions of the users on the DER and the environmental conditions as parameters. Provided that an identification mechanism is installed in the building under consideration and some of the occupants are equipped with RFID cards, two types for the Occupancy and Flow Model are defined: the first one (Overall) refers to the occupants as a group, while the other one (Individual) refers to specific individuals. Using Radio Frequency Identification (RFID) equipment for some occupants will improve prediction accuracy and user profiling, as it will be possible to track occupants' location at any time having more specific information concerning their habits and schedules. With data provided from occupancy , INERTIA will provide short-term (near real time) and mid-term (next day) occupancy and flow prediction allowing for more efficient management of building's energy resources and providing the essential information for optimal local demand side management strategies. The estimate method will be based upon comfort parameter expressed as a discomfort probability, and based

 ISBN: 978-1-61208-418-3

on an analysis of the past history of the user's interactions with DER.

C. Key Performance Indicators

Key performance indicators are becoming a common instrument in public and private organizations used to analyze and monitor performance, and finally to drive informed decisions. In general, measuring the total performance of a system is used to learn, improve and cover the goal settings while also to provide the tools for the extraction of the optimal policy. Thus, the importance of well-defined performance indicators is even higher when dealing with optimization frameworks as the one examined within the INERTIA Project. Since these different coexisting sub-systems pose different performance constraints (energy related constraints, users comfort related constraints, business & flexibility related constraints) which are most often conflicting, INERTIA will adopt a holistic approach that will equally address and balance those aspects within a single integrated performance framework.

D. User Interface system in IoT context

The User Interface system of an application based on IoT is the synthesis of the interconnection of active entities. It should represent in a human friendly mode the modeling behind these entities and these interconnections. So in IoT, we have a lot of active elements and the User Interface system should connect all these active elements with the last and probably most important active element one: the final user. For this reason, the User Interface system should not only merely shows entity's data but also especially gives to user the possibility of being an active actor of its IoT world.

For INERTIA project, being an active user means that the user should be involved in profiling and forecasting in order to address the "structural INERTIA" of the Distribution Grid. How? Sharing his habits, his preferences about devices and comfort and his activities. This surely leads to privacy issues that must be treated in a suitable way (e.g.,: with anonymous data collecting, *Privacy by Design* [8]). Furthermore, since we are talking about IoT, User Interface system must fit security and scalability requirements.

1) Architecture

Each "thing" feature in IoT is uniquely identifiable through its embedded computing system and is capable to interoperate within the existing Internet infrastructure in order to coexist actively with others entities. This means the User Interface system should be able to communicate, in scaling way, with all these things, by filtering and synthetizing their data and also routing user's inputs. The solution identified and explained in this paper uses the Model View Presenter design pattern in order to:

- Makes *View Model*-independent
- Moves application logic outside de *View*
- Fits best scaling and security policies

The three main roles (*Model*, *Presenter* and *View*) can be summarized as seen in Figure 4.:

- *Model*: all devices, entities and systems connected through the internet forming what we call Internet of Things; e.g., data base storing sensors data, devices, sensors, controller, etc.;
- *Presenter*: the "middle man" transforming all Model data into information. It also connects the user with others IoT entities – systems, devices, sensors – according to the idea of user as active element like all other "things";
- *View*: the graphical component that displays data and routes user commands to the presenter. It gives the user the chance to "touch with hands" the IoT entities;

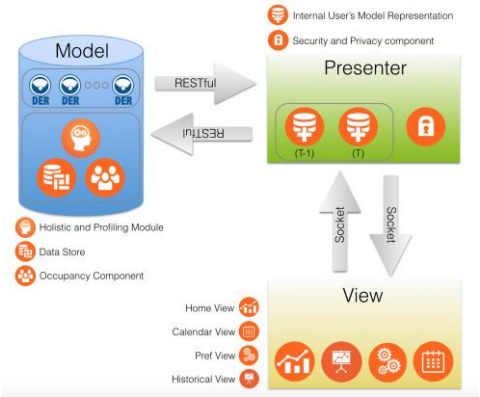


Figure 4. Model view presenter

2) Presenter – a deeper analisys

Presenter is the system's back end. It collects data from the *Model*, transforming these data into information and presents information to the *View*. Since we are talking about a system working in real time mode, the Presenter should fits this requirement. Anyway, an effective real time is not feasible, so it might be better talking about to call it "near-real time" flows of information.

How we do this? Essentially the Presenter builds an internal intermediate representation of user's model. This intermediate model stores information such as for example user's DERs, occupancy area, comfort values and consumptions. Periodically the Presenter queries the entities of IoT (or the store in which the IoT entities send data) to instantiate and populates the intermediate model. At each time (T) the Presenter compares this intermediate model with the previous one at time (T-1). All changed values of the intermediate model will be sent to the *View*. In this way the *Presenter* sends to *View* information changed only when they are changed, minimizing traffic, minimizing payload and diminish view's responsibility.

 ISBN: 978-1-61208-418-3

3) View – functionalities

Thanks to the *View*, the user is an active part of INERTIA world (see Figure 5). The View component of the design pattern is organized into different subcomponents:

a) HOME view: user monitors status and consumption of all DERs (personal and zonal), taking under control his comfort status. He also monitors his position in the building;

b) PREFERENCES view: user submits to INERTIA world his preferences about his devices, such as for example, HVAC winter or summer temperature, elettrical vehicle time-in and SoC, personal devices start and stop;

c) CALENDAR view: user submits to INERTIA world his habits and commitments such as for example "surgey in operating room from 11:00 am to 3:00 pm" or "training every Monday from 6:00 pm to 9:00 pm in the workout room";

d) HISTORICAL view: user monitors the historical data about consumption at different time;

Figure 5. Personal user interface

E. Communication technologies

All the components of the user interface architecture must communicate with each other. The connection between Presenter and View uses a socket connection. A socket is an end-to-end link over a single TCP connection, having the following features:

- Full duplex
- Bidirectional
- Always on
- Rapid data transformation due to a header much smaller then HTTP header

These features fit very well with an application working in (near) real-time mode, in which the user is an active element that should be always connected with the INERTIA world, possible submitting inputs, examining status of DERs and devices frequently.

The issue with this communication architecture is the limit of the number of connections in relation with the message rate per second [9]. For these reason it's important to have the possibility in scaling to multiple servers and minimizing the messages sent from server to client.

Communication between Inertia IoT Entities (*Model*) and the *Presenter* is made by the Linksmart [10]: this is an Open Source Middleware allowing developers to incorporate heterogeneous physical devices into their applications through easy-to-use web services for controlling any device. So for each entity we have a RESTful service that can be used to access data and control devices.

III. BOTTOM UP APPROACH

In the INERTIA concept, the DER will constitute active and flexible components carrying contextual knowledge of their local environment.

To deploy the INERTIA strength, DER will form dynamic clusters comprising self-organized networks of active nodes that will efficiently distribute and balance global and local intelligence. This aggregation is done in two steps. The first will be at the single building level and the second to the level of cluster of building.

A. Local control and automation hub (Building Automation System)

A whole tertiary building can be represented by a building automation system named Local Control Hub (LCH).

Thanks to *Building Automation Systems*, all individual building subsystems DER can become part of a single central system, also able to learn users' needs and behavior, to anticipate solutions or provide recommendations. These systems make use of forecasting, optimization and evaluation algorithms that acquire real-time data from smart sensors and meters, placed in strategic points of the building, capable of detecting internal microclimate parameters, space and ICT infrastructure use, attendance, weather data and energy quality. Based on the input data analysis, the system suggests actions to support building management optimizing energy consumption; users, on the other hand, as an active part of the system can monitor consumption instant by instant, by tablet or smartphone, and improve their behavior accordingly, becoming agents of saving themselves.

In this respect, the objects become self-recognizable and acquire intelligence due to the ability to communicate information about them and gain access to aggregate information from any other devices, allowing these systems to operate in real time.

Some ambient user interfaces (UIs) continuously collect data resulting from the occupant's interaction with existing traditional building hub devices and provide the necessary incentives through different interaction GUIs - e.g. through mobiles, monitors - , driving them to more energy- efficient choices.

B. Aggregator Control Hub.

The aggregator is an energy stakeholder that in a scale of aggregation following immediately the building and therefore manage different clusters of Local Control Hub using their portfolio, trading with the market stakeholders on behalf of small customers.

Aggregators gather, analyze and efficiently organize their customer load portfolio's and define specific active

 ISBN: 978-1-61208-418-3

demand (AD) strategies and services based on market needs. They act as an intermediary between suppliers and network operators and the different commercial and industrial (C&I) prosumers belonging to their portfolios.

IV. DEMO

Tests are underway to validate the proposed solution faced by INERTIA and the validation is done by means of simulations and laboratory tests. We have three field tests, which are a combination of actual field-testing and developed prototypes sited in Sweden, Greece and Spain.

It should be pointed that within the pilot we use one real Local Hub and a portfolio of simulated one that in their turn are two different kind of models.

The first kind is formed by simulated DERs, occupancy profiles, user preferences that all together constitute local hub set. This is a complex simulation approach since it implies deploying full LCH systems with all its components.

In light of this we made a second kind of building consumption profiles obtained from typical penetration in the pilot and based upon real DERs. Consequently it has been created a set of several building stereotypes.

The high level simulation then starts by obtaining a set of real consumption profile data (hourly or 15 minutes data) coming from building consumption measurements made in real pilot or from available typical building profiles. Then, the mix of controllable DERs is used together with the consumption profiles in order to calculate the flexibility made available at the simulated LCH. At the end by means of the portfolio of simulated LCHS it is created a mix of local control hub with different consumption and flexibility characteristics.

Then, we developed one test related to the network operation scenarios that involve the whole INERTIA chain stating form the DSO simulating some problem in the network seeding the corresponding DR signal to the aggregator, that deploying the needed control actions over the LCHs in its portfolio and finally the LCHs operating the final DERs that offer flexibility (Figure 6).

The test is considered successful under the fulfillment of the following conditions:

- The simulation of the congestion problem in the network provokes the calculation of DR signals (demand reduction) by the DSO Control Hub.
- These signals are sent to the Aggregator Control Hub that in turns sends the required control requests to the LCHs (real and simulated)
- The real LCH and micro-level simulated LCHs receive the control requests and operate the final DERs
- The demand at the MV supply level corresponds to the DR signals that were delivered
- The ACH generates demand forecasts updated according to the control actions taken
- The DSO Control Hub considers the updated demand forecasts and verifies that the congestion problem is solved

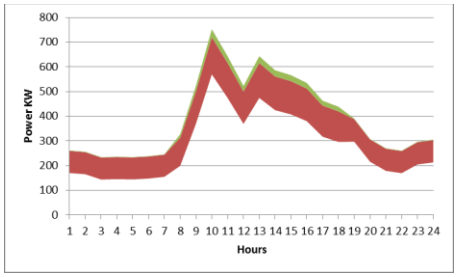


Figure 6. Flexibility requested

V. CONCLUSION

Demand side management has the capacity to overcome some of the major barriers of controlling and balancing supply and become a powerful tool at the hands of distribution grid. Moreover, as DERs continuously set an indispensable part of the EU Grids infrastructure, demand capacity and flexibility becomes a key performance factor with shared profit opportunities for all stakeholders involved. In addressing the "structural inertia" of existing Distribution Grids by introducing more active elements combined with the necessary control and distributed coordination mechanisms our project is an attempt in linking Internet of Things/Services principles to the Distribution Grid Control Operations.

REFERENCES

[1] H. Farhangi, "The path of the smart grid." Power and Energy Magazine, IEEE 8.1, 2010, pp 18-28.

[2] H. Lund, A. N. Andersen, P. A. Østergaard, and B. V. Mathiesen, "From electricity smart grids to smart energy systems–a market operation based approach and understanding." Energy 42.1, 2012, pp 96-102.

[3] A. H. Mohsenian-Rad, V. W. Wong, J. Jatskevich, R. Schober, A. Leon-Garcia, "Autonomous demand-side management based on game-theoretic energy consumption scheduling for the future smart grid."Smart Grid, IEEE Transactions on 1.3, 2010, pp 320-331.

[4] R. Belhomme, R. C. R. De Asua, G. Valtorta, A. Paice, F. Bouffard, R. Rooth, A. Losi, "Address-active demand for the smart grids of the future." SmartGrids for Distribution, 2008. IET-CIRED. CIRED Seminar. IET, 2008.

[5] http://www.inertia-project.eu/inertia/

[6] J. Froehlich, E. Larson, S. Gupta, G. Cohn, M. Reynolds, S. Patel, "Disaggregated end-use energy sensing for the smart grid." IEEE Pervasive Computing 10.1,2011, pp 28-39.

[7] N. Lu, Y. Zhang, "Design considerations of a centralized load controller using thermostatically controlled appliances for continuous regulation reserves." Smart Grid, IEEE Transactions on 4.2, 2013, 914-921.

[8] https://www.privacybydesign.ca "accessed June 2015"

[9] http://drewww.github.io/socket.io-benchmarking/ "accessed June 2015"

[10] https://www.linksmart.eu/redmine "accessed June 2015"

 ISBN: 978-1-61208-418-3