UBICOMM 2012

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

ISBN: 978-1-61208-236-3

September 23-28, 2012

Barcelona, Spain

UBICOMM 2012 Editors

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain

Pascal Lorenz, University of Haute Alsace, France

UBICOMM 2012

Forward

The Sixth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2012), held on September 23-28, 2012 in Barcelona, Spain, was a multi-track event covering a large spectrum of topics related to developments that operate in the intersection of mobile and ubiquitous technologies on the one hand, and educational settings in open, distance and corporate learning on the other, including learning theories, applications, and systems.

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

The goal of UBICOMM 2012 was to bring together researchers from the academia and practitioners from the industry in order to address fundamentals of ubiquitous systems and the new applications related to them. The conference provided a forum where researchers were able to present recent research results and new research problems and directions related to them.

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

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

Educational services will become more customized and personalized, and more frequently subjected to changes. Learning and teaching are now becoming less tied to physical locations, co-located members of a group, and co-presence in time. Learning and teaching increasingly take place in fluid combinations of virtual and "real" contexts, and fluid combinations of presence in time, space and participation in community. To the learner full access and abundance in communicative opportunities and information retrieval represents new challenges and affordances. Consequently, the educational challenges are numerous in the intersection of technology development, curriculum development, content development and educational infrastructure.

We take here the opportunity to warmly thank all the members of the UBICOMM 2012 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to UBICOMM 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the UBICOMM 2012 organizing committee for their help in handling the logistics and for their work that is making this professional meeting a success. We gratefully appreciate to the technical program committee co-chairs that contributed to identify the appropriate groups to submit contributions.

We hope the UBICOMM 2012 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in ubiquitous systems and related applications.

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

UBICOMM 2012 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, Osaka University, Japan
Ruay-Shiung Chang, National Dong Hwa University, Taiwan

UBICOMM Research Chairs
Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Carlo Mastroianni, CNR, Italy
Sergey Balandin , FRUCT, Finland

Juong-Sik Lee, Nokia Research Center - Palo Alto, USA
Ann Gordon-Ross, University of Florida, USA
Michele Ruta, Politecnico di Bari, Italy

UBICOMM 2012

Committee

UBICOMM Advisory Chairs

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Zary Segal, UMBC, USA
Yoshiaki Taniguchi, Osaka University, Japan
Ruay-Shiung Chang, National Dong Hwa University, Taiwan

UBICOMM 2012 Research Chairs

Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Carlo Mastroianni, CNR, Italy
Sergey Balandin , FRUCT, Finland
Juong-Sik Lee, Nokia Research Center - Palo Alto, USA
Ann Gordon-Ross, University of Florida, USA
Michele Ruta, Politecnico di Bari, Italy

UBICOMM 2012 Technical Progam Committee

Afrand Agah, West Chester University of Pennsylvania, USA
Rui Aguiar, Universidade de Aveiro, Porutgal
Tara Ali-Yahiya, Paris Sud 11 University, France
Timothy Arndt, Cleveland State University, USA
Mehran Asadi, West Chester University of Pennsylvania, USA
Sergey Balandin, FRUCT, Finland
Matthias Baldauf, FTW Telecommunications Research Center Vienna, Austria
Michel Banâtre, IRISA - Rennes, France
Matthias Baumgarten, University of Ulster-Belfast, Northern Ireland, UK
Aurelio Bermúdez Marin, Universidad de Castilla-La Mancha, Spain
Daniel Bimschas, University of Lübeck, Germany
Carlo Alberto Boano, University of Lübeck, Germany
Jihen Bokri, ENSI (National School of Computer Science), Tunisia
Sergey Boldyrev, Nokia, Finland
Diletta Romana Cacciagrano, University of Camerino, Italy
Jose Manuel Cantera Fonseca, Telefonica Investigacion y Desarrollo, Spain
Juan-Vicente Capella-Hernández, Universidad Politécnica de Valencia, Spain
Davide Carboni, CRS4 Research Center - Sardinia, Italy
Rafael Casado, Universidad de Castilla-La Mancha, Spain
José Cecílio, University of Coimbra, Portugal
Bongsug (Kevin) Chae, Kansas State University, USA
Sung-Bae Cho, Yonsei University - Seoul, Korea

Mhammed Chraibi, Al Akhawayn University - Ifrane, Morocco
Michael Collins, Dublin Institute of Technology, Dublin, Ireland
Kyller Costa Gorgônio, Universidade Federal de Campina Grande, Brazil
Claudio de Castro Monteiro, Federal Institute of Education, Science and Technology of Tocantins, Brazil
Jerker Delsing, Lulea University of Technology, Sweden
Steven A. Demurjian, The University of Connecticut, USA
Gianluca Dini, University of Pisa, Italy
Jianguo Ding, University of Luxembourg, Luxembourg
Yulin Ding, Defence Science & Technology Organization Edinburgh, Australia
Roland Dodd, Central Queensland University, Australia
Josu Etxaniz, University of the Basque Country, Spain
Andras Farago, The University of Texas at Dallas - Richardson, USA
Sana Fathalla, Sophia Antipolis / CNRS, France
Ling Feng, Tsinghua University - Bejing, China
Gianluigi Ferrari, University of Parma, Italy
George Fiotakis, University of Patras, Greece
Rita Francese, Università degli Studi di Salerno, Italy
Korbinian Frank, German Aerospace Center - Institute of Communications and Navigation, Germany
Kaori Fujinami, Tokyo University of Agriculture and Technology, Japan
Crescenzio Gallo, University of Foggia, Italy
Junbin Gao, Charles Sturt University - Bathurst, Australia
Shang Gao, Norwegian University of Science and Technology, Norway
Chris Gniady, University of Arizona, USA
Paulo R. L. Gondim, University of Brasília, Brazil
Ann Gordon-Ross, University of Florida, USA
Dominic Greenwood, Whitestein Technologies - Zürich, Switzerland
Frederic Guidec, Université de Bretagne Sud / Université Européenne de Bretagne, France
Fikret Gurgen, Isik University - Istanbul, Turkey
Norihiro Hagita, ATR Intelligent Robotics and Communication Labs, Kyoto, Japan
Jason O. Hallstrom, Clemson University, USA
Arthur Herzog, Technische Universität Darmstadt, Germany
Shaohan Hu, UIUC, USA
Xiaodi Huang, Charles Sturt University - Albury, Australia
Javier Alexander Hurtado, University of Cauca, Colombia
Marko Jaakola, VTT Technical Research Centre of Finland, Finland
Faouzi Kamoun, University of Dubai, UAE
Fazal Wahab Karam, Norwegian University of Science and Technology (NTNU), Norway
Rehana Kausar, Queen Mary, University of London, UK
Hamed Ketabdar, Deutsche Telekom Laboratories / TU Berlin, Germany
Subayal Khan, VTT, Finland
Inas Khayal, Masdar Institute of Science and Technology - Abu Dhabi, United Arab Emirates
Donnie H. Kim, Intel, USA
Cornel Klein, Siemens AG/Corporate Research and Technologies - Münich, Germany
Reinhard Klemm, Avaya Labs Research-Basking Ridge, USA
Jeffrey Tzu Kwan Valino Koh, National University of Singapore, Singapore
Jakub Konka, University of Strathclyde, UK
Dmitry G. Korzun, Petrozavodsk State University, Russia / Helsinki Institute for Information Technology - HIIT, Finland

Natalie Kryvinski, University of Vienna, Austria
Frédéric Le Mouël, INRIA/INSA Lyon, France
Nicolas Le Sommer, Université de Bretagne Sud - Vannes, France
Juong-Sik Lee, Nokia Research Center, USA
Pierre Leone, University of Geneva, Switzerland
Jian Liang, Cork Institute of Technology, Ireland
Kai-Wen Lien, Chienkuo Institute University - Changhua, Taiwan
Bo Liu, University of Technology - Sydney, Australia
Qiong Liu, FX Palo Alto Laboratory, USA
David Lizcano Casas, Universidad Politécnica de Madrid, Spain
Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Juan Carlos López, University of Castilla-La Mancha, Spain
Jeferson Luis Rodrigues Souza, University of Lisbon, Portugal
Elsa María Macías López, University of Las Palmas de Gran Canaria, Spain
Sathiamoorthy Manoharan, University of Auckland, New Zealand
Teddy Mantoro, University of Technology Malaysia, Malaysia
Sergio Martín Gutiérrez, UNED-Spanish University for Distance Education, Spain
Oscar Martinez Bonastre, University Polytechnic of Valencia, Spain
Carlo Mastroianni, ICAR-CNR - Rende, Italy
Natarajan Meghanathan, Jackson State University, USA
Andreas Merentitis, AGT Group (R&D) GmbH - Darmstadt, Germany
Kathryn Merrick, University of New South Wales & Australian Defence Force Academy, Australia
Elisabeth Métais, CNAM/CEDRIC, France
Moeiz Miraoui, Gafsa University, Tunisia
Nikolai Nefedov, Nokia Research Center, Switzerland
Rui Neves Madeira, New University of Lisbon, Portugal
Quang Nhat Nguyen, Hanoi University of Technology, Vietnam
Gregory O'Hare, University College Dublin (UCD), Ireland
Kouzou Ohara, Aoyama Gakuin University, Japan
Akihiko Ohsuga, The University of Electro-Communications (UEC) - Tokyo, Japan
George Oikonomou, University of Bristol, UK
Ian Oliver, Nokia, Finland
Carlos Enrique Palau Salvador, University Polytechnic of Valencia, Spain
Agis Papantoniou, National Technical University of Athens (NTUA), Greece
Ignazio Passero, Università degli Studi di Salerno - Fisciano, Italy
Serena Pastore, INAF- Astronomical Observatory of Padova, Italy
Jyrki Penttinen, Nokia Inc. - Parsippany, USA
Jorge Pereira, European Comission, Belgium
Yuliya Ponomarchuk, Far Eastern Transport University, Russia
Daniel Porta, German Research Center for Artificial Intelligence (DFKI) - Saarbrücken, Germany
Iván Pretel García, DeustoTech - Deusto Institute of Technology, Spain
Elmano Ramalho Cavalcanti, Federal University of Campina Grande, Brazil
Juwel Rana, Luleå University of Technology, Sweden
Peter Reiher, UCLA, USA
Hendrik Richter, LMU - University of Munich, Germany
Jose D. P. Rolim, University of Geneva, Switzerland
Michele Ruta, Politecnico di Bari, Italy
Luis Sanchez, Universidad de Cantabria, Spain

Copyright Information

Table of Contents

Compass and WLAN Integration for Indoor Tracking on Mobile Phones

Moritz Kessel, Martin Werner, Claudia Linnhoff-Popien
Mobile and Distributed Systems Group
Ludwig-Maximilians-University Munich
Munich, Germany
{*moritz.kessel,martin.werner,linnhoff*}*@ifi.lmu.de*

***Abstract*—Indoor positioning with smartphones in ubiquituos computing scenarios still poses some problems with respect to accuracy and precision as well the need for a calibrated infrastructure and map data. This paper presents a method for indoor positioning based on the combination of 802.11 WLAN fingerprinting using weighted kNN and a simple indoor movement model based on digital compass information. We introduce a novel computation scheme for the distance in WLAN signal space, additionally considering the Euclidean real world distance between each fingerprint and the position predicted by the movement model. A detailed evaluation in a test environment at our site demonstrates a performance gain of more than 10% as compared to a classical Kalman filter. Moreover, we also show that the Kalman filter offers a slightly better capability to correct the accumulation of errors over time when accurate movement information in form of step detection is available.**

Keywords-Indoor Location Systems; 802.11 Fingerprinting; Mobile Phone Tracking; Location-Based Services.

I. Introduction

Location-based Services (LBS) [1], such as navigation and information services, are among today's most popular mobile services in ubiquitous computing scenarios. Furthermore, the position of objects or persons is essential in many ubiquitous computing applications since information about the surroundings is often more important than information on places far away. Fortunately, the increasing market penetration of modern smartphones, small devices with high processing power and localization capabilities, boosts the availability of location-dependent information. An example is the Global Positioning System (GPS) [2] offering accurate positioning free of charge in outdoor environments. However, GPS is (for the moment) not able to provide accurate indoor position information for ubiquitous services. Especially in the case of indoor navigation the position error should not exceed a few meters to be able to distinguish between several rooms and floors to provide step-by-step guidance and generate navigation instructions.

In the past few years, multiple indoor positioning techniques have been developed. Systems based on ultra wideband or ultrasonic systems utilize an expensive dedicated infrastructure, but offer high accuracy position estimates. Those are often only available in a small area, where higher accuracy compensates the high cost. An advantage of these precise systems is that tracking functionality can easily be integrated by sampling position estimates over time. Since the error of each estimate is small, a realistic track can be observed over a period of time.

Other systems make use of existing infrastructure with little or no additional expenses and therefore are limited in their accuracy. These kinds of positioning systems are often based on WLAN, Bluetooth or inertial measurement units (IMU). In contrast to expensive systems, they are often used to offer localization services in large indoor areas, especially in complex buildings such as museums, shopping malls, airports, hospitals, or university buildings. Adding high quality tracking functionality to these systems is considerably more complex, due to the possibility of unrealistic jumps of consecutive position estimates over a large distance (WLAN) and the accumulation of errors over time (IMU). But exactly the high quality tracking information is of great importance to a large number of services such as navigation, tracking of users and goods, or automated quality control in factory settings. Zheng and Xiaofang offer a good overview on use-cases and techniques for computing and working with spatial trajectories in [3].

In this paper, we expand SMARTPOS [4], an indoor positioning system for smartphones based on WLAN fingerprinting and a digital compass, to support continuous positioning and tracking. While the unmodified system achieves a high accuracy within few meters for positioning it is prone to jump between consecutive estimates and therefore offers low quality tracks. We show that the user's orientation measured by the digital compass of the smartphone can be used for a more stable position estimation. The key contribution is the inclusion of the predicted position based on the previous position and a movement model into the fingerprint nearest neighbor calculation. Basically, this favors fingerprints near the last position estimate resulting in smoother paths, which are more realistic and more accurate as we demonstrate in this paper. The applied movement model is either simply defined by constant velocity or step detection with constant step-length, both enhanced with a direction measured by a digital compass.

The remainder of this paper is structured as follows: In the next section, a short overview of existing indoor positioning and tracking systems is given and differences to our system

ISBN: 978-1-61208-236-3

mentioned. In Section III, the original SMARTPOS system [4] is presented, while in Section IV the enhancements of the system for continuous positioning are explained in detail. In Section V, their impact on the tracking performance is analyzed and discussed. Section VI concludes the paper and gives hints on future work.

II. Related Work

The topic of indoor positioning and tracking is deeply investigated in academic and industrial research and one of the most active research topics concerning LBS. A vast variety of technologies and algorithms have been proposed and still no satisfactory solution exists that offers satisfactory position information for every use case.

Many pedestrian indoor positioning systems rely on WLAN fingerprinting algorithms [4], [5], [6], [7], which offer position estimates with sufficient accuracy (i.e., 1-3m) while utilizing the existing WLAN infrastructure and therefore avoiding high expenses. These algorithms belong to the area of pattern matching and work in two phases: The first phase is called the calibration phase, where a database is created by the collection of received signal strength indicator (RSSI) at certain reference positions from the surrounding access points (AP). The accumulated information of RSSI, AP and reference position at a specific time/interval is called a fingerprint. In the second phase, the positioning is carried out by comparing current RSSI measurements with the previously stored values from the database. Different algorithms calculate the position as the reference position of the nearest fingerprint in signal space [5], the average of the k-nearest neighbors (kNN) with or without the distance in signal space as additional weight [4]. Some algorithms also utilize Bayesian methods [6], [7] based on probability distributions derived by multiple measurements over a length of time. COMPASS [7] is one of the first fingerprinting systems that addresses the problem of attenuation effects caused by the human body by adding a digital compass to the system. In the calibration phase, fingerprints for several selected orientations (typically each $45°$ or $90°$) are collected at reference positions. In the positioning phase, the user's orientation is measured by a digital compass and only the fingerprints with a similar orientation estimate are used for the positioning algorithm. However, none of the above-mentioned approaches considers the user's movement. A good overview of other existing positioning systems using radio frequency (RF) technologies such as radio frequency identification (RFID), ultra wide band, WLAN and Bluetooth is given in [8].

Another class of pedestrian indoor positioning systems is based on IMUs such as accelerometer or magnetometer. These systems offer only relative positioning capacities as they measure position changes rather than an absolute position. This results in an accumulation of sensor errors over time, which is the reason why most systems consider additional information, e.g., WLAN fingerprinting or map information, for recalibration. Woodman and Harle show in [9] that a building model can compensate the drift of inertial sensors and add WLAN positioning for obtaining an initial position fix in their Dead Reckoning system. However, WLAN positioning is not considered for the correction of position data. Evennou and Marx compare a Kalman and a particle filter for fusing location estimates of a WLAN fingerprinting algorithm with high quality accelerometer data [10]. They report a high increase in accuracy compared to individual systems, but do not yet include the target's or its predicted position in the WLAN fingerprinting. Ruiz et al. utilize a tightly coupled Kalman filter to fuse foot-mounted IMU-based position estimates with additional RSSI information by RFID tags [11]. Their approach is similar to ours, since they include the estimated position by IMU in the position estimation by RFID, but they use pathloss instead of fingerprinting techniques and foot-mounted sensors instead of cell phones. Chan et al. describe a system [12], which utilizes orientation information to predict the region, in which the next matching fingerprint is to be expected. Fingerprints outside that predicted region are completely ignored for the nearest neighbor estimation, so that wrong orientation information can lead to wrong location assumptions. Our system is not as restrictive and weights several factors applying the distance calculation on fingerprints all around instead of a regional limited choice.

In the last few years, many indoor positioning systems have been adopted or developed for deployment on cell phones. These offer additional sensors, which can directly be integrated to enhance the position accuracy of existing systems by combining the information of different sources to a position estimate of higher accuracy than any single source could provide. Martin et al. present one of the first WLAN positioning systems, which integrates both calibration and positioning on a mobile phone [13]. They use various nearest neighbor algorithms, but do not use additional sensors nor a prediction model. Perttula et al. show in [14] that modern smartphones have enough processing power to support Bayesian location estimation methods for WLAN fingerprinting, but no further sensors are added to the system to improve the localization capabilities. In [15], Liu et al. propose a particle filter based on a hidden Markov model to combine the inertial sensors of a smartphone with WLAN fingerprinting and report a significant increase in accuracy. Nevertheless, they do not include the predicted position in the choice of position candidates by the measurement model.

In our approach, we integrate WLAN fingerprinting with IMU data by the means of a prediction model based on the IMU data and a measurement model based on the predicted position and WLAN fingerprinting. The novelty of our approach is given by the direct inclusion of the predicted position in the WLAN fingerprinting approach by combining the Euclidean distance in map and signal space.

III. The SMARTPOS System

In this section, we give a short overview of SMARTPOS (for more details see [4]), an indoor positioning system based on 802.11 fingerprinting and a digital compass. The focus is on the functionality and the previously gathered results, which are helpful to understand enhancements for tracking functionality described in the following section.

A. Description

SMARTPOS runs stand-alone on a mobile phone and consists of a management module for the creation and maintenance of a fingerprint database and a module for location determination. In the database creation step, RSSI is stored for several reference positions (RP) in the building. At each reference position four fingerprints are recorded, each time aligning the phone along one of the four main axes of the building. A single fingerprint consists of

- a list of the average RSSI values of five consecutive measurements for each visible access point (AP) along with the MAC-address of the specific AP
- the pixel coordinates of the reference position on a bitmap of the floor
- and the average of the measurements of the phone's compass during the sampling time, which is an indicator for the user's orientation

B. Previous Results

We evaluated the impact of several variations of deterministic fingerprinting, using a k-nearest-neighbor (kNN) algorithm with varying parameter k. In a testbed consisting of a wing of a university building (approximately 200m^2) and with a database consisting of 79 RPs and thus 316 individual fingerprints, we evaluated the impact of weighted kNN, the treatment of missing values, and the performance gain by including the user's orientation in the position estimation step. The evaluation was carried out with respect to the accuracy and the precision of SMARTPOS with the following results:

- Weighted kNN results in a slightly higher accuracy than non-weighted kNN
- Ignoring missing RSSI causes higher accuracy and precision for larger k as compared to assigning a fixed minimal RSSI value (which was -100dBm) below the phone's sensibility for all missing AP information
- Considering orientation information greatly increases accuracy and precision for a smaller k

IV. Continuous Positioning

To increase the accuracy of SMARTPOS in continuous positioning scenarios, we enhance the position estimation with a prediction step. This step is computed by the means of a movement model supported by the smartphone's compass. There are several ways to combine WLAN position updates with a movement model and measurements from inertial sensors. Most commonly either a Kalman or a particle filter is used to fuse the data from different sources (compare [10]). While a particle filter is a discrete approximation of a probability density function, a Kalman filter is an approximation of a linear dynamic system from noisy measurements. For indoor positioning, a Kalman filter is usually a linear combination of the position estimated by a measurement model $p_{mes,i}$ and the estimate of a prediction model $p_{pre,i-1}$ based on the previous state vector and the direction and velocity of movement (see equation (1)).

$$p_{i|i-1} = p_{mes,i} \cdot (1-\gamma) + p_{pre,i-1} \cdot \gamma \qquad (1)$$

This algorithm works well, but we see the possibility of optimization using kNN. Since a new estimated position (or one of the k nearest neighbors) can be at an unrealistically large distance from the last estimate, we want to modify the choice of nearest neighbors. Therefore, we propose a novel approach similar to a Kalman filter, applying a measurement and a prediction model, but instead of interpolating the predicted position estimate and the measured estimate, we include the interpolation into the nearest neighbor algorithm. We influence the choice of nearest neighbors in signal space by the deviation (Euclidean distance) of the position predicted by the movement model and the position of each neighbor candidate in the real world. Hence, we include a probability of presence into the distance to neighbors in signal space.

A. Measurement Model

The measurement model consists of a basic weighted kNN algorithm with a refined distance definition (between a current measurement and a fingerprint) allowing to include the map distance directly into the computation of the nearest neighbors: Let s_i be the Euclidean distance of the current RSSI measurement (at time t) to the recordings of the i'th fingerprint in the database. Let m_i be the Euclidean distance on a map between the predicted position p_t at time t and the position of the i'th fingerprint. We define the distance $d_i = (1-\alpha) \cdot s_i + \alpha \cdot m_i$ between the current measurement and the i'th fingerprint as a linear interpolation of s_i and m_i, assigning a fixed constant weighting factor α. This definition of a distance between a measurement at a given predicted position and a previously recorded fingerprint leads to smoother paths favoring fingerprints near to a predicted position over fingerprints further away. Obviously, the value of α needs to be configured in order to balance the prediction model and the measurement model.

B. Prediction Model

Having obtained a position estimate at least once, we can apply a prediction model to continuously estimate the current position of a moving target until another measurement can be utilized to recalibrate the predicted position. We apply either a constant movement model or a step counter mechanism (similar to the one from Link et al. [16] without map correction) based on accelerometer data, both enhanced with the readings of the smartphone's compass. In the first case, the position is updated according to the constant speed, the elapsed time and the current orientation of the phone between two consecutive compass readings. In the case of the step counter, a step is detected whenever the vertical acceleration drops by more than 2m/s^2 within five consecutive accelerometer readings (i.e., approx. 1s). Those readings are not considered for step detection twice, so when a step is detected, the previous readings are discarded. This mechanism ensures that a single large drop in vertical acceleration measured by consecutive accelerometer readings is not interpreted as multiple steps. A constant step-length is assigned and every detected step is mapped in the direction derived by the compass at the same time. Note that the smartphone needs to be held in approximate horizontal fashion for either the step detection algorithm and the compass. For the moment, we do not apply any map matching techniques to clarify the impact of the prediction model on the measurement model.

V. Evaluation

For the evaluation, a test database of fingerprints in a wing of our site was created. All information was gathered with a HTC Desire smartphone. The 57 reference positions are arranged in an approximate grid with fingerprints measured in the direction of all four main axes of the building, which results in 228 fingerprints in total (the gray dots in Figure 1).

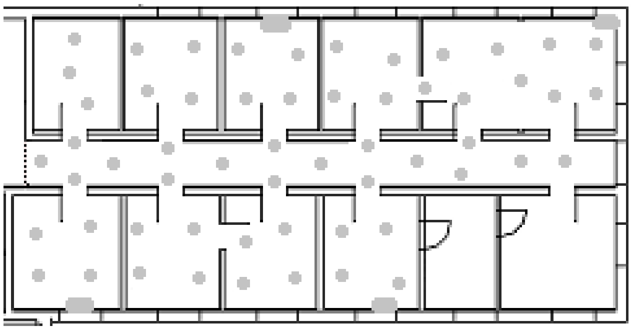


Figure 1: Reference positions (gray dots) and access points (gray rounded rectangles).

We then recorded two tracks (see Figure 2) storing the RSSI values of consecutive active scans (approx. sample rate of 1 Hz) together with the MAC-address of the APs and the readings of the compass as well as the accelerometer (both with an approx. sample rate of 5 Hz). All data is enriched with a timestamp of the specific measurement time and saved in a file. This ensures that our results originate from an identical setting for all the different positioning methods. The quality of the positioning method is evaluated in respect to two criteria: the accuracy indicated by the mean position error and the precision indicated by the standard deviation. In the following the results from a detailed evaluation in the described setting are presented and discussed.

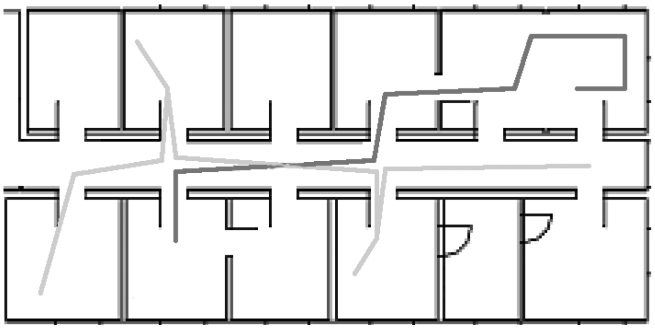


Figure 2: Two test tracks (T1 in light gray with a length of 42m, T2 in darker gray with a length of 27m) starting from right to left.

SMARTPOS is evaluated as follows: First the tracking performance of the deterministic kNN approach and the impact of using orientation information is analyzed. Afterwards, a movement model is added and the Euclidean distance between the reference position and the user's currently estimated position is introduced into the distance calculation of the current RSSI readings and each fingerprint as described. We show that the right choice of the movement model is crucial to the accuracy of the system. Furthermore, we demonstrate the accuracy of the compass by combining an experimentally calculated movement model to the compass readings without considering RSSI. Finally, our proposed algorithm is then compared to a classical Kalman filter.

A. Tracking Accuracy of the original SMARTPOS

While standard kNN with $k = 3$ (3NN) is based on RSSI measurements only, we include the user's orientation into our algorithm in the oriented kNN with $k = 3$ (O3NN). In the latter case, only those fingerprints that have a maximal deviation of 50° of the user's orientation are considered for the nearest neighbor algorithm. Table II shows the results for both tracks. As assumed, the consideration of the user's orientation leads to a significant reduction in error (above 10%) and a even larger precision gain (approx. 33%).

Table I: Tracking Accuracy of the original SMARTPOS:

	3NN T1	3NN T2	O3NN T1	O3NN T2
Average error	2.01 m	1.74 m	1.72 m	1.52 m
Standard dev.	1.52 m	1.00 m	1.02 m	0.67 m

B. Combining Euclidean Distance in Signal and Map Space

In this section, the effect of the combination of the Euclidean distance in map and signal space (see Section IV) on the tracking accuracy is further investigated. We experimented with different movement models, alternating the fixed velocity value, and different measurement models for the empirical determination of the parameter α in our test environment.

Small values of alpha result in a smaller influence of the prediction model on the measurement model while larger values add weight to the prediction model. In our test environment $\alpha = 0.7$ (with a speed of 1m/s (A07V10)) lead to the best results in both tracks, resulting in a mean tracking error of 1.33m (T1) and 1.00m (T2) with a standard deviation of 0.85m (T1) and 0.62m (T2). It has to be noted that the lower accuracy of the first track has its origin in the varying walking speed (zero velocity at turn-points), which the movement model is unable to take into account.

With the empirically solved assignment of a fixed value $\alpha = 0.7$, different velocities for the movement model were evaluated. In our model, velocities between 0.8m/s and 1.0m/s are suitable for tracking indoor movement of pedestrians, even in the case of non-constant walking speed. Even so, the mean error can vary to a maximal amount of approximately 10% within this speed interval. The minimal measured mean error was 1.22m (T1) and 0.90m (T2) with a standard deviation of 0.81m (T1) and 0.65m (T2) for a velocity of 0.8m/s.

Finally, the accuracy of the combination of the proposed measurement model with a prediction model based on step detection is evaluated. Each time a step is detected, the predicted position is moved 1.0m in the direction of the current compass heading. The results with our testdata show that the influence of the measurement model should be reduced and $\alpha = 0.85$ leads to a minimal mean error in both tracks, which was 1.33m (T1) and 0.88m (T2) with a standard deviation of 0.78m (T1) and 0.46m (T2). In this case, one can use the measurement model to overcome the accumulation of errors of the step detection algorithm such as wrong step-length, heading, or unreliable step detection.

C. Accuracy of Orientation Information

It is crucial for the prediction model to obtain realistic data readings. While accelerometer readings of todays smartphones are assumed to be too noisy for double integration or even simple integration for the assignment of a velocity to a movement model (a claim which is supported by our data), they can be used for step detection. Furthermore, we observe a high reliability in the compass information. There is an information lag in cases of abrupt turning, but the overall accuracy of the compass values in indoor environments is astonishing. Figure 3 shows an estimated track, which was computed by entering the real starting position and using solely the movement model with an empirically solved constant movement speed applied on the second track (i.e., a track which was recorded while moving with a constant speed). The average error to the real track is 0.82m with a standard deviation of 0.55m and therefore even better than the advanced SMARTPOS system. This is clearly due to the over-fitting of the model to the data, however, it remarkably supports the claim that compass readings contain valuable information for indoor positioning even for low cost smartphone sensors. The accumulated error over time in this experiment was 2.96m.

When using the step detection algorithm instead of a movement model, the average error was 1.20m (T1) and 1.10m (T2) with a standard deviation of 0.77m (T1) and 0.61m (T2) for a step-length of 1.0m. Moreover, the error accumulated over time at the end of each track (i.e. 1.90m in T1 and 0.85m in T2) was less than 5% of the track-length. These results indicate that hand-held step detection with smartphone accelerometer and measuring the direction of the step with a smartphone compass is possible and even without any additional correction schemes quite accurate. However, even a small percentage of error can result in large errors after some time, which makes correction methods indispensable.

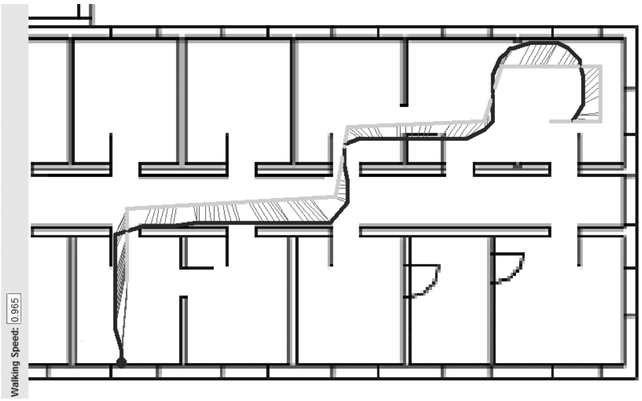


Figure 3: The real track in light gray, the estimated track in dark gray, error vectors are the thin lines in between.

D. Comparison with a Kalman Filter

We compare our results with a classical Kalman filter computing the position estimate as a linear combination of the position estimate computed by either the prediction and the measurement model. In this case, the measurement model does not use the novel measurement scheme, but a standard kNN approach for WLAN position estimation. A weighting factor $\gamma \in [0, 1]$ is applied to equation (1), with $p_{mes,i}$ being the position estimate of the measurement model at time i and $p_{pre,i-1}$ being the position estimate of the prediction model, based on the last estimated position p_{i-1}. The minimal mean error achieved by the Kalman filter (with a speed of 1.0m/s) is 1.51m (T1, $\gamma = 0.7$) and 1.14m (T2,

$\gamma = 0.8$) with a standard deviation of 0.89m (T1) and 0.53m (T2), which means the best case error is still $14-18$cm (i.e., more than 10%) larger than that of our proposed algorithm using the same movement model.

When using the step detection algorithm, however, the contrary is the case. Since step detection also models non constant speed and is therefore in many cases more accurate as a constant speed movement model, the prediction model becomes also more accurate than the measurement model. The system needs only minor corrections to compensate for the accumulation of errors, which is better achieved by a classical Kalman filter. The minimal mean error with $\gamma = 0.85$ in either track is 1.24m (T1) and 0.83m (T2) with a standard deviation of 0.69m (T1) and 0.42m (T2), meaning that the classical Kalman filter is about 7% better than our proposed algorithm.

Furthermore, both algorithms can easily be combined, by utilizing our algorithm for determining the nearest neighbors in the measurement model and the Kalman filter for the interpolation with the predicted position. Since this procedure obviously favors the prediction model, the interpolation parameters need to be applied to compensate for this effect. With $\alpha = 0.75$ and $\gamma = 0.5$ the minimal mean error was 1.14m (T1) and 0.87m (T2) with a standard deviation of 0.81m (T1) and 0.50m (T2), which is a error reduction in the first track, but an increase in the second track. Table II shows an overview on the evaluated parameters. The table also includes accuracies of related research systems for reasons of comparability with respect to the order of magnitude of the expected position error. A direct comparison is not possible due to different set ups and test environments.

Table II: Overview of evaluated tracking accuracies:

T1	**avg. err.**	**std. dev.**	**end err.**
Alpha0.7Velocity1.0	1.33 m	0.85 m	-
Alpha0.7Velocity0.8	1.22 m	0.81 m	-
Alpha0.85Steplength1.0	1.33 m	0.78 m	-
PredictionSteplength1.0	1.20 m	0.77 m	1.90 m
Gamma0.7Velocity1.0	1.51 m	0.89 m	-
Gamma0.85Steplength1.0	1.24 m	0.69 m	-
Alpha0.75Gamma0.5Steplength1.0	1.14 m	0.81 m	-
T2			
Alpha0.7Velocity1.0	1.00 m	0.62 m	-
Alpha0.7Velocity0.8	0.90 m	0.65 m	-
Alpha0.85Steplength1.0	0.88 m	0.46 m	-
PredictionVelocity0.975	0.82 m	0.55 m	2.96 m
PredictionSteplength1.0	1.10 m	0.61 m	0.85 m
Gamma0.8Velocity1.0	1.14 m	0.53 m	-
Gamma0.85Steplength1.0	0.83 m	0.42 m	-
Alpha0.75Gamma0.5Steplength1.0	0.87 m	0.50 m	-
Other Systems			
Woodman and Harle [9]	<0.5 m	-	-
Evennou and Marx [10]	1.53 m	-	-
Ruiz et al. [11]	1.35 m	-	<1.5 m
Chan et al. [12]	1.82 m	-	-

VI. Conclusion and Future Work

This paper presents enhancements for SMARTPOS, a positioning system running stand-alone on smartphones based on deterministic WLAN fingerprinting and a digital compass. The key concept is the addition of a prediction model to the system. We tested both a simple indoor pedestrian movement model with constant speed and a model based on step detection. For orientation information the digital compass of the smartphone is used. The position estimate p of the prediction model is utilized to include the map distance from p and a fingerprint's reference position into the nearest neighbor search in signal space.

We evaluated different weighting factors for the combination of map and signal distance and researched the effect of the walking speed on our model. All experiments were carried out in a testbed of approximately 200m^2 and evaluated on two different tracks. One track was recorded while walking with a constant speed, while the other tracks included turn points with zero velocity. In the track with constant speed the mean error was reduced to 0.90m with a standard deviation of 0.65m, while with the more complicated track we were able to reduce the mean error to 1.22m with a standard deviation of 0.81m. The results show that our approach outperforms a classical Kalman filter, using a linear combination of the prediction and the measurement model, when no accurate information about the movement of the target is available. If high quality position prediction is possible (e.g., with a sophisticated step detection algorithm using the smartphone accelerometer), the Kalman filter with a heavy weight on the prediction model is a better option.

References

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

[2] E. Kaplan, *Understanding GPS: Principles and Applications*, ser. Artech House Mobile Communications. Artech House Publishers, 2006.

[3] Y. Zheng and X. Zhou, *Computing with Spatial Trajectories*. Springer, 2011.

[4] M. Kessel and M. Werner, "SMARTPOS: Accurate and precise indoor positioning on mobile phones," in *Proceedings of the International Conference on Mobile Services, Resources, and Users (MOBILITY'11)*, 2011, pp. 158–163.

[5] P. Bahl and V. N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *Proceedings of the Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM'00)*, vol. 2, 2000, pp. 775–784.

[6] M. Youssef and A. Agrawala, "The horus WLAN location determination system," in *Proceedings of the International Conference on Mobile Systems, Applications, and Services (MobiSys'05)*, 2005, pp. 205–218.

[7] T. King, S. Kopf, T. Haenselmann, C. Lubberger, and W. Effelsberg, “COMPASS: A probabilistic indoor positioning system based on 802.11 and digital compasses,” in *Proceedings of the International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization (WiNTECH’06)*, 2006, pp. 34–40.

[8] H. Liu, H. Darabi, P. Banerjee, and J. Liu, “Survey of wireless indoor positioning techniques and systems,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 37, no. 6, pp. 1067–1080, 2007.

[9] O. Woodman and R. Harle, “Pedestrian localisation for indoor environments,” in *Proceedings of the International Conference on Ubiquitous Computing (UbiComp’08)*, 2008, pp. 114–123.

[10] F. Evennou and F. Marx, “Advanced integration of WIFI and inertial navigation systems for indoor mobile positioning,” *EURASIP Journal on Advances in Signal Processing*, vol. 2006, pp. 1–11, 2006.

[11] A. R. J. Ruiz, F. S. Granja, J. C. P. Honorato, and J. I. G. Rosas, “Accurate pedestrian indoor navigation by tightly coupling foot-mounted imu and rfid measurements,” *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 1, pp. 178–189, 2012.

[12] E. C. L. Chan, G. Baciu, and S. C. Mak, “Orientation-based wi-fi positioning on the google nexus one,” in *6th International Conference on Wireless and Mobile Computing, Networking and Communications*, ser. WiMob 2010, pp. 392–397.

[13] E. Martin, O. Vinyals, G. Friedland, and R. Bajcsy, “Precise indoor localization using smart phones,” in *Proceedings of the International Conference on Multimedia (MM’10)*, 2010, pp. 787–790.

[14] A. Perttula, H. Leppäkoski, S. Tikkinen, and J. Takala, “WLAN positioning on mobile phone,” in *IAIN World Congress, Stockholm, Sweden*, 2009.

[15] J. Liu, R. Chen, L. Pei, W. Chen, T. Tenhunen, H. Kuusniemi, T. Kroeger, and Y. Chen, “Accelerometer assisted robust wireless signal positioning based on a hidden markov model,” in *Proceedings of the Position Location and Navigation Symposium (IEEE/ION PLANS’10)*, 2010, pp. 488–497.

[16] J. A. B. Link, P. Smith, N. Viol, and K. Wehrle, “FootPath: Accurate map-based indoor navigation using smartphones,” in *Proceedings of the 2011 International Conference on Indoor Positioning and Indoor Navigation (IPIN’11)*, 2011.

Performance Evolutions of Velocity-Aware Routing Protocol for Mobile Ad hoc Networks

Muneer O. Bani Yassein, Alaa N. Alslaity
Department of Computer Science
Jordan University of Science and Technology,
Irbid 21100, Jordan
masadeh@just.edu.jo, analslaity08@cit.just.edu.jo

Ismail M. Ababneh
Department of Computer Science
Al al-Bayt University
Mafraq 25113, Jordan
ismael@aabu.edu.jo

Abstract— **In recent years, Mobile Ad hoc Networks (MANETs) became an attractive research target. Several protocols were proposed to facilitate the vital operations of such type of Networks. Many routing protocols have been proposed in the literature, as routing operation is considered as one of the most important procedures used in MANETs. Among the so many protocols, on-demand routing protocols are of major contribution to handle the routing operations effectively. Ad hoc On-demand Distance Vector Routing protocol (AODV) stands to provide an excellent example of the on-demand routing protocols. AODV corresponds to the unique nature of MANETs by incorporating several features for discovering and initiating paths on an on-demand fashion, reducing both control and processing overhead, providing a multi-hop routing capability and maintaining the dynamic topology. Nevertheless, many opportunities for further improvements are still possible. In this paper, we attempt to incorporate mobility-aware features along with the AODV routing protocol features so as to handle mobility encountered by mobile nodes, to improve the performance and to add some promising capabilities. Our suggested protocol computes the node mobility periodically and uses this computed value to make useful routing decisions thereafter. Simulations are done using GloMoSim 2.03 simulator. According to the results, our proposed protocol proves its superiority over the original AODV protocol in terms of the reduced overhead and the increased packet delivery ratio.**

Keywords-Mobile Ad hoc Networks (MANETs); Routing; AODV; Mobility; Velocity Awareness.

I. INTRODUCTION

A Mobile Ad hoc Network (MANET) is a group of mobile nodes that communicate and collaborate with each other without the need to any means centralization or pre-existing infrastructures. Due to the lack of the centralized access points, the mobile nodes are required to act as both hosts and routers and the same time to perform the routing process properly. Mobile ad hoc networks can be used in numerous situations and can provide tremendous opportunities, particularly if there is a need for establishing a network for a limited period of time in a location where wired infrastructure is nonexistent or very difficult to deploy. The applications of MANETs include search and rescue operations, academic and industrial applications, and Personal Area Networks (PANs).

Compared with the other types of networks, MANETs have the following exclusive characteristics: bandwidth and transmission rate limitations, energy constraints and dynamic topology [9].

Mobile ad hoc networks, as the name indicates, are mainly characterized by the dynamic topology due the mobility of nodes, hence, the name “Mobile”. There are no restrictions on nodes mobility and nodes are free to move any time towards any direction and at any speed [2]. In addition to mobility issues, a MANET has security and energy constraints as well as bandwidth limitations.

In our work, we concentrate basically on mobility considering it as a major factor in MANETs that affects the overall performance of the network. This is because the frequent and high mobility of nodes can cause frequent link breakages, resulting in a less reliable routes and a more route re-initiation. The extra route discovery process requires more Route Request Packets (RREQ), Route Reply Packets (RREP), and Route Error Packets (RERR) [3], this in turn, leads to more control packets overhead.

The primary objective of this paper is to take the previously mentioned limitations into consideration to design and implement a stable and overhead efficient routing protocol. The proposed protocol concerned mainly on the network overhead caused due to the usage of uncontrolled flooding and that caused due to the mobility nature of MANETs. In this protocol, nodes calculate their mobility periodically and use it mainly to establish a reliable route towards the destination during route discovery process. Simply put, a reliable route is the one with low mobility, yet; low probability of failure.

The rest of this paper is organized as follows: Section 2 overviews the state of the art works in mobility aware protocols. Section 3 illustrates our proposed protocol and the methodology of quantifying mobility. The simulation environment and the experimental results are discussed in Section 4. Finally, Sections 5 concludes the paper and provides our future directions.

II. RELATED WORK

Mobility is considered as one of the main challenges in MANETs. In [4], link duration is proposed as mobility metric to evaluate the overall network mobility status. The authors defined link duration as the time period during which two mobile nodes remain within the transmission range of each other. They used this metric as a major clue to indicate network performance, because long-lived links increase network stability.

Idrees et al. [5], proposed a mobility-aware scheme such that Hello Packets were used to enhance mobility awareness in the AODV. Upon receiving a Hello Packet, along with the assistance of the GPS coordinates of the source node, a lightweight mobility aware agent on each node of the network compares these coordinates with previous ones and then can determine information about the mobility of the originator node. When a node receives a RREQ packet and needs to send a RREP (it is either the destination or it has an active route to the desired destination), it will use the mobility awareness to choose the best neighbor which is not moving frequently. This process of selecting a best neighbor is done at each intermediate node. As a result, a path with the maximum number of low mobile nodes is established between source and destination.

In their proposed work, Qin et al. [6] considered three parameters that are used for monitoring the mobility status by individual nodes. These parameters are: node degree, average link duration and number of link breakages. These parameters can be obtained by "hello" messages exchange and they assist each node in sensing the status of its neighbors. In addition, a node can know how many links may be broken if it has not received "hello" messages from the previously connected nodes within some period of time, and then it can calculate the link duration for each broken link, and the average link duration at the moment. In order to examine the effect of the three proposed parameters, they are deployed and monitored at different mobility levels and with different mobility models.

Liang Q and Thomas [6] observed that the number of link breaks obtained by a node has nearly linear relationship with node mobility, which is defined as the relative speed between two nodes. The correlation is based on the average of all the nodes in the network, and the value of this metric fluctuates significantly for each node during the simulation.

In [7], Enneya et al. proposed a mobility-aware method to improve the performance of AODV. They define mobility metric and used it in both route discovery and route maintenance. In route discovery, the hop-count metric that is used in standard AODV is dropped, and it is replaced with a combination of two mobility parameters: average and mean of the "calculated mobility" along the path between any source node and destination. Consequently, more stable routes were obtained. In route maintenance, the local repair mechanism was extended in order to avoid the RERR packets by allowing the node that detects a broken link to choose an alternative route based also on the mobility metric. This affects the overall overhead of re-initiating the route discovery process and also reduces the use of RERR packets.

III. METHODOLOGY

Mobility is of a major importance factor in the ad hoc networks environments. Depending on the nodes' mobility level, the overall network topology can be described. That is, if nodes are of low mobility and change their physical location seldom, then the network topology is said to be stable (or semi-stable). However, if nodes move very rapidly, then no expectation can be made on the network topology because what holds true for a specific period of time cannot be guaranteed to still true at the time after. Through the literature, it is shown that the majority of ad hoc routing protocols are incapable to handle high mobility.

In the literature, there are many mobility metrics that are used to quantify nodes mobility [7]. In our approaches, we depend solely on the locally available topological information, such that the change in the (x, y) coordinates for a particular node provides a good indication of the network movement pattern and mobility.

This section provides a detailed discussion of our proposed protocol and the contribution it adds over the traditional AODV protocol.

A. Our Protocol

In our work, we propose a Velocity-aware Ad hoc On-demand Distance Vector (VA-AODV) protocol that is capable to periodically compute mobility and make useful routing decisions accordingly. Our VA-AODV protocol offer major contributions and improve the performance of the original AODV protocol.

Unlike the AODV, wherein, the source node broadcasts the RREQ message to all its neighboring nodes (regardless to their mobility status) for the sake of finding the intended destination node, our VA-AODV protocol takes into consideration the mobility of neighboring nodes and picks the nodes with lower mobility to perform the route discovery process. In other words, in our VA-AODV, each node computes its own mobility periodically (i.e., every HELLO_INTERVAL). Then, broadcasts the value of its own mobility along with the HELLO message to inform its neighbors about its mobility status. Each node in turn, updates its neighbor table by adding ascending-ordered entries of (node ID, velocity) pairs for all neighbors, such that the ascending order is based on nodes' velocity.

In VA-AODV, when a source node wishes to communicate with a destination and it does not have a route to that destination, it initiates a route discovery process by referring to its neighbor table and picking a set of nodes with lower velocities to participate in the route discovery process (instead of choosing the whole neighbors, as the case in AODV).

We refer to the selected set of neighbors, which will participate in the route discovery process as the *CoveringSet*, and it is defined as the set of 1-hop neighbors that cover the overall 2-hop neighbors. The *CoveringSet* should satisfy two conditions; it should ensure full coverage for the 2-hop neighbors, and it should consist of the neighbors with lower velocities as much as possible. Building *CoverinngSet* is a distributed process in that each node builds its own *CoveringSet* independently.

The process of VA-AODV is done as follows: when a node (S) wants to communicate with a destination node (D) that is not within the transmission range of S, it firstly creates its *CoveringSet* using its neighbor table. Starting from the first entry in the neighbor table (remember that this table is sorted in an ascending order based on the velocity), S checks whether the current neighbor add additional coverage for some 2-hop nodes or not. If so, current neighbor is inserted to the *CoveringSet*, otherwise S continues with the next neighbor. This process repeated until achieving full coverage for the entire 2-hop neighbors regardless the number of nodes that are in the CoveringSet. Hence, the number of nodes participated in the CoveringSet are not defined in advance, rather it depends on the coverage condition (i.e., the CoveringSet should covers the entire 2-hop neighbors). Once node S finished building its CoveringSet, it appends this set to the RREQ packet and broadcast it to its neighbors, only those neighbors who's IDs included in the CoveringSet will relay the packet. The same applies for the intermediate nodes where they look their neighbor table up and decides which neighbors are allowed to relay the RREQ further. Therefore, the overall selected route is stable and more reliable.

B. *Velocity Quntfication*

In our VA-AODV protocol, we assume that each node is equipped with a GPS device from which it obtains its own (x, y) coordinates. The availability of position information as well as the continuous tracking of the changes in this information within a specific period of time t provides each node with the ability to calculate its own distance crossed during that time t, which can be used for the purpose of speed calculation.

To explain our velocity quantification methodology, let us denote the position of node i at time t as Pi,t which is actually obtained from the coordinates pair (xt, yt). Further, let the position of the same node at time t+α be denoted as Pi, t+ α which corresponds to (xt+ α, yt+ α), then the crossed distance for this node during the time period T = (t+α) – t is denoted as DT and is computed as given in equation 1:

$$Dt = \sqrt{(x_{t+\alpha} - x_t)^2 + (y_{t+\alpha} - y_t)^2} \qquad (1)$$

Because each node sends hello messages to its neighbors every HELLO_INTERVAL, it can calculate its velocity (or speed) at the end of each HELLO_INTERVAL and append the value of speed with the hello message. In other words, let ε be the HELLO_INTERVAL time, and given the crossed distance DT, then the velocity Vε of node i during the period of time T can be calculated as follows:

$$V_{\varepsilon} = \frac{Dt}{\varepsilon} \qquad (2)$$

Upon receiving the hello message, each recipient node updates its neighbor table such that a new entry will be added for the originator of the hello message if it does not already exist. The added entry will be of the form <nbrAddr, Vε>, where Vε is the velocity (speed) at which the distance DT was crossed.

C. *Our Contribution*

Our Velocity-Aware Ad hoc On-demand Distance Vector (VA-AODV) routing protocol is designed to work in mobile ad hoc networks as an adaptive, decentralized and mobility-aware protocol that outperform the original AODV in the following aspects: First, the VA-AODV controls the route discovery process by selecting a set of nodes (with low velocity) to send (or relay) the RREQ messages, this in turn will reduce the control overhead associated with the traditional AODV. In addition, the nodes perform mobility quantification in a simple and distributed manner based on the locally available information about position changes. This in fact, provides very precise information about velocity. Our mechanism of mobility aware routing guarantees more stable and reliable routes since each node chooses only stable routes, this will decrease the number of broken links, and thus, reduces the number of reinitiating route discovery trials and reduces the number of dropped packets, as consequent, the packet delivery ratio is increased and the network overhead is decreased. In particular, our velocity-aware approach contributes mainly in terms of reducing the overall control overhead (since the number of relayed RREQ packets by intermediate nodes is reduced).

IV. Performance Evaluation

In order to evaluate the performance of our VA-AODV protocols, the proposed mechanism is simulated using GloMoSim 2.03 simulator [8]. The simulation environment and parameters are clarified in the subsequent sections.

A. *Simulation Environment*

The simulation area that is considered for simulations is 600 m × 600 m. The mobility of nodes is represented by the choice of a uniform speed between a minimum speed, vmin=0 and a maximum speed vmax, where vmax = 2, 5, 10, 20 and 40 m/s. this wide range of speeds used (from 2m/s up to 40m/s) is selected carefully to show us the behavior of the proposed protocol for any speed. The Mobility model used through simulation is the Random way point and the channel capacity is 2 Mbps. We aim to assess the behavior of VA-AODV in the dense networks, so that, and through empirical, all the experiments done using 40 nodes with 250m transmission range. Also we used the Constant Bit Rate (CBR) traffic generator, and the number of sources is set to be 24 nodes selected randomly and send to a randomly chosen different receivers. Each source generates 1 and 5 packets/seconds for different scenarios. The time for simulation is 300s and bidirectional link between each pair of adjacent nodes is considered. In the MAC layer (i.e., Data Link layer), we used the IEEE 802.11 communication protocol.

B. *Simulation Parameters*

We evaluate the performance of our proposed protocol using the following simulation parameters [11]:

- **Packet Delivery Ratio (PDR)**: the packet delivery ratio is a ratio of the correctly delivered data packets.
- **Routing Overhead**: the routing overhead ratio is the ratio of the network control packets sent to the correctly delivered data packets.
- **Saved Rebroadcasts (SRB)**: the saved rebroadcast represents the ratio of the number of route request (RREQ) packets retransmitted to the total number of route request (RREQ) packets received by any node [10].

C. *Simulation Results*

In this section, we provide a performance comparison between the AODV protocol and our proposed protocol, VA-AODV in terms of control overhead, PDR and SRB. The following scenarios show us the effects of speed with number of nodes equal 40 nodes; in the first scenario, each source node sends 1 packet/second (i.e., the traffic load= 1Pkt/s), while in the second one we used traffic load = 5Pkts/s.

Figures 1, 2 and 3 show the performance results for the control overhead, PDR and SRB, respectively for a number of nodes =40 and a traffic load of 1 packet/second. Figure 1 shows the superiority of our protocol over the AODV in terms of reducing the average control overhead. This is due to the fact that our protocol tends to control flooding by selecting only a subset of nodes with low mobility to retransmit packets. This reduction of retransmissions saves a lot of control packets (RREQ, RREP, and RERR) from being sent, and this reduces the overall routing overhead. The figure shows also that as the maximum speed of nodes increases, the overhead encountered by AODV increases as well. This is because the faster the node's movement speed, the less stable the links are, and the more the link breakages. The instability caused by high node speed requires sending more control packets (RREQs) needed for route re-initiation and (RERR) needed for local repair.

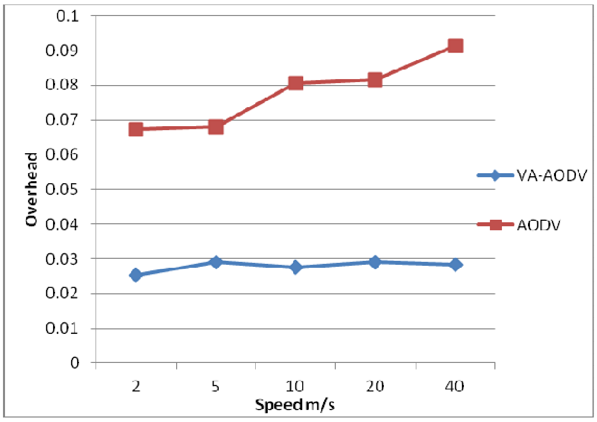


Figure 1. Average Overhead vs. Speed

The results in Figure 2 show that the Packet Delivery Ratio achieved by VA-AODV is much better than that of the AODV, especially for high speeds (20 and 40 m/s). This is expected because the velocity awareness of our protocol reduces the number of broken links by choosing the only stable nodes. This in turn guarantees a better delivery of packets.

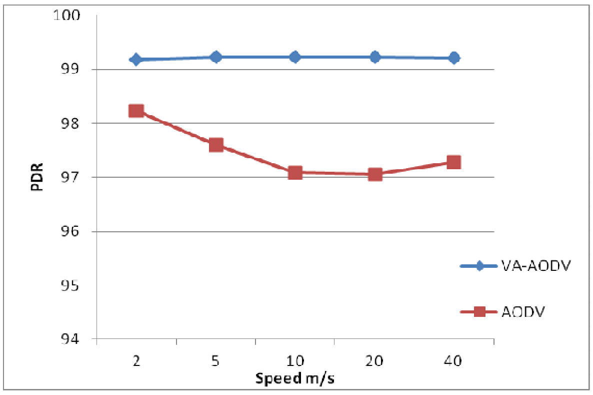


Figure 2. Average PDR vs. Speed

Figure 3 depicts the saved rebroadcasts achieved by our protocol in comparison with that achieved by AODV. As the figure shows clearly, our protocol significantly outperforms the AODV in terms of avoiding redundant retransmissions of the received packets. In addition, our protocol proves its stability and ability to save rebroadcasts even with high speed values, whereas the AODV protocol degrades clearly as the nodes speed increases.

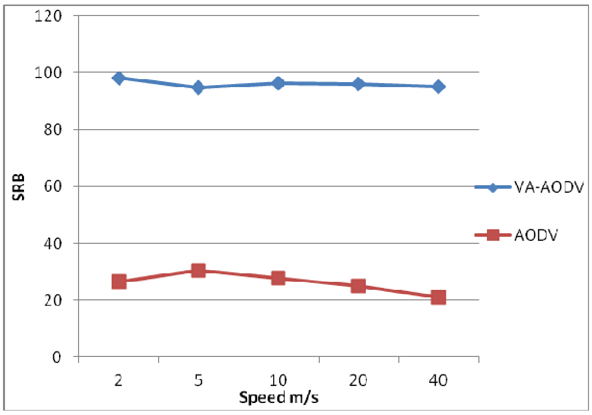


Figure 3. SRB vs. Speed

Figures 4, 5 and 6 show the performance results for the control overhead, PDR and SRB, respectively for a number of nodes =40 and a traffic load of 5 packet/second.

Figure 4 illustrates the superiority of our protocol over the AODV for all speed values. It can be inferred that with very low speed value (i.e., speed= 2m/s), the performance of both AODV and VA-AODV are almost similar, while the performance enhancement becomes evident for higher speed values.

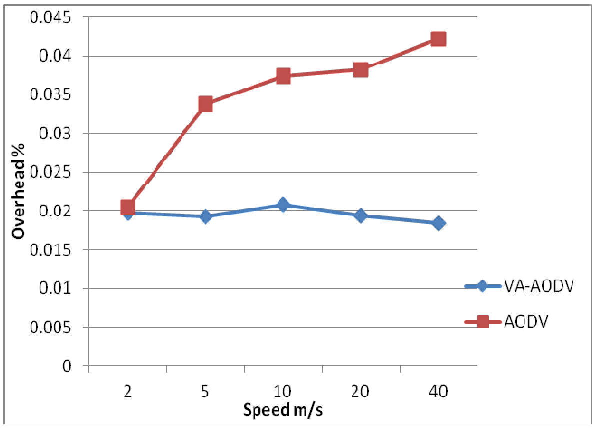


Figure 4. Average Overhead vs. Speed

Figure 5 shows that for different speed of nodes, and as the number of packets transmitted increases, the average packet delivery ration decreases for the AODV while it remains stable for our protocol.

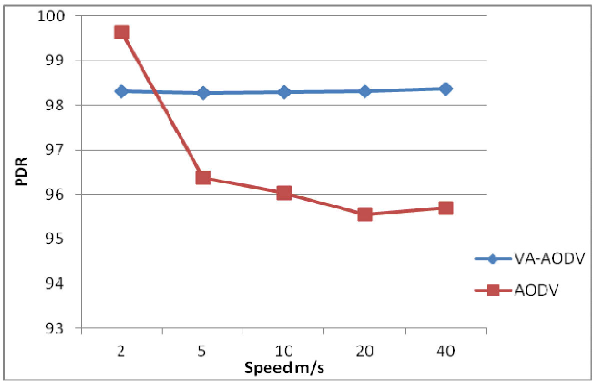


Figure 5. Average PDR vs. Speed

By varying the maximum nodal speed over a range of 2, 5, 10, 20, and 40 m/s and having and having a traffic load of 5 packets/second, it can be shown in Figure 6 that VA-AODV can achieve higher SRB when compared against AODV which uses blind flooding as a main mechanism for route discovery, thus redundant retransmission of packets occurred frequently.

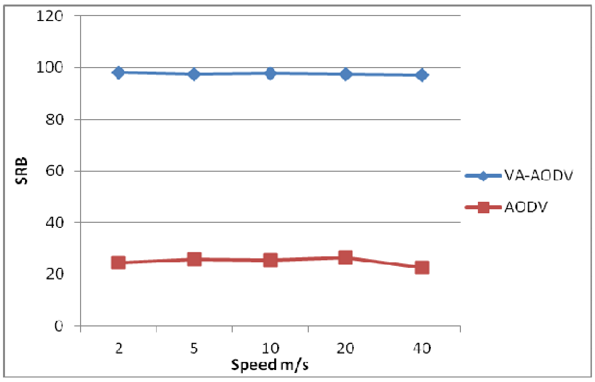


Figure 6. SRB vs. Speed

Based on the simulation results illustrated in this section, it is clear that our proposed VA-AODV protocol enhances the performance of the original AODV protocol in terms of reducing the control overhead; increasing the packet delivery ratio and increasing the saved rebroadcast. The VA-AODV significantly outperforms the AODV protocol in terms of reducing overhead by 69%. Regarding packet delivery ratio, the experiments show that our protocol outperforms AODV by 2.79%. Finally, our protocol achieves substantial improvement of the saved rebroadcast performance metric, such that the VA-AODV outperforms AODV by 77.86%. Moreover, the results show that VA-AODV ensures stability, in that it gives stable results for different speeds.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an ad hoc on-demand, velocity aware routing protocol that achieves significant enhancement over the AODV in terms of reducing the average control overhead, increasing the average packet delivery ratio and increasing the saved rebroadcast of packets. The proposed VA-AODV protocol depends on the change of a node's position during a specified period of time to calculate average node's velocity as mobility indicator in order to assist in making a proper routing decision. Taking nodes velocity (as a mobility metric) into consideration ensures a better routing performance in terms of decreasing control packets overhead, increasing packet delivery ratio and increasing the number of save rebroadcasted packets.

Although taking nodes velocity as a major factor for routing decisions gets better performance, it is not enough to depend on the node's absolute speed. There are three main parameters of the mobility; speed, position, and direction. In general, only one of these parameters is considered in selecting the next hop during the route discovery process [34]. Indeed, it is not sufficient to consider only one of these parameters as the only parameter for route discovery process. Thus, we should add other parameters (In addition to the velocity) to the algorithm in order to make it more precise and more reliable. Moreover, the proposed protocol needs more evaluation methods and simulations to ensure its superiority over other protocols.

References

[1] M Abolhasan, T Wysocki, and E Dutkiewicz, "A review of routing protocols for mobile ad hoc networks", Ad Hoc Networks, 2(1), pp. 1-22, 2004.

[2] E. M. Royer and C. E. Perkins, "Evolution and future directions of the ad hoc on-demand distance-vector routing protocol," Ad Hoc Networks, vol. 1, no. 1, pp. 125-150, July 2003.

[3] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing", Mobile Computing Systems and Applications, 1999. Proceedings. WMCSA '99. Second IEEE Workshop on Digital Object Identifier, pp. 90-100, 25-26 Feb 1999.

[4] J. Boleng, W. Navidi, and T. Camp, "Metrics to Enable Adaptive Protocols for Mobile Ad Hoc Networks", Proceedings of the International Conference on Wireless Networks (ICWN '02), Las Vegas, Nevada; pp. 293-298, 2002 June.

[5] M. Idrees , M. M. Yousaf, S. W. Jaffry, M. A. Pasha, and S. A. Hussain, "Enhancements in AODV Routing Using Mobility Aware Agents", IEEE International Conference on Emerging Technologies, ICET 05, pp. 98-102, 17-18 September 2005.

[6] L. Qin and T. Kunz, "Mobility Metrics to Enable Adaptive Routing in MANET", IEEE International Conference on Wireless and Mobile Computing, Networking and Communications,. (WiMob'2006), pp. 1–8, 2006.

[7] N. Enneya, M. Al Koutbi, and A. Berqia, "Enhancing AODV Performance based on Statistical Mobility Quantification", The IEEE International Conference on Information & Communication Technologies: from Theory to Applications, ICTTA06, pp. 2455-2460, April 2006.

[8] X. Zeng, R. Bagrodia, and M. Gerla, "GloMoSim: a Library for Parallel Simulation of Large-Scale Wireless Networks", Proceedings of the 12th workshop on parallel and distributed simulation, Bnaff, Alberta, Canada, pp. 154-161, July 1998.

[9] M. Bani Yassein, S. Nimer, and A. Al-Dubai, "A new dynamic counter-based broadcasting scheme for Mobile Ad hoc Networks," Journal of Simulation Modelling Practice and Theory, vol. 19, no. 1, pp. 553-563, 2011.

[10] B. Sun, C. Gui, Q. Zhang, B. Yan, and W. Liu, "A Multipath on-Demand Routing with Path Selection Entropy for Ad Hoc Networks", The 9th International Conference for Young Computer Scientists, ICYCS, pp. 558-563, Nov. 2008.

[11] A. Hanashi, A. Siddique, I. Awan, and M. Woodward, "Dynamic Probabilistic Flooding Performance Evaluation of On-Demand Routing Protocols in MANETs", Proceedings of the 2008 International Conference on Complex, Intelligent and Software Intensive Systems; pp. 200-204, 2008.

The Context Manager: Personalized Information and Services in Mobile Environments

Pablo Curiel, Ana B. Lago
Deusto Institute of Technology - DeustoTech
MORElab - Envisioning Future Internet
University of Deusto
Avda. Universidades 24
48007 - Bilbao, Spain
Email: {pcuriel, anabelen.lago}@deusto.es

***Abstract*—In this paper, we present a context management infrastructure for mobile service environments. Due to the remarkable advances the mobile technologies have experimented in the last years, mobile devices have become one of the most promising scenarios for the deployment of context-aware systems. For this reason, the aim of the proposed infrastructure is to provide context information to applications and services, both executed in the end-user terminals or in the network, enabling them to adapt their behaviour to each user and situation. The solution here exposed relies on semantic technologies and open standards to improve interoperability, and is based on a central element, the context manager, which acts as a central context repository and carries out demanding tasks in behalf of mobile devices.**

***Index Terms*—context management; semantic technologies; pervasive computing; mobile computing; context-aware services**

I. INTRODUCTION

Context awareness is a subject which has attracted interest since it was introduced by Schilit and Theimer in [1]. The reason for this growing interest is that, by using context information, context-aware systems are capable of adapting their working behaviour, as well as providing information and services more relevant in the situation of the end user. This way, context is defined as any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves [2].

One of the most promising scenarios for the deployment of context-aware systems are mobile phones. Since its appearance, mobile technology has undergone remarkable changes. First of all, mobile phones have become an everyday-use device, an almost essential element in our daily lives and which we carry with ourselves every time. But with the outstanding step forward in the technology of these devices in the last few years, this relevancy has become more noticeable. Nowadays, mobile phones are powerful hand-held computers capable of carrying out plenty of tasks and remaining online at all times. In addition, they include numerous sensors, and consequently have access to a great amount of personal and environmental information. At the same time, mobile phones are generally used for short and rapid interactions and in distractive environments, so providing information and services adapted to each situation is of great value for the end users. Therefore, mobile devices can greatly benefit from context-aware systems.

Taking this into account, in this paper we present a context management infrastructure for mobile environments. The aim of this infrastructure is to provide context information to applications and services which are either executed in the end-user terminals or in the network. To deal with the changing nature of this kind of environments and the wide range of devices present in them we propose a solution which relies on semantic technologies, easing the knowledge sharing and interoperability, and which provides generic mechanisms to deal with context information. Also, as mobile devices have limited computational capabilities, the context management infrastructure takes most of the computational burden of dealing with context, enabling them to simply request the information they need to carry out their tasks.

The remaining of this papers is structured as follows: in Section 2, related work is reviewed. In Section 3, the context management infrastructure is described. In Section 4, the implementation details are explained along with a preliminary validation scenario. Finally, in Section 5, discussion and future work are exposed.

II. RELATED WORK

Numerous context-aware systems for mobile environments have been developed in the last years.

The work in [3] proposes a Context Managing Framework whose aim is to provide context information to mobile applications in order to adapt their behaviour according to that context. As our system, it uses a blackboard approach, based on a central context server, to decouple source and consumer communication. But in contrast to our solution, it places this central server inside the mobile phone and limits the context information usage to each mobile phone's own applications.

CoBrA project [4] proposes an architecture for supporting context-aware systems in smart spaces. It is based in a central agent, a broker which maintains a shared model of the context representing all the entities in a given space. It also uses a policy language to allow users to define rules to control the use and the sharing of their private contextual information. However, the system is not explicitly designed to operate in mobile environments.

In [5], CASS, a middleware for context-aware mobile applications is introduced. Its main goal is to give service to devices with low processing capabilities, like mobile phones. For this purpose, as we do in our system, it delegates demanding tasks regarding context processing to external entities in the network. It also resembles our solution in making usage of a central context repository. But, whereas we adopt a semantic representation for that central repository, which enables a more flexible context processing and sharing, CASS uses a database for this purpose, which follows no semantic representation and relies mainly on a rule-based engine to process that context.

The MOBE architecture by Coppola et al. [6] intends to send applications to mobile devices depending on the context the user is in. MOBE differs from our solution in that it delegates most of the computing burden in the mobile phones, making them responsible for both gathering and processing context information, as well as deciding which applications to request in each case. It also restricts context information usage to select and adapt mobile applications, while in our system both mobile applications and applications executed in the network take advantage of this information.

Finally, in the context dissemination middleware [7], a peer-to-peer approach, based on web services is used to share context information between mobile devices. It uses a rule-based publish/subscription paradigm to enable this context information distribution. However, in contrast to our system both the context information both the rules are structured in an ad hoc format based on XSD (XML Schema Definition Language) [8], instead of a standard semantic representation format, which makes it difficult aspects like interoperability with other systems.

III. Context Management Infrastructure

The context management infrastructure is responsible for dealing with context information during its whole life cycle, from the provisioning of this type of information to the usage of it in benefit of the user, including all the intermediate processing needed to present that information in the way needed by the entities which make use of it. Therefore, it is a critical element in every context-aware system, and its design must be carefully carried out to guarantee the right operation of the whole system.

A. Context management requirements

In this section the requirements defined for the context management infrastructure are detailed.

First of all, to ease knowledge sharing and interoperability between the different entities in the system, the managed context information must be represented following a semantic model. Among the existing approaches to model context information, ontologies are used, because, as Strang and Linnhoff-Popien point out, they are the best solution for this task in ubiquitous computing applications [9]. Semantic technologies allow computing entities to better understand the meaning of the information they are working with, enabling them to perform reasoning and thus problem solving and decision taking. They also provide more flexible and expressive logical connections an relationships between data, even among those coming from different sources. And due to they embracing an open-world approach, they have the ability to better deal with the uncertainty and incompleteness of data in the real world. Therefore, all the entities in the system share a common ontological model and exchange context information according to it. Moreover, all the interactions between entities of the system that involve context information follow standards designed to work with semantic data, like RDF [10] to represent and share context information, and SPARQL [11] to query this kind of information.

Due to the limited computational capabilities of mobile devices, the context management infrastructure should take most of the computational burden of dealing with context information. For this purpose, we introduce a central element, the context manager. This element acts as a central repository, receiving context information from the sources, processing it as required and storing it, allowing consumers to access it. This blackboard approach enables resource-limited devices to only act as context source or consumers, relieving them from executing demanding tasks with context information. Indeed, it enables a data-centric approach granting independence between context sources and providers, as the context manager receives context information from the sources and stores it, responding consumers' queries about that information. Thus, it prevents consumers from asking directly to the sources and at the same time enables them to only think about what information they need, not where it comes from.

However, the context manager is not an atomic element, but consists of a series of independent and reusable components, which carry out different tasks with context information, facilitating system scalability, as they are even able to operate in different machines. Nonetheless, the context manager must provide a unique entry point to the context providers and consumers, known as the context manager API, in order to provide its functionality in a standard an unified way. This separation between the context manager logic and the access to it also enables providing different communication protocols to expose its functionality.

Finally, in order to meet the requirements regarding context information accessing of the different kinds of consumers present in the environment, both synchronous and asynchronous access are be provided. Thus, the context manager must be able of responding synchronous queries as well as registering queries, checking when those queries match and asynchronously notifying the corresponding consumer of this event.

B. The Context Manager

As detailed in the previous section, the context manager is the central element of the context management infras-

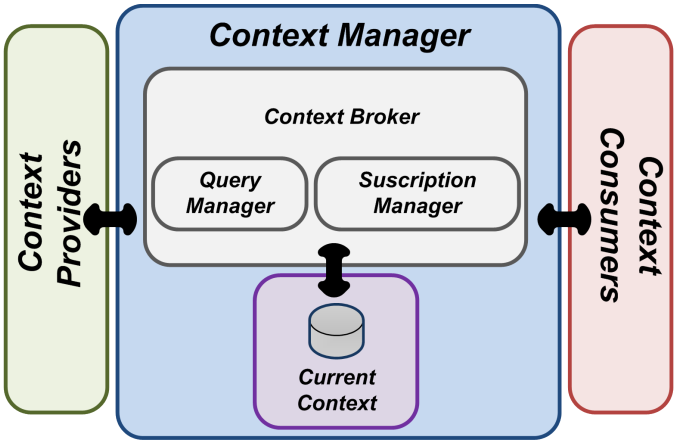


Fig. 1. Logic Architecture

tructure. It is comprised of various components which carry out different tasks with context information and exposes a unified interface for context sources and consumers to access its functionality, relieving them from carrying out demanding tasks with this type of information. In this section, the different components of the context manager, shown in Figure 1, are introduced and detailed.

The **current context** is the element which stores the context information which is valid in each moment, that is, the one that represents the current status of the entities which are considered part of the context in every moment. This information is stored as an RDF triplestore following the ontological model shared by the rest of the system and can be kept either in memory or disk. The **context broker** is the component responsible for managing this context repository. This way, it receives the context information from the sources, stores it in the current context and attends the consumer requests querying these repository. At the same time, it has two subcomponents in charge of answering the consumers' queries: the **query manager**, which deals with synchronous queries, and the **subscription manager**, responsible for registering consumers' subscriptions for context information changes and notifying them asynchronously when those changes take place.

C. The Context Management API

As defined, the context manager, even if it is composed of various independent components, exposes a unique entry point to its functionality. This API exposes several methods to manage context information in a model-independent way and relying on standard technologies. Indeed, the methods here detailed are merely an access layer, as the operational logic belongs to each corresponding component, so several communication protocols could be implemented to support as many devices as needed.

The methods exposed in the context manager API are the following:

- **Add Context Info**. This method enables a context source to add or update the current context space by providing information in RDF format.
- **Remove Context Info**. Using this method enables a context source to delete a context instance from the context space given its identifier.
- **Get Context Info**. This method enables a context consumer to retrieve a known instance from the current context given an identifier.
- **Query**. By calling this method, a context consumer can synchronously access the current context space providing an SPARQL query. The context manager will execute this query and return the corresponding context information to the context consumer in RDF format.
- **Subscribe**. This method enables a context consumer to asynchronously access the current context space. The consumer provides an SPARQL query which the context manager will register and a callback address which the second will use to asynchronously notify the first when the query is matched.
- **Unsubscribe**. To delete a subscription created with the previous method, a context consumer needs to invoke this method providing the Subscription ID which corresponds to the subscription that wants to remove.
- **Notify**. This method is not exposed by the context manager itself, but by those consumers which make use of the Subscription system. This way, each time an asynchronous query is satisfied, the context manager invokes this method of the corresponding consumer, providing the context information in RDF format.

IV. Implementation and Validation Scenario

As a first validation step of our proposal, we have developed a prototype of the context manager which fully implements the API described in the previous section, as well as an end-user application and a service, which act as context source and consumers.

The context manager is developed in Java using the OSGi component framework. To work with semanticized context information, we use the well known Jena2 [12], semantic web toolkit, and Jenabean [13], a library which bridges the gap between working with RDF graphs and Object-Oriented Programming. The context manager API is exposed as a RESTful [14] interface.

To test the functionality, a validation scenario is proposed, which involves an end-user application developed for the Android mobile OS and a service which tracks both user location and Twitter accounts to infer user status and availability, in order to suggest plans to nearby friends. The scenario goes by as described. John, Mike and Greg are friends. The two first live in Madrid, while the third lives in Barcelona. The three of them are tech-savvy, and therefore both social network and smartphone heavy users. They also use a social alerts service, which helps them to keep in touch with their friends. One day, Greg travels to Madrid, and the context manager, which periodically receives users' locations, reported by their smartphones, detects that the three friends are nearby and notifies the social alerts service. However, Mike has recently updated his Twitter account, informing his followers of his busy day,

’What a day! I’ve got 3 meetings in a row!’. Consequently, the social alerts service, who periodically checks the three friends’ Twitter accounts, infers that Mike is not available, but sends alerts to John and Greg suggesting them to arrange a meeting. Finally, if both agree, it checks their user profiles to select a restaurant they both like to organize a lunch.

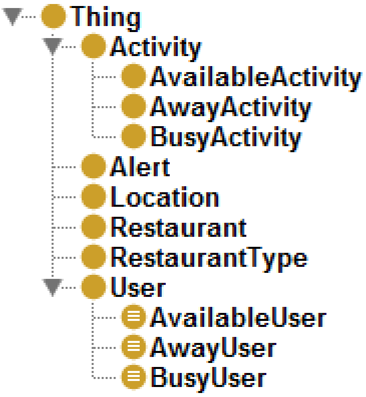


Fig. 2. Ontology of the validation scenario

To enable context information sharing between the context manager, the mobile application and the social alerts service, the ontology show at Figure 2 is used, which models information about Users (their personal data, preferences, current location and current activity), Locations, Restaurants and Alerts sent to users.

V. Conclusion and Future Work

In the present article, we have proposed a context management infrastructure for mobile environments, in which applications and services both executed in the end-user terminals or in the network use this context information to become more relevant for the users. As this kind of environments are of a very changing nature an a wide range of devices coexist in them, our proposal offers generic and abstract methods to work with context information, and relies on semantic technologies and open standards, trying to offer a solution as interoperable and extensible as possible.

On to other matters, even if the computational capabilities of mobile devices have noticeably increased in the last years, their ability to carry out demanding tasks with context information is limited, so our proposal delegates this heavy tasks in a central element, the context manager, which acts as a context information repository, exposing an interface to provide and consume context information. This also enables context source and consumer independence and a data-centric approach, in which consumers only have to worry about what information they need, not where to retrieve it from.

Our next steps will involve including the privacy and security policies required for this kind of systems, which grant both preventing unsolicited access to sensitive data while enabling legitimate access to those entities which need it to successfully carry out their tasks. In addition, implementing support for a context history component, which keeps track of the context information changes, could be a subject of interest as long-term user behaviour and trends can be inferred using this kind of information. Finally, a more rigorous validation will be carried out, involving performance tests with more demanding real-life use cases, in order to detect possible weakness and assess the validity of the proposed solution.

Acknowledgment

This work is supported by the CENIT Research Program, as part of the INGENIO2010 Spanish National Fund.

References

[1] B. Schilit and M. Theimer, “Disseminating active map information to mobile hosts,” *IEEE Netw.*, vol. 8, no. 5, pp. 22–32, 1994.

[2] A. K. Dey, “Understanding and using context,” *Personal Ubiquitous Computing*, vol. 5, no. 1, pp. 4–7, Jan. 2001.

[3] P. Korpipää, J. Mantyjarvi, J. Kela, H. Keranen, and E. Malm, “Managing context information in mobile devices,” *IEEE Pervasive Comput.*, vol. 2, no. 3, pp. 42–51, 2003.

[4] H. Chen, T. Finin, and A. Joshi, “An intelligent broker for context-aware systems,” in *Proc. of Ubicomp*, Seattle, Washington, USA, 2003, pp. 183–184.

[5] P. Fahy and S. Clarke, “CASS-a middleware for mobile context-aware applications,” in *Proc. Workshop on Context Awareness, MobiSys*, 2004, pp. 304–308.

[6] P. Coppola, V. Della Mea, L. Di Gaspero, S. Mizzaro, I. Scagnetto, A. Selva, L. Vassena, and P. Riziò, “MoBe: context-aware mobile applications on mobile devices for mobile users,” in *Proc. International Workshop on exploiting context histories in smart environments, ECHISE*, Munich, 2005.

[7] G. Gehlen, F. Aijaz, M. Sajjad, and B. Walke, “A mobile context dissemination middleware,” in *Proc. International Conference Information Technology, ITNG’07.*, 2007, pp. 155–160.

[8] XML Schema Working Group. (2010, Jan.) XML schema. http://www.w3.org/XML/Schema/. [Last accessed on August 10, 2012].

[9] T. Strang and C. Linnhoff-Popien, “A context modeling survey,” in *Proc. Workshop on Advanced Context Modelling, Reasoning and Management*, Nottingham, England, 2004.

[10] RDF Working Group. (2004, Feb.) RDF - semantic web standards. http://www.w3.org/RDF/. [Last accessed on July 5, 2012].

[11] E. Prud’hommeaux and A. Seaborne. (2008, Jan.) SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query/. [Last accessed on July 5, 2012].

[12] Apache Software Foundation, “Apache jena,” http://jena.apache.org/, Apr. 2012, [Last accessed on August 10, 2012].

[13] “Jenabean,” http://code.google.com/p/jenabean/, Feb. 2010, [Last accessed on August 10, 2012].

[14] L. Richardson and S. Ruby, *RESTful web services*. O’Reilly Media, 2007.

An Ontology-based Context Management System for Smart Environments

Laura M. McAvoy[1], Liming Chen[*], Mark Donnelly [*]
University of Ulster, UK
[1] McAvoy-L3@email.ulster.ac.uk
[*]{l.chen, mp.donnelly}@ulster.ac.uk

***Abstract* – This paper proposes an ontology-enabled system for context management for smart environments. Central to the system is ontological sensor modelling, which attaches metadata and meaning to sensor data, thus supporting data repurposing and high-level content recognition. In addition, semantic sensor descriptions allow sensors to be automatically identified whenever they are put into an environment. Based on this, a novel plug-n-measure data acquisition mechanism has been developed to automatically detect and recognise new devices and update the contextual data relating to these devices on a real-time basis. The context management system has been developed based on the latest semantic technologies and deployed in an intelligent meeting room. The paper describes an experiment and presents initial results, which has demonstrated that the system is working.**

Keywords-context management; ontology-based context modelling; inference and reasoning; smart environments; plug-n-measure mechanism.

I. INTRODUCTION

In a smart environment, computer systems interact seamlessly on a continual basis with occupants through the use of context-aware enabling technology [1]. Context is a term used to define any information which can be used to characterise the state of an entity. An entity is defined as “a person, place or object that is considered relevant to the interaction between a user and an application, including the user and applications themselves” [2]. Context-awareness is the ability to use this context in a relevant manner to update information, which may be of use to a user based on what they are doing at that time.

It is possible to make an environment smart and context aware by monitoring an inhabitant’s behaviours and environment in real-time along with other aspects. These other aspects include; fusing and interpreting the multiple modalities of signals and features, inferring the behaviour, anomalies or changes of the inhabitant continuously and providing application specific functions to the inhabitant’s needs and requirements [3]. Context is therefore very important within Smart Environments.

Various context management systems exist, including “The context toolkit” [4], the framework for managing context information within mobile devices [5] and the framework [6] which provides a graphical user interface (GUI) to the user.

However, whilst progress has been made within the area of smart environments and context management systems, some challenges still remain. This paper proposes an integrated context modelling and management system to deal with some of these challenges. The challenges are performing the repurposing of data [21], temporal reasoning [22] and multiple levels of context information [23]. The system which is proposed in this paper uses semantic technologies in particular, ontologies to model sensors and smart devices across a number of smart environment domains and exploit semantic reasoning to reason, store and share this information for various applications.

In previous work, the W4H model [23] is based on disaster specific domains and the context information associated with them. Unlike the W4H system which is domain specific, the system proposed in this paper can deal with context information and the multiple levels of it across a wide variety of domains through the use of a diverse ontology. The system is also able to handle large amounts of data which may be associated with certain larger domains due to the semantic repository that is used. Whilst temporal reasoning has been previously researched [22], the ability to perform temporal reasoning and access context histories remains a challenge. This system allows applications to access current context, temporal reasoning and context histories which may be of use to them. Finally the repurposing of data can be achieved through the use of the various architecture components.

A novel plug-n-measure mechanism for data acquisition is also introduced and discussed within this paper. The mechanism can detect new devices which are added to the environment, on a real-time basis. The context management system was also tested on a real-time basis within an intelligent meeting room as part of an experiment and application scenario. What the paper does not address is data extraction and analysis, which still has to be performed on the collected data.

The rest of this paper is organised as follows; Section II examines related work, opportunities and challenges within this area. Section III discusses the system architecture and the components that reside within. Ontology-based context management is described and a plug-n-measure mechanism is introduced and discussed. Section IV presents the system implementation and evaluation before discussing the experiment design within an intelligent meeting room. Data collection and the application scenario are also discussed in Section IV, before the paper concludes in Section V.

II. RELATED WORK

There are three leading approaches to context modelling. These are key-value [7], object-oriented [8] and ontology-based [9]. The current paper focuses on extending existing work in ontology-based context modelling. Both context modelling and context management have been areas of great interest within the research community, with a lot of work being done by many researchers within these areas.

Shih *et al.* [10] created a framework that used an upper level ontology to infer information about the user. The ontology used context deduced from sensor data and background information on the user, eliciting relationships between the two. Situation awareness was realised within the possible scenarios through the use of this ontology. First order logic rules were applied for inference and used alongside privacy aspects to deal with the privacy problem associated with situation awareness. The framework was aimed mainly at mobile devices and did not use real-time data.

A full context life cycle management was investigated by Jih *et al.* [11]. They used a context repository, ontology and reasoner to collect raw data, applying rules to the data and categorising the context information into one of three states; active, suspended and terminated. The raw data updates the context repository before being sent to the ontology to reason the data. Rules are then applied to infer about the data; two things happen from this point, a service is delivered and context inconsistencies are resolved before the data is sent to update the repository with the new consistent context data.

Ramparany *et al*. [12] used a central component within their work, a context manager. This manager collects context information from sensors, builds and maintains a situation model. As well as doing this, the manager can subscribe to any changes which are reported by each sensor which enables the updating or modification of the situation model. The sensors send information in the form of Resource Description Framework (RDF) [13] to the context manager.

Each research group creates and maintains a framework in a different way, usually aimed at the situation that is being dealt with and whether the information will be received on a real-time basis or not. All of them are aimed at storing and managing the information in a consistent manner.

Many opportunities and challenges exist within context modelling and management; the main opportunities which exist are within the scenario area, as a multitude of scenarios could be defined, using a context management system. Another opportunity however lies in the utilisation of applications from the deployment of the system. How other applications react can be extensive and it can range from updating mobile phone devices to creating an online booking system. The challenges which have been dealt with in the above work include the privacy problem associated with situation awareness [10], scalability and openness [12] and challenges associated with context i.e., information representation, context inconsistency and converting low level data into high level context [11]. The remaining challenges were discussed in Section I.

In this respect, this paper proposes a context modelling and management system for smart environments, alongside a novel context data acquisition mechanism. The system captures the sensor data, updates the repository and can be queried on a real-time basis by an application. The context data mechanism automatically detects new devices within the smart environment and updates the system accordingly. The ontology-based context modelling and management system which is proposed plans to overcome the challenges associated with context i.e., the repurposing of data, temporal reasoning and multiple levels of context information.

III. THE SYSTEM ARCHITECTURE

The ontology-enabled architecture (Fig. 1) illustrates the components which are required for this context modelling and management system. It consists of five main components; the devices component (contained within the Environment layer), sensor ontology component, enrichment component, semantic repository component and the API query and retrieval component. Each component relates to different parts of our system and environment; they are necessary to collect sensor data and change it from low-level into high-level context which can be re-used and shared.

The raw data comes from the devices and sensors which are located within the smart environment. This raw data is passed via the ontologies component to be stored as semantic data within the semantic repository. How this raw data is semantically enriched is discussed in more detail in Section IV C. The applications can access the semantic data which is stored in the semantic repository via the API query and retrieval component to update, retrieve or query the stored information. What is presented below is an outline of the main system components as illustrated in Fig. 1.

Devices component: This component deals with all sensors and devices which are deployed throughout the smart environment. When a sensor is activated, a signal is sent out to a receiver which is then displayed as data via a Java program running on a local computer. This data is saved into the context management system (CMS) automatically, both in the form of processed semantic metadata and raw data. The data which is outputted by the device component is low-level data, which can sometimes be ambiguous. By sending it to the CMS and passing it via the sensor ontology component, this ambiguity can be dealt with and corrected accordingly.

Semantic repository component: The semantic repository is where data is stored, both in raw and processed semantic data format. The data which has been outputted by the sensor device is passed to the enrichment component where it is made semantically rich. This enriched data gets saved to the semantic repository in the form of triples (subject, predicate and object). Metadata and meaning is added to the data within this component.

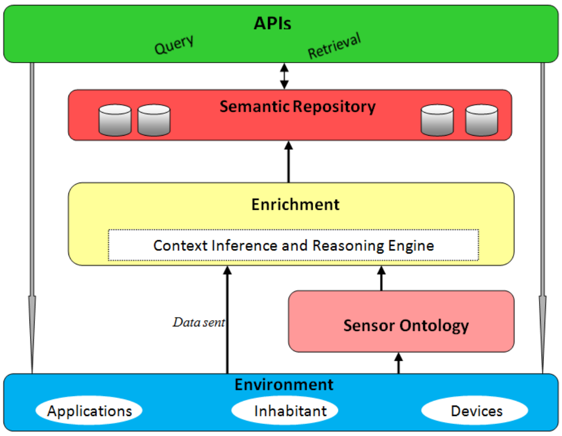


Figure 1. Architecture of the proposed context management system.

Sensor ontology component: This component is made up of ontology-based context models, which model the environment and the sensing devices within it. The context models can be applied to a wide variety of domains and devices. The ontologies are used like a schema to define the data as entities, individuals and properties. New devices can be added to the model using this component. The output from the sensor ontology and enrichment component is processed semantic data, which can be stored in the semantic repository and used by applications.

Enrichment component: The enrichment component consists of the context inference and reasoning engine, which is where the data gets sent after being passed to the sensor ontology component. Inference and reasoning rules are used against the data to process it and then the processed data is sent to the repository component.

Applications component: Applications use the processed data from the semantic repository through the use of the API component. This component uses the data queried or retrieved from the API component to update a user on a real-time basis.

API query and retrieval component: The API query and retrieval component provides facilities for the retrieval, query and updating of the information stored within the semantic repository.

All of these components placed together form the ontology-enabled system, which takes low level primitive sensor data and changes it into high-level context information. The CMS keeps the context information up to date and ensures that the information available to requesting applications is correct and relevant.

A. Sensor Modelling

The sensing devices within the smart environment have to be modelled to be semantically enriched. An ontology represents data in a formal manner by using entities and the relationships which link them together. Due to an ontology-based context model's use in knowledge sharing, reuse and reasoning, it is ideal for use within a smart research environment, various aspects such as multiple sensor data and heterogeneity can be dealt with through ontology-based context modelling. It is able to deal with context on all levels required within a smart environment as it can be divided into upper-level and lower-level ontologies.

By first creating a basic sensor ontology for the Smart Environments Research Group (SERG) domain, modelling the various sensors that exist along with their attributes the ontology could then be expanded upon. Once a basic sensor ontology had been designed using some of the devices from within the research environment it was then possible to expand the basic ontology by importing a larger, more diverse ontology which is discussed more in Section IV A. This ontology gave a wide overview of classes, subclasses and properties which could be applied to a wide variety of domains. By extending the original ontology with the developed sensor ontology [16], it was then possible to edit some of the properties and subclasses to meet desired specifications, an example of the sensor section can be seen in Fig. 2; this created an ontology which can be applied to multi-sensor domains which may include an array of smart devices. This extended ontology should now cover the majority of domains and devices within the smart environment world, with only a very small amount of future devices or domains needing to be created from the beginning. As well as being able to be used across a diverse set of domains, the extended ontology can also deal with multiple levels of context information found within smart environments.

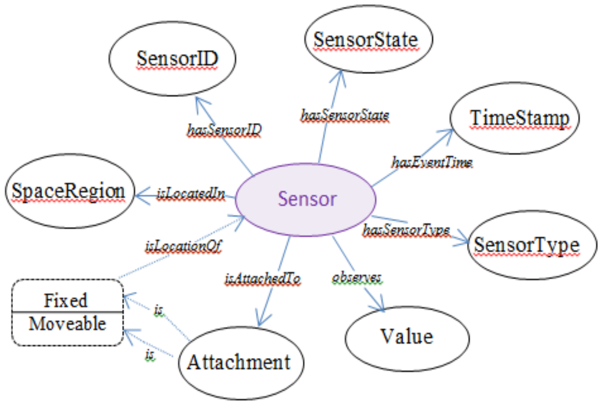


Figure 2. Example of the sensor ontology representation.

By using ontologies, it is possible to collect the data from the devices and add semantically rich metadata to this data. New devices can also be added in at this stage, ensuring that the ontologies are up to date with the devices that are in the environment. The ontology can be added directly to the semantic repository and used to semantically enrich the data which is added to the repository via a Java program. The reason that the environment is modelled in this way is that data can be ambiguous and ontology-based context modelling is one of the strongest types of context models available with the ability to deal with incomplete, ambiguous and informal data [14]. It ensures that contradictory information does not exist e.g., a device cannot be both a passive infra-red (PIR) sensor and a pressure sensor. Using this type of context model shall lower the amount of ambiguous data which is stored within the data repository. Ontology-based context modelling can also easily be applied to existing environments, which can be an advantage when it comes to sharing information across various domains and amongst numerous people.

The data which is obtained from the smart research environment is stored within the semantic repository, as previously discussed. This data can then be managed via the API component, using a query language. Finally it is difficult to use and distribute the same data amongst different applications and platforms due to the heterogeneous nature of different devices. By modelling the devices within the ontology, it makes it easier to disseminate across different platforms.

B. Plug-N-Measure Mechanism

A unique context data acquisition tool, which shall be referred to as a plug-n-measure mechanism throughout this paper, was created. A plug-n-measure mechanism as defined in this paper is similar to plug-n-play however the device does not have to be connected to the computer. A plug-n-measure mechanism can be defined as a context data acquisition tool which automatically detects new devices within a smart environment on a real-time basis. The mechanism then updates the system accordingly with the relevant context information pertaining to this new device.

Normally, when a new device is added to a smart environment, a person has to install the device, the drivers and manually update any information pertaining to that device. This can be a time-consuming effort, especially as the person may have to go through saved data and update information relating to an unknown device manually. As well as being time-consuming it can also lead to ambiguous or incomplete data being saved. By using alpha-numeric information associated with each device, the mechanism can add the device and information relating to the device to the system. The mechanism can also update any related information and finds the drivers, where possible, to install them automatically. By having the plug-n-measure mechanism in place, the likelihood of ambiguous or incomplete data is vastly lowered.

Using identifiers which are added to the incoming sensor data from the smart environment, the sensor type can be established. Other factors can then be used to determine where the sensor is located and what it is attached to. By modelling the sensor information within the ontology, semantic metadata pertaining to the new sensor can be added to the semantic repository. As new information about the sensor is determined, this can be updated within the repository in the form of semantic metadata. When new data is received from any device, the data is appended to previous information within the repository which relates to that device, saving anyone from having to go through context histories and manually update the sensor information for the new device.

IV. SYSTEM IMPLEMENTATION AND EVALUATION

An ontology-based context management system was developed. In order to validate the approach a controlled experiment was set up, data was collected and an application scenario was designed. The following sub-sections discuss the system implementation and use within the smart environment.

A. *System Implementation*

This section discusses how the system was implemented, which technologies were used in the implementation process and why.

The ontology-based context management system consists of a range of components some were newly created, i.e., the ontologies component which was further expanded by importing an existing ontology, whilst others were created or used from existing sources i.e., the semantic repository component, Sesame.

To create the ontology, Protégé [15] was used for the development environment. A basic sensor ontology was created within Protégé, depicting a wide range of sensing devices which may be available within a smart environment. A date and time entity was then added, along with other entities, such as possible locations within a smart environment. To further expand and enrich this basic ontology, the semantic sensor network (SSN) ontology [16] was imported enabling the use of the ontology across a wider number of domains. By using web ontology language (OWL) [17] and RDF [13], formal support for logical reasoning is provided along with expressivity and the ability to share information more readily amongst applications.

Once the ontology had been created, semantic repositories were researched, before deciding upon Sesame [18] for use within this system. Sesame was chosen due to the ability to work well with Java programming, which was essential to link in to the server side software. Sesame also offers the opportunity to use a number of different types of repositories, for this system a native RDF repository with RDF Schema inferencing was used. However, the system can also work as a native RDF repository without RDF Schema inferencing or as an in-memory RDF repository with or without RDF Schema inferencing. The semantic repository is one of the main components within this system and it was therefore necessary to have a repository which could store metadata, be updated on a real-time basis and be easy to query from an application. Sesame handles data in an easy to understand format and can be queried both from the Java program and from Sesame’s own interface, workbench.

Once Sesame was added to the system, the ontology was further refined and sensors from the smart environment were updated and added to the ontology. For testing purposes the ontology was then uploaded to Sesame and triples were manually added to Sesame using their workbench interface. This ensured that the ontology worked alongside Sesame and any incoming data. After this had been established the Java program was connected to Sesame and run; real-time data was semantically enriched and stored in Sesame as triples.

Now that the main components within the system worked on a real-time basis, the functions component was assessed. Due to Sesame working well with SPARQL Protocol and RDF Query Language (SPARQL) [19], the ongoing research work with SPARQL and the fact this query language was developed for use with RDF, using this as the APIs component made sense. Diverse queries can be made using SPARQL and Sesame offers an API to use with this query language. By being able to use the API in the Java program, connecting the APIs component to the semantic repository was easy. The applications component which is programmed in Java could use SPARQL to query, retrieve and update data from Sesame.

Once all of the implementation had been completed, the next step was to test the system on a larger scale real-time basis, as part of an experiment.

B. *Experimental Design*

Based across one entire floor of a building, SERG utilises the smart research environment to undertake research which could assist people with their everyday lives [20]. The main devices are located within a dedicated smart kitchen, living-room and meeting room where experiments frequently take place. These devices are modelled in an ontology-based context model to make sense of the environment.

The meeting room contains a small table, 6 chairs, an interactive white board, a projector and a computer (Fig. 3). There are also a variety of sensors deployed within this room, which can help when implementing a scenario. The sensors which are of main interest are the six pressure sensors positioned on the chairs, the six accelerometers attached to each chair, the three PIR sensors on the walls, the three contact sensors on the doors and the audio sensors which detect the noise level within the room

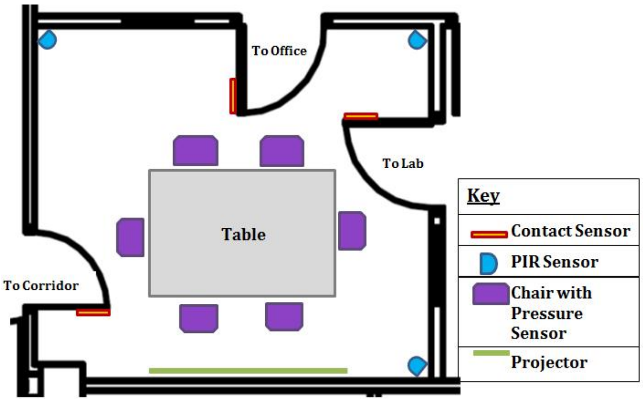


Figure 3. The meeting room layout.

without recording sound. The data from the pressure sensors can be used to determine when a person is sitting in the seat and when they have stood up. Along with the pressure sensors, the accelerometer devices are also attached individually to each chair, these devices determine when the chair is moved in or out from the table. Used together, all of these sensors can depict whether a participant is in the meeting room and if a meeting is taking place. All of this information is also saved as raw data and includes a timestamp to be cross-referenced against the semantic metadata at a later date.

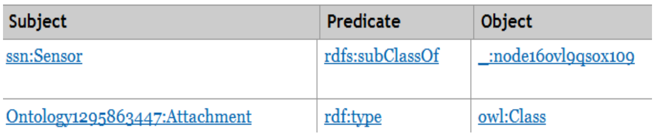
C. Data Collection

When a sensor is triggered, numerical information is sent to an interface logger, which is connected to a computer. Identifiers are also added to the numerical information within the Java program, such as sensor status, sensor ID, sensor type and a timestamp. All of this information is modelled in the previously discussed ontology, and semantically enriched with additional metadata pertaining to each sensor. The sensor ID and/or the sensor type identifiers are matched with the correlating information within the ontology and additional metadata, which has been pre-modelled, is added to the raw data. The semantically enriched data is stored in Sesame in the form of triples (Fig. 4). This data can then be queried, retrieved and updated using SPARQL.

The sensor ontology resides within the semantic repository and incoming sensor information is appended to the ontological information relating to that sensor. A context identifier can also be added to the SPO triples to create subject, predicate, object, context (SPOC). The context referred to in this manner relates to context identifiers and if used can group related context information together, making it easier for a user to query a group.

D. Application Scenario

Whether a meeting is taking place or when a meeting room is available for use can be established by using the sensors in the meeting room. If a meeting is taking place, the contact sensors on the doors should be activated, followed by at least two of the pressure sensors and accelerometers on the chairs being activated. It may also be possible to track numerous peoples' movement within the room through the use of passive infra-red (PIR) sensors; finally, audio sensors could pick up amplitude within the room. All of these sensor activations placed together form a scenario which suggests that a meeting is taking place. As each sensor is activated, the information is sent to the system and updated to processed semantic metadata, before being passed to the applications within the room.



Subject	Predicate	Object
ssn:Sensor	rdfs:subClassOf	_:node16ovl9qsox109
Ontology1295863447:Attachment	rdf:type	owl:Class

Figure 4. Triples: as stored within the Semantic Repository.

To evaluate the system, it was run on a real-time basis. The running of the system was to establish if it worked correctly and also if it was possible to ascertain whether a user was passing through the meeting room, a meeting was taking place or one person was using it for a conference call. By querying the repository, which sensors were activated and in which order could also be established. The system was able to save the information on a real-time basis and it was possible to query and retrieve the information to determine what had happened within the meeting room. The raw sensor data had semantic metadata added to it and was saved accordingly within the system. The plug-n-measure mechanism could also detect when a new sensor had been added to the environment and was able to update the data accordingly. The application scenarios discussed in the rest of this section are only an example of what could be achieved.

Dr. Smith, Dr. Jones and Dr. Bloggs have arranged to have a meeting together. As they enter the intelligent meeting room, their mobile phones automatically switch to meeting mode. This is handled by the devices, enrichment, API and applications components within the CMS. As they are seated the LCD display outside the meeting room displays the message "meeting in progress". Detecting that Dr. Jones is in the meeting room via the devices and enrichment component, the CMS sets up the projector and computer accordingly, allowing Dr. Jones to make the presentation that he had previously noted down in his electronic calendar. The calendars for Dr. Smith, Dr. Jones and Dr. Bloggs change their statuses to "in a meeting" and update their appointments for the rest of the day accordingly via the applications and API components. If other users want to use the meeting room whilst it is occupied, they can query it via the API component and will be notified via mobile or email that the meeting room is occupied. They will also be updated on free timeslots that have not been pre-booked for the selected room.

Ms. Doe wishes to use the meeting room, but she does not want to pre-book it as she does not have a definitive meeting time. By using the CMS on an "as and when required basis" she can retrieve context information. Ms. Doe uses the application component to input her preferred room. The component infers from the CMS through the retrieval and reasoning on sensor data, that the room Ms. Doe wishes to book is in use. She receives this information, along with a selection of possibly free rooms and she repeats the same process via the CMS to find a vacant room. Once she finds a vacant room, she can use it and the CMS is updated with relevant information to notify other users that, that room is now occupied.

Another way in which the CMS can be used is by management personnel. Using the applications component, they query and retrieve the stored data which can then be visualised using a visualisation tool. Based on this information, management can see how each room is used and set up additional rooms accordingly, e.g., a room specifically for conference calls etc.

V. CONCLUSION AND FUTURE WORK

The study presented in this paper addressed the challenges with regard to the ambiguity of collected data, temporal reasoning and sharing and re-using data amongst various applications. Temporal reasoning is also a problem as many application scenarios within smart environments closely relate to a temporal sequence of events. These events usually form a situation which is modelled via the low-level sensor data which can be collected from a smart environment; not only do applications want to access the current information they may also want to access context histories. These context histories usually contain the high-level information.

An ontology-based context management system was created along with an adaptive contextual data acquisition mechanism for use within a smart environment. Through the implementation and testing of this system on a real-time basis within a meeting room scenario, results were obtained which showed that low-level sensor data could be captured and semantically enriched with metadata and stored within a semantic repository. The plug-n-measure mechanism was able to recognise new devices which were deployed within the environment and update the system with contextual data. By adding semantic metadata to low-level sensor data, it makes it easier to disseminate the high level context across multiple domains and platforms. The semantic repository and functions component of the system make it possible for applications to access the current information as well as context histories.

The study has shown that the presented context management system can be applied to numerous scenarios and leads to an ample amount of opportunities for proposed applications. Low-level sensor data can be semantically enriched to high-level context enabling the sharing, re-use and reasoning of the context by different applications, platforms and domains.

The implementation of the system shows that the system can recognise sensors and other devices as they are added to the smart environment. It enables the repurposing of contextual data as well as a high level context inference based on temporal reasoning and domain heuristics. The future work includes analysis and extraction of the meeting room data for high level context use, i.e., activity recognition within the smart environment.

REFERENCES

[1] D. J. Cook and S. K. Das, “How smart are our environments? An updated look at the state of the art”, *Pervasive and Mobile Computing*, vol. 3, pp. 53–73, 2007.

[2] A. K. Dey, "Understanding and Using Context," vol. Personal and Ubiquitous Computing (2001), pp. 4-7, 2001.

[3] L. Chen, C. D. Nugent, and H. Wang, “A Knowledge-Driven Approach to Activity Recognition in Smart Homes”, *IEEE transactions on knowledge and data engineering*, vol. 24, no. 6, pp. 961-974, 2010.

[4] A.K. Dey, G. D. Abowd, and D. Salber, “A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications', *Human-Computer Interaction*, vol. 16, no. 2, pp. 97-166, 2009.

[5] P. Korpipää, E.Malm, I. Salminen, T. Rantakokko, V. Kyllönen, and I. Känsälä, “Context Management for End user Development of Context-Aware Applications”, MDM 2005, Ayia Napa Cyprus, pp. 304-308, 2005.

[6] H. van Kranenburg, M. S. Bargh, S. Lacob, and A. Peddemors, “A context management framework for supporting context-aware distributed applications”, *IEEE Communications Magazine*, August 2006, pp. 67-74, 2006.

[7] M. Samulowitz, F. Michahelles, and C. Linnhoff-Popien, "CAPEUS: An architecture for context-aware selection and execution of services." in *Third International Working Conference on New Developments in Distributed Applications and Interoperable Systems*, The Netherlands, pp. 23-40, 2001.

[8] K. Cheverst, K. Mitchell, and N. Davies, "Design of an object model for a context sensitive tourist GUIDE." vol. Computers and Graphics 23, pp. 883-891, 1999.

[9] H. Chen, T. Finin, and A. Joshi, “Using OWL in a Pervasive Computing Broker”, In Proceedings of Workshop on Ontologies in Open Agent Systems (AAMAS’03), pp. 9-16, 2003.

[10] F. Shih, V. Narayanana, and L. Kuhn, “Enabling Semantic Understanding of Situations from Contextual Data in a Privacy-Sensitive Manner”, Activity Context Representation – Techniques and Languages: Papers from the 2011 AAAI Workshop (WS-11-04), pp. 68-73, 2011.

[11] W. Jih, C. Huang, and J. Hsu, “Context Life Cycle Management in Smart Space Environments”, Proceedings of the 3rd workshop on Agent-oriented software engineering challenges for ubiquitous and pervasive computing (*AUPC’09)*, London, UK, pp. 9-14, 2009.

[12] F. Ramparany, Y. Benazzouz, L. Chotard, and E. Coly, “Context Aware Assistant for the Aging and Dependent Society”, Workshop Proceedings of the 7th International Conference on Intelligent Environments 2011, Nottingham, UK, pp. 798-809, 2011.

[13] RDF – Semantic Web Standardds, “Resource Description Framework (RDF)”, W3C 2004 [Online], Available: http://www.w3.org/RDF/ [retrieved: July, 2012].

[14] T. Strang and C. Linnhoff-Popien, "A context modeling survey", Proc. First International Workshop of Advanced Context Modelling, Reasoning and Management at UbiComp 2004, Nottingham, UK, 2004, pp. 33-40.

[15] The Protégé Ontology Editor and Knowledge Acquisition System, “Welcome to Protégé”, 2012 [Online], Available: http://protege.stanford.edu [retrieved: July, 2012].

[16] W3C Semantic Sensor Network Incubator Group, “*Incubator Activity > W3C Semantic Sensor Network Incubator Group*”, W3C 2011 [Online], Available: http://www.w3.org/2005/Incubator/ssn/ [retrieved: July, 2012].

[17] OWL - Semantic Web Standards, “*Web Ontology Language (OWL)*”, 2010 [Online], Available: http://www.w3.org/2001/sw/wiki/OWL [retrieved: July, 2012].

[18] OpenRDF.org: Home, “openRDF.org … home of Sesame”, 2012 [Online], Available: http://openrdf.org [retrieved: July, 2012].

[19] SPARQL Query Language for RDF, “SPARQL Query Language for RDF”, 2008, Available: http://www.w3.org/TR/rdf-sparql-query/ [retrieved: July, 2012].

[20] C.D. Nugent, M.D. Mulvenna, X. Hong, and S. Devlin, “Experiences in the development of a smart lab”, *The International Journal of Biomedical Engineering and Technology* (IJBET ’09), vol. 2, no.4, pp.319-331, 2009.

[21] L. Chen, C. D. Nugent, M. Mulvenna, D. Finlay, and X. Hong, "Semantic smart homes: Towards knowledge rich assisted living environments," in Springer Berlin/Heidelberg, 2009, pp. 279-296.

[22] L. Hsien-Chou and T. Chien-Chih, "A RDF and OWL-Based Temporal Context Reasoning Model for Smart Home," *Information Technology Journal*, vol. 6, pp. 1130-1138, 2007.

[23] H. Truong, A. Manzoor, and S. Dustdar, "On modeling, collecting and utilizing context information for disaster responses in pervasive environments" 24th August 2009, vol. CASTA '09, pp. 25-28, 2009.

Extending Moodle Services to Mobile Devices: The Moodbile Project

María José Casany, Marc Alier, Enric Mayol, Jordi Piguillem, Nikolas Galanis, Services and Information Systems Department
Universitat Politècnica de Catalunya Barcelona-tech
Barcelona, Spain
mjcasany@essi.upc.edu, marc.alier@upc.edu, mayol@essi.upc.edu, jpiguillem@essi.upc.edu, ngalanis@essi.upc.edu

Francisco J. García-Peñalvo, Miguel Ángel Conde, Informática y Automática Department
Universidad de Salamanca
Salamanca, Spain
fgarcia@usal.es, mconde@usal.es

***Abstract*—Learning Management Systems (LMS) are widespread among most education and training institutions. Even though LMS are a mature technology, they have left the vanguard of innovation in e-learning to mobile devices and tablets. Mobile Learning (M-learning) may enhance e-learning by increasing communication and conversation opportunities to convents the learning process more collaborative and learner-centred. This paper describes a way to integrate mobile devices and educational applications with a LMS as Moodle through web services: The Moodbile Project. Rather than just creating mobile apps that replicates LMS functionalities on a mobile device, Moodbile provides to m-learning developers with the necessary tools to allow mobile devices to interact with the LMS. In this paper, we describe our proposal of an open specification of web services to support the integration of mobile external applications with Moodle.**

Keywords-e-Learning; m-Learning; Moodle; LMS.

I. Introduction

E-learning has experienced an extraordinary growth over the last years; learning paradigms, technological solutions, methods and pedagogical approaches have been developed. We have reached a point in time when most of learning institutions have adopted Learning Management Systems (LMS). LMS are systems that organize and provide access to online learning services (such as access control, provision of learning contents, communication and administration of users and groups tools) for students, teachers and administrators [15].

Nowadays, the success of LMS is so great: over 90% of Spanish universities and colleges use a LMS [16], 95% of the learning institutions in the US also use an LMS [10], and 79,5% of large companies use these systems during their training program [21]. In LMS, the standard organizational unit is the course, and this structure restricts students to the content designed for a particular course and to interact only with the other participants of the course. Therefore, student's engagement in LMS is lower in contrast to their engagement in other environments or tools such as mobile devices, Web 2.0 tools or game consoles. These environments provide opportunities for customization, communication and a sense of ownership impossible in current LMS [12], [17].

In this sense, LMS must evolve to adapt to new user requirements, technologies and opportunities. For example, LMS must be able to interact with external applications such as social networks, blogs, mobile applications, virtual environments etc. [17], and it must allow networked learning throw easy collaboration and communication tools. This interaction between the LMS and other tools will require flexibility and interoperability.

The expansion of mobile devices with new browsing capabilities and touch interfaces provide new ways to learn (mobile learning or m-learning). M-learning is a new learning approach to support personal learning demands that may happen anywhere and/or at any time; or in response to the process of coming to know, by which learners in cooperation with their peers and teachers, construct transiently stable interpretations of their world [18]. M-learning puts the control of the learning process in hands of the learner itself [8] and enhances collaboration and flexibility.

One possible way to promote m-learning applications and to overcome some of their limitations may be the integration with the LMS. This integration will also facilitate LMS evolution, interoperability improvement and its adaption to new social needs and new technologies.

In this paper, we propose a first step of an interoperability solution to extend LMS to the world of mobile devices. Section two describes the research scenario with a description of the problem of interoperability between m-learning applications and LMS, and a brief state of the art of the current approaches to the problem. Section three presents the Moodbile Project, and some details of our proposal to extend the Moodle Services Architecture to allow m-learning. Finally, section four summarizes conclusions and future work.

ISBN: 978-1-61208-236-3

II. RESEARCH FRAMEWORK

In this section, we describe opportunities of m-learning and two approaches to integrate m-learning applications to LMS.

A. Mobile Learning: opportunities and challenges

Portability of mobile devices and their ability to connect to the Internet makes such device ideal tools for the storage of learning experiences and reference materials, as well as, to be a general tool to enhance the learning process. The Horizon Report [9] identifies mobile phones as a priority technology for next years. Mobile technologies have a special role, because they increase our communication and conversation opportunities.

M-learning introduces some opportunities and challenges in the learning process. Some of the contributions of m-learning are: 1) it is learner-centered [13], 2) it is a new alternative for information delivery and 3) it enhances collaborative learning [18]. Therefore, we may state that m-learning increases learning flexibility by customizing learning to be more personalized and learner-centered [3], [20].

On the other hand, m-learning faces several challenges such as: 1) lack of teacher confidence, training or technical difficulties with mobile devices [6], [22], 2) lack of institutional support [6], [22], 3) interoperability problems with LMS which usually are designed as monolithic systems [2] and 4) limited impact because many initiatives are isolated from the rest of the learning process [11].

One possible solution to overcome these challenges and to avoid LMS extinction is the integration of m-learning initiatives with LMS. This approach has several advantages. From the technological point of view, the LMS can be a tool to spread learning innovation, and m-learning projects can be more than isolated experiments because they would be integrated with the learning dynamics of educational institutions. From the student's point of view, they could personalize their learning process thanks to mobile devices as well as collaborate with peers. From the teacher's point of view, they could continue to use LMS as their working platform, leaving mobile devices for students.

B. Integrating m-learning with LMS: Related work

Integration between m-learning applications and LMS is not an easy task, because LMS do not usually include interoperability standards to communicate with external applications; they are usually designed as monolithic or layered systems [17]. The integration of m-learning applications with LMS has two scenarios:

1) Extending the LMS to the mobile world

The first scenario is based on the creation of m-learning applications that extend the scope of the LMS. Such mobile applications usually follow one of two different approaches: The first approach focuses on engagement with mobile devices and mobile native applications. The benefits of this approach include access to engaging design, free use of hardware features and fast and lightweight technology. However, the main limitation is that applications are device specific, which usually requires high development costs. The second approach focuses on the interaction with a browser, so the technology is ubiquitous and device-independent. However it may also be slower and it may be harder to access for some smart phones. For example, the LMS Blackboard is focused on native applications for mobile devices, while Moodle follow both approaches.

Usability and online/offline work are important issues when extending LMS to the mobile world. Specific restriction on mobile devices to display information and to interact with the user, must be also taken into account. M-learning applications also allow offline work when network coverage is not available, so it is required to synchronize local data of the device with LMS data [19]. To sum up, extending LMS to the mobile scenario transforms the LMS into a web platform that must provide services to mobile devices usually using web service technology.

Some of the proposals in this scenario try to create a clone mobile LMS allowing most of the LMS functions in mobile devices, without taking into account limitations of mobile devices such as data input or screen size.

2) Integrating external m-learning applications into the LMS

The second scenario is based on the integration of external m-learning applications into the LMS. Since most LMS are not service oriented, any attempt to integrate external applications with the LMS must be done ad hoc. This approach has important disadvantages such as the difficulty in maintaining and extending the new integrated system or the limited impact of these solution [1], [2].

III. MOODBILE

This section describes the main goals of the Moodbile project, an initiative designed to open up the LMS Moodle to the mobile scenario. First, we define the analysis requirements of the project. Second, we describe a layered architecture to adapt Moodle to the service oriented paradigm. Finally, we describe services we designed to allow mobile devices access to the LMS.

A. Project definition and motivation

The Moodbile project [23] aims to enable mobile learning applications to work together with the LMS. Moodbile is an open source project. Moodle is the host LMS platform. Rather than just creating mobile applications that replicate the LMS functionalities on a mobile device, Moodbile provides the developers of applications for education with the necessary tools to interact with the LMS.

The motivation of the Moodbile project is to open up the most commonly used e-learning platforms and LMS, originally designed as monolithic or layered systems, to the service paradigm and to mobile world.

To communicate Moodbile with the LMS, some Moodle functions are redesigned as services and they may also be used to integrate external applications into Moodle.

B. Moodle Web Services Architecture

Moodle Arquitecture is designed following the classic three-tiers architecture. Like a big amount of PHP applications, Domain Tier implements very atomic functions, and the major part of business logic is located at Presentation Tier.

With the design of Moodle Web Services Architecture, two main problems had been addressed. How to solve that the main business logic is located at the Presentation tier instead of the Domain Tier, and how to design the architecture to support a variety of web services protocols without coupling business logic and such protocols.

The solution proposed to Moodle Head Quarter was the architecture shown in Fig.1.

Domain and Presentation tiers have not been changed with respect to Moodle Architecture. Instead of refactoring both tiers, the External tier was created. This tier is an extension of the standard Moodle External tier and can basically access methods from the standard Moodle External tier and the Moodle core The External tier is responsible for the specification and implementation of services that Moodle provides to external consumers or apps. However, this tier replicates some minor business logic and control statements.

To provide full support to the most widespread Moodle web services protocols, a Connectors tier was designed. The Connector tier contains specific components that adapt service's specifications of the External tier to the provided protocols. At the same time, this tier handles authentication and session management. Protocols supported are REST, SOAP, XML-RPC, AMF among others.

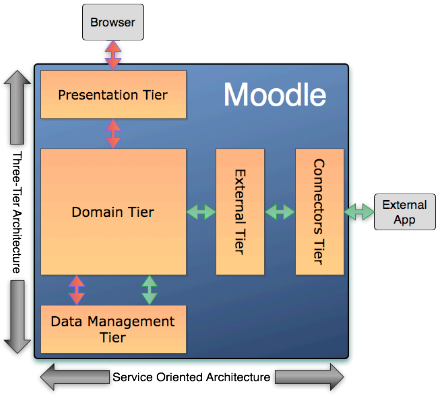


Figure 1. Moodle Web Services Architecture

The extension presented in Fig. 1 was created to solve the problem of integrating mobile applications with Moodle. Even though Moodle 2.0 already had a collection of web services, these web services focused on developing an API suitable for massive batch actions like user or course creation and inscriptions. These methods are not suitable for the integration of mobile learning applications. Moreover, issues as security management are not properly provided by these web services.

Therefore, Moodbile project was initiated to design a Moodle extension that would turn Moodle into a web services provider for mobile applications, with the design of a web-service layer to access most suitable Moodle features for mobile applications

C. *Moodbile Requirements*

Moodle's *External tier* only provides basic services. Therefore, it must be extended to provide additional services to allow mobile application interact with the LMS. These features were identified and selected from related work (some of them described in section II.B) and from the analysis of the log of the Moodle server of our university, where we found the most used features accessed from mobile devices [5].

In the first step of the Moodbile project we have considered to include the features like: view course activities, view course participants, view student's grades, view resources, view and upload assignments, access forums and discussions, read and reply posts, view upcoming calendar events and view user profile.

D. *Moodbile's Architecture*

This section describes the proposed extensions to the *Moodle Web Services Architecture* to provide additional web services for mobile applications [4], [7].

This extension involves extending the tree tiers: The *Domain* tier extension re-implements some features that Moodle does not provide in the proper way to be used by Web Services, and, some additional services and features specific are provided to support Moodbile functionality. The *External* tier is where the actual services for mobile integration are defined and implemented. The *Connectors* tier is used to provide additional web services protocols and authentication methods more suitable for mobile devices, such as Oauth [14].

This extension is based in the following additions:

1) *New authentication method.*

Moodle Web Services are oriented to academic management operations such as course creation, new users registration and students enrollment. Moreover, Moodle services consumers are under control of the same organization, so complex authentication methods are not required. Moodle users authenticate with their username/passwords or with special access tokens. To secure the connection, it may be enough to restrict IPs that will be able to invoke this web services.

In a mobile environment, such mechanisms are not enough, since user/password or token authentication is not secure since this information is sent in each web service call from a mobile device, and it is not possible to establish a white list of authorized IPs. To solve this problem, Moodbile project applies an OAuth authentication mechanism allowing teachers and students to access Moodle using third-party applications in a secure way.

2) *New Connectors*

Moodle offers several web services protocols but some (or none) of them are not suitable to a mobile environment. For this reason, Moodbile purposes several new connectors. To give support to HTML 5 applications, an AJAX compatible connector with JSON messages was implemented. This connector shares Moodle session, so user must be logged in to use an HTML 5 client. In a similar way, JSONP allows HTML 5 applications to access to Moodle Web Services using native authentication methods.

JSON-RPC connector was implemented to add a lightweight protocol that can be used effectively form a mobile phone. This connector uses Moodle Web Services authentication methods. Finally, a JSON-RPC connector and a new REST connector are implemented using OAuth authentication. These connectors are the most commonly used for Moodbile clients for its ease to use and security improvements.

3) *New Services*

New services are defined into the *External* tier and grouped in packages [24]. Packages are classified into Basic Services, Course Content Services, Personal Content Services or Platform Services.

a) *Basic Services*

Course package provides basic services to retrieve user courses information. *User* package provides services to list course enrolled users and their profiles.

b) *Course content services*

Services in this group allow the interaction with course content and activities. Thus, *Assignment* package provides services to get assignments information and to submit homework. *Forum* package makes possible to manage forums information, discussions and posts. *Resource* package provides all the services related to digital resources access.

c) *Personal content services*

Services in this group provide access to users personal content. *Blog* package allows read, write and access to Moodle blogs. *Calendar* package give access to user calendar, scheduling and events. *Grade* package allows to retrieve users grades and outcomes. *Message* package allows to use the internal messaging.

d) *System Services*

Lang package provides I18N services to develop clients that use the same terminology of the host. *File* package allows upload and download files form Moodle in a secure way. And *System* package provides host and system information services

4) *Moodbile Clients*

Several mobile clients have developed based on HTML 5, Android and iOS. They are being used to test Moodbile Architecture and Services and to research in m-Learning topics.

IV. CONCLUSIONS AND FURTHER WORK

M-learning enhances collaborative learning and increases learning flexibility by allowing it to be more personalized and student-centred. But on the other hand, m-learning faces interoperability problems with LMS (the basic e-learning infrastructure of many educational institutions). The Moodbile project aims to propose an interoperability solution to integrate m-learning applications with the LMS, incorporating m-learning applications into the learning process of educational institutions. This will allow m-learning applications to widen their scope instead of being isolated from the learning process. It also will allow LMS to be more flexible e-learning platforms.

Moodbile aims to propose an open specification of web services to support the integration of external applications with the LMS. The initial version of the specification works for Moodle, but authors are working to adapt this specification to other LMS (such as Sakai and Olat) to create an LMS-independent specification. This specification is open-source and available for developers of third-part applications to use this specification. Authors are also designing special m-learning activities inside the LMS, that is, activities used from mobile devices but created inside the LMS.

ACKNOWLEDGMENT

This work has been partially funded by the MiPLE project (TIN2010-21695-C02-02.8) of the Spanish Ministry of Science and Innovation.

REFERENCES

[1] Alier, M., Casany, M.J. and Piguillem, J., "Towards mobile learning applications integration with learning management systems," Multiplatform e-learning systems and technologies: mobile devices for ubiquitous ICT-based education, 2007, pp. 182.

[2] Alier, M., Casany, M.J., Conde, M.A. and García-Peñalvo, F.J., "Interoperability for LMS: the missing piece to become the common place for e-learning innovation," International Journal of Knowledge and Learning, **6**(2), 2010, pp. 130-141.

[3] Bull, S. and Reid, E., "Individualised revision material for use on a handheld computer," In J. Attewell & C. Savill-Smith (EDS) Learning with mobile devices, Learning and skills development agency 2004, pp. 35-42.

[4] Casany, M.J., Alier, M., Conde, M.A. and García-Peñalvo F. J., "SOA initiatives for eLearning: a Moodle case," In Advanced Information Networking and Applications Workshops*, 2009* (AINA'09*)*, IEEE, pp. 750-755.

[5] Casany, M.J., Alier, Mayol, E., "Using LMS activity logs to measure mobile access," In Proceedings of third International Conference of Technology Enhaced Learning 2012.

[6] Cobcroft, R.S., Towers, S. and Smith, J., Bruns, A., "Mobile learning in review: Opportunities and challenges for learners, teachers, and institutions," In Proceedings Online Learning and Teaching (OLT) Conference 2006*,* pp. 21-30.

[7] Conde, M.A., García-Peñalvo F. J., Casany, M.J. and Alier, M., "Adapting LMS architecture to the SOA: an architectural approach," Fourth International Conference on Internet and Web Applications and Services 2009, IEEE, pp. 322-327.

[8] Downes, S., "E-learning 2.0," eLearning magazine: education and technology in perspectives, 2006, pp. 21-29 http://elearnmag.org/subpage.cfm*,* (*retrieved: 04/2012).*

[9] Johnoson, L., Smith, R., Levine, A. and Haywood, K., "Horizon Report 2010," Available at: http://wp.nmc.org/horizon2011/ (*retrieved: 04/2012).*

[10] Lonn, S. and Teasley, S.D., "Saving time or innovating practice: Investigating perceptions and uses of Learning Management Systems," In Computers & Education*,* **53**(3), 2009, pp. 686-694.

[11] Martín, S. et al., "Middleware for the development of context-aware applications inside m-Learning: Connecting e-learning to the mobile world," In Proceedings of the Fourth International Multi-Conference on Computing in the Global Information Technology, ICCGI'09*,* 2009, pp. 217-222.

[12] Mcloughlin C. and Lee, M.J.W., "Social software and participatory learning: Pedagogical choices with technology affordances in the

Web 2.0 era," *ICT:* Providing choices for learners and learning, Proceedings ascilite Singapore *2007*.

[13] Naismith, L., Lonsdale, P., Vavoula, G. and Sharples, M., "Literature review in mobile technologies and learning," FutureLab Report, 2004.

[14] OAUTH, "The OAuth 1.0 protocol RfC," Internet Engineering Task Force (IETF*)*. 2010. EHammer-Lahav E. Editors. available at: http://tools.ietf.org/html/rfc5849 *(accessed 30/04/2012*

[15] Paulsen, M.F., "Online education Systems: Discussion and definition of terms," In G.Web-Education Systems in Europe, Hagen: Zentrales Institut für Fernstudienforschung, FernUniversität, 2002, pp. 23-28.

[16] Prendes, M.P., "Plataformas de campus virtuales de software libre: Análisis comparativo de la situación actual de la Universidades Españolas," Informe del Proyecto EA-2008-0257 de la Secretaría de Estado de Universidades e Investigación, 2009.

[17] Sclater, N., "Web 2.0, personal learning environments, and the future of learning management systems," Research Bulletin*,* **13**, 2008, pp. 2008-2009.

[18] Sharples, M., Taylor, J. and Vavoula, G., "Towards a theory of mobile learning," In Proceedings of mLearn 2005, 1(1), pp. 1-9.

[19] Trifonova, A. and Ronchetti, M., "A general architecture to support mobility in learning," Advanced Learning Technologies, 2004. Proceedings IEEE International Conference on 2004, pp. 26-30.

[20] Vavoula, G.N. and Sharples, M., "KLeOS: A personal, mobile, knowledge and learning organisation system," In Proceedings of the IEEE International Workshop on Mobile and Wireless Technologies in Education Aug 29-30, Vaxjo, Sweden, 2002, pp. 152-156.

[21] Wexler, S., Grey, N., Miller, D., Nguyen, F. and Barnevelda, A., "Learning Management Systems: The good, the bad, the ugly... and the truth," E-learning Guild, 2008.

[22] Zawacky-Richter O., Brown, T. and Delport, R., "Factors that may contribute to the establishment of mobile learning in institutions," Results from a Survey. Interactive mobile technologies*,* **1**(1), 2007, pp. 40-41.

[23] Moodbile website: http://moodbile.org (*retrieved: 04/2012).*

[24] Services created for Moodbile online documentation. http://docs.moodbile.org/index.php/Moodbile_WS_Documentation_v0.2 (*retrieved: 04/2012).*

Exploring Efficient Methods to Extract Pedestrian Flows on a Mobile Adhoc Network

Ryo Nishide, Hideyuki Takada
Faculty of Information Science and Engineering
Ritsumeikan University
Kusatsu, Japan
nishider@fc.ritsumei.ac.jp, htakada@cs.ritsumei.ac.jp

***Abstract*—This paper explores possibilities to extract pedestrian flows from Bluetooth detection logs in a distributive manner. Bluetooth devices are installed in mobile equipments such as laptops, tablet PCs, cell phones, and PDAs, which pedestrians carry with them in daily life. If these devices can be detected and logged, it may be possible to analyze the movements and density of surrounding pedestrians in real world. Moreover, we aim to build this system on a server-less adhoc network, in which the network can be built autonomously, and the data can be managed between mobile devices distributively. Our goal is to build this system as simple as possible, while avoiding the initial preparations to deploy sensors in real world. The results of experiments have revealed that detection logs implicitly record traces of surrounding pedestrian flows, which might provide possibilities to analyze and distinguish pedestrian flow patterns based on different situations. The paper has also discussed the related issues on network construction including methods to interpolate missing detections.**

***Keywords*-Distributive Database; Bluetooth; Social Context; Pedestrian Flows; Mobile Devices; Adhoc Network.**

I. Introduction

In recent years, according to the increase of urban population and the expansion of social activities, we cannot avoid sharing the same public spaces with other people when travelling outdoors as well as in daily life. These situations include such occasions as transporting by train, bus, or walking, eating lunch at restaurants, meeting at appointed places, and going to a particular site such as a historical spot, amusement park, festival and business show. In any occasions, it will be one of the major concerns for people whether the area is crowded or less-crowded, and sometimes, it is necessary to know what is actually going on in such places, including the changing flow of pedestrians. For example, though people tend to look for less-crowded places to pass by or stay, they might occasionally choose a crowded place as a popular spot, pondering or being curious about what is occurring there.

On the other hand, many location-based services have appeared on market owing to the enhancement of computational ability, wireless communication technology such as wifi and Bluetooth, and GPS technology deployed in mobile devices. These advancements have paved way to explore methods for detecting pedestrian flows or social activities using high performance mobile devices [1], [2], and extend the applications for location-based systems such as recommendation system [3], navigation system [4], information sharing system, and so on. In most of these systems, users are not only information viewers, but also the information providers who send information as user check-in data, queue length to wait in line, user comments or evaluation of a place, and congestion information to the system, so that these information can be shared between users or applied to computation to enhance the results. However, most of these systems require users to send data manually, which is inconvenient for pedestrians, otherwise, collecting data automatically from mobile devices, which may leak their privacy information. Some of them require abundant initial preparations to construct the system.

This paper proposes methods to extract the density and flows of pedestrians using the Bluetooth detection logs on a mobile adhoc network. This adhoc network can be generated from connection between mobile devices to work as a distributive database, which can be managed and updated the detection log data, or modified the log data by accessing to geometrically adjacent devices to check for missing detections. The policy of this work is to avoid initial preparations such as installing a large number of expensive immovable sensors and high performance computational equipments in real space, in order to minimize cost, time and effort. In this research, we focus on the attempt to extract pedestrian flows in real world, while the specific services to utilize the detection results are left for future work.

To begin with, the related researches and comparable studies are reviewed in Section II. Then, the methods for extracting the density and flows of pedestrians, and the data management scheme are explained in Section III. Section IV discusses the architect and mechanism of the proposed system. In Section V, based on the results of experiments, the pedestrian flows are examined in different situations, by the analysis of detection patterns. Further study and additional experiments will be needed to cope with the interpolation of missing data and the deployment on an adhoc environment.

II. Related Research

Several researches have emerged in the attempt to extract social contexts owing to the development of mobile equipment and adhoc communication.

O' Neill et al. [5] and Nicolai et al. [6] examined the correlation between Bluetooth detecting and pedestrian movement by deploying stationary Bluetooth sensors in the environment and analyzing the logs. Eagle et al. [7] have shown methods to analyze social patterns of user's activity in a daily routine. These studies show that Bluetooth scanning and analysis of detection logs have possibility to extract the flow of pedestrians, however, not every Bluetooth device can be guaranteed to be detected depending upon the performance of the device and situation of the space. Thus, their methods may not be able to cope with too many incoming data caused by crowded pedestrians.

To cope with such problems, Kim et al. [8] examined the detection pattern of Bluetooth device logs, and employed clustering algorithm and Gaussian blur to remove noises caused by inquiry fault of undetected Bluetooth devices. They inferred the transition time of events from multiple device detections. However, inquiry fault for devices cannot be detected individually. As there are many complicated situations in real world, this method may not be enough to cope with various situations. Weppner et al. [9] estimated crowd density through collaboration with multiple devices to improve the accuracy of detections. Users were assigned to carry multiple devices to perform Bluetooth scanning together, which might be troublesome for users.

Our research is contemplated to extract social context by scanning Bluetooth device of surrounding environments, with consideration to the user's location and the communication range of Bluetooth devices. The method is proposed to work autonomously and distributively with the users' devices on an adhoc network, avoiding such troubles as installing fixed sensors or carrying multiple devices. It also enables to correspond with inquiry faults by performing computation collaboratively with nearby devices.

III. Proposed Method

To build a system to extract pedestrian flow without initial preparations, two points must be considered; (i) an autonomous method to determine the movements of pedestrian without mounting cameras or sensors in real world, and (ii) the server-less infrastructure to maintain the system and manage data on mobile devices distributively.

A. Extraction of Pedestrian Flows

We attempt to grasp social contexts such as changes of pedestrian flows and density by detecting the surrounding electronic equipments. Recent handheld electronic equipments like cell phone, smart phone, PDA, and laptop are installed with wireless devices such as wifi and Bluetooth, which pedestrians carry with them in their daily lives. If these devices surrounding the user are detected and logged continuously, it may be possible to detect not only the density of crowd, but also the changes of social contexts such as the pedestrian flow. In fact, the detection pattern differs depending upon the situations of the surrounding pedestrians (Fig. 1). Thus, by analyzing the detection patterns, it might be possible to assume the social contexts or trends and changes of surrounding situations. We avoid extracting the personal information of pedestrians, such as locations or user's name, since this kind of information might violate the privacy issue of pedestrians. Instead, we examine the detection patterns (e.g. numbers and changes of simultaneous or continuous detections) of devices carried by pedestrians surrounding the user.

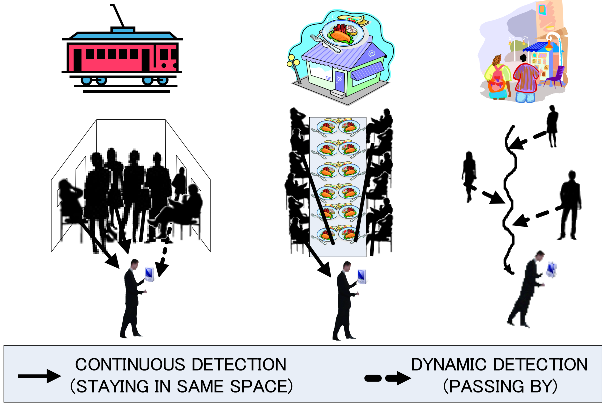


Figure 1. Detections in Different Situations

We have conducted a preliminary investigation to examine the statistics of detectable types of terminal (mobile phone, PC, etc.) at various places [10]. We have compared two wireless technologies: wifi and Bluetooth. Wifi was detected from many types of electronic equipments, either carried by pedestrians or fixed in the environment. Therefore, wifi seems difficult to discriminate the types of equipments, whether they are carried by the pedestrians or not. On the other hand, by the investigation of Bluetooth signals, most of the detected Bluetooth radios were from mobile devices. Especially, Softbank mobile phones were highly detected, probably because several models with Bluetooth functions were sold in Discovery mode (a configuration option to enable the surrounding terminals to discover the user's terminal) as a default setup. In this paper, we focus on Bluetooth devices, in order to detect the flows and movements of pedestrians, as most of Bluetooth devices are installed in equipments to be carried by users.

The method we have proposed can be performed only by utilizing the mobile device carried by the user, without installing additional equipments such as mounting fixed sensors or video cameras in the environment.

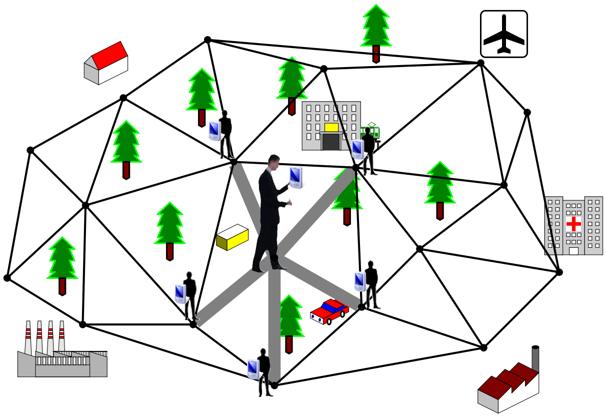
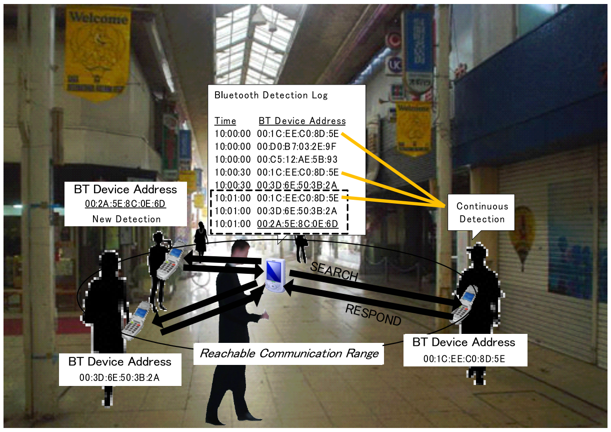


Figure 2. Delaunay Network with Mobile Devices



Bluetooth Detection Log

```
Time      BT Device Address
10:00:00  00:1C:EE:C0:8D:5E
10:00:00  00:D0:B7:03:2E:9F
10:00:00  00:C5:12:AE:5B:93
10:00:30  00:1C:EE:C0:8D:5E
10:00:30  00:3D:6E:50:3B:2A
10:01:00  00:1C:EE:C0:8D:5E
10:01:00  00:3D:6E:50:3B:2A
10:01:00  00:2A:5E:8C:0E:6D
```

Figure 3. Detection of Pedestrian Flows

B. *Mobile Adhoc Network*

Another issue of concern is the management scheme of pedestrian flow data obtained from each mobile device. It is not efficient to collect and manage the entire data sent from mobile devices on a server. Therefore, a mechanism is necessary to manage data and perform computation between mobile devices cooperatively.

To build such mechanism, it is contemplated to construct an adhoc network using mobile devices so that the network can be utilized to manage the data and communicate with other devices. In this network, each device builds connection directly with other devices without communicating to the base station. In generating connections, it is important to employ an efficient scheme to choose mobile devices to connect with, considering their location and limited communicable distance of mobile devices. Note that not all surrounding pedestrians with mobile devices are users to generate connections on adhoc network. Some of their devices can be cell phones or the kinds with less or no computational capability.

In this paper, we propose to employ P2P Delaunay network, which is a geometry-based P2P network whose topology is defined by the geometric adjacency of mobile devices (Fig. 2) [11], [12]. These devices are connected in a geometrical structure Delaunay Diagram, which is well-known in computational geometry. It has the features as (i) each device connects to a close-by devices based on its geographical distance, (ii) the degree of connection for each device is low (approximately six), (iii) the network can correspond with join/leave of device only affecting the surrounding devices to reconstruct and update the connection, and (iv) the data is reachable to distant device through multi-hop communication.

P2P Delaunay Network enables to construct a networked environment in which the mobile devices are connected to each other autonomously and distributively. It is not necessary to prepare a server in order to maintain the system or manage pedestrian flow data on it. Moreover, it also provides possibilities to perform collaborative computation or processing to work with set of mobile devices nearby. Delaunay Network works effectively for accessing to the data in geometrically adjacent devices, which can expose the missing detection by observing at the detection logs of surrounding devices.

IV. System Structure

In this section, we discuss the architect and mechanism of the system.

A. *Detection of Surrounding Bluetooth Devices*

During the manufacturing process of Bluetooth device, each device is assigned with its individual ID expressed in 48-bit MAC address. This address is described as Bluetooth Device Address (BDA) and is used for communicating with other devices by sending their BDAs to identify each other. Thus, BDAs are sent constantly without requiring authentication to build connection with other Bluetooth devices. Our target is on class 2 Bluetooth devices embedded in handheld mobile equipments as cell phones, laptops, PDAs, etc, with their communication range reachable to approximately 10 meters distance. The protocol for the Bluetooth inquiry first receives the BDA of surrounding Bluetooth devices, and then inquires the names of these devices. A combination of BDAs and timestamps are stored in the log file for every constant time interval.

Figure 3 shows an example of the pedestrian's Bluetooth Device which has entered the reachable communication range of user's device. User's device continuously sends inquiry to search for the surrounding pedestrians' devices, and logs the time and BDA of devices which have responded to user's inquiry. From the log, different types of detection patterns can be verified, such as continuously detected, newly detected, undetected or disappeared, and so on, which might be the key to determine the dynamic flows of pedestrians in real world.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-236-3

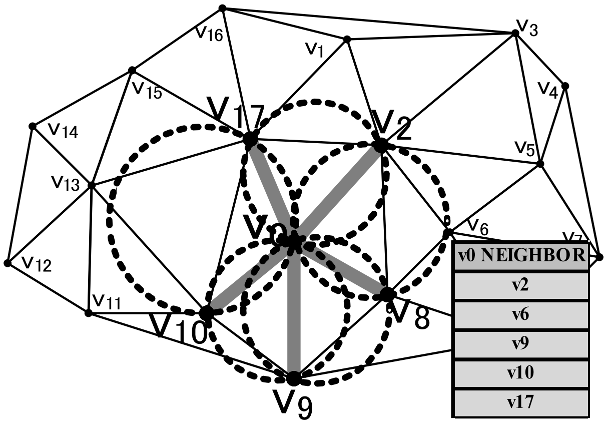


Figure 4. Determination of Nodes to Connect

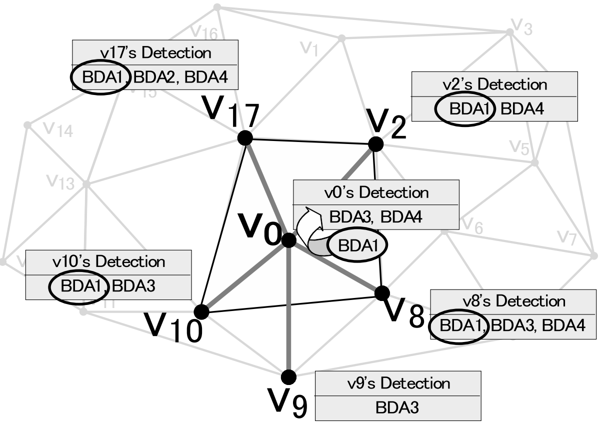


Figure 5. Interpolation of BDA Data (BDA1)

B. Network Construction

In our research, we contemplate to apply the method proposed in the past work [11] to generate P2P Delaunay Network with mobile devices. We assume that each mobile device only has the location information of other devices, but not the knowledge of how the other devices are connected. Thus, each mobile device must choose the appropriate mobile devices to connect, referring to their location information to generate a P2P Delaunay Network.

To build connections of a P2P Delaunay Network under such situations, each node (referring to the geographical location of mobile device) draws an inscribed circle with two other nodes on a plane, with a property of Delaunay Network that no other nodes shall be included in the internal side of the circle. Figure 4 shows an example of v_0 determining the node to generate connections on a plane. Inscribed circles are generated connecting three nodes each, namely (v_0, v_i, v_j) $\{0 \leq i \leq 17, 0 \leq j \leq 17, i \neq j\}$, which any of these circles has no nodes in the internal. The nodes (v_2, v_8, v_9, v_{10}, v_{17}) are assigned as the neighbors of v_0 to generate connection with. If rest of the nodes (v_1 - v_{17}) does the same processes v_0 has done, a Delaunay Diagram can be generated. The detail algorithm for generating and maintaining connections are discussed in the past work [11]. Delaunay Network can be used not only to generate or maintain connections with adjacent nodes on a plane, but also to perform collaborative computation with adjacent nodes described in the following section.

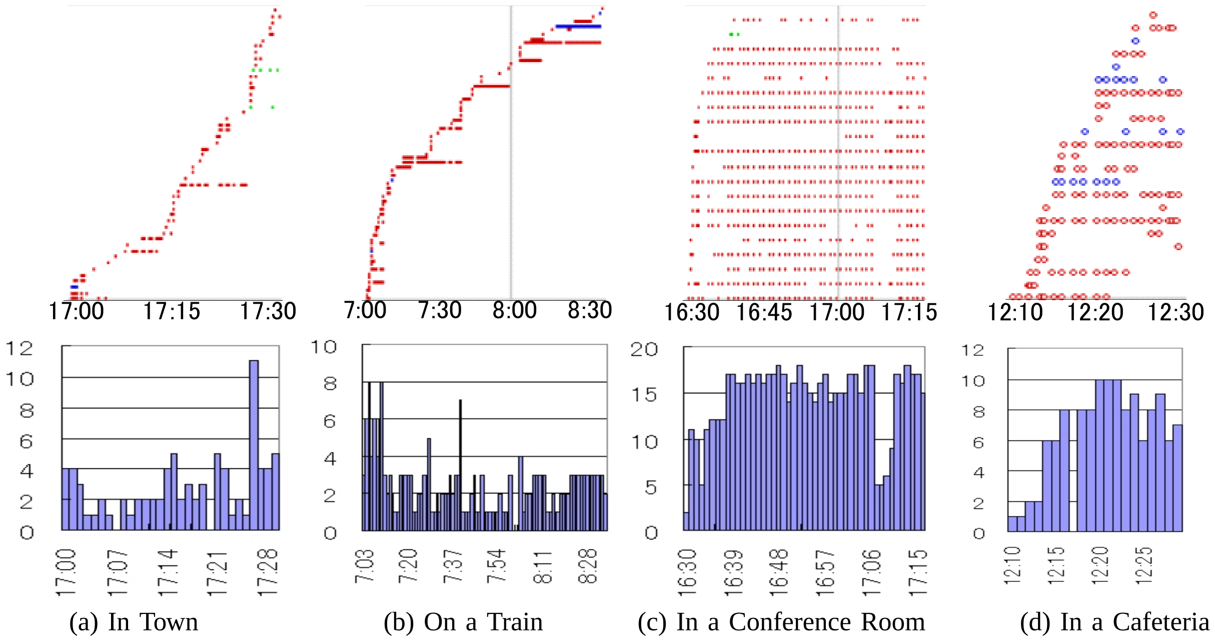
C. Interpolation of Missing Detection

We have described methods to extract and manage the Bluetooth detection logs on an adhoc network. However, there are false-negative cases that some devices within the communication range may not be detected. That is, abundant BDA data pours in at once especially at a crowded place, and the device cannot handle them all within the limited time interval while scanning for the surrounding Bluetooth devices.

To deal with such problems, we consider methods to check the detection logs of adjacent nodes on Delaunay network, and interpolate the BDA data which is definitely within the communication range of Bluetooth device. Initially, each node sends a copy of its own detection logs to adjacent nodes, and receives their copy of detection logs. Then, it extracts the BDA data which is not detected from its device, but detected from other adjacent nodes' devices. These BDA data will be the target data to perform interpolation, and the location of these adjacent nodes will be the criterion to determine whether or not to perform interpolation.

We validate only the BDA data owned by more than three adjacent nodes to perform interpolation. That is, a polygon is drawn using the location of adjacent nodes with the target BDA data as vertices. If the location of its own node is within the polygon, then the target BDA data shall be the one to be interpolated. We have chosen polygonal shape to determine the interpolation, because it is obvious that the entire polygonal region is covered from the communication range of Bluetooth device. The purpose of this interpolation method is to deal with missing detection, and the deformation of communication range caused by walls, buildings, and other obstacles are beyond our focus.

Figure 5 shows the interpolation process using the same Delaunay Network in Fig 4. Node v_0 has five adjacent neighbor nodes, namely v_2, v_8, v_9, v_{10}, v_{17}, and has the copy of their BDA detection logs. Among the BDA on detection logs, BDA1 is the only one that v_0 does not have, but more than three adjacent nodes (v_2, v_8, v_{10}, v_{17}) have. Using these nodes as vertices, a polygon is drawn starting from the upper node in clockwise direction. Finally, BDA1 can be determined to be included in v_0's detection data, as it is allocated within the polygon area.



(a) In Town (b) On a Train (c) In a Conference Room (d) In a Cafeteria

Figure 6. Detection Pattern of BDA (upper), Detected Number of BDA (lower)

Table I
CHARACTERISTICS OF PEDESTRIAN FLOW BY SITUATIONS

		User		Surrounding ppl		Sharing Space
		Stay	Move	Stay	Move	
Town		△	○	△	○	△ or ×
Lecture room		○	×	○	×	○
Cafeteria		○	×	○	○	○ or ×
Train	Moving	×	○	×	○	○
	Stopping	○	×	○	○	○ or ×

notations: (○)many; (△)some; (×)few/none

V. VERIFICATION AND DISCUSSION

The authors have done several investigations to observe surrounding Bluetooth devices in various situations, such as daily commuting and working at a university, or special events as conferences, sight-seeing tour, festival, and so on. To collect data, we used HP iPAQ 112 Classic Handheld PDA, which has been setup to record BDA with a timeout interval of 6 seconds after sending inquiry signal for every 30 seconds cycle.

Figure 6 shows BT detection logs in different situations. In this paper, four cases have been examined, namely strolling in town, transporting by train, attending the conference, and taking lunch at a cafeteria. The specific movements of the user and the surrounding people are described in Table I. The results of examination of detection logs are summarized in the points later. The upper diagram of Fig. 6 shows the detection pattern of Bluetooth devices, with the time-line expressed on the horizontal-axis, and the device ID assigned in chronological order of the incoming BDA on the vertical-axis. The mobile phones are colored in red, and PCs and devices other than mobile phones in green, and the unidentified devices in blue. The lower diagram of Fig. 6 shows the number of detected devices, with the time-line expressed on the horizontal-axis, and the quantity of BDA on the vertical-axis.

(a) Strolling in Town: Fig. 6(a) shows the changes of multiple detection logs encountered while strolling in town. The number of BDAs is not constant as the number of passers-by is always changing. Even if the pedestrians are walking in the same direction, their devices disappeared occasionally probably because their directions coincided only for a while or their walking speed was different. On the other hand, the same BDA was continuously identified in some places while the user was shopping or dropping in stores.

(b) Transporting by Train: Fig. 6(b) shows the detection in the train during rush hours. From the log, we can verify such situations as: (i) devices were continuously detected from passengers in the same car, (ii) many incoming and outgoing devices were detected when changing train; and (iii) a large number of people got on/off the train at major stations such as Osaka and Kyoto. The passenger's devices can be continuously detected even when the train is moving. However, due to the limited size and rectangular shape of the car, the quantity of detections has been low even if the train is crowded. Because of these observations, the authors have considered that it is necessary to identify the situation from detection patterns or some other methods, as it cannot be verified merely by the quantity of devices.

(c) Attending the Conference: Fig. 6(c) shows that many BDAs were detected continuously in the same room. As

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-236-3

most of the participants are staying in the room during the conference, the number of BDAs is almost constant (counting around 14 to 18 devices), except the time for coffee break. Because the room was wide enough to hold many people, the quantity of detections has been kept high. **(d) Taking lunch at a Cafeteria:** Fig. 6(d) shows that many devices have been detected during lunch time, as customers enter, take lunch and leave the cafeteria one after another. Some devices are detected continuously with long duration, and others are divided into several times with short duration, because two types of situations are mixed together: people sitting and eating lunch, and people walking around to look for seats or friends.

Based on the results of the experiments, the pedestrian flow can be assumed by the analysis of the data of detection logs as follows:

- **The number of BDA detection log:** crowdedness of people (requiring reference to the scale of space)
- **Time length of BDA detection:** people staying in same space or duration of the event
- **Appearance/Disappearance in BDA detection:** people staying, entering, leaving, or passing by

The detection logs show that there are several undetected devices even among those staying in the same space. Therefore, the interpolation of the missing detection is also needed in order to utilize the detection results.

VI. Concluding Remarks

We have shown possibilities to extract pedestrian flows by examining the detection patterns of surrounding Bluetooth devices, and proposed to apply methods to generate mobile adhoc network and manage detection data on the network, adhering to our policy to avoid initial preparations to install cameras or sensors on the environment, or manage data on a one point server. For deployment in actual environment, issues of energy consumption with mobile device battery must be considered in advancing the research. Moreover, privacy issue is also another concern because such personal data as user name and location should not be exposed to others. For future work, we plan to perform detailed analysis on Bluetooth device logs, examine the applicability with other sensory data, and provide location-based application using social contexts as pedestrian flows. On the other hand, we plan to continue further study on Delaunay networks, explore efficient ways or possibilities to manage social contexts data and log files, and evaluate our methods to interpolate missing data caused by inquiry faults.

Acknowledgment

This work was partially supported by MEXT/JSPS KAKENHI Grant Number 23650033.

References

[1] P. Lukowicz, A. Pentland, and A. Ferscha, “From Context Awareness to Socially Aware Computing,” IEEE Pervasive Computing, 11 (1), pp. 32–41, 2012.

[2] A. Campbell et al., “The Rise of People-Centric Sensing,” IEEE Internet Computing, 12 (4), pp. 12–21, 2008.

[3] N. Ratipanichvong, J. Yamamoto, R. Nishide, and H. Takada, “Place Recommendation for Pedestrian Reflecting Real-Time Situation and User’s Preferences”, IPSJ Symposium Kansai Branch, F-27, 2011.

[4] M. Kawabata, R. Nishide, M. Ueda, and S. Ueshima, “The Context-adaptable Pedestrian Navigation System and Usability in Practical Settings”, IEEE Pacific Rim Conf. on Communications, Computers and Signal processing, 1(05CH37690), pp. 368–371, 2005.

[5] E. O’Neill et al., “Instrumenting the city: Developing methods for observing and understanding the digital cityscape”, UbiComp, pp. 315–332, 2006.

[6] T. Nicolai and H. Kenn. “About the relationship between people and discoverable bluetooth devices in urban environments”, 4th int’l conf on mobile technology, applications, and systems, pp. 72–78, 2007.

[7] N. Eagle and A. Pentland. “Reality mining: sensing complex social systems”, Personal Ubiquitous Computing, 10, pp. 255–268, 2006.

[8] D. Kim and D.-K. Cho, BlueSense: Detecting individuals, locations, and regular activities from Bluetooth Signals, http://urban.cens.ucla.edu/cs219/images/0/0b/BlueSense.pdf (Retrieved March 9, 2012)

[9] J. Weppner and P. Lukowicz, “Collaborative Crowd Density Estimation with Mobile Phones,” 9th ACM Conf. Embedded Networked Sensor Systems, 2011.

[10] R. Nishide, T. Ushikoshi, S. Nakamura, and Y. Kono, “Detecting Social Contexts from Bluetooth Device Logs”, Supplemental Proc. Ubiquitous Computing (UbiComp), pp. 228–230, 2009.

[11] M. Ohnishi, R. Nishide, and S. Ueshima, “Incremental Construction of Delaunay Overlaid Network for Virtual Collaborative Space”, 3-rd Proc. Conf. on Creating, Connecting and Collaborating through Computing (C5’05), (IEEE CS Press), pp. 77–84, 2005.

[12] Y. Sun, Q. Jiang, and M. Singhal, “An Edge-Constrained Localized Delaunay Graph for Geographic Routing in Mobile Ad Hoc and Sensor Networks”, IEEE Trans. Mob. Comput. 9(4), pp. 479–490, 2010.

Analyzing Moodle/LMS Logs to Measure Mobile Access

María José Casany, Marc Alier, Nikolas Galanis, Enric Mayol, Jordi Piguillem
Department of Service and Information System Engineering
Universitat Politècnica de Catalunya
Barcelona, Spain
e-mail: {mjcasany, ludo, ngalanis, mayol, jpiguillem}@essi.upc.edu

Abstract—**Most Educational Institutions worldwide have deployed web based Learning Management Systems (LMS) as a means to provide support for their presence-based lectures and offer online-exclusive learning. These LMSs were designed and developed for users accessing the system through web browsers on desktop computers or laptops. However, over the last years, an increasing percentage of the registered accesses to various LMS platforms have been from mobile devices such as smartphones. While tackling the problems arising through the design of a mobile client for the Open Source LMS Moodle called Moodbile, the question of how to decide which services of Moodle could be accessed from smartphones became very relevant. This paper presents a data analysis study conducted on the Moodle server logs of the Universitat Politècnica de Catalunya - Barcelona Tech (UPC) virtual campus, Atenea, and the insight gained regarding the particular characteristics of the accesses from mobile devices. The main achievement of this study is that it provides insight of the use of the university LMS from mobile devices.**

Keywors-M-Learning; LMS; Moodle; Web analysis; Activity Logs.

I. INTRODUCTION

With the vast majority of learning institutions around the world embracing e-learning as a valid avenue for knowledge dissemination, there is an ongoing struggle of a number of Learning Management Systems (LMS) for a piece of the market. Moodle is one of the largest open source LMSs with a registered user base of more than 57 million people and more than 66,600 registered and verified sites as of December 2011 [1].

Universitat Politècnica de Catalunya UPC – Barcelona Tech is one of the institutions that opted to move to Moodle from a proprietary LMS. The migration to this new Moodle-based platform named Atenea, was carried out in stages, starting in 2004 and completing in 2007. Since then, there has been a dramatic rise in the usage of the services provided by UPC's virtual campus both on-site and off-site and currently Atenea gives service to more than 30.000 students.

Parallel to the establishment of Atenea and the surge of activity in this new platform the smartphone market emerged from its shy beginnings to become an important part of the mobile device market [2], [3]. This soar in the number of users owning a smartphone over the past few years has inevitably been noticed in the general usage statistics of Atenea, where we detect that a small but not negligible percentage of all user accesses is done from smartphones.

This observation poses a series of interesting questions as to the nature of these smartphone accesses, their success rate and any special requirements they impose upon the LMS compared to traditional desktop usage. The predominant concern is whether Moodle in general, and Atenea in particular, are ready to cater effectively to the needs of smartphone users. This concern arises not only from a software development perspective, but also from the learning design of the activities.

Tackling the issue of the emergence of mobile users in the Open Source LMS Moodle, the Moodbile project proposes a Service Oriented Architecture (SOA) in order to provide mobile learning applications with a set of Web Services that give access to a set of Moodle functionalities, while keeping these services as abstract and decoupled as possible, so that they could easily be ported to other LMSs if necessary [4]. Once the project started, the question of how to prioritize the services of Moodle became very relevant.

To study this question, this paper presents a data analysis study conducted on the server logs of Atenea. This study tries to answer several research questions: What tasks do mobile and desktop users perform on Atenea? Are there any significant differences between the tasks performed by mobile users and desktop users? Are short access time tasks the best ones for mobile users? And, what tasks are more suitable to be adapted for mobile access?. One of the results of this study shows that mobile sessions are shorter than desktop sessions. In fact mobile users usually access the Atenea/Moodle server to do only one task. This and other results are discussed in this paper.

The organization of the paper is the next: since our general goal is to integrate m-learning applications with Moodle, the related work regarding this integration is summarized in section two. After that, to decide which services of the LMS may be adequate for mobile devices, the data analysis of the Atenea server logs is presented and organized in two sections; section three presents the research questions and method to analyze the data and section four presents the results we found. Section five presents a discussion of the most relevant issues and finally section six presents the conclusions and further work.

II. RELATED WORK

The integration between m-learning applications and LMS is not an easy task, because LMS do not usually include interoperability standards to communicate with external applications; they are usually designed as monolithic or layered systems [5]. This section analyses some of the previous projects that have extended the LMS to the mobile scenario.

This extension is based on the creation of m-learning applications that extend the scope of the LMS. Such mobile applications usually follow one of two different approaches. The first approach focuses on engagement with mobile devices and mobile native applications. The benefits of this approach include access to engaging design, free use of hardware features and fast and lightweight technology. However, the main limitation is that applications are device specific, which usually requires high development costs. The second approach focuses on the interaction with a browser, so the technology is ubiquitous and device-independent. However, it may also be slower and it may be harder to access for some smart phones.

Usability and online/offline work are important issues when extending LMS to the mobile world. Specific restriction on mobile devices to display information and to interact with the user must be taking into account, and properly adapted. Some m-learning applications allow offline work when network coverage is not available or expensive. Offline work also implies that mobile applications must, at some point, synchronize the data stored locally on the device with the data stored on the LMS [6].

Lehner and Nosekabel [7] did one of the first studies about mobile devices that interact with virtual campuses. In this study, m-learning complements traditional learning. The *Welcome* system was developed to offer access to certain contents and services (such as calendars or events) of the virtual campus of the *Regensburg University* using mobile devices. The communication between the virtual campus and the mobile device is done mainly using SMS messages.

A classification of the services and functionalities of a LMS are presented in [6] and [8]. LMS functionalities are separated in four groups: data resources, e-learning specific services, common services (such as authentication, authorization or event management) and presentation of contents. They also identify the main issues of a LMS's architecture that may be resolved to offer these services to a mobile device. These architectural issues are: 1) context discovery (the system must check automatically the mobile device features and decide which services may be provided), 2) adaptation of contents and 3) synchronization between the mobile device and the LMS. They present a custom-made LMS developed in the University of Trento that follows this architecture in order to support mobility.

Hinkelman [9] developed, in Japan, a module of Moodle 1.6 to do testing using mobile devices. This version mainly offered testing services and feedback to students. Due to technological issues, this project was developed to work with Japanese mobile phones (because the tool is based on CHTML and 98% of the Japanese mobile phones supported this language). Afterwards, a study to adapt Moodle to mobile devices centered in the adaptation of contents is presented in [10].

The *Open University* has been working on Moodle extensions to mobile devices for quite long time. At 2009 they presented Mobile VLE for Moodle, a m-learning application to access Moodle from mobile devices. This application provides a subset of Moodle functionalities to be accessed by means a mobile device. This selection was done by popular polls to students. Students rated very high the following LMS functionalities as candidates to be the mobile services: assessment scores, messages (read course messages and unread forum posts), tasks, planning (see current week and its tasks, also the following weeks and the whole course) and resources (read resources from mobile devices and download if it is supported by the mobile phone) [11].

Momo [12] (*Mobile Moodle*) and MLE [13] (*Mobile Learning Engine*) projects developed m-learning applications to access some Moodle 1.9 functions. The Momo m-learning application is based on J2ME (Java 2 micro edition, a java version for mobile devices) while the MLE project developed a J2ME client application and an additional web version to access Moodle courses from mobile browsers. Some of the Moodle modules/activities supported by this project are the following: lesson, quiz, task, resource, forum, survey, choice, wiki (read only), database (search and query) and message.

Project MPage [14] develops a Moodle 1.9 client for iPhone. Some of the Moodle modules/activities supported by this project are the following: view course categories, access MyMoodle, edit events, access to resources in different formats, chat, choice, forum and Quiz.

Moviltest [15] is a J2ME application to download Moodle 1.9 tests and execute them in the mobile phone. After finishing the test, the results can be sent back to the Moodle server.

Moodle.org [16] has published a list of functionalities for an iPhone client for Moodle. The main functionalities they want to offer are the following: 1) To upload video, audio and other file formats to the user's private space in the Moodle server. 2) To view courses where the user is enrolled as well as to view other users enrolled in the same courses. 3) To view activities and content of a course and to download these contents to the mobile client. 4) To view user grades for students. 5) To receive notifications from the Moodle server, as well as to create and send new internal email messages. 6) To view forums, discussions and create and reply posts. 7) To view calendar events and assignments deadlines.

The current version of the prototype designed by *Moodle.org* only allows uploading files to the user's private space in the Moodle server, viewing course participants and view the list activities and contents of a course.

The related work is summarized in Table I. The table contains the studies that extend LMSs to the mobile scenario. The respective functions of the LMS involved in the study are listed.

TABLE I. SUMMARY TABLE OF THE DIFFERENT STUDIES INVOLVING EXTENDING THE LMS TO THE MOBILE WORLD.

LMS Functionalities from Related Work	
Source	***LMS functionality***
[8]	Upcoming Calendar events
[7][9]	Create an LMS adapted to info mobility from the scratch
[10]	Quiz
[12]	Resource, assessment, assignment, messages, posts
[13][14]	Quiz, lesson, assignment, resource, forum, survey, choice, message, wiki
[15]	Course activities, myMoodle, event, resource, chat, choice, forum, quiz
[16]	Quiz
[17]	View course participants, upload files, list course activities

III. RESEARCH AND DESIGN METHOD

A. *Research questions*

This work tries to analyze the characteristics of mobile users who access the LMS (Moodle) trough mobile web browsers and has been focused in answering the following research questions:

1. Which are the tasks performed by users from mobile browsers and from desktop or laptop browsers (referred to as desktops for the remainder of the paper for brevity)?
2. Is there any significant difference between the tasks performed from mobile devices and the tasks performed from desktops?
3. Are short access time tasks the best-suited ones for mobile devices?

Which tasks are more suitable to be adapted for mobile devices?

B. *Related work*

Several sources can be used to identify the basic patterns of mobile users accessing the LMS and which activities are more used from such small devices. The most popular approach is to make a survey for students and teachers [11], [19].

However, recently, new less intrusive and less subjective approaches are being adopted to gather data or requirements [20], [21]. These approaches include data analysis from different sources such as web server logs or LMS logs. Web server logs are vast collections of data about accesses to specific web pages. The main limitation of analyzing web server logs is that they contain only low-level data. LMS log files are perhaps the most promising source of automatic gathered online learning data. Since students typically login on such systems, the LMS logs keep track of users and sessions. These logs also gather a range of relatively high-level student data such as grades, posts in a forum etc. These data are more focused on student activity than web server logs. In [22], there is a summary of several alternative approaches to automatically analyze e-learning data as well as the different data sources used for the analysis.

Nevertheless, in the above approaches, only one data source was used. The challenge with respect to data gathering is the interrogation of several data sources. If the LMS data were correlated with additional information gathered from other systems, a richer picture of student learning process could be generated [23]. In our study, the data from the LMS and web server has been merged in order to gather information about the client operating system. Another limitation of the previous approaches is that none of them are specifically designed to analyze mobility and LMS.

C. *Data sources and analysis*

This study was conducted using Moodle/Atenea logs and web server logs of the first academic semester of 2011. More than 15 million entries/registers were analyzed. We have addressed the analytical process in the following three phases, shown in Fig. 1.

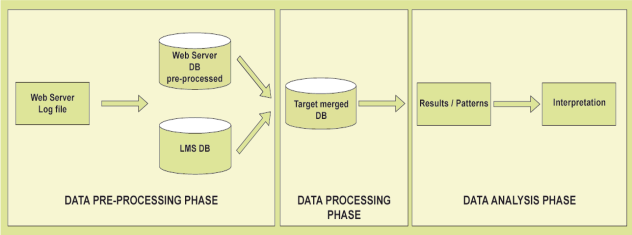


Figure 1. Data analysis phases.

1. Data Pre-Processing: that includes selection and capture of data. During this step, data is cleaned from empty or useless web server log entries. Some derived information is calculated or aggregated from web server log entries (see Table II). All this information is stored in a relational DB and merged with the entries provided by the LMS log. The merging criteria take into consideration data and time, IP-address and Moodle module accessed or type of action performed (view, add, update, etc.).

TABLE II. EXAMPLE OF PRIMARY AND DERIVED VARIABLES FORM WEB SERVER AND LMS LOG.

Variable name	Description	Type (Primary/Derived)
Course	Moodle course id	Primary
Module	Moodle module accessed	Primary
Action	Moodle action performed	Primary
Operation System	Type of operating system	Derived
Year period	Exams or lectures	Derived
Day slot	Morning, midday, afternoon, evening, night	Derived

2. Data Processing. In this phase, data of the database is processed and aggregated accordingly to facilitate the generation of partial reports to support the analysis and to answer our research questions.
3. Data Analysis**.** In this phase, data is analyzed based on the previously generated reports and conclusions of the analysis are presented.

IV. Results and findings

The analysis of the data retrieved from both sources points out that most of the accesses to the LMS (96,21%) are performed from desktop or laptop computers, while only 3,48% are from mobile devices and 0.28% from tablets. Three distinct types of accesses to the LMS have been identified: queries, updates and logins/logouts. Fig. 2 shows the relative percentages of the three types of accesses for desktops and mobile devices.

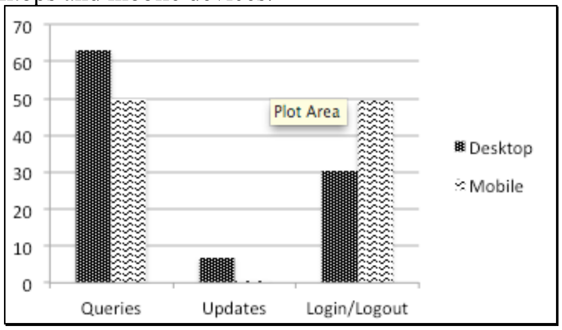


Figure 2. Relative percentages of the three types of accesses for desktops and mobile devices.

Regarding mobile Operating Systems, more than half (58.49%) of the registered accesses were from an iOS device followed by Android devices (18,67%) and Blackberry OS (12%). Fig. 3 presents a detailed breakdown of the various OS percentages.

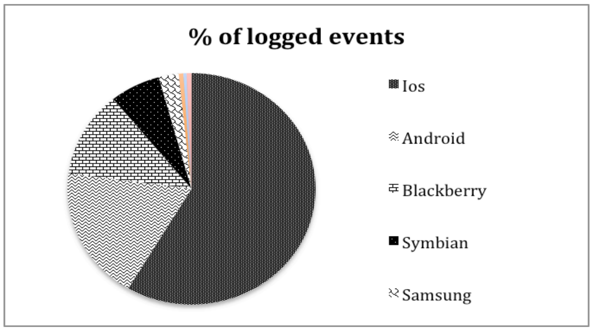


Figure 3. Breakdown of logged events by mobile OS.

The number of logins from mobile devices is very high compared to the total number of mobile operations. Therefore, we have isolated the average % of mobile logins and we detected that the mobile sessions are very short, almost atomic: 45,15% of the average events recorded from mobile devices are logins, while only 23,74% of the logged events from desktop computers are logins. This situation is similar in almost all the mobile operating systems (except from Windows CE and Android mobiles where the average % is a little lower). Therefore, we can state that many times the mobile users try to login to the LMS without success and when they do succeed; they only do one action (the average number of actions per session from mobile devices is 1,12 compared to 3,21 in desktop). In this sense, we hypothesize that Mobile users usually access the LMS from a link to do one single action. Additionally, logs state that the usual entrance point to the LMS is not the main course page, because only 20% of the logged events correspond to the "course view" action.

Finally, we have analyzed which actions mobile users and desktop users perform in the LMS. In general, the most frequently used LMS activity modules are quiz, assignment, forum, course, resource and the access to the user profile, as it is shown in Fig. 4. Among update actions the most attempted action from mobile devices are "answer the quiz" followed by "post in a forum".

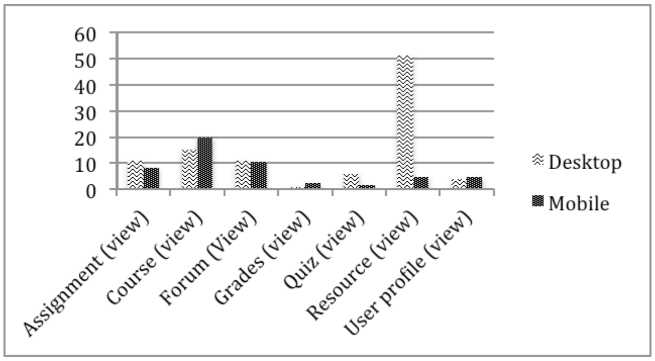


Figure 4. Percentages of query actions carried out.

To analyze the actions with a higher percentage of mobile users, actions have been divided in two groups: updates and queries. Action-updates with a higher percentage for mobile users compared to desktop ones are quiz attempts with 86,85% from mobile devices compared to 67,12% from desktop users, followed by forum accesses for discussion/post creation with 10,51% vs. 5,27%. Action-queries with a higher percentage include course view with 20% vs. 15 user profile view 4,87% vs. 3,96% and finally, consulting grades with 2,44% vs. 0,96%. Another observation is that "view resources", which is the second most queried activity, has a similar percentage across both platforms (50,5% for mobile compared to 51% for desktop).

Furthermore, we have analyzed and compared the data retrieved during spring lecture season (February - May) of 2011 and of the exams season of autumn semester (January 2011) of 2010 and the spring semester (June) of 2011. 61,86% of the total registered accesses were during the spring lecture season and the 38,14% were during the two exam periods. Table III shows the relative percentages of the accesses during these two periods. We have also included activity from tablets for completion purposes.

TABLE III. Breakdown of accesses registered during lecture and exam periods.

	Exams (E)	Lectures (L)
Desktop	95,34%	96,75%
Mobile	4,29%	2,98%
Tablet	0,32%	0,25%

From the table data, we observe that although the activity from desktops remains almost unchanged during the two periods, there is a roughly 50% increase in activity from mobile devices.

Figs. 5 and 6 present the variations in activity for mobile devices and desktops across these two periods.

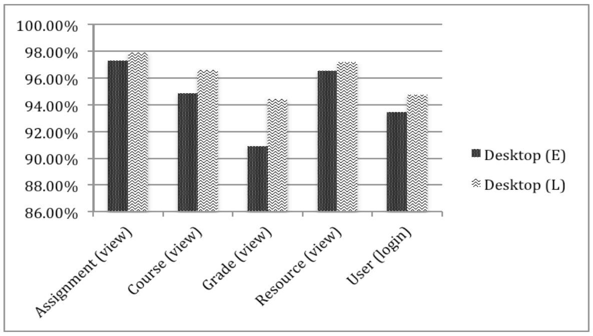


Figure 5. Desktop activity during exams (refered as E) and lectures (refered as L).

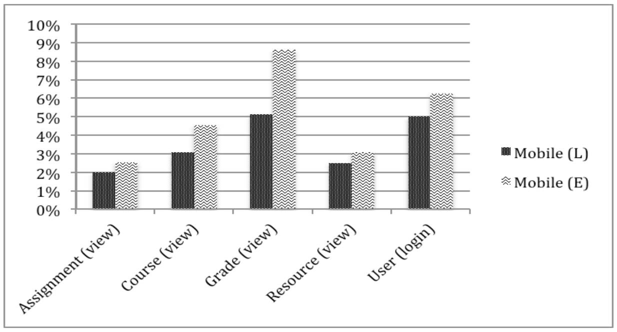


Figure 6. Mobile activity during exams (E) and lectures (L).

From Fig. 6, we see that during the exams period, mobile users mainly consult grades followed by course view and resource view.

Finally, we divided the day into 5 time periods:

- Early morning (0:00 – 7:00)
- Morning (7:00 – 13:00)
- Midday (13:00 – 16:00)
- Afternoon (16:00 – 20:00)
- Night (20:00 – 24:00)

In general, we notice that activity is higher during the afternoon (34,47%) followed by morning (28,58%), midday (18,28%), night (13,33%) and finally, early morning (5,33%). Fig. 7 shows a rise in mobile activity during the night hours while desktop activity drops during the same hours.

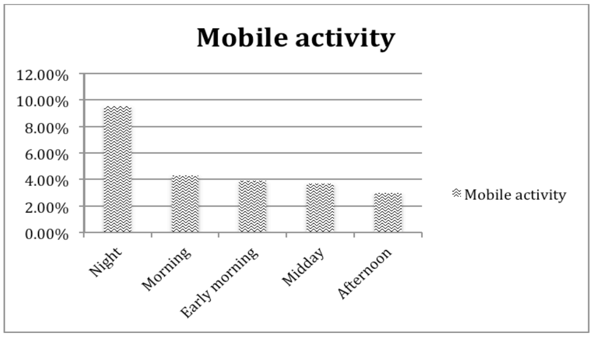


Figure 7. Breakdown of mobile activity during the different time periods.

Analyzing the mobile activity in more detail, we notice that queries and updates increase considerably during night hours (7,15% of the updates and 9,07% of the queries), followed by morning activity (0,49% of the updates and 3,45% of the queries).

During night hours the most accessed activities are: view grades, view course, view wiki, view user profile, view choice, enroll into course, add posts or discussions in a forum, and view task.

V. DISCUSSION

The first interesting observation is that most of the Atenea modules that are accessed using computers seem to be used as well from mobile devices. One explanation could be that the Atenea has been modified by UPCnet to improve accessibility and usability. These improvements include adding caption fields to tables, links and figures, adding explanations to popup menus, etc. [24].

Another issue is the high percentage of login activity from some mobile devices. The login activity represents approximately 45% of the mobile activity. From this we deduce that many times the mobile user cannot log in to Atenea. Fig. 8, shows the percentage of mobile activity dedicated to login attempts.

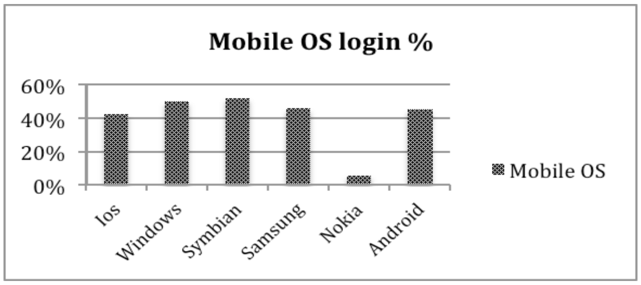


Figure 8. Percentage of mobile login activity.

From the related work and the Atenea log analysis, we have found that the following Moodle features are the most accessed from mobile devices.

TABLE IV. MOODLE FEATURES TO BE USED FROM MOBILE DEVICES.

	From Related Projects	**From Atenea/Moodle log Analysis**
Internal Message	x	
Forum posts and discussions	x	x
Task /assignment	x	x
Resource (view)	x	x
Choice and quiz	x	x (quiz only)
Course activities (view)	x	x
Course participants (view)	x	
Grade (view)	x	

So, we have considered the following Moodle features as necessary to be includes in the Moodbile project development pipeline: view course activities, view course participants, view student's grades, view resources, view and upload tasks, access forums and discussions, read and reply posts, do quizzes, view upcoming calendar events and view user profiles.

VI. CONCLUSIONS

In the analysis of the logs of Atenea, only 3,48% of accesses came from mobile devices; but in spite of this, it is fair to assume that this percentage is going to grow significantly. Relative large screens (4,3 inches and above) that replace the cheap feature phones are starting to show up on the smartphone market. Some market studies show how tablets (with screens from 7 to 10 inches) are cannibalizing the market of cheap netbooks that students used to buy during the last four years [25].

From the results of the study, we find especially relevant our hypothesis that mobile users usually access the LMS from a link to do one single action. This hypothesis is based on the fact that mobile sessions are very short (about 1.2 logged actions per session) and that from the logs we know that students do succeed in doing this action. From this we conclude that the navigation design of the LMS needs to be tailored for this quick usage pattern.

Another issue is the fact that almost half of the actions performed from mobile are to log in and out. Mobile LMS front-ends should automatically login the student/teacher, cache the contents of the LMS, and make it available offline when connection is unavailable, slow or expensive.

Finally, teachers need to be aware that students access their online courses through mobile devices, and make their online courses more mobile friendly. Learning design has to take this issue into deep consideration.

ACKNOWLEDGMENT

This work has been funded by the "Spanish Ministry of Science and Innovation" in project MiPLE code TIN2010-21695-C02-02.8.

REFERENCES

[1] Wexler, S., Grey, N., Miller, D., Nguyen, F. and Barnevelda, A., Learning Management Systems: The good, the bad, the ugly... and the truth. E-learning Guild, 2007.

[2] IDC Press release: Smartphones sales worldwide. IDC Press. http://www.idc.com/getdoc.jsp?containerId=prUS23123911 (retrieved: may 2012)

[3] Gartner newsroom: Smartphones sales worldwide. Gartner Inc. http://www.gartner.com/it/page.jsp?id=1924314 (retrieved: may 2012)

[4] Casany, M.J., Alier, M., Mayol, E., Piguillem, J., Galanis, N., Conde, M.A. and García-Peñalvo F. J., "Moodbile: A Framework to integrate m-learning applications with the LMS," In Journal of Research and Practice in Information Technology (2012) in press.

[5] Sclater, N., "Web 2.0, personal learning environments, and the future of learning management systems," In Research Bulletin, **13**, pp. 2008-2009.

[6] Trifonova, A., Ronchetti, M., "A general architecture to support mobility in learning," In Proceedingsof the IEEE International Conference of Advanced Learning Technologies, 2004, pp. 26-30.

[7] Lehner, F., Nosekabel, H, "The role of mobile devices in E-Learning first experiences with a wireless E-Learning environment," In Proceedings of the IEEE International Workshop of Wireless and Mobile Technologies in Education, 2002, pp. 103-106.

[8] Colazzo, L., Molinari, A., Ronchetti, M., Trifonova, A., "Towards a multi-vendor mobile learning management system," In Proceedings of the ED-Media, 2003, pp. 121-127.

[9] Hinkelman, D., "Moodle for Mobiles Project", 2005. Available at: http://moodle.org/mod/forum/discuss.php?d=33033 (retrieved: may 2012)

[10] Cheung, B., Steward, B., Mcgreal R., "Going mobile with Moodle: First steps," In Proceedings of IADIS International Conference on Mobile Learning, Dublin. International Association for the Development of the Information Society, 2006.

[11] Thomas.,R.C., "Mobilizing the Open University: case studies in strategic mobile development," In Journal of the Research Center for Educational Technology*,* **6**(1), 2010, pp. 103-110.

[12] Mobile Moodle. http://www.mobilemoodle.org/momo18/ (retrieved: may 2012)

[13] Mobile Learning Engine. http://mle.sourceforge.net/ (retrieved: may 2012)

[14] Mpage. http://massmedia.hk/moodle/course/view.php?id=2 (retrieved: may 2012)

[15] Cosme, C.A., Pedrero, A., Alonso, V., "Moviltest: adaptación de cuestionarios de Moodle para dispositivos móviles," V Simposio Pluridisciplinar sobre Diseño y Evaluación de Contenidos Educativos Reutilizables (SPDECE'08), 2008.

[16] Moodle.org mobile. http://docs.moodle.org/dev/Mobile_app (retrieved: may 2012)

[17] Alier, M., Casany, M.J. and Piguillem, J., "Towards mobile learning applications integration with learning management systems," Multiplatform e-learning systems and technologies: mobile devices for ubiquitous ICT-based education, 2009, pp. 182-194.

[18] Alier, M., Casany, M.J., Conde, M.A. and García-Peñalvo, F.J.G., "Interoperability for LMS: the missing piece to become the common place for e-learning innovation," International Journal of Knowledge and Learning, 2010, **6**(2), pp. 130-141.

[19] Mills, K., "M-Libraries: Information use on the move," A report from the Arcadia Programme. Cambridge, UK: University of Cambridge, 2009.

[20] Sinickas, A., "Keeping Keeping score: Making performance data more compelling Part 1," In Strategic Communication Management, 11(4), 2007, pp. 32–35.

[21] Gofton, K., "Data firms react to survey fatigue," In Marketing, 3, 1999, pp 29-30.

[22] Black, E.W., Dawson, K., Priem, J., "Data for free: Using LMS activity logs to measure community in online courses," In The Internet and Higher Education, vol. 11(2), 2008, pp. 65-70, Elsevier.

[23] Dawson, S., "'Seeing' the learning community: An exploration of the development of a resource for monitoring online student networking," In British Journal of Educational Technology, vol. 41(5), 2010, pp 736-752.

[24] UPCnet, "Experiencias de mejora en usabilidad y accesibilidad en Moodle," In Proceedings of the MoodleMoot-Spain 2011, San Sebastian, Spain.

[25] Gartner newsroom: Tablets market share forecast. Gartner Inc. Available at:. http://www.gartner.com/it/page.jsp?id=1800514 (retrieved: may 2012)

Elckerlyc goes Mobile
Enabling Technology for ECAs in Mobile Applications

Randy Klaassen, Jordi Hendrix, Dennis Reidsma, Rieks op den Akker
Human Media Interaction
University of Twente
Enschede, Netherlands
PO Box 217, 7500AE Enschede, Netherlands
{r.klaassen, j.k.hendrix, d.reidsma, h.j.a.opdenakker}@utwente.nl

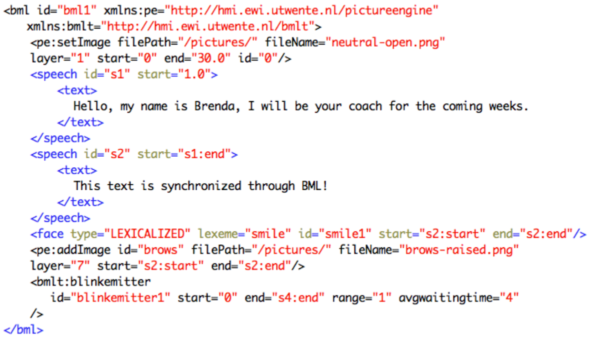
***Abstract*—The fast growth of computational resources and speech technology available on mobile devices makes it possible for users of these devices to interact with service systems through natural dialogue. These systems are sometimes perceived as social agents and presented by means of an animated embodied conversational agent (ECA). To take the full advantage of the power of ECAs in service systems, it is important to support real-time, online and responsive interaction with the system through the ECA. The design of responsive animated conversational agents is a daunting task. Elckerlyc is a model-based platform for the specification and animation of synchronised multimodal responsive animated agents. This paper presents a new light-weight PictureEngine that allows this platform to embed an ECA in the user interface of mobile applications. The ECA can be specified by using the behavior markup language (BML). An application and user evaluations of Elckerlyc and the PictureEngine in a mobile embedded digital coach is presented.**

***Keywords*-Mobile User Interfaces**

I. Introduction

Advances in user interface technology — speech recognition, speech synthesis and screen capacities — allow more and more people to engage in spoken interaction with services on their mobile phones. Examples of these services are intelligent personal assistants in search applications, persuasive systems or characters in games. It is well known from user studies that the use of a talking head or an embodied conversational agent (ECA) has a positive effect on user experience when using these kind of services [1].

The presentation of a service agent by means of a persona supports the idea of the computer as a social actor. Research has shown that animation of human-like social behaviours and expressions by means of a virtual human or embodied conversational agent strengthens the impression that the agent is present and engaged in the interaction. In human-human conversations the one who has the speaker role is monitoring his addressees while speaking. Listeners give backchannels, short comments, and may also interrupt the speaker. By his gaze behaviour the speaker shows his interest with the addressee. By adjusting or stopping his speech he shows being responsive to the listeners comments and that he is really engaged in the conversation. Gaze behaviour in conversations is important for interaction management, in particular for signaling that one wants to have the floor, that the speaker wants to keep the floor or is willing to yield the floor. Expressions of emotion are prime indicators of engagement in what is going on in the conversation [2]. In designing virtual humans that are able to show these social signals and responsiveness one needs well designed model-based specification languages and tools.



```
<bml id="bml1" xmlns:pe="http://hmi.ewi.utwente.nl/pictureengine"
   xmlns:bmlt="http://hmi.ewi.utwente.nl/bmlt">
    <pe:setImage filePath="/pictures/" fileName="neutral-open.png"
    layer="1" start="0" end="30.0" id="0"/>
    <speech id="s1" start="1.0">
        <text>
          Hello, my name is Brenda, I will be your coach for the coming weeks.
        </text>
    </speech>
    <speech id="s2" start="s1:end">
        <text>
          This text is synchronized through BML!
        </text>
    </speech>
    <face type="LEXICALIZED" lexeme="smile" id="smile1" start="s2:start" end="s2:end"/>
    <pe:addImage id="brows" filePath="/pictures/" fileName="brows-raised.png"
    layer="7" start="s2:start" end="s2:end"/>
    <bmlt:blinkemitter
      id="blinkemitter1" start="0" end="s4:end" range="1" avgwaitingtime="4"
    />
</bml>
```

Figure 1. An example of a BML specification for an ECA.

The SAIBA framework [3] provides a good starting point for designing interactive virtual humans. Its Behaviour Markup Language (BML) defines a specification of the form and relative timing of the behaviour (e.g. speech, facial expression, gesture) that a BML realizer should display on the embodiment of a virtual human. An example of a specification in BML can be found in figure 1. Elckerlyc is a state-of-the-art BML realizer. In [4] its mixed dynamics capabilities are described as well as its focus on continuous interaction, which makes it very suitable for virtual human applications requiring high responsiveness to the behaviour of the user.

The Elckerlyc platform can act as a back-end realizer for different embodiments, like physical robots or realistic 3D full kinematic virtual humans. Using a full 3D virtual human on a mobile phone is too heavy in terms of processing power and battery usage. To be able to use the Elckerlyc platform on a mobile phone a light-weight animation embodiment is needed. This paper presents the PictureEngine, a light-weight animation embodiment that enables our SAIBA-based BML realizer to be implemented and run on mobile applications. Section II describes the Elckerlyc platform in more detail. The PictureEngine will be discussed in Section III, the Android implementation of the platform and the PictureEngine in Section IV.

Research by e.g. Bickmore [1] showed that personification of the user interface of coaching systems can have positive effects on the effectiveness of the coaching program. Real-time animations do have a positive effect on the user experience. Compared to static pictures or prerecorded movies, real-time animations are able to react immediately to the user. Responsiveness increases the experience of engagement of the agent. In section V a personalised context-aware multi-device coaching application will be discussed. The coaching application makes use of the mobile Elckerlyc platform. The ECA developed for this application presents feedback from the digital coach by animated spoken interaction. We conclude with with a description future work on the development of the mobile embodied coach and user evaluations of a coaching application that is using the PictureEngine.

II. The Elckerlyc platform

In behaviour generation, at least two main aspects can be distinguished. The first aspect is the planning of the actions and movements as means to a certain goal that the agent intends to achieve. The second one is the actual detailed realisation of the verbal and non-verbal behaviours in terms of “embodiments” of the (graphical) virtual human - including the generation of the speech by a text-to-speech synthesizer. This distinction between intent planning, behaviour planning and behaviour realisation is the basis of the SAIBA[1] framework [5]. According to this framework the detailed behaviours are specified in the Behaviour Markup Language (BML)[6].

The Elckerlyc platform is a BML realizer for real-time generation of behaviours of virtual humans (VHs). The Elckerlyc platform has been described and compared with other BML behaviour realizers (for example EMBR [7] and Greta [8]) in various papers [9], [10], [4].

Dependent on the application and task that the intelligent system has, the virtual human presents for example the character of a tutor, an information assistant, or a conductor. The goal is to make these embodied conversational agents look like believable and convincing communicative partners while interacting with humans. This requires the generation and coordination of “natural” behaviours and expressions.

Reidsma and Welbergen [10] discusses several features of the modular achitecture of Elckerlyc and relates each of them to a number of use requirements. A general overview of the Elckerlyc system can be found in Figure 2. The input of the Elckerlyc platform is a BML specification. “BML provides abstract behaviour elements to steer the behaviour of a virtual human. A BML realizer is free to make its own choices concerning how these abstract behaviours will be displayed on the embodiment. For example, in Elckerlyc, an abstract ‘beat gesture’ is by default mapped to a procedural animation from the Greta repertoire. The developer may want to map the same abstract behaviour to a different form, i.e., to a high quality motion captured gesture.”[10]. Different Engines will handle their own parts of the behaviour specification and generate synchronised instructions for realising i.e. speech output, body gestures, postures and facial expression. The output of all the engines is displayed on one embodiment, like a realistic 3D full kinematic virtual human, the Nabaztag or a graphical 2D cartoon like picture animation. Figure 3 shows three types of embodiments supported by the Elckerlyc platform.

Not every embodiment is able to render all the behaviours that can be specified in BML. This depends on what the embodiment offers e.g. a robot that is not able to smile or a picture animation that lacks a picture showing the smiling face cannot render the requested smiling behaviour. The interface between the output of Elckerlyc and the embodiment occurs in a Binding. A Binding is an XML description to achieve a mapping from abstract BML behaviours to PlanUnits that determines how the behaviour will be displayed in the embodiment. Bindings can be customized by the application developer.

This paper discusses how these Bindings were exploit. A light-weight PictureEngine was developed that makes it possible to run Elckerlyc on mobile Android platforms. Elckerlyc allows for a transparent and adjustable mapping from BML to output behaviours (rather than the mostly hardcoded mappings in other realizers), and allows for easy integration of new modalities and embodiments, for example to control robotic embodiments, or full 3D embodiments. To run Elckerlyc on mobile platforms a light-weight PictureEngine was developed that allows rendering of behaviours and expressions using layering of pictures.

III. The PictureEngine

A realistic 3D full kinematic virtual human embodiment is not suitable for use on mobile devices for multiple reasons. Not only do such devices lack the processing power to render this kind of environment, but displaying a full scene including a full body ECA on the relatively small screen of a mobile device is quite impractical. The displayed size of

[1] www.mindmakers.org

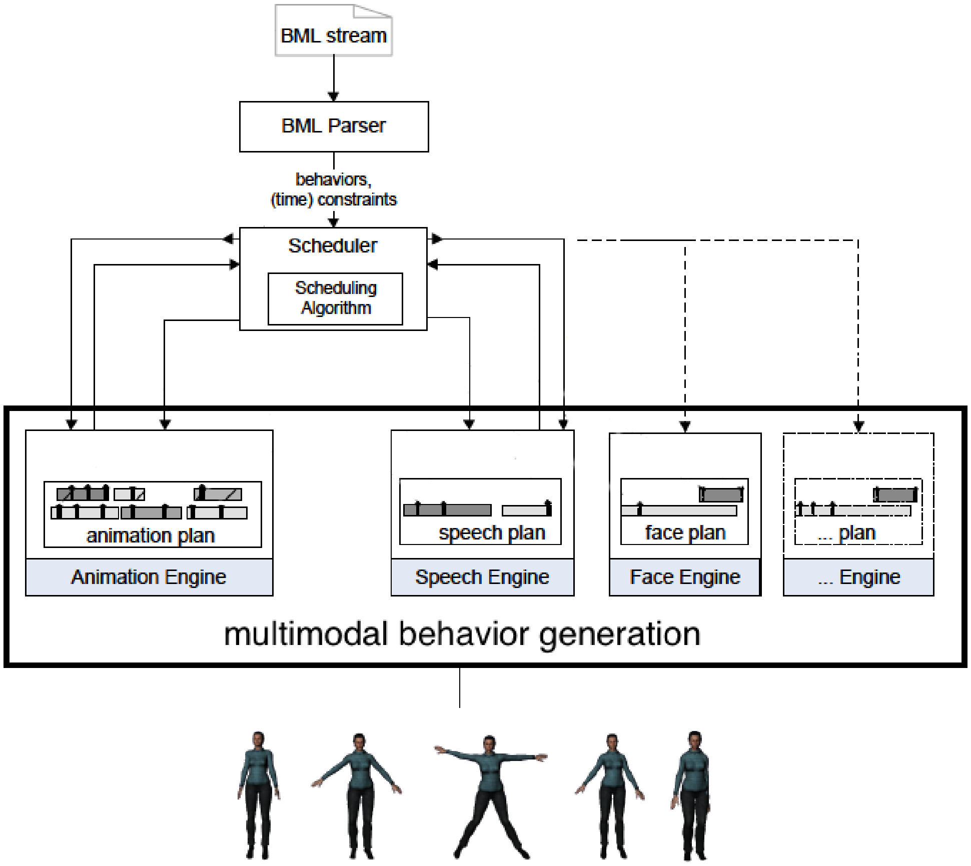


Figure 2. Overview of the Elckerlyc architecture. BML input is processed by the Elckerlyc system by different engines. The result is combined into one embodiement.

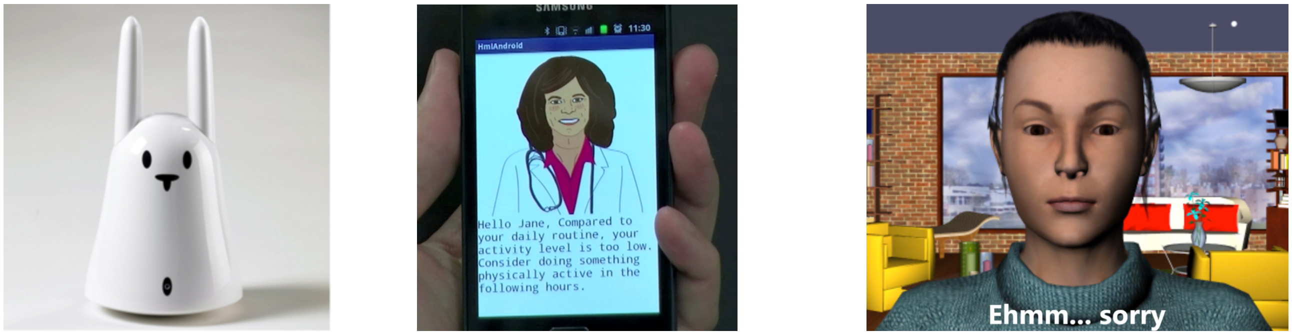
the ECA would make it so small that its expressions would hardly be visible. The high processing demands would also drain the device's battery quickly. In order to avoid all these problems, Elckerlyc uses a different graphical embodiment on the Android platform, the PictureEngine.

The PictureEngine is a lightweight graphical embodiment that uses a collection of 2D images in order to display the ECA. While having a 2D image embodiment does present some limitations, it also has its advantages. First of all, it has low demands in terms of processing and memory. It also allows for great variation in the design of ECAs. One could for example design a cartoon figure ECA, an ECA based on more lifelike illustrations, an ECA based on prerendered 3D images, or even an ECA based on photographic images of a real person. This section discusses the most important aspects of the PictureEngine.

A. Layers

In order to generate a dynamic ECA from a collection of images, the PictureEngine uses a layer-based approach. Different parts of the ECA are displayed on different layers of the final image, and can thus be in different states. For example, one layer may contain the eyes, while another contains the mouth. The base layer normally contains the ECA in a base state, meaning that when the ECA is in a neutral or passive state, the user sees only this base layer. That means that while each (facial) feature of the ECA does have its own layer, they are also present in the base layer. This means that the base layer contains for example a full face with a neutral expression, even though the eyes and mouth may have their own layers. There can also be layers containing features that are not visible in the base state, such as hands that only move into the frame when executing a gesture. By using this layer approach, different parts of the ECA can be manipulated independently and combined in order to generate different expressions. This also allows the ECA to do several (connected or unconnected) things at once, such as blink while also speaking and pointing at something.

As noted earlier, the layer approach does present some limitations. Because the features of the ECA are in separate layers, the base onto which these features are displayed



Hello Jane, Compared to your daily routine, your activity level is too low. Consider doing something physically active in the following hours.

(a) The Nabaztag rabit (b) a 2D cartoon like picture animation (c) a realistic 3D full kinematic virtual human

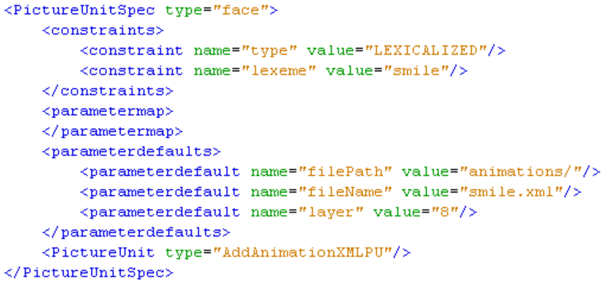
Figure 3. Three types of embodiment used as back-end for the Elckerlyc platform

(usually a face, and possibly part of the body) is generally static. This means that any movement of the entire ECA poses a problem. When an ECA has facial features on different layers, the layered structure prevents it from moving around. This also applies to smaller movements such as nodding, shaking and tilting of the head. However, because the PictureEngine is designed to be used on smaller screens, the ECA will generally be displayed as a talking head, a closeup of a face covering most of the available screen space. In this kind of environment, having the ECA perform locomotion is already impractical and, since there is hardly any room for the ECA's environment to contain anything but itself, arguably unneccessary.

B. Animations

While single images may suffice for portraying expressions in many cases, there are other cases where an ECA simply has to display some motion in order to come across as believable. To make this possible, the PictureEngine also allows the use of animations instead of single images. These animations are defined by using a simple XML format that allows a number of images to be listed, together with the duration for which they are to be displayed. While these durations are specified in seconds, the nature of the BML scheduler allows the duration of animations to be adjusted according to the BML code that is being realised, causing the animation to play faster or slower depending on the timespan determined by the scheduler.

These animation XML files have an additional feature that provides an advantage over using an already established format for image animations: the possibility to include synchronization information in the animation specification. This allows a synchronization point to be included in the specification between any two frames of an animation. These synchronization points are available for use in the main BML code. In this way, it is possible to e.g. synchronize the stroke of a beat gesture animation with a certain word within a speech element.



```
<PictureUnitSpec type="face">
    <constraints>
        <constraint name="type" value="LEXICALIZED"/>
        <constraint name="lexeme" value="smile"/>
    </constraints>
    <parametermap>
    </parametermap>
    <parameterdefaults>
        <parameterdefault name="filePath" value="animations/"/>
        <parameterdefault name="fileName" value="smile.xml"/>
        <parameterdefault name="layer" value="8"/>
    </parameterdefaults>
    <PictureUnit type="AddAnimationXMLPU"/>
</PictureUnitSpec>
```

Figure 4. PictureBinding entry for a smile.

C. PictureBinding

Like other Elckerlyc embodiment engines, the PictureEngine uses a Binding. This PictureBinding allows a combination of a BML behaviour class and (optionally) several constraints to be mapped to a certain PictureUnit (i.e. an image or animation). It is possible to include anywhere from zero constraints to all the constraints defined for the corresponding BML behaviour type. This allows the designer of a PictureEngine ECA to refine those behaviours that are most relevant to the ECA, and implement any others in a more general fashion.

The actual PictureBinding itself is defined in an XML file containing the behaviour classes and constraints and the PictureUnits and parameters they are to be mapped onto (see Figure 4 for an example). The accessibility of this format allows an ECA to be designed or modified by someone who does not have knowledge of the inner workings of Elckerlyc. Only knowledge of BML and the available PictureUnits and their parameters is required to be able to build a complete PictureBinding.

D. Lipsync

In order to visually display the fact that the ECA is speaking, the PictureEngine provides a rudimentary lipsync facility. This lipsync feature is implemented in the same way as the lipsync provided by the default AnimationEngine.

However, where the AnimationEngine provides a full mapping from visemes to animation units, the PictureEngine lipsync currently does not make use of such a mapping (although it could be added in the future). In its current state the lipsync allows a single animation to be specified which is played whenever the ECA is speaking. This animation is repeated for the number of times it fits into the duration of the speech unit (and slightly adjusted so that the amount of repetitions becomes a round number).

IV. Android implementation

Since the Elckerlyc platform is implemented almost entirely in Java, all of its core elements run on Android without any modification. However, since Android has its own environment for visual and audio output, some additions are required. This does not mean that the Android application uses a modified version of the core Elckerlyc platform. The fact that Elckerlyc uses an XML format to define the loading requirements for a specific ECA allows the Android application to simply load its own versions of a few key components. This allows the core Elckerlyc system to be used in the Android application as-is, so any changes to the Elckerlyc core can be directly used in the Android application without having to modify or port it first. The subsystems for which the Android application contains its own versions are discussed here.

A. Graphical Output

Because the Android platform has its own graphical environment, the engines which provide graphical output use a modified component for printing their output in the Android application. This goes for both the PictureEngine, which handles the graphical display of the ECA, and the TextSpeechEngine, which outputs speech elements to a text area. Since PNG images can be handled without problems by the Android graphical environment, the additional code needed to replace the PictureEngine's default output subsystem with a version that works on Android is minimal.

B. SpeechEngine

In the case of the SpeechEngine (for the rendering of spoken text using text-to-speech (TTS)) the differences with Android are unfortunately more severe. The TTS engines used in the PC version of the SpeechEngine contain several dependencies on native PC systems and cannot be used on Android without significant changes. However, Android does offer an internal TTS system. Using this internal system avoids the costly process of porting a TTS engine and any possible efficiency issues this may bring. In order to make use of the internal Android TTS system, an Android adaptation of the Elckerlyc SpeechEngine is needed. This includes the module that loads and initializes the engine, as well as the parts of the system dealing with the actual TTS operations.

The main problem with the Android TTS system is that it is not possible to obtain timing information for utterances, meaning there is no way to find out exactly at what time a word is spoken. This causes the BML scheduler to be unable to use synchronization points within utterances. This makes it hard to precisely synchronize other behaviours with specific words being spoken. A partial solution is that utterances are presynthesized to a file in order to find the total duration of the utterance. This provides the crucial information for the Elckerlyc scheduler. This ”preloading” of utterances causes a delay at startup before the ECA starts playback of the requested BML code.

Furthermore, the TTS also does not offer any viseme information, making it impossible to use real lipsync on Android. This is the main reason the PictureEngine does not currently support true lipsync.

C. Subtitles

Because the PictureEngine can run on a mobile device, the chances of the user having trouble hearing the text spoken by the TTS on the Android system are quite high. This could be caused by factors such as environment noise, low volume or bad speakers. In order for the user to still be able to interact with the ECA in these situations, the Android application also offers an on-screen representation of any spoken text, comparable to subtitling. The TextSpeechEngine (on-screen text display) receives the text handled by the SpeechEngine and displays this in a text area, synchronized (per utterance) with the TTS.

V. Application

With the growing availability of online services and ubiquitous computing capabilities it becomes easier to develop systems that can support people in changing their behaviour or lifestyle [11]. Sensor data and context information is available anywhere. Many of these systems support people in their daily life by providing support by means of a human or digital coach. These systems can support users in coping with chronic diseases like COPD [12] and diabetes, but also to be more physically active [13] [14]. Persuasive systems [15], and especially behaviour change support systems, are information systems designed to form, alter or reinforce attitudes, behaviours or an act of complying without using deception, coercion or inducements [16].

In the EU Artemis project Smarcos we developed a personal digital health coach that supports users in attaining a healthy lifestyle by giving timely, context-aware feedback about daily activities through a range of interconnected devices. The two targeted user groups of the coaching system are office workers and diabetes type II patients. Office workers will receive feedback about their physical activity level, while diabetes type II patients also receive feedback about their medication intake. Physical activity is measured by a 3D accelerometer and medication intake is tracked by a smart pill dispenser. The pill dispenser uses the mobile network to connect to the internet. The system is context-aware and multi-device which means that the (digital) coach can support the users in various contexts and on different devices. GPS information is provided by the mobile phone of the user. The system sends feedback to the mobile phone of the user (iOS or Android), their laptop or PC, and their television.

All input and output devices are connected to the Smarcos cloud. User profiles and preferences, contextual information and sensor data is uploaded to the cloud and stored in a central database. The digital coach continuously keeps track of all user data and contains coaching rules. When the coach receives a trigger it starts to evaluate the coaching rules. When one of the rules is true, it will select a suitable message from the coaching content database and send the message to the user through one of the available output devices and through one of the available modalities. Feedback can be presented using different output modalities Feedback can be sent as a text message, can be presented in a graph or can be presented by animated spoken interaction with an ECA.

Personalisation of the user interface by means of ECA may affect the effectiveness of the behaviour change program and the user experience. Results from other studies indicate that the use of an ECA in a persuasive system has a positive effect on how the feedback is received by the user and on the results of the coaching program [17], [18], [19].

A first user evaluation with a basic version of the Smarcos personal digital health coach compared two alternatives for providing digital coaching to users of a physical activity promotion service. Participants in the study (n=15) received personalised feedback on their physical activity levels for a period of six weeks. Feedback was provided weekly either by e-mail or through an embodied conversational agent. The messages by the ECA were prerecorded video messges. Users' perceptions of the digital coaching was assessed by means of validated questionnaires after three weeks and at the end of the study. Results show significantly higher attractiveness, intelligence and perceived quality of coaching for the ECA coach.

VI. Conclusion and future work

To take full advantage of the known benefits of personification of the user interface of service systems, a mobile platform that is able to present embodied conversation agents in mobile applications is presented. The platform makes use of the Elckerlyc system. Because it is too heavy to render realistic 3D virtual humans on mobile devices a light-weight PictureEngine was developed. The PictureEngine makes it possible to use the Elckerlyc system on the Android platform and generate realtime animations of embodied conversational agents. The PictureEngine is used in the Smarcos coaching application as a mobile embodied coach.

Long term user evaluations with the Smarcos coaching platform, including the mobile emdodied coach, are planned to investigate the effects of personified coaching feedback on user experience, quality of coaching and effectiveness of the coaching program.

During a six weeks user experiment 80 participants will use the Smarcos coaching platform for physical activity. Every participant has to meet the daily personal activity goal. The participants will get feedback about their progress on their mobile phone. The system will send feedback to the users with reminders to be more physically active or to upload the activity data, motivational message, tips and a weekly overview of their coaching program.

The design of the user evaluation will be a between subject design. The participants will be divided into two groups of 40 participants each. One group will receive the feedback presented in text, while the other group will receive the feedback presented by the ECA.

The effects on the coaching program of the way of presenting the feedback will be measured by means of questionnaires and by observation of the performance of the participants. During the experiments the progress towards the goal of the user, the actual amount of physical activity and the times the participant uploads their activity data is logged by the system.

At the start, at the end and halfway the experiment the participants will be asked to complete a questionnaire to measure the user experience, the credibility towards the system and the quality of coaching. User experience will be measured by the AttrakDiff2 questionnaire [20], credibility will be measured by the Source Credibility Questionnaire [21] and the quality of coaching will be measures by the Quality of Coaching questionnaire [22].

Although it is shown that the PictureEngine can run on mobile Android devices it would be worth exploring options for using a different TTS system in the future. This would allow the application to regain the speech-related functionality that is currently unavailable on Android, such as synchronization within utterances and viseme-based lipsync. A next step in the development of the PictureEngine is looking for techniques to allow small movements by the ECA, such as nodding and shaking of the head.

Acknowledgment

This work was funded by the European Commission, within the framework of the ARTEMIS JU SP8 SMARCOS project 100249 - (www.smarcos-project.eu).

References

[1] T. Bickmore, D. Mauer, F. Crespo, and T. Brown, "Persuasion, task interruption and health regimen adherence," in *Proceedings of the 2nd international conference on Persuasive technology*, ser. PERSUASIVE'07. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 1–11.

[2] M. Argyle, *Bodily Communication*, MethuenEditors, Ed. Methuen, 1988, vol. 2nd.

[3] D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt, “Multi-party, multi-issue, multi-strategy negotiation for multimodal virtual agents,” in *Intelligent Virtual Agents*, ser. Lecture Notes in Computer Science, H. Prendinger, J. Lester, and M. Ishizuka, Eds. Springer Berlin / Heidelberg, 2008, vol. 5208, pp. 117–130.

[4] H. van Welbergen, D. Reidsma, Z. M. Ruttkay, and J. Zwiers, “Elckerlyc - a BML realizer for continuous, multimodal interaction with a virtual human,” *Journal on Multimodal User Interfaces*, vol. 3, no. 4, pp. 271–284, August 2010.

[5] E. Bevacqua, K. Prepin, E. de Sevin, R. Niewiadomski, and C. Pelachaud, “Reactive behaviors in SAIBA architecture,” in *Proc. of 8th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, S. Decker, Sichman and Castelfranchi, Eds., May 2009.

[6] H. Vilhjalmsson, N. Cantelmo, J. Cassell, N. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A. Marshall, C. Pelachaud, Z. Ruttkay, K. Thórisson, H. van Welbergen, and R. van der Werf, “The behavior markup language: Recent developments and challenges,” in *Proceedings of the 7th International Conference on Intelligent Virtual Agents*, C. Pelachaud, J.-C. Martin, E. André, G. Collet, K. Karpouzis, and a. p. D. Pelé, pages=90–111, Eds., 2007.

[7] A. Heloir and M. Kipp, “Real-time animation of interactive agents: specifciation and realization,” *Applied Artificial Intelligence*, vol. 24(6), pp. 510–529, 2010.

[8] E. Bevacqua, M. Mancini, R. Niewiadomski, and C. Pelachaud, “An expressive ECA showing complex emotions,” in *AISB’07 Annual convention, workshop “Language, Speech and Gesture for Expressive Characters”*, April 2007, pp. 208–216.

[9] D. Reidsma, I. de Kok, D. Neiberg, S. Pammi, B. van Straalen, K. Truong, and H. van Welbergen, “Continuous interaction with a virtual human,” *Journal on Multimodal User Interfaces*, vol. 4, no. 2, pp. 97–118, 2011.

[10] D. Reidsma and H. v. Welbergen, “Elckerlyc in practice on the integration of a BML realizer in real applications,” in *Proc. of Intetain 2011*, 2011.

[11] M. Kasza, V. Szücs, A. Végh, and T. Török, “Passive vs. active measurement: The role of smart sensors,” in *UBICOMM 2011, The Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies*, 2011, pp. 333 – 337.

[12] H. op den Akker, V. Jones, and H. Hermens, “Predicting feedback compliance in a teletreatment application,” in *Proceedings of ISABEL 2010: the 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, 2010.

[13] O. Ståhl, B. Gambäck, M. Turunen, and J. Hakulinen, “A mobile health and fitness companion demonstrator,” in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Demonstrations Session*, ser. EACL ’09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 65–68.

[14] G. Geleijnse, A. van Halteren, and J. Diekhoff, “Towards a mobile application to create sedentary awareness,” in *In Proceedings of the 2nd Int. Workshop on Persuasion, Influence, Nudge Coercion Through Mobile Devices (PINC2011)*, vol. 722, 2011, pp. 90–111.

[15] B. Fogg, *Persuasive Technology. Using computers to change what we think and do*. Morgan Kaufmann, 2003.

[16] H. Oinas-Kukkonen, “Behavior change support systems: A research model and agenda,” in *Persuasive Technology*, ser. Lecture Notes in Computer Science, T. Ploug, P. Hasle, and H. Oinas-Kukkonen, Eds. Springer Berlin / Heidelberg, 2010, vol. 6137, pp. 4–14.

[17] O. A. B. Henkemans, P. van der Boog, J. Lindenberg, C. van der Mast, M. Neerincx, and B. J. H. M. Zwetsloot-Schonk, “An online lifestyle diary with a persuasive computer assistant providing feedback on self-management,” *Technology & Health Care*, vol. 17, p. 253–257, 2009.

[18] D. Schulman and T. W. Bickmore, “Persuading users through counseling dialogue with a conversational agent,” in *Persuasive ’09: Proceedings of the 4th International Conference on Persuasive Technology*. New York, NY, USA: ACM, 2009, pp. 1–8.

[19] D. C. Berry, L. T. Butler, and F. de Rosis, “Evaluating a realistic agent in an advice-giving task,” *Int. J. Hum.-Comput. Stud.*, vol. 63, no. 3, pp. 304–327, 2005.

[20] M. Hassenzahl, “Funology,” M. A. Blythe, K. Overbeeke, A. F. Monk, and P. C. Wright, Eds. Norwell, MA, USA: Kluwer Academic Publishers, 2004, ch. The thing and I: understanding the relationship between user and product, pp. 31–42.

[21] J. C. McCroskey, “Scales for the measurement of ethos,” *Speech Monographs*, vol. 33, no. 1, pp. 65–72, 1966.

[22] J. Cote, J. Yardley, J. Hay, W. Sedgwick, and J. Baker, “An exploratory examination of the coaching behaviour scale for sport,” *AVANTE*, vol. 5, pp. 82–92, 1999.

User-Centric Personalization and Autonomous Reconfiguration Across Ubiquitous Computing Environments

Kevin O' Mahony, Jian Liang, Kieran Delaney
NIMBUS Centre
Cork Institute of Technology
Cork, Ireland
kevin.omahony, jian.liang, kieran.delaney@cit.ie

Abstract—**In the era of Ubiquitous Computing (UbiComp), during our typical daily routines we may encounter multiple, shared and heterogeneous UbiComp environments across various locations. As these environments are meant to be shared between multiple users, interaction control methods (gestural, touch, voice commands, etc.) and the contexts (light, temperature, sound, informational services, etc.) of the environment are not personalized for individual users naturally. Typically, users are required to manually configure interaction preferences and conditions each time they encounter UbiComp systems. This however, refers to a tedious and redundant reconfiguration procedure, which is against the concept of UbiComp. In this paper, we present our work targeting on improving the personalization and reconfiguration procedure. Firstly, a user-centric personalization approach is proposed for facilitating users in determining how an UbiComp environment should adapt to their own preferred configurations. Then, an autonomous reconfiguration procedure is proposed, ensuring that a user's preferences are maintained and accessible across multiple ubiquitous computing environments seamlessly.**

Keywords-Ubiquitous Computing; personalization; reconfiguration; human computer interaction.

I. INTRODUCTION

Current approaches for personalization in Ubiquitous computing (UbiComp) environments are mostly based on determining a user's context in order to provide customized content or services. Dominated by developer-centric approaches, users are provided with content and services, which are statistically or commonly meant to be suitable for their current situation. However, a user's preferences can often conflict with these provided services. Therefore, even though applications may enable many proactive services, users still expect to personalize applications further to suit their own particular preferences [1, 2, 3, 4]. Conventional developer-centric approaches towards personalization in UbiComp environments constrain users in taking an active role in determining their own preferences for interaction controls and conditions. We believe that in UbiComp environments it is necessary that users should be supported in taking a dominant role in further personalizing how the environment needs to be interacted with and how the conditions of the environment should suit their preferences accordingly.

There is another issue that so far, personalization approaches have been quite isolated, and demonstrated partially without considering maintaining the consistency of personalized interaction controls and conditions across multiple environments. UbiComp envision a world where implicit interactions between humans and computers naturally take place across multiple interfaces of multiple environments [5, 6]. As UbiComp environments proliferate in size and diversity across many locations, taking a holistic approach in designing personalization approaches turns out to be more and more important [7]. Ensuring personalized interaction control and condition preferences are consistent and memorable across multiple UbiComp environments are also reflecting the importance for adhering to existing usability design principles as founded from the HCI research domain [8, 9, 10, 11, 12]. Meanwhile, the usability standards need to be extended further and applied towards personalization approaches for UbiComp, as they will be integral for ensuring effective usability and user experience in the future. However, conventional personalization approaches demonstrated are impractical for users within a shared UbiComp environment. For instance, once a system is personalized, user's preferences are mostly stored with it, which potentially may be shared and used by many other users, to whom the UbiComp system may have to be manually reconfigured each time. Many other issues have been highlighted revealing how the identification and ownership of users' profiles may become susceptible to privacy and security concerns as users' preferences are left stored on UbiComp systems.

In this work, we primarily focus on a user-centric personalization approach for personalizing and reconfiguring conditions and interaction controls across UbiComp environments. The proposed approach indicates a way where personalization is determined by the user, who takes an active role in defining how UbiComp environments need to be interacted with, and how the environments should be configured based on their own preferences. An autonomous reconfiguration procedure has also been proposed as to seamlessly maintain consistency of users' personal preferences across multiple shared UbiComp environments. Environments are automatically reconfigured according to current users' preferences each time they encounter them. This ensures that a user does not have to explicitly and manually reset each system each time. To demonstrate more practically, we have developed a testbed with a personal wearable device, which stores a user's preferences and provides them to the UbiComp environments.

ISBN: 978-1-61208-236-3

Approaches suitable to resolve privacy and security concerns, where UbiComp environments resources and users profiles are protected from unauthorized access, have been considered as separated research topics in regards to the focus of our work. The provisioning of user profiles between the UbiComp environments and the user is only available when authentication is firstly accomplished. The personal device for keeping user's preferences can be integrated with many authentication subsystems, such as biometric checking, to prevent unauthorized usage. Therefore, a user is not required to repeat authentication through the more typical approaches. Another consideration we have made is the management of concurrency issues of an UbiComp environment, which can be modeled separately and solved systematically on the backend so that end users are protected from the complexity of potential interactivities. Finally, it is proposed that systematically, ownership and privacy of user's preferences are protected; as user's "profiles" are only shared during the session when the user is active in the UbiComp environment.

The rest of the paper is organized as follows: we position our work against other related research in Section II. Then, the user-centric methodology is presented in Section III. Technical implementation details are explained in Section IV. To demonstrate, a user case scenario is described in Section V, where evaluations from two experiments are presented. At the end of the paper, our findings and future work are discussed in Section VI.

II. Related Work

In this work, we primarily focus on user-centric personalization and autonomous reconfiguration of interaction control and condition preferences across UbiComp environments. This is distinctive from most personalization research in the UbiComp domain, which has concentrated primarily on personalization of information (graphical user interfaces and informational content) displayed on interface screens.

Research has explored how user-centric personalization approaches can support users in configuring user preferences further due to the versatile characteristics of UbiComp. Such work from Atia and Tanaka [13] has demonstrated this where context parameters affect gesture-based interaction in UbiComp environments. Their experiments show positive results of how user interaction performance and experience is impacted when context parameters in UbiComp change. They also highlighted that people like to customize hand gestures accordingly when context parameters change in a situation. Kawsar and Nakajima have presented Persona [4], which is a portable tool that enables existing proactive systems applications to become extended to support multimodal personalization. These user centric approaches have illustrated the core concept of personalization and its necessity. However, once a user specifies one personalized configuration on one system, it is typically inaccessible across multiple systems, as the personalized settings are only stored on local system repositories. The user, therefore, has to manually personalize each system they encounter, new or shared interactive systems. Protecting ownership and access to personalized settings are in potential risk in this way also.

We focus on related research, which advocate user-centric personalization approaches, which address the challenges users experience interacting across multiple UbiComp environments. A key challenge addressed in the related research focuses on how the UbiComp environments should also automatically adapt enabling users to access their preferences across multiple UbiComp environments. However, to adhere to the ethos of UbiComp this transition should happen seamlessly without interrupting users natural activities within environments. Adapting UbiComp environments is primarily based on the provisioning of user preferences (user profiles) to UbiComp systems. As mobility support is fundamental to the principles of UbiComp, approaches towards provisioning and accessing user profiles across multiple UbiComp systems has become an issue for contention. Some approaches [4, 13] have stored user profiles on individual stationary systems; however, as mentioned previously user profiles become inaccessible and more vulnerable towards privacy and security concerns in this way. Other approaches propose storing user profiles on remote centralized repositories where access is dependent on Internet connectivity [14]. Microsoft has submitted a patent titled "Gesture Personalization and Profile Roaming" [15] which details how Microsoft's Kinect system learns a user's movements and stores this information as personalized gestures in a roaming profile. This ensures that when a user is using any Kinect System regardless of locations, personalized gestures can be accessed remotely over a network (Gesture Roaming). Since the user's personalized gestures are recognizable to the system, the system can perform more responsively and accurately. However, gesture profiles are inaccessible if there is no connection to the remote profile repository.

More related approaches to our work focuses on storing user preferences as user profiles on users' personal mobile devices. Personal devices such as mobile phones are now equipped with supporting technology and carried with the user, therefore it seems like a suitable approach for provisioning user profiles to UbiComp systems practically. Such work by [7, 14, 16] has used personal mobile phones, where users profiles are shared between the user's mobile phone and a UbiComp System. In our work, we have considered the impracticalities of this approach from a usability and user experience perspective. The mobile phone is used as a peripheral device as it is carried and held by the user for the provisioning of user profiles to UbiComp systems. The natural interaction flow is therefore interrupted as the user is required to engage with the mobile phone device in order for the UbiComp system to be reconfigured suiting users preferences, however this apporach conflicts with the principles of UbiComp which seek to maintain seamless and natural interaction. Also, a mobile device itself is a peripheral device to a user; it is susceptible to risk like other peripherals, like being stolen and unintentional exposure of private information to others. Therefore, we considered an autonomous reconfiguration approach, which would be more seamless, and natural, unhindering implicit interaction between the user and the UbiComp environment.

A major aspect of UbiComp is mobility, as users naturally interact with a variety of shared systems across many locations, user profiles should be seamlessly and safely accessable where UbiComp systems they encounter do not require explicit manual reconfiguration by the user. In our work, we believe that a user's profiles should be stored, maintained and utilized with minimum effort from the user. For this work, a specifically designed device is dedicated to the role of keeping profiles and exchanging profiles with the UbiComp environments when user encounters them. Ideally, any carriable devices, such as mobile and PDA can fulfill the tasks as long as they are not interrupting user's natural interaction with environments.

III. User-Centric Methodology

A user-centric methodology describes standardized approaches for user-centric personalization and autonomous reconfiguration of interaction control and conditions across UbiComp environments.

A. *User-Centric Personalization*

User-centric personalization is embodied through enabling users to determine themselves how conditions such as (light, temperature, sound, information services, etc.) in a UbiComp environment, should respond concurring with their own personal preferences for particular contexts. Also, through enabling users to map interaction control techniques (gestural, touch, voice commands, etc.) with actions to be performed, conditions such as (light, temperature, sound, information services, etc.) can be explicitly controlled based on users interaction control preferences.

A primary form factor intrinsic to supporting our user-centric personalization approach will be a wearable NFC device, which can be carried by a user. This device will be used for bridging the gap between the user and the UbiComp environment for the purpose of personalization of environmental conditions and control interactions. We apply a user-centric personalization procedure where the user is provided with a personalization management capability through an application deployed on a touchable interface (NFC enabled Smartphone).

1) *User Personalization Profile (UP Profile)*

Through the personalization management application, the user can create new or update existing "User Personalization Profiles" for UbiComp environments and particular contexts. A "UP Profile" consists of two parts, the User Profile and the Personalization Profile, which are shown in Figure 1. A User Profile consists of a unique ID, user's name/title, location, visual profile and access control permissions. A Personalization Profile consists of defined parameters of data that pertains to users personal preferences for environmental conditions (light, temperature, sound, etc.) and interaction controls (interaction/action pairs). Individual "UP Profiles" are created to determine how conditions in the UbiComp environment should react for particular contexts such as working, relaxation or sleeping. For example, an instance of "UP Profile"-"working profile"-is created through the personalization manager of the application, parameters such as light intensity, light colour, which are comfortable for user's working context are configured and documented.

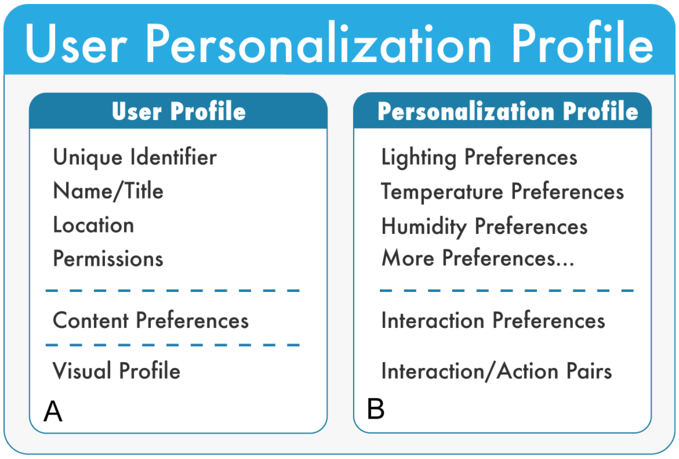


Figure 1. A User Personalization Profile

2) *Interaction/Action Mapping*

Although not practically implemented in this work, our user-centric personalization procedure incorporates a method of mapping interaction techniques with actions to be initiated. To support consistency and memorability of interaction control across multiple shared UbiComp environments our user-centric personalization procedure describes a method of pairing interaction techniques with actions, which will be practically implemented in our future work. Through this method users can create logical mappings between interaction techniques and actions to be initiated. This method affords users to make more cohesive pairings, which are compatible with their own personal habitual interpretation of interactions for controlling UbiComp environments [8, 9, 10, 12]. Through the personalization management application they can select an explicit interaction technique from a menu and map it with an action, for example (light on/off). Once the user explicitly performs the selected interaction technique in the UbiComp environment the selected mapped action is initiated. For example, a user could select a gesture technique such as a hand wave up/down motion from the personalization manager application menu and pair it with the light intensity increase/decrease action.

Once a user has completed configuring their preferences for interaction control and conditions these are then saved as a “UP Profile” called ‘working profile” by the user. This “working profile’ is locally stored on the user’s wearable device, where it can be automatically queried in the UbiComp environment again and in the future and across other shared UbiComp environments, which are capable and compatible to load this format of profile.

B. *Autonomous Reconfiguration*

Autonomous Reconfiguration describes a procedure following which multiple shared UbiComp environments automatically reconfigure interaction control methods and conditions to suit a user’s preferences. The activity logic of our approach for Autonomous Reconfiguration is illustrated in Figure 2.

Algorithm 1 Workflow of User-centric personalization

```
Require: User A has a carry-on device D which is able to
  store profiles P(s). UbiComp U is able to be configured.
  if A has a profile then
    U loads a configuration C1 that is stored as a P1 from
    D
  else
    U prompt to configure its contexts
    A specifies the configuration C2
  end if
  U adjusts the contexts to the configuration
  if Update the C1/C2 then
    U prompts the save option
    if A wants to save then
      U prompts to confirm to save
      Updated C1/C2 is saved as a P2 on D
    end if
  end if
  if A leaves U then
    U removes C1/C2
  end if
```

Figure 2. Activity Logic

Through wireless local communications, “UP Profiles” are uploaded from the user’s carrier device to the UbiComp system. Existing configurations on the UbiComp system are substituted with user’s “UP Profile”. Based on the configuration saved as the “UP Profile” the UbiComp environment responds accordingly adapting to suit the user. As the “UP Profile” is stored on the device which is constantly with the user, it is roaming together with him/her; then connectivity and the provisioning of “UP Profiles” to UbiComp systems depends on users physical location and movement as they share their “UP Profile” with UbiComp systems. We call this “User Profile Roaming”, due to the profile being linked with the user. Subsequently, a “UP Profile” is readily available as the user encounters multiple shared UbiComp environments across other locations. Therefore, the user does not have to manually reconfigure shared UbiComp systems each time they encounter them to suit their preferences for interaction control and conditions. Also, to ensure that ownership and privacy concerns are maintained, “UP Profiles” are only shared during the session when the user is active in the UbiComp environment. Once the user has completed the session within the UbiComp environment, “UP Profiles” are automatically cleared from the UbiComp System.

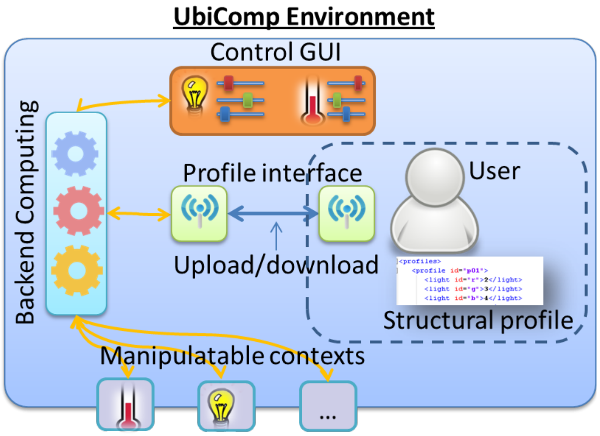
IV. Implementation

There are various ways to practically build up a prototype system which can realise the proposed methodology. In this section, we present both the structure and the implementation components of our testbed system.

A. *System Architecture*

The options for building a UbiComp system are only limited by people’s imagination nowadays. System developers are provided with an abundant of technologies and also technical commodities which aim to be integrated into something useful requiring minimum effort. Therefore, simplicity is the main characteristic of the system structure we have decided. As described in Section V, two physical small scale models representative of two UbiComp environments are transformed as our UbiComp testbed where a few major components are embedded as catagorised below and illustrated in Figure 3.

- Profile interface, through which user’s preferred settings, stored as a structural profile file, are uploaded to the environment; also when user adjusts the settings and decides to keep it for a longer term, the updated settings are downloaded to the user through this interface. Different communication methods can be applied in practical implementation.
- Control GUI is a visual interface that enables the user to explicitly manage controlling the settings of the encountered environment.
- Manipulatable contexts are typical adjustable utilities, which normally function differently according to various users, in the environment, such as lights and heating.
- Backend computing, which is invisible to users, orchestrates all other components, such as loading current user’s profile, displaying the information on the control GUI and interpreting the user profile into manipulation commands to the contexts. Linkages can be both wired and wireless.
- Structural profile is the file stored on the user carry-on device, recording user’s preferences of the context setting.



```
<profiles>
   <profile id="p01">
      <light id="r">2</light>
      <light id="g">3</light>
      <light id="b">4</light>
```

Figure 3. System structure of an UbiComp system

B. Testbed implementation

In building our testbed, we choose LAMP (Linux, Apache, MySQL and PHP) based web services and python based interface program as the backend computing. NFC (Near Field Communication) is selected as the communication method between profile interfaces; one interface is embedded within the UbiComp environment, while the other is on the user's carry-on wearable device. Lighting is set as the demonstrative manipulatable context. And wired connection is used to deliver control signals to the individual lighting device. Control GUI is implemented with a game engine – Unity3D – which generates a virtual 3D indoor environment with intuitive control functions that enable users to adjust the context lighting and save their preferences. "User Personalization Profiles" are the structural profile, where components of the "User Personalization Profile" are defined in an XML schema.

V. Case Study

In this section, we firstly describe a possible user-case scenario, which we use to indicate how our methodology could be applied more practically across remote UbiComp environments in the future. We then validate our methodology with two sets of experiments; firstly the user-centric personalization experiment; and secondly, the autonomous reconfiguration experiment. The experiment testbed setup is described also and illustrated in Figure 4.

A. User Case Scenario

"Mr. Jones's job requires him to travel abroad quite frequently. He spends much time staying and working from different offices and hotel rooms across many locations. When Mr. Jones is working or relaxing he prefers the ambient lighting conditions of his accommodation to suit his preferences for such contexts. He would prefer the ambient lighting conditions in all environments he stays in to automatically adapt to suit his preferences for working, relaxation and sleeping, rather than having to adjust all the lights manually each time."

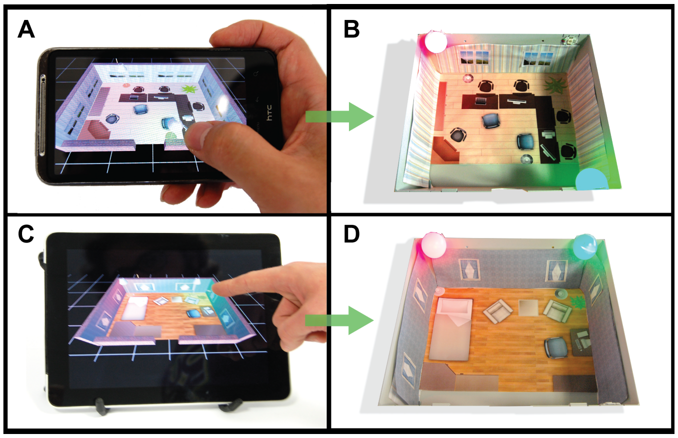


Figure 4. Experimental Testbed Setup

B. Experiment Testbed Setup

The possible user-case scenario described is demonstrated in a way more feasible for our experimentation purposes. The experiment testbed setup consists of two small physical scale models representative of two UbiComp environments. For descriptive purposes, the physical scale model as illustrated in Figure 4(B) is used as the first location representative of a user's local office environment, whereas the physical scale model illustrated in Figure 4(D) is used as the second location representative of a hotel room. In Figure 4(A, C) we show screenshots of the control GUIs taken from the personalization management application. Control GUIs provide virtual 3D environments indicating the users current physical environment, as to comprehend more meaning to the user. In both Figure 4(A) and Figure 4(C) the control GUI is deployed and accessible from touchscreen platforms, which are already part of both UbiComp environments.

C. User-Centric Personalization Experiment

To demonstrate the user-centric personalization procedure as shown in Figure 4(A, B), we firstly share a default UP Profile with the UbiComp System (Local Office), when there is a pairing between the wearable carrier device and the UbiComp system (Local Office), this is achieved through the profiling interface as illustrated in Figure 3. In our testbed setup, we consider the profiling interface to be already a part of the UbiComp system. In this experiment we firstly consider the user, Mr. Jones, to be initially carrying a default UP Profile, which has not been previously configured by him beforehand. When he encounters the first UbiComp System (Local Office), once a pairing between his wearable device and the UbiComp System (Local Office) is initiated, the default UP Profile is uploaded to the UbiComp System (Local Office). Once uploaded, the UbiComp environment's lighting conditions adapt according to the information stored as the default UP Profile. This is also comprehended in a meaningful way through a virtual 3D environment as displayed to Mr. Jones on the Control GUI, see Figure 4(A). The default

ISBN: 978-1-61208-236-3

lighting conditions may not be suitable towards his preferences; through the Control GUI interface he reconfigures the light intensity by directly selecting each light from the 3D virtual environment interface. User interface control buttons are mapped according to each light, which are used to increase/decrease the light intensity, see Figure 4(A). As described in the user-case scenario, Mr Jones is enabled to configure the light intensity to suit his contexts such as for working, relaxation or sleeping. Therefore, once satisfied with the light intensity he is then enabled to save these as UP Profiles where they can be updated, saved and uploaded to the his carrier device for future usage.

D. Autonomous Reconfiguration Experiment

To demonstrate autonomous reconfiguration, in our experimental testbed setup we use a second small-scale model representative of a remote hotel UbiComp environment, see Figure 4(D), as it is demonstrates how UP Profiles can be reused again to maintain consistency of Mr. Jones's preferences across multiple UbiComp environments. When there is a pairing between the user's wearable carrier device and the second UbiComp system (Hotel Room), Mr. Jones's working UP Profile as created previously is uploaded from the wearable device to the UbiComp system (Hotel Room). The light intensity in this second UbiComp environment automatically dimmers to the exact parameters matching the light intensity configured previously for a working context in the first environment (Local Office). This ensures that Mr Jones does not have to manually adjust the light intensity again to suit his working context, as his preferences for environmental conditions automatically remain consistent across multiple UbiComp environments. When he has completed interacting with an UbiComp environment, his personal UP Profiles are removed from the UbiComp system repository, as UP Profiles are only permanently stored on the user's wearable device.

VI. CONCLUSION

In this paper, we have presented a new methodology of user-centric personalization and autonomous reconfiguration across multiple shared ubiquitous computing environments. Through this methodology, it is emphasized that users should play a dominant role in deciding their own preferences for interaction control and the environmental context. Another aspect highlighted in this paper is the seamless maintenance of consistency of user experience across multiple UbiComp environments. Scenario based experiments have shown how the methodology is practically implemented and the effectiveness it brought into the real-life scenarios. For this stage of our work, room models have been used for simplicity in demonstrating the concept. In the following work, deployment within a real-life environment will be carried out, including optimization of GUI designs and reliability of the pairing procedure on the profile interface. More modalities of interaction controls (gestural, touch and voice) and more contexts (temperature, sound and humidity) will be included into the real-life deployment.

ACKNOWLEDGMENT

This work is sponsored through an Enterprise Ireland Project – eGo: easily connect the world (IR-2010-0005). Juan Francisco Martinez integrated the hardware.

REFERENCES

[1] D.A. Norman, Emotional Design, NY: Basic Books, 2005.

[2] M.R. Morris, J.O. Wobbrock and A.D. Wilson, "Understanding users' preferences for surface gestures," Proc. of Graphics Interface (GI '10), Canadian Information Processing Society, Toronto, Ont., Canada, pp. 261–268, 2010.

[3] J.O. Wobbrock, M.R. Morris and A.D. Wilson, "User-defined gestures for surface computing", Proc. of the 27th International Conference on Human Factors in Computing Systems, ACM, New York, NY, USA, pp. 1083–1092, 2009.

[4] F. Kawsar and T. Nakajima, "Persona: a portable tool for augmenting proactive applications with multimodal personalization support," Proc. of the 6th international conference on Mobile and ubiquitous multimedia. New York, NY, USA: ACM, 2007.

[5] M. Weiser, The Computer for the 21st Century, Scientific American, September 1991.

[6] D. Salber, A.D. Dey, and G.D. Abowd, "Ubiquitous Computing: Defining an HCI Research Agenda for an Emerging Interaction Paradigm, GVU Technical Report; GIT-GVU-98-01, Georgia Institute of Technology, 1998.

[7] D. Chalmers, M. Chalmers, J. Crowcroft, M. Kwiatkowska, R. Milner, E. O' Neill, T. Rodden, V. Sassone, and M.Sloman, "Ubiquitous Computing: Experience, design and Science", 2006.

[8] R.D. Vatavu, "Nomadic Gestures: A technique for reusing gesture commands for frequent ambient interactions," Journal of Ambient Intelligence and Smart Environments 4, pp. 79–93. 2012.

[9] D.A. Norman, Natural user interfaces are not natural, ACM Interactions 17(3), pp. 6–10, 2010.

[10] D.A. Norman and J. Nielsen, Gestural interfaces: A step backward in usability, ACM Interactions 17(5) (2010), 46–49. International Standards for HCI and Usability, 2012.

[11] N. Bevan, "International standards for HCI and usability," International Journal of Human-Computer Studies, v.55 n.4, pp. 533-552, October 2001.

[12] D. Saffer, In Interactive Gestures: Designing Gestural Interfaces, O Reilly Media, Inc, 1005 Gravenstei Highway North, Sebastopol, CA 95472, 2008.

[13] A. Atia, S. Takahashi, K. Misue, and J. Tanaka, "UbiGesture: Customizing and Profiling Hand Gestures in Ubiquitous Environments, " Proc. 13th International Conference on Human-Computer Interaction. Part II: Novel Interaction Methods and Techniques, Springer-Verlag, July. 2009.

[14] J. Bohn, User-Centric Dependability Concepts for Ubiquitous Computing. PhD thesis, No. 16653, ETH Zurich, Zurich, Switzerland, May 2006.

[15] Z. Zhang et al, "Gesture Personalization and Profile Roaming," US Patent 2011/0093820 A1, October 19th, 2009.

[16] D. Retkowitz, I. Armac, and M. Nagl: "Towards Mobility Support in Smart Environments," Proc. of the 21st International Conference on Software Engineering and Knowledge Engineering (SEKE 2009), Boston, Massachusetts, USA, pp. 603-608. Knowledge Systems Institute Graduate School, 3420 Main Street, Skokie, Illinois 60076, USA, 2009.

An Aspect-Based Resource Recommendation System for Smart Hotels

Aitor Almeida[1] , Eduardo Castillejo[1], Diego López-de-Ipiña[1], Marcos Sacristán[2], Javier Diego[3]

[1] DeustoTech –Deusto Institute of Technology, Universidad de Deusto
Avenida de las Universidades 24, 48007, Bilbao (Spain)
{aitor.almeida, eduardo.castillejo, dipina}@deusto.es
[2] Treelogic
Parque Tecnológico de Asturias, 30, E33428, Llanera (Spain)
marcos.sacristan@treelogic.com
[3] Logica
Avda. Manoteras, 32, Madrid 28050 (Spain)
javier.diego@logica.com

***Abstract*—The number of resources (services, data, multimedia content, etc) available in Smart Spaces can ver overwhelming. Finding the desired resource can be a tedious and difficult task. In order to solve this problem, Smart Spaces contain much information that can be employed to filter these resources. Using the user context-data available in Smart Spaces can help refining and enhancing the recommendation process, providing more relevant results. To help users finding the most suitable resource we have developed a recommendation system that takes into account both user and resource features and context data like the location or current activity. This recommendation system is flexible enough to be applied to different types of resources and domains. In this paper we describe the resource aspects identified to be used in the recommendation system and how they are combined to create a metric that allows us to select the best resource for each situation.**

Keywords-Smart environments; resource recommendation; context aware; accessibility.

I. INTRODUCTION

Spain is one of the top 5 tourism destinations along France, United States, China and Italy [1]. As a consequence, tourism is an important part of the Spanish economy [2]. In order to maintain this leading position, the hospitality sector must continue evolving and improving, including new technologies and enhancing the user experience. The objective of the THOFU (http://www.thofu.es/) project is to work on the technology that will enable us to create more intelligent and reactive hotels. One of the research areas within the project is the creation of a recommendation system that will enable us to provide the user with the most suitable resources for his current needs and context. In the project's vision resources are anything that a user can consume: apps, hotel services (both physical and virtual), multimedia content, data, etc. The smart hotel must be proactive, helping its users with their needs. In order to do this the smart hotel must know the user preferences, tastes and limitations. It must be capable of analyzing the different aspects that define a resource to offer the most appropriate one to the user. To do this, we have developed an aspect based resource recommendation system. To be able to do this recommendation we have identified the aspects of a resource that can be used to describe it in a smart environment. These aspects take into account both the resource and user features and the current context (as formulated in [3]). Our recommendation system approach has several advantages:

- It is applicable to all the resource types identified in the intelligent hotels domain: digital and physical services, multimedia content and data.
- We evaluate different aspects of the resource taking into account the characteristics of the content, the needs and capabilities of the user and data from the current context. This allows us to create a comprehensive picture of the current situation to recommend the most suitable resource.
- The process can be configured by modifying the weight of each individual aspect in the final metric. This allows us to adapt the recommendation system to specific domains.
- Our system not only analyzes the current situation of the user, it also takes into account what his next actions can be to anticipate future needs.

In this paper, we present the proposed system. In Section 2, we analyze the current state of the art in recommendation systems, in Section 3, we describe the system architecture, in Section 4, we describe some use cases of the proposed system and finally in Section 5, we discuss the conclusions and future work.

II. RELATED WORK

Since the mid-1990s, recommender systems have become an important research area attracting the attention of e-commerce companies. Amazon [4], Netflix (http://www.netflixprize.com/) and Yahoo! Music [5] are widespread examples on making recommendations to its users based on their tastes and previous purchases. Although these systems have evolved becoming more accurate, the main problem is still out there: to estimate the rating of an item which has not been seen by users. This estimation is usually based on the rest of items rated by the current user or on the ratings given by others where the rating pattern is

ISBN: 978-1-61208-236-3

similar to the user's one. Therefore, the problem consists on extrapolating somehow the utility function (which measures the usefulness of an item to a user) to the whole rating space. This utility function is represented by all the ratings made by the user. This way, recommendation engines have to be able to predict or estimate the ratings of the not yet rated items for users.

The research in this area has, as a result, a different classification based on the way item recommendation is made [6]:

- *Content-based:* recommendations are made just by looking in the history of the already rated items by the user.
- *Collaborative filtering:* past recommendations for users with the same preferences generate recommendations for the current user.
- *Hybrid techniques:* as a combination of content-based and collaborative recommendation approaches.

Content-based systems recommend items which are similar to those that a user rated positively in the past [7]. Shardanand *et al.* [8] state some of the problems of this approach, as the vagueness in the description of an item, which clearly affects the whole system. Items need to have enough descriptive features to enable the recommendation engine to recommend them accurately. The problem is that different items with the same features can be indistinguishable to the system.

Collaborative filtering techniques deal with the concept of similarity between users. The utility of an item is predicted by those items which have been rated by similar users. Sarwar *et al.* [9] defend this approach by defining collaborative filtering as the most successful recommendation technique to date. In [8] a personalized music recommendation system is presented, namely Ringo, which is a social information filtering system which purpose is to advise users about music albums they might be interested in. By building a profile for each user based on their ratings, it identifies similar users so that it can predict if a not yet rated artist/album may be to user's liking. LikeMinds [10] defines a closeness function based on the ratings for similar items from different users to estimate the rating of these items for a specific user. It considers a user which has not already rated the item and a so-called mentor who did it. Introducing two new concepts (horting and predictability) horting is a graph-based technique in which users are represented as nodes and the edges between them indicate their similarity (predictability) [11]. The idea is similar to nearest neighbor, but it differs from it as it explores transitive relationships between users who have rated the item in question and those who have not.

In order to reduce the limitations of previously reviewed methods, hybrid approaches combine both of them [12]. Others have introduced new concepts to this area, such as semantics [13] and context [13].

However, one of the most important improvements in the recommendation systems field is the definition of measures (or aspects) to describe the utility and relevance of the items. Aspects play an important role in data mining, regardless of the kind of patterns being mined [14]. Users' ratings are a good way to trace the interestingness and the relevance of items. Despite of the ratings, there are many measures which allow us to go into these items taking into account the use of them (their consumption) by the users. In other words, we look into the behavior of users for measuring their interestingness for these "items" (for now on we will refer items as resources). From our point of view a resource could be a product, an application or any kind of service (e.g., multimedia, news and weather or connectivity infrastructure services). We have studied several measures from the literature to evaluate those which best fit in our recommender system, such as minimality [15, 16], reliability [17], novelty [18], horting, predictability and closeness [14], and utility [9]. Location is one of the most important measures in many context-aware systems. Several authors has worked in location based recommendation systems [20][21][22]. In these papers, authors use location data captured with GPS and mobile devices to create timely and targeted recommendations for users.

III. System Architecture

To be able to evaluate the suitability of the resources for a given user we have identified a series of aspects that define the identified resource types:

- Physical services: Those services that are used in the real world (e.g the hotel restaurant, pool, gymnasium, etc)
- Virtual services: Services accessed using a device.
- Multimedia content (e.g. video, music, etc)
- Information (e.g. maps, news, etc)

These aspects must be generic enough to be able to use them to describe all the type of resource and expressive enough to capture the different facets of the resources. In the current implementation, we have considered four of them, but we discuss the other ones in the future work section. The four aspects that we currently take into account are the following:

1. Predictability;
2. Accessibility;
3. Relevancy;
4. Offensiveness;

Each one of those aspects is used in the calculation of the suitability value (see Formula 1). The weight of each aspect on the final value can be modified to better adapt the recommendation system to the specific domain of each smart environment. The values of the weights will depend on the requirements of the specific scenarios. For example, if a smart space has a considerable number of users with disabilities, the *accessibility* of the resources will be especially significant. On the other hand, if the scenario is composed by a single space, the *relevancy* will not be as

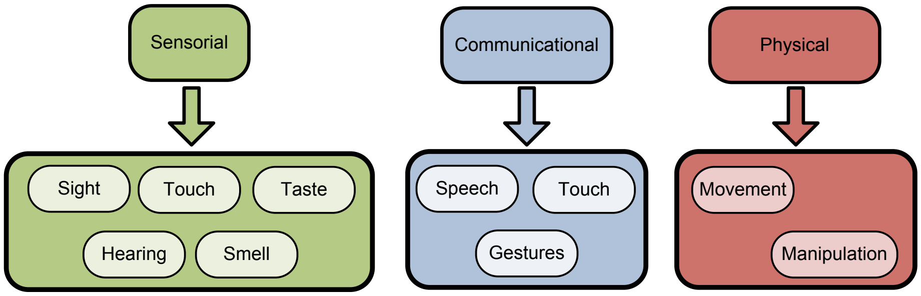


Figure 1. Taxonomy of the user abilities taken into account in the *accessibility* aspect. Disabilities are classified in this three categories.

important as the rest of the aspects. The suitability value is always personalized to a specific user and can change over the time along the preferences of the user.

$$M_{tot} = \sum \omega_i f_i \tag{1}$$

where:

- M_{tot} is the value of the suitability of each resource.
- ω_i is the weight for an aspect.
- f_i is the value of the aspect of a resource. The values of the aspects are normalized

A. *Predictability*

The first aspect we evaluate is the predictability. This aspect reflects how likely a resource is to be used based on the resources consumed previously. This likeliness is expressed as a probability value between 0 and 1. We use Markov Chains to create the model of the user's resource usage. This model allows us to ascertain patterns in the user behavior. E.g. When one user stays on the hotel his morning routine consists in using the "Press Digest" to recover the headlines of the day, the "Room Service" to order breakfast and the "Transport Service" to call a taxi. With the generated model, we will able to predict that after using the "Room Service" the most probable service to be consumed is the "Transport Service" (see Figure 2).

To build the transition matrix for the Markov Chains, we use the previous history of the user's resource consumption as the training set. This transition matrix can be retrained with the new data recovered from the user with each visit to the hotel, adapting itself to the changes in the user preferences. As we discuss in the future work section, one of the main problems with using Markov Chains is that we only take into account the last consumed resource to predict the next one due to the Markov Property.

B. *Accessibility*

One of the most important aspects is the accessibility features of the resource. Users of intelligent environments possess a wide variety of abilities (sensorial, cognitive and so on) that must be taken into account to assess the suitability of the resources. Whatever the resource is, users must be able to consume it. We have used the user abilities taxonomy proposed in [19]. We have restricted the user abilities to three groups (see Figure 1):

- *Sensorial abilities:* Those abilities related to the user input.
- *Communicational abilities:* Those abilities related to the user output.
- *Physical:* Those abilities related with the capability of the user to move his extremities.

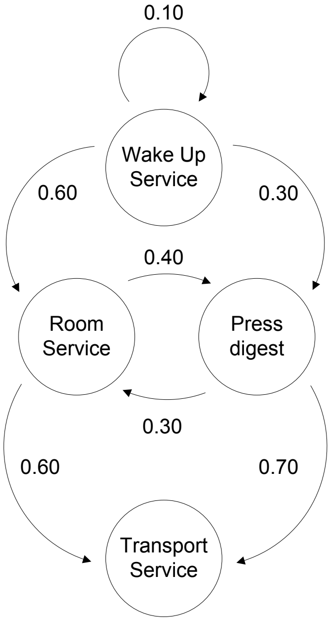


Figure 2. One of the Markov Chains created with the resource consumption data for the *predictability* aspect. Using the created model the recomender system can predict the likeness of one resource to be the next to be consumed.

Each resource has two types of abilities associated, the required and recommended user abilities. If the user does not have one of the required abilities the value of the aspect is automatically set to 0. This is done to reflect the fact that the user can not consume the resource, thus being completely useless for that user. If the user does not have a recommended ability the accessibility value receives a penalization (see Formula 2).

$$A_{acc} = 1 - \omega|Rec_{not}| \tag{2}$$

where:

- A_{acc} is the accessibility value for the resource.
- ω is the penalization weight.
- $|Rec_{not}|$ is the number of recommended abilities not met by the user.

C. Relevancy

This aspect measures the importance of a given resource [20] to the user's current context [3]. For example, a user jogging may be interested in the location of parks and running routes but a user having breakfast in the hotel may be interested instead in the public transports available in the city. One of the main problems we encountered evaluating this aspect was the selection of the context variables. The selected variables must be significant enough to be applicable to all the resource types described previously. We have identified three context variables that meet these requisites. We have analyzed the variables to identify the most common values within the Smart Hotel scenario. In this scenario the most important values are those that are closely related with the hotel, but it also takes into account those were the hotel can offer some service to the users:

1. *User location.* In the tourism domain, we have considered the following locations: client's room, hotel's lobby, hotel's restaurant, hotel's swimming pool, hotel's gymnasium and outside the hotel.
2. *Time of the day.* We have divided the day in twelve periods of two hours.
3. *Current activity.* In the tourism domain we have identified seven activities: sleeping, hygiene routine, eating, exercising, working, shopping and visiting tourist attractions.

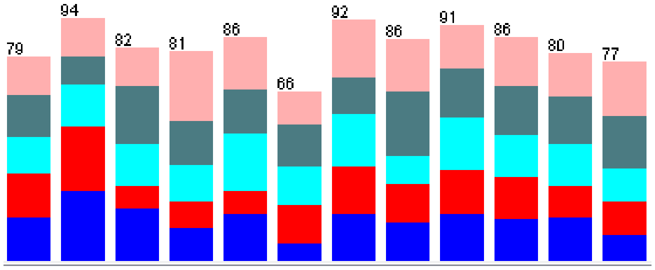


Figure 3. Distribution of the resource consumption in the different periods of time in the used training set for the *relevancy* aspect.

The context information is provided by other modules of the THOFU project that are out of the scope of this paper. Using the usage data recollected from the users we have trained a soft classifier that, given those three context variables, calculates the relevancy of a resource.

For the classifier we have used a nearest neighbor search. KNN (k-nearest neighbor) is a supervised (the training data is labelled), non-parametric (the model does not take a predetermined distribution form but it is in inferred from the data), lazy learning (there is no specific training phase) classification method. KNN assumes that the instances are distributed in a feature space. Since the instances exist in a multidimensional space, there is a computable distance between them. The most commonly used distance is the Euclidean distance. The algorithm takes a user-defined *k* constant. The instances are classified taken the *k* nearest training examples in the feature space.

To implement this classifier we have used the libraries included in the Weka framework [29]. We have used LinearNNSearch as the nearest neighbor search algorithm, with a *k* value of 3 and the Euclidean distance as the distance function.

D. Offensiveness

This aspect measures the suitability of a resource based on a rating system. We use the age categories (3, 7, 12, 16 and 18) and the content descriptions (violence, bad language, fear, sex, drugs, gambling, discrimination and online) developed for the PEGI (Pan European Game Information)[31] rating system. To evaluate it we use a similar system that the one used in Section 3.1 to calculate the accessibility, but taking the age categories as required constraints and the content descriptions as the recommended ones.

IV. Use Case

To better illustrate how the developed system works we will explain how the system works taking two different users as examples. The first user is a 27 year old male with a hearing impairment. The second one is a 6 year old child. The users have five resources available to them in this example: The wake up service (R1), the room service (R2), the press digest (R3), the multimedia system (R4) and the transport service (R5). For this example, the weights for the metric calculation are:

- *predictability* and *relevancy* have a weight of 1
- *accessibility* and *offensiveness* have a weigh of 0.5

We assume that both users are in their rooms and that the wake up service has just been activated by an alarm. The first user uses the Markov Chain model described in

Figure **2**. The wake up service and multimedia system both have hearing requirements, but offer alternative means to use them. The first user has not stated any content restriction. The results are shown in Table I.

TABLE I. RESULTS FOR THE FIRST USER

	Predictability	Accessibility	Offensiveness	Relevancy
R1	0.10	0.9	1	0.8
R2	0.60	1	1	0.7
R3	0.30	1	1	0.4
R4	0	0.9	1	0.2

R5	0	1	1	0.3

The second user uses the Markov Chain model described in

Figure 4. The user has not any disability, so every resource attains the maximum score in *accessibility*. The press digest has a minimum age category of 7 and it receives a score of 0 in *offensiveness*. The results are shown in Table II.

TABLE II. RESULTS FOR THE SECOND USER

	Predictability	**Accessibility**	**Offensiveness**	**Relevancy**
R1	0.45	1	1	0.2
R2	0.05	1	1	0.1
R3	0	1	0	0.1
R4	0.50	1	1	0.9
R5	0	1	1	0

Using the Formula 1 the recommended resource for the first user will be the room service (R2) in this scenario.

$$M_{tot} = 1 \times 0.6 + 0.5 \times 1 + 0.5 \times 1 + 1 \times 0.7 \qquad (3)$$

In the case of the second user, the selected resource will be the multimedia system (R4).

$$M_{tot} = 1 \times 0.5 + 0.5 \times 1 + 0.5 \times 1 + 1 \times 0.9 \qquad (3)$$

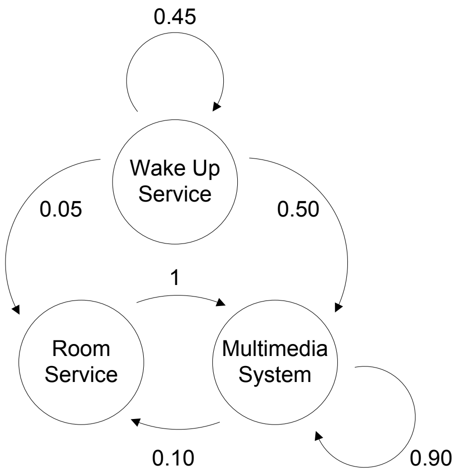


Figure 4. Markov Chain user for the second user

V. CONCLUSION AND FUTURE WORK

In this paper, we have described a resource recommendation mechanism for smart environments based on the evaluation of different aspects of the resources. Our approach provides several advantages:

- The proposed mechanism is generic enough that it can be applied to any type of resource (services, multimedia content, etc). To achieve this we have identified those aspects that are not specific for a given domain or resource.
- We take into account several aspects of a resource, providing a holistic approach to the problem of the resource recommendation.
- The importance of each individual aspect can be tailored for each domain and specific problem modifying their weights in the metric. This allows us to adapt the mechanism to the requirements of specific smart spaces.

One of the problems identified in this approach is the use of Markov Chains to evaluate the predictability aspect. With the use of Markov Chains we only evaluate the current event and not the previous events that preceded it. In order to tackle this problem we plan to explore the use of time series to improve the forecasting algorithm.

We are also analyzing a more extensive set of aspects that will give us a better picture of the evaluated resources. We are currently studying the inclusion of the following aspects:

- *Timeliness* [24]: evaluates how up to date is the information of a resource.
- *Satisfaction* [25][26]: measures the opinion of the users about a resource.
- *Attention* [27][28]: The average number of interactions per time unit with a consumed resource.
- *Closeness* [11]: Evaluates what resources are consumed by similar users.

By adding these new aspects we aim to create more significant resource recommendations that meet better the user’s needs. Finally we would like to include in the context data information about the vagueness and uncertainty of the model. To do this we plan to use the ambiguity assessing techniques we described in [30]. This will allow us to model the context more realistically and will improve the overall preciseness of the system.

ACKNOWLEDGMENT

This work has been supported by project grant CEN-20101019 (THOFU), funded by the Spanish *Centro para el Desarrollo Tecnológico Industrial (CDTI)* and supported by the Spanish *Ministerio de Ciencia e Innovación*.

REFERENCES

[1] Interim Update. UNWTO World Tourism Barometer (UNWTO). URL: http://www.unwto.org /facts/eng/pdf/barometer/UNWTO_Barom11_iu_april_excerpt.pdf, retrieved: July, 2012.

[2] Economic report. Bank of Spain. URL: http://www.bde.es/informes/be/boleco/coye.pdf, retrieved: July, 2012.

[3] Dey, A.K. Understanding and using context. Personal and ubiquitous computing, vol 5, no. 1, pp 4-7, 2001.

[4] G. Linden, B. Smith, and J. York, “Amazon. com recommendations: Item-to-item collaborative filtering,” Internet Computing, IEEE, vol. 7, no. 1, pp. 76–80, 2003.

[5] P. L. Chen *et al*., “A Linear Ensemble of Individual and Blended Models for Music Rating Prediction.”. KDDCup 2011 Workshop, 2011.

[6] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible

extensions," Knowledge and Data Engineering, IEEE Transactions on, vol. 17, no. 6, pp. 734–749, 2005.

[7] D. M. Pennock, E. Horvitz, S. Lawrence, and C. L. Giles, "Collaborative filtering by personality diagnosis: A hybrid memory- and model-based approach," in Proceedings of the 16th conference on uncertainty in artificial intelligence, 2000, pp. 473–480.

[8] U. Shardanand and P. Maes, "Social Information Filtering: Algorithms for Automating 'Word of Mouth'," 1995.

[9] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in Proceedings of the 10th international conference on World Wide Web, 2001, pp. 285–295.

[10] D. Greening, "Building consumer trust with accurate product recommendations," LikeMinds White Paper LMWSWP-210-6966, 1997.

[11] C. C. Aggarwal, J. L. Wolf, K. L. Wu, and P. S. Yu, "Horting hatches an egg: A new graph-theoretic approach to collaborative filtering," in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 201–212.

[12] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combining content-based and collaborative filters in an online newspaper," in Proceedings of ACM SIGIR Workshop on Recommender Systems, 1999, p. 60.

[13] S. Kim and J. Kwon, "Effective context-aware recommendation on the semantic web," International Journal of Computer Science and Network Security, vol. 7, no. 8, pp. 154–159, 2007.

[14] L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," ACM Computing Surveys (CSUR), vol. 38, no. 3, p. 9, 2006.

[15] B. Padmanabhan and A. Tuzhilin, "Small is beautiful: discovering the minimal set of unexpected patterns," in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 2000, pp. 54–63.

[16] Y. Bastide, N. Pasquier, R. Taouil, G. Stumme, and L. Lakhal, "Mining minimal non-redundant association rules using frequent closed itemsets," Computational Logic—CL 2000, pp. 972–986, 2000.

[17] P. N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 32–41.

[18] S. Sahar, "Interestingness via what is not interesting," in Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 1999, pp. 332–336.

[19] A. Almeida, P. Orduña, E. Castillejo, D. López-de-Ipiña, M. Sacristán. Imhotep: an approach to user and device conscious mobile applications. Personal and Ubiquitous Computing. vol 15, no. 4, pp 419–429 January 2011.

[20] V.W. Zheng. Y .Zheng, X. Xie, Q. Yang. Collaborative location and activity recommendations with gps history data. Proceedings of the 19th international conference on World Wide Web. 2010.

[21] Y. Zheng, L. Zhang, Z. Ma, X. Xie and W.Y. Ma. Recommending friends and locations based on individual location history. ACM Transactions on the Web, vol 5, no. 1, pp 5. 2011.

[22] V.W. Zheng, B. Cao, Y. Zheng, X. Xie and Q. Yang. Collaborative filtering meets mobile recommendation: A user-centered approach. Proceedings of the 24th AAAI Conference on Artificial Intelligence. 2010

[23] Leo Pipino, Yang Lee, and Richard Wang. Data Quality Assessment. Communications of the ACM, 45:211-218, 2002. PLW02

[24] Ballou, D.P., Wang, R.Y., Pazer, H. and Tayi, G.K. Modeling information manufacturing systems to determine information product quality. Management Science 44, 4 (1998),462–484.

[25] Bo Pang, Lillian Lee. Opinion Mining and Sentiment Analysis. Journal of Foundations and Trends in Information Retrieval vol 2, no. 1-2, 2008.

[26] Beatson, A. and Coote, L.V. and Rudd, J.M. Determining consumer satisfaction and commitment through self-service technology and personal service usage. Journal of Marketing Management, vol 22, no. 7-8, pp 853-882, 2006.

[27] S. Kim and J. Kwon, "Attention-Based Information Composition for Multicontext-Aware Recommendation in Ubiquitous Computing," Smart Sensing and Context, pp. 230–233, 2006.

[28] T. H. Davenport, "May we have your attention, please?," Ubiquity, vol. 2001, no. June, p. 3, 2001.

[29] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, Issue 1. 2009

[30] A. Almeida and D. López-de-Ipiña. Assessing Ambiguity of Context Data in Intelligent Environments: Towards a More Reliable Context Managing System. Sensors. vol 12, no. 4, pp 4934-4951. April 2012.

[31] Pan European Game Information (PEGI). Available online: http://www.pegi.info/, retrieved: July, 2012.

A Metadata Monitoring System for Ubiquitous Computing

Caio Batista, Gustavo Alves, Everton Cavalcante,
Frederico Lopes, Thais Batista
UFRN – Federal University of Rio Grande do Norte
Natal, Brazil
{caiosergiobatista, gustavoalvescc, evertonranielly,
fred.lopes}@gmail.com, thais@ufrnet.br

Flávia C. Delicato, Paulo F. Pires
UFRJ – Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
{fdelicato, paulo.f.pires}@gmail.com

Abstract—**In the highly dynamic context of ubiquitous systems, applications need to be continuously aware of QoS and QoC metadata to ensure their required level of quality. We present *QoMonitor*, a metadata monitoring system that receives synchronous and asynchronous requests from clients (a middleware system that supports ubiquitous applications), recovers metadata from several context providers, and sends them to the clients. We also present an evaluation of *QoMonitor* under a quantitative perspective, which aims to address the time for assessing QoS and QoC parameters and the time to completely reply to synchronous and asynchronous requests in the context of a health care application. The proposed monitoring system enables ubiquitous applications to focus on addressing the business requirements of the application and abstract away the burden of dealing with the complexities related to synchronous and asynchronous metadata monitoring.**

Keywords – metadata; monitoring; Ubiquitous Computing; health care application.

I. Introduction

Ubiquitous Computing [1] uses a variety of devices, sensors and networks to form a distributed, highly heterogeneous environment integrated to daily activities of users. Typically, ubiquitous applications are composed of *services* and use *context information* from several sources to perform their tasks. In this scenario in which applications encompass contextual data and services from different sources, it is essential to know the quality of the provided information and services so that applications can use those that satisfy their requirements. Therefore, the selection of the proper services among those provided by several available providers is performed according to the quality of context information, called *Quality of Context (QoC)* [2] and/or the quality of the provided services, called *Quality of Service (QoS*). During the execution of the applications, it is also necessary to ensure that services and context information continue to satisfy the QoS/QoC application requirements.

Both QoS and QoC quality parameters are typically described by *metadata*, which contain information about observable variables regarding services and/or context, such as resolution, precision, and freshness, for QoC, and error rate, uptime, and response time, for QoS. Ubiquitous applications are inherently dynamic since they use: (i) mobile devices, which can often be or not be in the area covered by a given network; (ii) wireless connections, which are subjected to interruptions and fluctuations in the intensity of the transmitted signal, and; (iii) physical parameters, such as temperature, pressure, location, which can frequently change. In this highly dynamic context, applications need to be continuously aware of QoS and QoC metadata to ensure their required level of quality. For instance, in health care applications, vital data from patients (context information) need to be provided at a high refresh rate (QoC parameter) and by a service with a low response time (QoS parameter).

In this scenario, an important challenge is to provide efficient means to monitor QoS and QoC metadata, thus enabling the application to periodically gather the monitored metadata and also to be asynchronously notified whenever a given metadata becomes available. In the literature, some works on monitoring metadata in ubiquitous applications focus just on QoS [12] or QoC monitoring [11, 13] and either on synchronous or asynchronous mode [11, 13]; however, it is important to support both monitoring modes and QoS and QoC metadata. In this perspective, this paper presents *QoMonitor*, a metadata monitoring system that receives synchronous and asynchronous requests from clients (ubiquitous applications and/or middleware), recovers metadata from context providers, and sends them to the clients. By using the proposed monitoring system, ubiquitous applications can focus on addressing fundamental problems of the application and abstract away the burden of dealing with the complexities related to synchronous and asynchronous metadata monitoring. Furthermore, metadata monitored by *QoMonitor* can be available to ubiquitous applications, besides it can be associated with a middleware that would be responsible for managing this information in order to select the services that will be used by an application, for example.

The *QoMonitor* monitoring system consists of three repositories: (i) a *metadata repository*, which persists all QoS and QoC metadata of the monitored services and the metadata provided by the services providers; (ii) a *service repository*, which stores information about all monitored services and the parameters needed to communicate with them, and; (iii) a *client repository*, which stores client information, thus enabling the monitor to communicate with its clients. In addition, *QoMonitor* contains: (i) an *ontology module*, which is responsible for specifying metadata using an ontology model to represent the concepts in an unambiguous way; (ii) a *requests handler*, which receives the requests from clients, gathers the metadata and replies to them, and; (iii) an *assessment module*, which is responsible for effectively monitoring and assessing QoS/QoC metadata of the services stored in the service repository.

ISBN: 978-1-61208-236-3

This paper is structured as follows. Section II briefly describes a health care application which serves as a running example used along this paper. Section III presents *QoMonitor*, the proposed metadata monitoring system. Section IV contains an evaluation of *QoMonitor*. Section V presents related work. Finally, Section VI contains the final remarks.

II. CASE STUDY

This case study is an application related to the health care context in a scenario inspired in the work of Hegering et al. [3]. The application considers as users patients with critical diseases, doctors, and ambulance staffs that use conventional or specific-purpose mobile devices, connected through heterogeneous wireless networks (e.g. Wi-Fi, 3G, Bluetooth, etc.) and/or wired infrastructures. Patients have their vital functions (e.g. blood pressure, cardiac beat rate, etc.) continuously monitored by body sensors. Besides the information provided by these sensors, medical information (called medical profile) about the patients (e.g. if a patient is smoker or not, if he has diseases and/or allergies, etc.) and previous events/medical diagnosis is available. In health care applications, if a patient has complications in his/her current health state, a set of actions must be taken, such as to trigger emergency staffs to give aid to the patient. This application was chosen because of its relevance in a real-world scenario and because it uses different types of context information deriving from different types of mobile or specific-purpose devices. In addition, many kinds of services can be considered, including services with the same functionality (such as GPS, 3G, or Wi-Fi service location), allowing us to monitor similar services implemented using distinct technologies and providing different levels of quality.

To exemplify a health care application, the blood pressure of a patient was chosen as a parameter to be monitored. To sum up, if the blood pressure of the current patient is higher than a specific limit, then he/she may be suffering a cardiac attack or other kind of complication. Thus, emergency staffs are triggered and information about the current health conditions and the patient's medical profile are provided, as well as information about his/her localization. At the same time, the application sends a message to the patient's doctor. Figure 1 gives an overview of the different types of users and the context information processed by the application.

Firstly, the patient's blood pressure is synchronously or asynchronously monitored by the *GetBloodPressure* (*S1*) and *SubscribeBloodPressure* (*S1*') services, respectively. If the value of the blood pressure exceeds the acceptable limit, then the patient's medical profile, which contains previous and current information about the patient and may influence his/her treatment or be correlated to the problem in question, is consulted by executing the *ConsultMedicalProfile* (*S2*) service. Next, available closest doctors are found by executing the *SearchClosestDoctorsCel* (*S3*) and *SearchClosestDoctorsGPS* (*S3*') localization services, so that one of them can be selected to be used by the application according to their quality (QoS/QoC) parameters. Afterwards, a SMS alert about the current medical state of the patient is sent to the doctors by executing the *SendSMS* (*S4*) service.

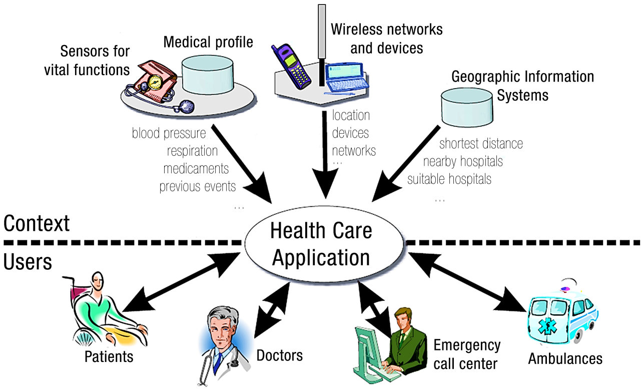


Figure 1. Users and context information in a health care application.

Together with data gathered by consulting the patient's medical profile, the application chooses and triggers emergency staffs through *SearchClosestAmbulancesGPS* (*S5*) and *CallAmbulance* (*S6*) by using the localization of these emergency staffs with respect to the patient. Next, the best route (the shortest or the faster course) between the patient's location and the emergency staff is determined by the *DetermineRoute* (*S7*) service, so that the emergency staff can reach the patient and carry him/her from the current location to the hospital, completing the aid action. Finally, the patient's medical profile is updated with the event that just took place by executing the *UpdateMedicalProfile* service (*S8*).

III. QOMONITOR: A METADATA MONITORING SYSTEM

QoMonitor is a metadata monitoring system that receives synchronous and asynchronous requests from clients (ubiquitous applications and/or middleware), recovers metadata from context providers, and sends them to the clients. This section presents the architecture and operation of *QoMonitor* (Section III.A) and how it can be used in the context of our running example (Section III.B).

A. Architecture and operation

Figure 2 illustrates the architecture of *QoMonitor*, which was specified with a modular design, so that each component can work in an independent way. *QoMonitor* provides two communication interfaces: *IClient* for communicating with clients and *IServer* for communicating with service providers. The *Server Façade* modularizes all monitor communication with service providers thus being responsible for registering new services in the *Service Repository* and communicating with the providers. When one of the service providers is registered in *QoMonitor*, the *Server Façade* receives the data provided by this provider and forwards them to the *Service Repository*. The *Service Repository* is responsible for storing information regarding all monitored services and the parameters required to communicate with them. There are two ways to add new services to the *Service Repository*. In the first one the client makes a request to retrieve QoS/QoC metadata, and if the service's data are not in the repository, then the *Client Façade* provides the data for storing the new service in the repository. In the second way, the service registers itself in the monitor through the interface provided by the *Server Façade*. Whenever a new service is added to the

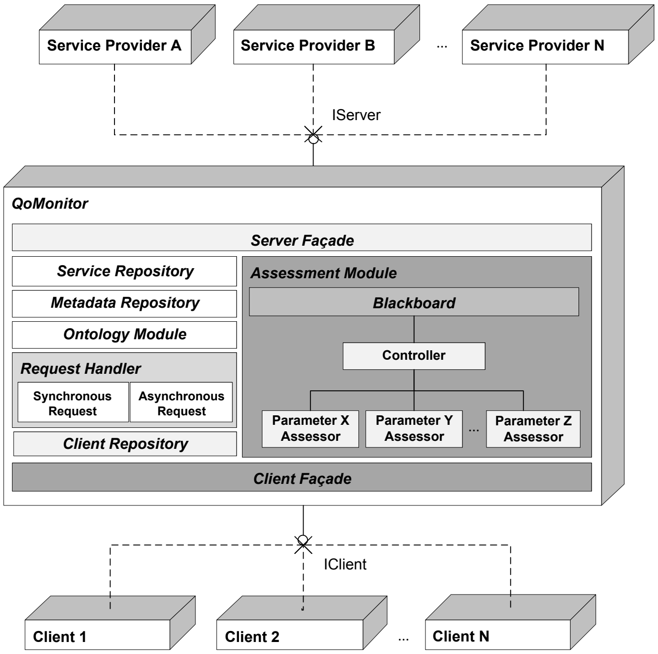


Figure 2. *QoMonitor* architecture.

repository, the *Assessment Module* is notified to start the monitoring and assessment of the services.

The *Client Façade* is responsible for allowing the communication of the clients with the monitor, which can be any ubiquitous application or middleware that needs to make use of QoS/QoC parameters. To perform this communication with clients, it was defined a simple API (summarized in Table 1) that implements the *IClient* interface. Through this API, clients can register themselves on the monitor and make synchronous and asynchronous requests. To register itself, the client calls the *register* method, which receives as parameters the client's name, IP address, and access port. Then, the *Client Façade* forwards these data to the *Client Repository* for storing them to be used when the monitor needs to reply the requests of this client. To perform a synchronous request, the client calls the *getServiceQuality* method, which receives as parameters the data regarding the monitored service and a list of quality parameters to be sent to the client, and returns the parameters with their respective values represented in the ontology format. For instance, considering the *GSMSystem* service from the case study, a call to the *getServiceQuality* method would receive the following parameters: <*GSMSystem*, *192.168.0.100*, *8080*> (service name, IP address, and access port), <*SMS*, *Person*> (a list of input parameters), and <*errorRate*, *uptime*, *responseTime*> (a list of the quality parameters to be monitored).

TABLE 1. METHODS OF THE QOMONITOR COMMUNICATION API.

Method	**Functionality**
register	clients register themselves in the monitor
getServiceQuality	clients make synchronous requests to get QoS/QoC metadata
subscribeServiceQuality	clients make asynchronous requsts to get QoS/QoC metadata
unsubscribeServiceQuality	the monitor stops the sending periodic responses to clients

If the client wants to make an asynchronous request, then it should call the *subscribeServiceQuality* method, which receives as parameters: (i) the data regarding the service and the client; (ii) a list of quality parameters to be sent to the client, and; (iii) a *return condition* in the form of a <*parameter*, *comparison*, *value*> triple. For instance, if the client wants to be informed when the parameter *errorRate* of the *GSMSystem* service is greater than 0%, a call to the *subscribeServiceQuality* method would receive the following parameters: <*GSMSystem*, *192.168.0.100*, *8080*> (service name, IP address, and access port), <*SMS*, *Person*> (a list of input parameters), <*Client1*, *192.168.0.199*, *8080*> (client name, IP address, and access port), <*errorRate*, *uptime*, *responseTime*> (a list of quality parameters to be monitored), and <*errorRate*, *greaterThan*, *0.0*> (the return condition). To perform this kind of request, the client must implement a method called *callback* in order to enable the communication between the monitor and the client, so that this method is responsible for receiving the response from the monitor regarding the asynchronous request. The monitor periodically checks if the return condition has been satisfied, and while it is true, the monitor replies to the client providing the parameters with their respective values represented in the form of the ontology, through the *callback* method. To stop the sending of responses from the monitor, the client calls the *unsubscribeServiceQuality* method.

With the data regarding the request, the *Client Façade* forwards them to the *Request Handler* using references to clients and service providers in the respective repositories. Finally, when the data regarding the performed assessments are available, the *Client Façade* receives the QoS/QoC metadata in the ontology form together with a reference to the current client. If the current service is not in the *Service Repository*, then the *Client Façade* calls the *Service Repository* for storing its data and the repository notifies the *Assessment Module* informing that a new service has been added. This module immediately starts the monitoring and assessment of data.

The *Metadata Repository* is responsible for persisting all QoS/QoC metadata assessed by the monitor and also QoS/QoC metadata provided by service providers. In turn, the *Ontology Module* is responsible for representing these data in the form of an ontology as depicted in Figure 3, in which QoS and QoC parameters respectively extend the *QoS Parameter* and *QoC Parameter* classes defined in the ontology. An ontology is a data model that represents a set of concepts within a domain and the relationships between them [14], thus providing formal expressiveness and avoiding ambiguity in the semantic interpretations of the same information. For instance, Dobson et al. [4] define the QoS parameter *ROCOF* (rate of failure occurrence), which has the same definition of the *error rate* parameter defined by Guo et al. [5] and that is used in this paper as the error rate in a given time interval. This situation can generate an interpretation problem that can be solved by using ontologies. When a monitor component wants to receive metadata in the ontology format, this component provides a reference to the service in the *Service Repository* and the *Ontology Module* forwards it to the *Metadata Repository*, which performs a search and returns the data of the current service. With these data, the *Ontology Module* performs operations to represent them in the ontology format used by the monitor.

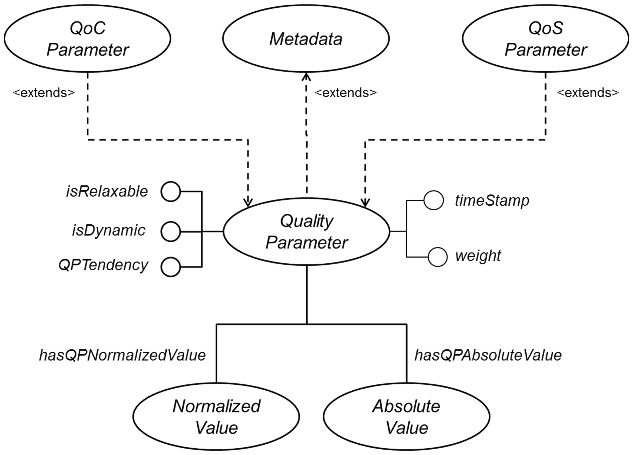


Figure 3. Ontology used by *QoMonitor* for representing metadata.

The main component of the monitor is the *Assessment Module*. This component is responsible for assessing QoS/QoC metadata of the services stored in the *Service Repository* and monitoring them and is composed by three types of elements: *assessors*, *Blackboard*, and *Controller*. Each assessor is responsible for assessing one specific quality (QoS/QoC) parameter from information gathered through requests to the monitored services by the *Assessment Module*. This information is: (i) the time spent to complete the request (*CompletedTime*); (ii) if the service was available or not (*isAvailable*); (iii) the instant in which the request was made (*TimeStamp*), and; (iv) the date and time of creation/sensing of the context information provided by the service (if it is a context service), so that this information is important because it enables inferring the age (*Age*) of the context information provided by the service. The *Blackboard* component incorporates the idea of a shared data repository, which is interesting since the assessors of different QoS/QoC parameters use the same aforementioned information to calculate the value of these parameters. Thus, the use of the *Blackboard* component avoids a large number of requests to the monitored services since, without this element, each one of the assessors would make isolated requests to the services in order to gather the metadata information, thus negatively impacting their performance. To avoid this problem, the *Blackboard* centralizes this information so that each assessor is able to receive it and calculate the value of the quality parameter to which the assessor proposes to measure. For instance, assessors regarding QoS parameters such as availability and error rate can make use of historical data stored in the *Blackboard* about the availability of the service to perform the assessment.

The idea of the *Controller* component is to control the access to the information stored in the *Blackboard* and the information gathered from the assessment of the parameters, so that the assessors do not know the source of the data that they use to make the assessment, thus modularizing the architecture. The monitoring of services works independently, by using threads, and starts at the time when the monitor is available, so that monitoring and assessment operations are executed while the monitor receives and replies requests. This continuous monitoring is intended to speed up the response time of requests since when a request is made, QoS/QoC metadata are already stored and can be accessed by the *Request Handler* to reply to the clients.

Finally, the *Request Handler* is responsible for retrieving QoS/QoC metadata through the *Ontology Module* and forwarding them to the clients. When a client makes a synchronous request, the *Client Façade* forwards it to the *Request Handler*, which retrieves the current data through the *Ontology Module* and replies to the *Client Façade*. When an asynchronous request is forwarded to the *Request Handler*, it monitors if the QoS/QoC data satisfy the return condition informed by the client; in this case, the *Request Handler* continuously monitors these data in order to identify whether the return condition is satisfied. When the return condition is met, the *Request Handler* immediately replies to the *Client Façade* that calls the *callback* (listener) method implemented by the client. As the monitoring operation is independent of the other operations performed by the monitor, the response time of a request is considerably small since the service has already been monitored and its parameters have been assessed before making the request. An exceptional situation happens when a client makes a request regarding a service that is not present in the *Service Repository*, so that this service must be added to the repository and then the monitoring is started.

B. Monitoring service providers

Before starting the monitoring of services, two time intervals need to be defined in the monitor. The first one is called *TimeToRequest* and is the time interval in which the *Assessment Module* makes requests to the service providers. The second time interval is called *TotalTime* and is the time in which information is considered recent. For instance, if the *TotalTime* is set to ten minutes, then information gathered more than ten minutes ago will be ignored since this information is considered outdated and can interfere in the assessment calculations of the quality parameters.

Next, the *Blackboard* receives a list of references to the available services from the *Service Repository* and makes periodic requests (according to *TimeToRequest*) through the *Server Façade* to the respective service providers using their data (address and list of parameters), thus returning the time spent to complete the request (*CompletedTime*), whether it has been performed successfully. If the request has not been successful, then the *Server Façade* throws an exception that is caught by the *Blackboard*. For the case study previously described, the context services provided by *GPSLocalizationMiddleware* and *CellularLocalizationMiddleware* use QoC metadata, so that the age of this information (*Age*) is also gathered. In the case study under consideration, no service provider provides the QoS/QoC metadata beforehand, so that after each request the *Blackboard* stores the request data (*CompletedTime*, *isAvailable*, *TimeStamp*, and *Age*). If the request has failed, then *CompletedTime* is equal to -1, *isAvailable* is false, *Age* is null, and *TimeStamp* remains the same. If the service provider itself provides the QoS/QoC metadata, then the *Blackboard* forwards these metadata to the *Ontology Module*, which builds the representation of these data in the ontology format and then sends them to the *Metadata Repository* for storage.

With the data stored in the *Blackboard*, the *Controller* is called to access the data history of the requests that are in the *Blackboard* and forward them to each of the assessors. After all assessors finished their assessment and returned the results to the *Controller*, it forwards the data to the *Metadata Repository* for storing them. This execution is repeatedly done with a time interval defined by *TimeToRequest* and is independent of the requests made by the clients since the idea here is that QoS/QoC metadata are already stored before client requests. Thus, the monitor will quickly reply to client requests and can share data whether two clients make requests to the same service. If for some reason the metadata are not available, e.g. when the first monitoring is performed, then the *Request Handler* remains on standby until the data are available.

The running example outlined in Section II clearly illustrates the importance of using a QoS/QoC metadata monitor system when deciding which service will be used by a ubiquitous application. For instance, the service providers *GPSLocalizationMiddleware* and *CellularLocalizationMiddleware* are responsible for providing services to localize the doctors of the monitored patient and ambulances, each one using different technologies and possibly different QoS/QoC parameters. Without monitoring data, the application will not know which service is best suited to be used, in terms of quality parameters. The client (a ubiquitous application and/or a middleware) can find out the QoS/QoC metadata of the services that are available by making synchronous requests to the monitor. It can also decide which is the best time to use a particular service provider by making asynchronous requests, e.g. in situations when the response time is smaller than fifty milliseconds or the freshness is smaller than two seconds, etc.

IV. EVALUATION

This Section presents an evaluation of *QoMonitor* metadata monitoring system proposed in this paper under a quantitative perspective, which aims to address the time for assessing QoS and QoC parameters (Section IV.B) and the time spent to completely reply to synchronous and asynchronous requests to the monitor (Section IV.C). For the purposes of this evaluation, we have used the health care ubiquitous application outlined in Section II. The services used in the case study were implemented as Web services using the Java programming language and the Apache Axis framework [6] and deployed on an Apache Tomcat application server [7] installed in a computer with an Intel® Core^{TM} i7 2.7 GHz processor, 6 GB of RAM memory and Linux Ubuntu 12.04 operating system, which worked as the server to which requests were performed. In the computational experiments, the *QoMonitor* monitoring system was executed in a computer with an Intel® Core^{TM} i5 2.3 GHz processor, 4 GB of RAM memory and Mac OS X operating system and has performed requests to the services deployed on the Apache Tomcat server installed in the remote server. Aiming to execute such experiments under similar conditions to those observed in a real-world scenario, *QoMonitor* and the server in which the services were hosted were placed on different networks, so as not completely disregard the influence of the network in the process. In the quantitative evaluations presented in the Sections IV.B and IV.C, we have performed fifteen independent executions for each service and for each of the six parameters listed in Section IV.A, namely: *error rate*, *response time*, *MTBF*, *MTTR*, *uptime* and *freshness*. In these executions, the values chosen for the *TotalTime* and *TimeToRequest* times were twenty minutes and five seconds, respectively.

A. QoS and QoC parameters

In Ubiquitous Computing, context information is gathered from several sources, e.g. it can be provided by users, sensed from sensor devices, derived from multiple origins, etc. Buchholz et al. [2] enumerate some QoC parameters, such as *precision*, *correctness*, *resolution*, *freshness*, etc. For the evaluation performed in this paper, we have considered the *freshness* QoC parameter, which expresses the information age, i.e. the time elapsed since the information was generated. Thus, if the information is recent, then it will be more reliable since old information may be outdated. We have chosen just this parameter to assess since other QoC parameters such as precision and resolution are provided by the sensors that measure them [8], so that they can not be properly monitored by our monitoring system. Although precision, resolution and other QoC parameters are not assessed by our *QoMonitor*, if they are published by the service provider, then the monitor can retrieve them and store these metadata in the ontology format.

Similarly, metadata for QoS parameters are associated with the services used by the ubiquitous applications and are intended to identify the quality of the service. Among the various QoS parameters enumerated in the literature [9, 10], we have considered in this evaluation the following five QoS parameters: (i) *response time*, which is the time elapsed from the instant in which the client performs a request to the instant in which it processes the response message sent by the server; (ii) *MTBF*, which is the mean time between system failures during its operation; (iii) *MTTR*, which is the mean time between a system failure and its return to operation (recovery); (iv) *error rate*, which measures the error rate for data transmission or service operation in a given time, and; (v) *uptime*, which refers to the operating time (i.e. availability) of a service.

B. Asessment of QoS/QoC parameters

Table 2 and Table 3 present the minimum, maximum and average assessment times (in milliseconds) for each of the quality parameters considered for each service of the case study enumerated in Section II (from *S1* to *S8*). Here, only the localization services *SearchClosestDoctorsCel* (*S3*), *SearchClosestDoctorsGPS* (*S3*') and *SearchClosestAmbulancesGPS* (*S6*) and the blood pressure monitoring services

TABLE 2. MINIMUM AND MAXIMUM ASSESSMENT TIMES OF THE CONSIDERED QOS AND QOC PARAMETERS.

Services / Parameters	***error rate***		***response time***		***MTBF***		***MTTR***		***uptime***		***freshness***	
	MIN	***MAX***	***MIN***	***MAX***	***MIN***	***MAX***	***MIN***	***MAX***	***MIN***	***MAX***	***MIN***	***MAX***
S1	0.173	0.269	0.049	0.043	0.023	0.043	0.021	0.040	0.025	0.041	0.088	0.145
S1'	0.190	0.260	0.087	0.071	0.035	0.071	0.038	0.049	0.034	0.102	0.035	0.151
S2	0.168	0.248	0.054	0.092	0.022	0.059	0.022	0.037	0.034	0.039	-	-
S3	0.226	0.266	0.103	0.128	0.041	0.063	0.040	0.046	0.035	0.149	0.036	0.155
S3'	0.173	0.193	0.062	0.078	0.025	0.035	0.025	0.040	0.025	0.043	0.089	0.109
S4	0.172	0.265	0.074	0.128	0.026	0.048	0.025	0.075	0.026	0.042	-	-
S5	0.172	0.261	0.079	0.129	0.026	0.047	0.025	0.045	0.026	0.041	0.089	0.148
S6	0.193	0.286	0.064	0.115	0.026	0.045	0.025	0.045	0.026	0.040	-	-
S7	0.189	0.261	0.084	0.126	0.036	0.045	0.035	0.044	0.035	0.041	-	-
S8	0.174	0.240	0.045	0.093	0.022	0.036	0.021	0.039	0.025	0.035	-	-

TABLE 3. AVERAGE ASSESSMENT TIMES OF THE CONSIDERED QOS AND QOC PARAMETERS AND RESPECTIVE STANDARD DEVIATIONS.

Services / Parameters	***error rate***		***response time***		***MTBF***		***MTTR***		***uptime***		***freshness***	
	AVG	***STD***	***AVG***	***STD***	***AVG***	***STD***	***AVG***	***STD***	***AVG***	***STD***	***AVG***	***STD***
S1	0.211	0.037	0.069	0.018	0.033	0.008	0.030	0.007	0.033	0.006	0.105	0.022
S1'	0.225	0.025	0.116	0.017	0.043	0.008	0.045	0.003	0.041	0.017	0.126	0.030
S2	0.201	0.028	0.074	0.012	0.036	0.007	0.034	0.004	0.036	0.002	-	-
S3	0.251	0.010	0.120	0.006	0.045	0.005	0.042	0.002	0.045	0.029	0.137	0.030
S3'	0.181	0.005	0.067	0.006	0.028	0.003	0.028	0.005	0.031	0.006	0.094	0.007
S4	0.257	0.024	0.118	0.013	0.044	0.005	0.041	0.011	0.037	0.003	-	-
S5	0.234	0.023	0.117	0.016	0.041	0.005	0.040	0.004	0.037	0.004	0.133	0.017
S6	0.225	0.034	0.088	0.021	0.034	0.008	0.034	0.008	0.032	0.006	-	-
S7	0.233	0.025	0.110	0.014	0.041	0.003	0.039	0.003	0.037	0.002	-	-
S8	0.184	0.017	0.053	0.012	0.025	0.005	0.025	0.006	0.028	0.004	-	-

GetBloodPressure (*S1*) and *SubscribeBloodPressure* (*S1*') are context services, so that only these services have the QoC parameter *freshness*. The assessment time of a given parameter is basically the time spent by the respective assessor to make the calculations of the values regarding this parameter after the necessary data are recorded in the *Blackboard* component.

In Table 2, minimum and maximum assessment times are reported in the columns labelled as *MIN* and *MAX*, respectively. Similarly, in Table 3 the average assessment times are reported in the columns labelled as *AVG* and the respective standard deviations are reported in the columns labelled as *STD*. As can be clearly seen in Table 2 and Table 3, all the assessment times do not exceed the order of 1 millisecond, which is very beneficial in the sense that the monitor does not promote a significantly impact in terms of the assessment of the parameters.

C. Synchronous and asynchronous requests

Table 4 presents the times spent (in milliseconds) by the monitor to completely reply to synchronous and asynchronous requests made by clients regarding the services of the case study, i.e. the time elapsed between the instant when the request is received by the monitor until the instant in which the monitor sends the response, thus encompassing all operations involved in handling this request. As can be observed in Table 4, the minimum, maximum and average response times (respectively reported in the columns labelled as *MN*, *MX* and *AV*) are small, so that the time spent by the monitor for receiving requests and replying to them is most influenced by the network than the monitor itself. Therefore, we can conclude that the monitor does not promote a significant impact regarding this issue.

TABLE 4. RESPONSE TIMES FOR SYNCHRONOUS AND ASYNCHRONOUS REQUESTS.

Services/ Requests	**synchronous requests**				**asynchronous requests**			
	MIN	***MAX***	***AVG***	***STD***	***MIN***	***MAX***	***AVG***	***STD***
S1	46	86	60	12	46	132	65	28
S1'	45	82	60	10	46	139	64	23
S2	43	108	70	19	43	111	57	18
S3	47	96	65	16	49	75	57	8
S3'	49	84	61	11	47	76	57	10
S4	44	71	57	8	41	123	61	23
S5	48	110	70	18	46	91	60	15
S6	50	117	69	19	48	87	68	13
S7	47	96	65	16	43	88	64	16
S8	50	102	64	15	40	96	53	17

V. Related Work

To the best of our knowledge, works found in the literature focus just on QoS or QoC monitoring and either on synchronous or asynchronous mode, as we have mentioned. However, it is important to support both monitoring modes and also QoS and QoC metadata. In this Section, we briefly present some of these proposals.

Huebscher and McCann [11] present a mechanism to choose context services according to the QoC application requirements. The proposed mechanism defines synchronous and asynchronous functions to be used by an application when querying the QoC metadata and uses a directory service to store metadata and services. In a different way of our proposal, it considers only QoC metadata and does not use ontologies, thus being limited to a proprietary context model.

Truong et al. [12] present a tool for monitoring and analyzing QoS metrics of grid computing services. QoS metadata regarding individual services are collected and sent to a middleware that stores these monitored data. A reasoning engine performs QoS analysis based on rules contained in a component called *QoS knowledge base*, which stores QoS historical data, so that it is possible to define automatic actions to react to changes in the parameters by sending alerts to the client. Although this tool is proposed to monitor and analyze QoS metadata at runtime as it is similarly done by our *QoMonitor*, it does not deal with QoC metadata neither handle synchronous and asynchronous requests. In addition, both works enable clients to retrieve monitored data since they have a storage module. Since our work focuses specifically on monitoring QoS and QoC metadata and making them available for ubiquitous applications or middleware, the *QoMonitor* system provides the subsides needed to serve as input to another component (or even a middleware) that would be responsible for triggering these automatic actions associated with changes regarding QoS and QoC parameters.

Finally, Zheng and Wang [13] propose a tool that supports QoC management. Requests are handled by a context reasoner, which filters context information and notifies the subscribed components about context changes, thus supporting asynchronous requests. In addition, services are stored in a context repository and metadata are represented by an ontology. Although this work is very close to our proposal, it does not support synchronous requests and focuses just on QoC monitoring.

VI. Final Remarks

In this paper, we presented *QoMonitor*, a metadata monitoring system that is in charge of handling synchronous and asynchronous requests for monitoring QoS and QoC metadata. *QoMonitor* recovers metadata from several context providers, uses an ontology to represent such metadata, and sends them to the clients. By using the proposed monitoring system, ubiquitous applications can abstract away the burden of dealing with the complexities related to synchronous and asynchronous metadata monitoring. In addition, these monitored metadata can be available to ubiquitous applications and/or a middleware that would be responsible for managing this information in order to select the services that will be used by an application, for example. We have implemented this system and used it in a health care application and the evaluation of *QoMonitor* showed that the average time for assessing QoS and QoC parameters and the time spent to completely reply to synchronous and asynchronous requests to the monitor are significantly small. As a future work we aim to evaluate the delay when a given service is not available at the service repository yet and the monitoring system needs to ask a third-party element (typically an underlying middleware) to discover which context provisioning system provides such service.

References

[1] M. Weiser, “The computer of the Twenty-First Century”, Scientific American, vol. 265, no. 3, Sep. 1991, pp. 94-104.

[2] T. Buchholz, A. Kupper, and M. Schiffers, “Quality of Context: What it is and why we need it”, Proc. of the 10th Workshop of the HP OpenView University, 2003, pp. 1-14.

[3] H.G. Hegering, A. Kupper, C. Linnhoff-Popien, and H. Reiser, “Management challenges of context-aware services in ubiquitous environments”, Proc. of the 14th IEEE/IFIP Workshop on Distributed Systems: Operations and Management (DSOM 2003), LNCS, vol. 2867, Germany, Springer Berlin/Heidelberg, 2003, pp. 321-339.

[4] G. Dobson and R. Lock, “Developing an ontology for QoS”, Proc. of the 5th Annual DIRC Research Conf., 2005, pp. 128-132.

[5] G. Guo, F. Yu, Z. Chen, and D. Xie, “A method for semantic Web service selection based on QoS ontology”, Journal of Computers, vol. 6, no. 2, Feb. 2011, pp. 377-386.

[6] Apache Axis. Available at: http://ws.apache.org/axis [retrieved: July, 2012]

[7] Apache Tomcat. Available at: http://tomcat.apache.org/ [retrieved: July, 2012]

[8] A. Manzoor, H. Truong, and S. Dustdar, “Quality of Context: Models and applications for context-aware systems in pervasive environments”, The Knowledge Engineering Review, 2004, pp.1-24.

[9] V. Trana, H. Tsujib, and R. Masuda, “A new QoS ontology and its QoS-based ranking algorithm for Web services”, Simulation Modeling Practice and Theory, vol. 17, no. 8, Sep. 2009, pp. 1378-1398.

[10] M. Sathya and M. Swarnamugi, “Evaluation of QoS based Web service selection techniques for service composition”, International Journal of Software Engineering, vol. 1, no. 5, Feb. 2011, pp. 73-90.

[11] C. Huebscher and A. McCann, “An adaptive middleware framework for context-aware applications”, Personal and Ubiquitous Computing, vol. 10, no. 1, Dec. 2005, pp. 12-20.

[12] H. Truong, R. Samborski, and T. Fahringer, “Towards a framework for monitoring and analyzing QoS metrics of grid services”, Proc. of the 2nd IEEE Int. Conf. on e-Science and Grid Computing,USA,IEEE Computer Society, 2006, pp. 1-8.

[13] D. Zheng and J. Wang, “Research of the QoC based middleware for service selection in pervasive environment”, International Journal of Information Engineering and Electronic Business, vol. 3, no. 1, Feb. 2011, pp. 30-37.

[14] T. Gruber, “A translation approach to portable ontology specifications”, Journal of Knowledge Acquisition, vol. 5, no. 2, Jun. 1993, pp. 199-220.

User-centric Complex Event Modeling and Implementation Based on Ubiquitous Data Service

Feng Gao
Unit of Service Oriented Architecture
Digital Enterprise Research Institute, DERI
Galway, Ireland
email: feng.gao@deri.org

Sami Bhiri
Unit of Service Oriented Architecture
Digital Enterprise Research Institute, DERI
Galway, Ireland
email: sami.bhiri@deri.org

***Abstract*—Current complex event processing systems are often implemented as standalone engines that produce business events and feed process execution environments. Event patterns are defined with rule-based languages. Logical programming and/or stream processing techniques are used to detect matchings for the patterns. However, tremendous technical efforts are required both for the pattern definition and implementation. In this paper, we present a novel framework that provides a user-centric way to define complex event patterns and implement the patterns automatically. We allow the business users to describe their complex events with graphical notations and transform the graphical pattern into a stream query, then, we evaluate the query over primitive sensor data streams to obtain results as complex events.**

Keywords-user-centric; complex event processing; stream reasoning; web service; BEMN.

I. MOTIVATION

Business Process Management (BPM) provides concepts, methodologies and tools to design, implement, execute and reengineer business processes. Today business are demanding more flexibility from BPM to adapt to the fast changing business environment in-time. To this end, the concept of ”BPM 2.0” has been brought up that aims to bring more flexibility into BPM. Among the methodologies used in BPM 2.0, an important idea is to provide automation support for the implementation of business processes [1]. This allows business analysts to test and run the processes they design with minimized technical effort. Current research focus on combining BPM with Service Oriented Architecture to provide automation support for process implementation. In particular, service discovery will find direct matching services for process tasks and simple events, service composition will try to create services for process tasks when there’s no direct matching.

However, Complex Event Processing (CEP), an indispensable technique for business process systems [2], has not yet get enough automation support. Most current CEP tasks are delegated to CEP engines. These engines are usually equipped with rule-based languages and engines to define and analyze complex event patterns, and require some programming skills to encapsulate the corresponding event data to communicate with event processing engines. Unfortunately neither rule languages nor programming APIs are friendly enough for business users. As such, companies need significant technical efforts to implement a business process that requires CEP.

On the other hand, development in sensor networks is gaining increasing interests from enterprises. Many efforts have been made to integrate sensor functionalities with enterprise systems to manage business processes more dynamically. A natural use of sensors is to monitor the state changes of the real-world and notify event driven processes to take actions. However business events are complex and sensor events are primitive, thus CEP techniques are crucial for the deriving business events from sensor reading events.

To give an example of complex events derived from sensor readings, let us consider the following scenario: in a supermarket, sensors are deployed on shelves to monitor the numbers of remaining products, when products left on the shelf are insufficient, a sensor will report an out-of-stock event and trigger a replenishment process. If 10 out-of-stock events are captured for the same product during the past week, or a direct request from the manager asks for increasing the storage for the product is received, a complex event will raise to notify the need of increasing the amount of the product in the next purchase order. In this paper, we will demonstrate our work that provides a user-centric and automatic way to define and implement complex event patterns.

The remainder of the paper is organized as follows. Section 2 elaborates our research problems in detail. Section 3 presents the related work. Section 4 gives an overview of the proposed system. Section 5 discusses the graphical notations for complex event definitions. Section 6 briefly describes the algorithm to transform the event patterns into stream queries before we conclude in Section 7.

II. RESEARCH PROBLEM DESCRIPTION

Our research intends to provide means to define and implement complex event patterns in a user centric way. Our goal can be further decomposed into the following.

ISBN: 978-1-61208-236-3

A. User-centric Complex Event Pattern Definition

We need an intuitive, friendly language for business users to create the complex events they need. Graphical notation is our first choice. Flow-based graphical structure can be used to model the control and data dependency between primitive events. The language should be expressive enough for various business event patterns. Meanwhile, to ensure that the event patterns are operating on the correct event sources, a user-oriented mechanism is required to facilitate primitive event service discovery.

B. Automated Implementation for Complex Event Patterns

Implementing complex event pattern requires evaluating event rules represented by event patterns over streaming data. A Data Stream Management System (DSMS) is usually implemented as a component in a CEP system to process streaming data. Comparing to conventional DSMS employed by current CEP engines that can only process syntactical data, an emerging research area of stream reasoning aims to process continuous semantic data. Our key idea of automated implementation is to transform complex event patterns into declarative stream reasoning queries, so that the complex event patterns defined by the process modelers/business users can be evaluated by the stream reasoning engine and are practically made executable with minimal technical efforts. There exist some stream reasoning languages and engines, our approach can reuse some of them with proper modifications/extensions to evaluate complex event patterns.

III. Related Work

Complex Event Processing (CEP) is a technique to detect complex events in (near) real-time. A complex event represents a set of correlated events called its member events. A complex event pattern describes the temporal relationships, data specifications and other conditions required to derive the complex event from its member events. Most CEP systems describe event patterns in SQL-like languages. Wang et al. [3] and Dunkel et al. [4] use such languages to specify complex events and feed the process engine with derived events. However these languages only offer textual representations. BEMN proposed by Decker et al. [5] is the first work that attempts to provide a graphical and executable event pattern language and is an inspiration to our work, it is able to describe various business event patterns. Still, the BEMN language have some limitations regarding to the difficulty of implementation, scalability and expressiveness, which we will discuss in details later. All the approaches described above can only process the dynamic data in the stream and do not go beyond syntactical processing.

Stream Reasoning is an emerging research area that tries to enable reasoning on continuous data and support processing for both dynamic data and static background knowledge. Anicic et al. [6], proposes a prolog-based framework is to transform continuous triples as logic facts and stream query as rules, so that the processing of complex events can make use of both dynamic event data and static background knowledge. Le-Phuoc et al. [7] use a “white box” approach and support native stream reasoning operators, they also provide means to optimize the query plan so that the evaluation of stream query can be more efficient. We will transform our user-centric complex event definition into the executable stream query language defined in [7] (with some extensions) to enable automated implementation of complex event patterns.

Semantic CEP is discussed in several works. Moser et al. [8]elaborate the benefits to extend syntactical event correlation to semantic event correlations. In the paper they define 3 kinds of semantic correlations on event attributes: equivalence, inheritance and relation-based. Li et al. [9] use an ontology to describe event rules as well as context-aware devices (sensors), and an event hierarchy is used to model the causal relationship between different levels of events. In the framework propsed by Taylor et al. [10], users can select and correlate sensors based on the semantic sensor description, then, a semantic middleware will translate the users’ requirements into the internal language used by the CEP engine (e.g. EPL) and program the sensors to prepare the streams. Semantic query and reasoning in this approach will not affect the run-time processing so that it can guarantee high throughput of the CEP engine. However, it does not give a formalization of event patterns and have limited expressiveness in terms of AND and OR patterns and aggregations. Moreover, it relies on programmable sensors. Hasan et al. [11] propose a dynamic enrichment of the stream sensor data so that the information sensed can be correlated to background knowledge based on the interested situation at runtime. This approach is different to ours, where we load the static knowledge before runtime.

IV. System Architecture

In this section, we will introduce the architecture of our proposed system by presenting the overview of the framework and functional descriptions of its components.

A. Overview

An architectural design of our system is depicted in Figure 1. The system aims to realize user-centric and automated complex event implementation for business users in a service-oriented environment based on publish/subscribe messaging paradigm and semantic web technology.

B. Functionalities of Components

Due to the limited space we will only introduce the crucial components in the system framework in the followings.

- **Process Modeling Environment** provides utilities for business process designers to create, update, retrieve and delete process models. The process modeling tool will support graphical notations for describing complex

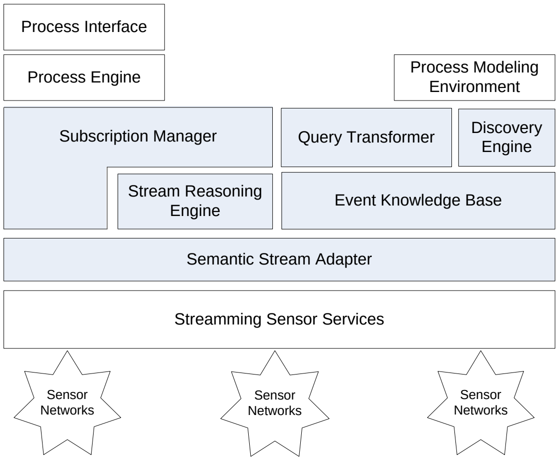


Figure 1. System Architectural Overview

event patterns. The graphical notations can be made compatible to the standardardized Business Process Modeling Notation (BPMN).

- **Event Service Discovery Engine** helps the process designer to select primitive sensor event services based on semantic service descriptions.
- **Query Transformer** will take the complex event descriptions as inputs to create stream queries over the sensor data streams. We will briefly describe the strategy of the algorithms used in the transformer later.
- **Event Knowledge Base** stores the top-level event ontology, domain-specific ontology and datasets (including sensor service descriptions).
- **Semantic Stream Adapter** is the direct consumer of the sensor services. It will subscribe to sensor services and receive messages from them. Then the adapter will convert these messages into RDF triples using mapping schemas defined in service descriptions and construct semantic streams for the stream reasoner.
- **Stream Reasoner** is the query engine for the RDF streams. It will evaluate a stream query over a specified window on semantic streams. Results of the stream query will be consumed by the process engine as business events.
- **Streaming Sensor Services** are the set of web services that produce notification messages used to construct streams. Currently we adopt the WS-Notification protocols to wrap functionalities of sensors into web services and produce primitive events.

V. Graphical Notations for Complex Events

BEMN [5] intends to provide a graphical representation for the event composition languages beyond conventional textual language. BEMN diagram can be integrated into BPMN process models seamlessly to facilitate complex event description in business processes. The complex event pattern in BEMN for the example given in Section 1 is shown in Figure 2.

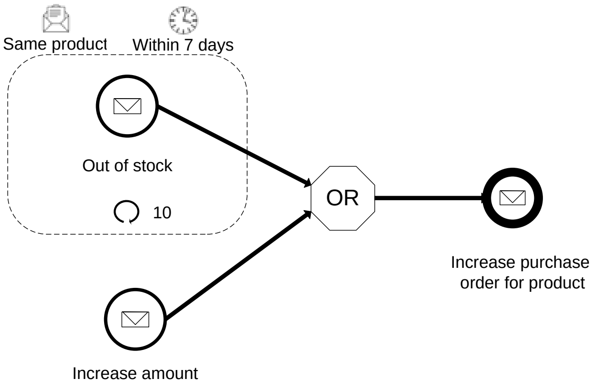


Figure 2. Example event pattern

Despite that the formal semantics of the language are defined and the execution environments are described, the original language did not take into account how current stream processing technique can be integrated to enable the execution of event models. As a result, some simplified matching functions are defined in the execution semantics to make only core event composition models executable. This will limit the expressiveness of executable event composition models and bring the overhead of translating general (non-core) models into core models. In our work, we align the semantics of event models with stream reasoning languages to make the models executable without such overhead.

Furthermore, BEMN execution environment exposes all the events to all the event patterns through a single event stream. This will bring efficiency issues as the matching functions need to filter out events from irrelevant sources. In our approach, we use the pub/sub messaging paradigm based on asynchronous web service interactions to create different event streams on-demand using WS-notification protocol, thus irrelevant events are filtered out before entering the event processing engine.

Moreover, BEMN language does not provide specification on data structure of event declarations. It is left to the programmers who implement the event streams. In this way, technical details are hidden from business users, with the price of compromising automation support for event pattern implementation. Also, it is difficult for the business users to create good event composition models without knowing what the primitive events really mean. We propose to use semantic web technology to help business users discover the primitive events they need, as well as create filters upon primitive event data. To this end, we refine the abstract syntax of event declarations and filters to allow more detailed definition of them.

VI. Pattern to Query Transformation

We will provide an algorithm based on Program Structure Tree (PST) to parse the event composition models and transform them into stream reasoning queries in a 'Divide-and-Conquer' style. First, we do not take parallel inhibition relation into consideration (an inhibition is considered parallel if when removed, source and target of this inhibition is still connected to start and end node, otherwise it is considered sequential), thus event patterns will be Directed-Acyclic-Graphs (DAGs), and we will traverse the DAG to find embedded Single-Entrance-Single-Exit (SESE) regions. Each SESE region can be sequential or branched. Each component in sequential SESE regions is connected with *SEQ* operators (as in [6]). Components in branched regions are translated into *GroupGraphPattern* (as in [7]) or *OPTIONAL* query patterns with filters on their occurrences (*bound()* filters). Then we deal with the inhibitions. Inhibition in sequential events will be transformed into optional patterns with *!bound()* filters. Inhibition in parallel will be defined using filters on timestamps. Finally, we will build aggregations and put filters to the correct scope. The query for the example in Figure 2 is listed in Listing 1.

Listing 1. Sample query for event pattern in Figure 2

```
SELECT ?pid, count(?x) as ?cnt
WHERE{
{STREAM <http://example.org/OutOfStock> [7 Days]
{?x a ces:OutOfStock;
                Evt: hasPayload [hasID ?pid]}}
UNION
{STREAM <http://example.org/IncreaseStorage> [NOW]
{?y a ces:IncreaseStorage;
                 Evt:hasPayload [hasID ?pid]}}}
GROUPBY ?pid HAVING (count(?x)>10)
```

VII. Conclusions and Future Work

In this paper, we present a novel framework to facilitate user-centric definition and automated implementation of complex events based on ubiquitous data service. The user-centricity is achieved by revising a graphical notation of complex events called BEMN [5] which can be seamlessly integrated with BPMN and is targeted to business users. The automated implementation is realized by allowing detailed description of primitive events in event patterns and translating event patterns defined by business users to a stream reasoning query, so that they can be evaluated immediately without further coding.

Apart from the implementation and evaluation of the proposed system, future works may be explored in the following 3 aspects: support for multiple stream reasoning systems, navigation of sensor functionalities and creation of complex event hierarchies. We intend to provide support for multiple stream reasoning systems by creating profiles for the BEMN revision to align its semantics with different query languages. Primitive event service discovery is one of the key enabling techniques of automatic implementation. We aim to provide a navigation based discovery by modeling service capabilities and their relationships to construct a service capability hierarchy/graph, which can be navigated by business users. Currently, we assume all member events are primitive sensor events and we are not able to create event causal hierarchy, we intend to break this assumption to support more comprehensive event models in the future.

Acknowledgments

This work is supported by the Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2),

References

[1] M. Kurz and A. Fleischmann, "Bpm 2.0: Business process management meets empowerment," in *Subject-Oriented Business Process Management*, 2011.

[2] D. Luckham, "The power of events: An introduction to complex event processing in distributed enterprise systems," in *Rule Representation, Interchange and Reasoning on the Web*, 2008.

[3] F. Wang, S. Liu, P. Liu, and Y. Bai, "Bridging physical and virtual worlds: Complex event processing for rfid data streams," in *Advances in Database Technology - EDBT 2006*, 2006, pp. 588–607.

[4] J. Dunkel, "On complex event processing for sensor networks," in *International Symposium on Autonomous Decentralized Systems, 2009.*, 2009, pp. 1 –6.

[5] G. Decker, A. Grosskopf, and A. Barros, "A graphical notation for modeling complex events in business processes," in *EDOC 2007*, 2007, p. 27.

[6] D. Anicic, P. Fodor, S. Rudolph, and N. Stojanovic, "Ep-sparql: a unified language for event processing and stream reasoning," in *Proceedings of the 20th international conference on World wide web*, 2011, pp. 635–644.

[7] D. Le-Phuoc, M. Dao-Tran, J. X. Parreira, and M. Hauswirth, "A native and adaptive approach for unified processing of linked streams and linked data," in *Proceedings of the 10th international conference on The semantic web - Volume Part I*, 2011, pp. 370–388.

[8] T. Moser, H. Roth, S. Rozsnyai, R. Mordinyi, and S. Biffl, "Semantic event correlation using ontologies," *On the Move to Meaningful Internet Systems OTM 2009*, pp. 1087–1094, 2009.

[9] Z. Li, C.-H. Chu, W. Yao, and R. a. Behr, "Ontology-Driven Event Detection and Indexing in Smart Spaces," *2010 IEEE Fourth International Conference on Semantic Computing*, pp. 285–292, 2010.

[10] K. Taylor, "Ontology-driven complex event processing in heterogeneous sensor networks," *The Semanic Web: Research and Applications*, pp. 285–299, 2011.

[11] S. Hasan, E. Curry, and M. Banduk, "Toward Situation Awareness for the Semantic Sensor Web: Complex Event Processing with Dynamic Linked Data Enrichment," *Semantic Sensor*, 2011.

Using Unsupervised Learning to Improve the Naive Bayes Classifier for Wireless Sensor Networks

Ardjan Zwartjes, Paul J.M. Havinga, Gerard J.M. Smit, Johann L. Hurink
PS, CAES, DMMP
University of Twente
Enschede, The Netherlands
g.j.zwartjes@utwente.nl, p.j.m.havinga@utwente.nl, g.j.m.smit@utwente.nl, j.l.hurink@utwente.nl

***Abstract*—Online processing is essential for many sensor network applications. Sensor nodes can sample far more data than what can practically be transmitted using state of the art sensor network radios. Online processing, however, is complicated due to limited resources of individual nodes. The naive Bayes classifier is an algorithm proven to be suitable for online classification on Wireless Sensor Networks. In this paper, we investigate a new technique to improve the naive Bayes classifier while maintaining sensor network compatibility. We propose the application of unsupervised learning techniques to enhance the probability density estimation needed for naive Bayes, thereby achieving the benefits of binning histogram probability density estimation without the related memory requirements. Using an offline experimental dataset, we demonstrate the possibility of matching the performance of the binning histogram approach within the constraints provided by Wireless Sensor Network hardware. We validate the feasibility of our approach using an implementation based on Arduino Nano hardware combined with NRF24L01+ radios.**

***Keywords*-Wireless sensor networks; Unsupervised learning; Classification algorithms.**

I. Introduction

Advancements in miniaturization and the declined cost of hardware have enabled the vision of Wireless Sensor Networks (WSN), where a network of tiny computers can monitor environments using sensors and wireless communication. The implementation of a practical WSN, however, is not a trivial task. Even on small WSNs, the amount of data that can be sampled by the sensor nodes is considerable. Simple micro-controllers can acquire samples at rates above 10kHz; this is more than what can practically be transmitted using current WSN radios.

For many applications the raw sensor data itself is not of interest. For example, in domestic fire detection [4] carbon-dioxide readings do not need to reach a human operator. The presence of a fire, however, is important information. In applications like this, online processing can be a valuable solution.

Online data processing comes in many forms, ranging from simple schemes to compress the data, to complex event recognition algorithms that draw intelligent conclusions. This last group of algorithms can result in considerable reductions in communication by removing the need to transmit the sensor readings. Considering that the energy needed to transmit a few bytes of data is significant [7], it is clear that online intelligent processing is a promising area of research.

A. Problem description.

The characteristics of WSN platforms limit the type of algorithms that can be used. Both memory and computational power are very limited [12], and the unreliability of individual nodes further complicates matters. The naive Bayes classifier is a classification algorithm that can be executed on simple hardware [11]. Its performance with regard to input unreliability and distributed execution make it an interesting algorithms for WSN applications [14], [15].

The naive Bayes classifier can be implemented in multiple ways. Some of which are unsuitable for WSN hardware, while others can show poor classification performance in certain circumstances. The goal of this research is to create a naive Bayes implementation that can run within the constraints provided by WSN hardware, provides excellent classification performance and has limited overhead caused by distribution.

B. Related work

An important part of the naive Bayes [13] algorithm is probability estimation. This part of the algorithm can be implemented in various ways. Perfect probability estimation requires complete knowledge of the data distribution of the measured data. For most scenarios, this is no feasible requirement. For practical purposes a sufficiently accurate probability estimation is needed, but the WSN platform limits the algorithms that can be chosen.

A straightforward choice for the probability estimation part of the naive Bayes classifier is the use of histograms [5]. This works by first dividing the input space for each input in a number of intervals. The second step is to determine for each interval how many samples belong to each class. These values give an estimate of the data distribution of the samples for each class over the intervals, which is needed for naive Bayes. Benefits of this approach for WSNs are: it does not require floating point operations, it is computationally inexpensive, no a-priori knowledge of the data distribution is required. These aspects make histogram based approaches suitable for WSN implementations.

The method in which the histogram borders are defined, however, is of great importance. Different methods can give very different results [10]. The most basic approach divides the input space in intervals of equal width. Given an uniform data distribution this works very well, with other data distributions like Gaussian distributions, however, sparsely populated intervals can lead to a decline in classification robustness [10]. Small random variations in training data can have a significant influence on the classification output.

A method that improves on this problem is the so called binning histogram approach [10]. In this approach, the input space is divided in equally populated intervals, thereby ensuring that each interval contains a relevant amount of samples. A drawback of this approach is that in order to obtain equally populated intervals, the training data needs to be stored and sorted. This is not a task that can be executed within the memory constraints provided by WSN hardware.

In this work, we propose the use of unsupervised learning techniques to create a suitable partitioning of the input space. Our hypothesis is that unsupervised learning techniques can obtain results similar to the binning histogram approach, without exceeding the limits of WSN hardware.

II. Method

The first step in our work was an investigation using an offline dataset. We looked into the performance of binning and fixed width histograms and compared these with a number of trained classifiers using unsupervised learning. We investigated two different unsupervised learning algorithms: Self Organizing Maps (SOM) [8] and K-Means [9]. These algorithms were chosen because of their suitability for WSN implementations. We do not claim that these two algorithms are the best choice, but they are well known and suitable to proof the concept of our approach.

To compare the different classifiers in multiple situations, we used the distance to the Receiver Operator Characteristic (ROC) center line for each classifier [6]. This distance provides a metric for a classifier's capability to discriminate between multiple classes, regardless of the bias between these classes [14]. We determined the mean ROC distance and the standard deviation for all classifiers over ten training runs.

In each training run we trained the classifiers using 5000 samples from each class, or if one of the classes was rather rare then we used the maximum amount of samples we could take without using more than half of the total samples of a class for training. We used this limit to ensure that there was enough data, not used during training, to validate the performance of the classifier.

For the K-Means algorithm we used the implementation provided by Matlab, for the SOM we created a custom implementation. We chose to implement the SOM algorithm ourselves in order to gain experience for the experimental validation described later in this section.

The dataset used in this investigation was made for previous research [14], [15]. It is a dataset from multiple sensors situated around a refrigerator and coffee machine in the social corner of our research group. Three different states were manually labeled for this dataset, namely: the state of the the cooler of the refrigerator, the state of the coffee machine and the state of the fridge door.

We trained classifiers for each of the different states, using all the different approaches for interval determination.

A. Experimental verification

After the experiments with the offline dataset showed promising results, we verified these results using an experimental implementation. The platform for this experiment provided all the complications found on WSNs. More specifically, we wanted to validate our approach on a platform with a cheap 8-bit microcontroller, a low power radio and a network topology where nodes could have a hop-distance of at least three.

On this platform, we demonstrated the feasibility of implementing an unsupervised learning algorithm to determine histogram intervals within the given constraints. Furthermore, we demonstrated that these intervals can be used as a base for a naive Bayes classifier running on such a network. We used the distribution scheme proposed in [15] to make multiple sensor nodes collaborate in a distributed naive Bayes classifier.

We chose a classification task for which we could easily provide automatic training information: the presence of people in an office. The capability to detect this kind of information is useful for many home-automation and safety applications.

The exact details of our implementation are described in Section III-B.

III. Results

This section describes the results gathered in this research.

A. Offline results

Figures 1 to 3 show the results from our tests on the offline dataset. In addition to the fixed width histogram approach and the binning histogram approach we tested the K-Means algorithm and SOMs.

Figure 1 shows the result for the fridge running event. This event was the most common in the used datasets, these results were all obtained using 10000 samples.

Figure 2 shows the result for the coffee machine running event. This event occurred less frequent then the fridge running event. The amount of positive samples limited our training set size to 5703 samples.

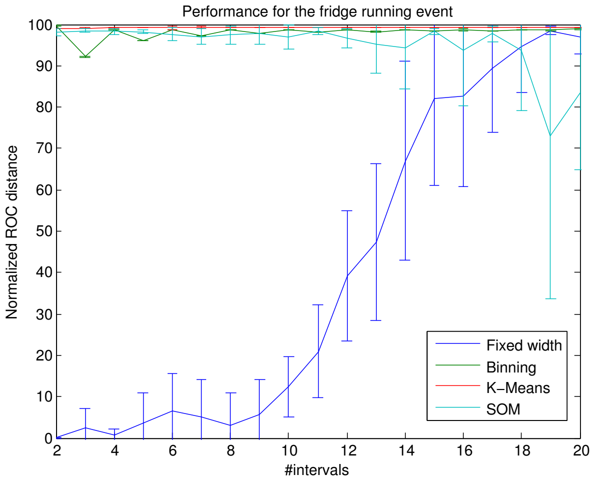


Figure 1. Fridge running performance

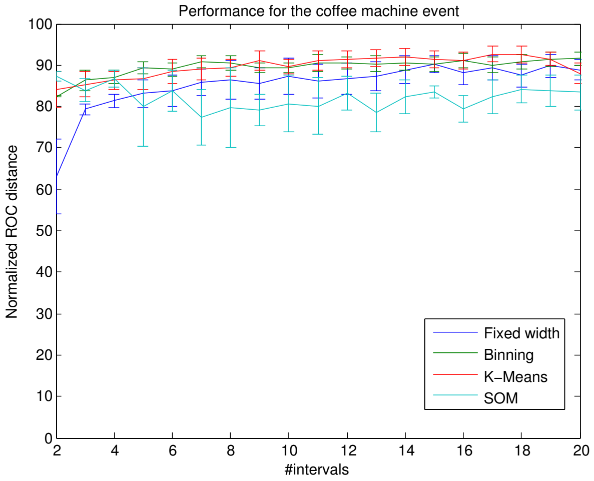


Figure 2. Coffee machine performance

Figure 3 shows the results for the fridge open event. This event was the rarest in our dataset. The amount of positive samples limited the training set size to 435 samples.

B. *Implementation*

As a platform for our implementation, we have chosen the Arduino Nano [1] experiment board. Arduino is a cheap platform for experimentation with electronic circuits. The Arduino Nano is equipped with an ATmega328 microcontroller, which has a limited set of 8-bit instructions and 2KB of SRAM (see Table I). We consider these specifications a realistic representation of WSN hardware.

For the radio, we have attached a NRF24L01+ [2] to the Arduino Nano (see Table II) . The NRF24L01+ is a low power 2.4Ghz radio that is well supported on the Arduino platform . We used the RF24Network library [3] to create a tree topology with a maximum hop distance of four. The topology of our network is shown in Figure 4. Also, the types

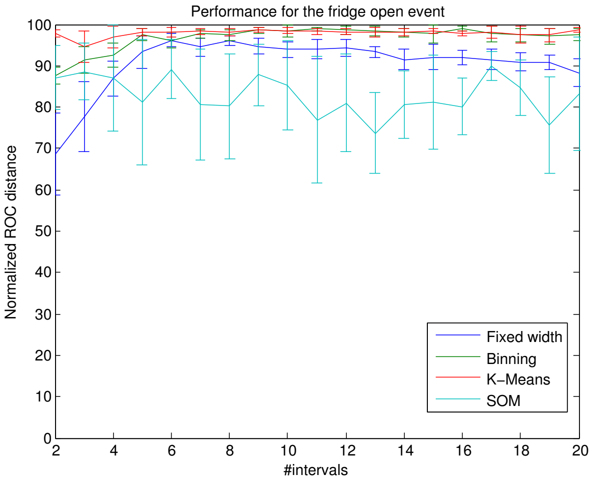


Figure 3. Fridge open performance

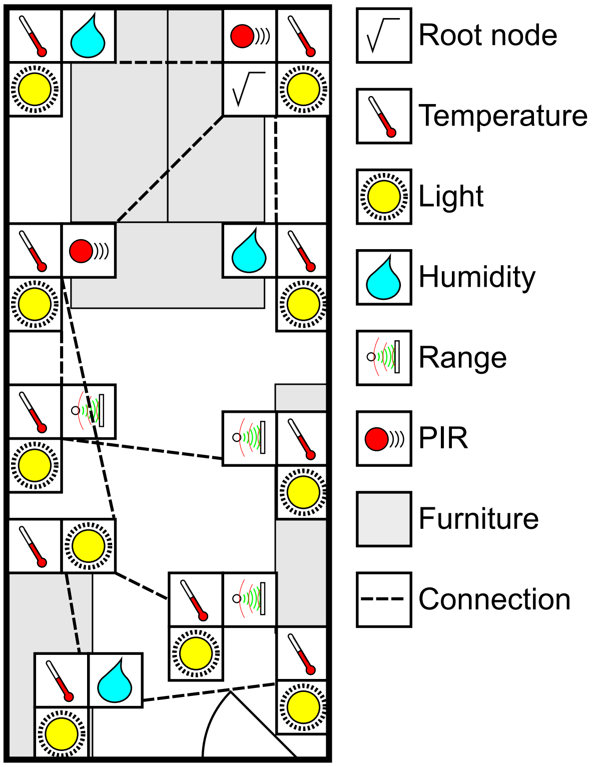


Figure 4. Network topology in office

of sensors with which each node is equipped are shown in Figure 4. The root node in this network is where all data is combined in a final classification. Each node transmits a combination of its local estimations and that of its children to its parent. Also, the types of sensors with which each node is equipped are shown in Figure 4.

We deployed ten sensor nodes around an office, using magnets to attach the nodes to white-boards and beams in the ceiling. This allowed for quick maintenance and

Type	ATmega328
Clock speed	16Mhz
Program memory	32KB
SRAM	2KB
EEPROM	1KB

Table I
MICRO-CONTROLLER

Type	NRF24L01+
Band	2.4Ghz
Data rate	2Mbps
Voltage	1.9-3.6V
Current	<13.5mA

Table II
RADIO

deployment. The ten sensor nodes were equipped with a number of different sensors. All nodes were equipped with LM35 temperature sensors and photo transistors to act as light sensors. Three nodes were furthermore equipped with ultrasonic range finders, three other nodes with humidity sensors. Finally two of the nodes were equipped with Passive Infra-Red (PIR) sensors to provide training information about the presence of movement in the office.

Each node was equipped with at most three sensors that were used by the Bayes classifier, the PIR sensors were just used as training feedback. We sampled each sensor at an arbitrary rate of 5Hz, and distilled three features from each sample stream, namely, the average value over the last second, the peak value over the last second and the slope over the last second.

We chose Self Organizing Maps (SOM) as the unsupervised algorithm to implement because these tend to divide the input space in equally populated partitions, making the expected results similar to the binning algorithm. Each node trained a SOM for each feature derived from each sensor, meaning there were up to nine SOMs per node. We chose four neurons as the size of the SOM, this number was a result of the memory constraints of the Arduino Nano.

For the distribution scheme, we used the method proposed in [15], meaning that each node had to send a single message to its parent in the network topology for each classification. We let the network make one classification per second.

Our implementation uses ten bytes of memory per sensor to buffer the sample data, 16 bytes per SOM, and 32 bytes to store each Bayes histogram. Given a typical node in our setup has three sensors with three features per sensor, our algorithm uses 462 bytes of memory. Each node uses the available EEPROM to periodically store the trained classifier.

We let our experimental setup run for a couple of days, where each node used its local sensors to create a SOM of its input space and used the feedback from the two PIR sensors to train its naive Bayes classifiers. We did not gather exact result of the accuracy of our platform, but we did periodically check if the networks classification output matched the real presence of people in the office. We noted a gradual improvement in classification performance as the training proceeded, a clear sign that our approach works as intended.

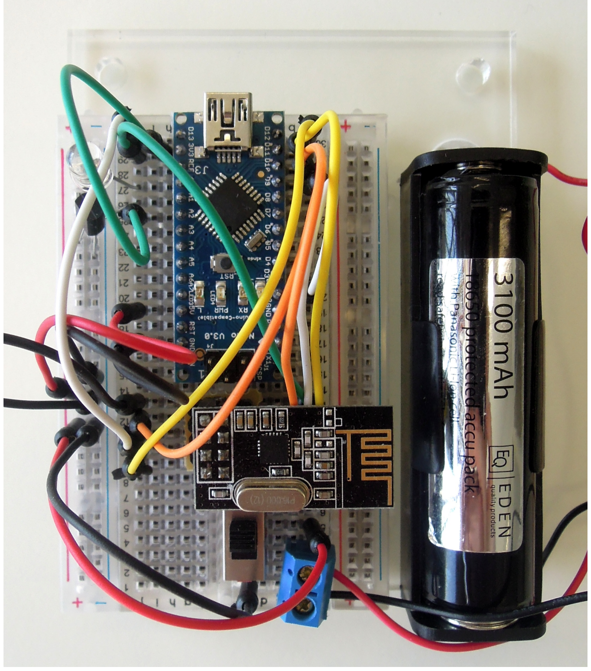


Figure 5. Sensor node

IV. ANALYSIS

In this section, we analyze the results provided in the previous section.

A. *Offline results*

As shown in Figures 1 to 3, binning histograms clearly give a better performance than the fixed width histograms. Only with a large amount of intervals, the performance of the fixed width histograms starts to improve. This is a clear indication that careful selection of the interval borders can result in improved classification performance.

For the fridge running event, both unsupervised learning algorithms show performance similar to the binning histogram approach and far superior to the fixed width histogram approach, especially with a low number of intervals. For the two rarer events, K-Means shows results similar to the binning approach, our SOM implementation however shows some less favorable results with a larger number of intervals.

This result can be explained by the way the Matlab version of K-Means is implemented. In Matlab K-Means works with an iterative process repeating the training until a nearly optimal result is achieved. Our SOM implementation, however, uses each sample only once, which is more realistic when considering WSN implementations. Given a large enough training set both solutions will converge on a correct partitioning, which explains the results for the fridge running event. For smaller training sets, however, the SOM has not yet converged to a stable solution which explains the decreased performance for these two events. The lower performance that can be seen for higher number of intervals can be explained in a similar way: the higher the number of intervals, the lower the number of samples that is used to train each interval. We expect that given enough samples, binning histograms, K-Means and SOMs will show the same behavior.

For practical implementations, this behavior is not problematic. Each sensor node only uses its local data to create the SOM, therefore each node can gather the information needed to create its SOM without using its radio. This means that the energy needed to create the SOM is minimal and all that is needed to determine the correct interval borders is patience.

A result that is clear for all events is that when using a low number of intervals, unsupervised learning algorithms can help create a much better naive Bayes classifier. This reduction in the number of intervals means that this approach can be implemented using even less memory than the fixed width interval approach.

B. Experimental verification

Our experimental implementation shows that our proposed solution is possible within the memory and computational constraints of WSN hardware. Our implementation runs on a simple micro controller with 2KB of RAM and an 8-bit instruction set. The used SOM algorithm automatically creates a suitable partitioning of the input space for each sensor. Given proper training data, we expect that the performance showed in the offline tests can be achieved in real life.

Our current implementation does not gather accuracy results, we could only visually check if the network was performing the desired task. As a proof that the algorithm was working as desired, this was sufficient. Especially considering that we did not investigate the optimal learning method for the SOM.

We consider the use of Arduino hardware for WSN experiments very successful. Arduinos are a versatile platform to which a variety of sensors can be attached. The addition of a radio allowed us to create a real sensor network. One limitation of the Arduino platform is energy usage. It was not designed with energy efficiency in mind and, for example, contains some LEDs that cannot be turned off.

V. Future work

Although this work gives some valuable insights in the application of clustering algorithms on naive Bayes classifiers, we have by no means created a system that can be directly applied on real life problems. This section describes some areas of research that need work before real life applications can be made.

Network maintenance and installation: While theoretical analysis of algorithms and practical experiments can provide valuable insight in the performance of algorithms on WSNs, real life deployments are far more dynamic. Deployment of the WSN and the training of a classification algorithm on such a network in an unaccessible environment combined with the complexities of replacing defective sensor nodes with new nodes in a trained classification network are matters that need to be investigated. We are working on research where we investigate the entire life cycle of a classification network and assess the complexities to use various algorithms during all the phases of this cycle.

Algorithm optimization: Although this paper shows promising results, we have not looked into the optimal settings for all the parameters. Different learning functions for the unsupervised algorithms, for example, can probably improve the speed with which the classifier learns, or the accuracy. Our offline results, for example, showed that the K-Means algorithm had a better performance than the SOM algorithm, especially for smaller training sets. Although this is probably caused by the fact that we used the standard Matlab toolbox for K-Means which uses the training data in multiple iterations, it is clear that it had better results with the same data. These are aspects that need to be investigated.

Time aspects: Evolution of conditions over time is an important consideration in the training of classifiers. In this research, we have looked into classifications from moment to moment. Feedback from previous classifications, however, could provide valuable information to improve classification performance. This feedback would change the structure of the algorithms. The effects of this change on the options for distribution is another direction of future research.

VI. Conclusion

In this work, we have demonstrated the merits of using unsupervised learning algorithms to determine histogram borders for naive Bayes. This approach can result in a significant improvement of the classifier over the traditional fixed width histogram approach, both on performance and on robustness. When given enough training data, this approach matches the performance of the binning histogram approach, without excessive memory or computational requirements. It should be noted that it is important to use enough data to train the unsupervised learning algorithms, if the unsupervised algorithm has not converged to a stable state undesirable results are possible.

We have demonstrated the validity of our offline results with an implementation on a realistic WSN platform. Next to the validation of our approach, our implementation demonstrates that the Arduino platform is an accessible platform for WSN experiments. The fact that our implementation was based on cheap and readily available components makes us recommend this approach for other research on WSNs.

REFERENCES

[1] Arduino. http://www.arduino.cc [retrieved: June, 2012].

[2] Nrf24l01+. http://www.nordicsemi.com/eng/Products/2.4GHz-RF/nRF24L01P [retrieved: June, 2012].

[3] Rf24network library. https://github.com/maniacbug/RF24Network [retrieved: April, 2012].

[4] M. Bahrepour, N. Meratnia, and P. J. Havinga. Automatic fire detection: A survey from wireless sensor network perspective. Technical report, Centre for Telematics and Information Technology, 2007.

[5] M. Bahrepour, N. Meratnia, and P. J. Havinga. Fast and accurate residential fire detection using wireless sensor networks. *Environmental Engineering and Management Journal*, 9(2):pp. 215–221, 2010.

[6] A. P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):pp. 1145–1159, 1997.

[7] N. Chohan. Hardware assisted compression in wireless sensor networks. 2007.

[8] S. Haykin. *Neural Networks: a comprehensive foundation*. Prentice Hall, 2 edition, 1999.

[9] A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):pp. 264–323, 1999.

[10] S. Kotsiantis and D. Kanellopoulos. Discretization techniques: A recent survey. *International Transactions on Computer Science and Engineering*, 32(1):pp. 47–58, 2006.

[11] E. Tapia, S. Intille, and K. Larson. Activity recognition in the home using simple and ubiquitous sensors. *Pervasive Computing*, 3001:pp. 158–175, 2004.

[12] M. Vieira, J. Coelho, C.M., J. da Silva, D.C., and J. da Mata. Survey on wireless sensor network devices. *Emerging Technologies and Factory Automation*, 1:pp. 537–544, 2003.

[13] H. Zhang. The optimality of naive bayes. *17th Florida Artificial Intelligence Research Society Conference*, 2004.

[14] A. Zwartjes, M. Bahrepour, P. J. Havinga, J. L. Hurink, and G. J. Smit. On the effects of input unreliability on classification algorithms. *8th International ICST Conference on Mobile and Ubiquitous Systems*, 2011.

[15] A. Zwartjes, P. J. Havinga, G. J. Smit, and J. L. Hurink. Distribution bottlenecks in classification algorithms. *The 2nd International Symposium on Frontiers in Ambient and Mobile Systems (FAMS)*, August 2012.

A Distributed Infrastructure for Real-time Continuous VOC Monitoring in Hazardous Sites

Gianfranco Manes, Giovanni Collodi
Francesco Chiti, Romano Fantacci
MIDRA Consortium
Via di S.Marta 3, 50100 Florence, ITALY
gianfranco.manes@unifi.it; giovanni.collodi@unifi.it
francesco.chiti@unifi.it; romano.fantacci@unifi.it

Rosanna Fusco, Leonardo Gelpi
Health, Safety, Environment and Quality, eni SpA
Piazzale E. Mattei, 1 00144 Rome, ITALY
rosanna.fusco@eni.com; leonardo.gelpi@eni.com

Antonio Manes
Netsens Srl
via Tevere 70, 50019 Sesto Fiorentino (FI), ITALY
antonio.manes@netsens.it

***Abstract*-The real deployment is described of a distributed point source monitoring system based on wireless sensor networks in an industrial site where dangerous substances are produced, used and stored. The system consists of a Wireless Sensor Network using Photo-Ionisation Detectors, continuously monitoring the Volatile Organic Compound concentration in the petrochemical plant at unprecedented time/space scale. Internet connectivity is provided via TCP/IP over GPRS gateways at a one-minute sampling rate; thus, providing plant management and, possibly, environmental authorities with an unprecedented tool for immediate warning in case of critical events. The platform is organised into sub-networks, each including a gateway unit wirelessly connected to the WSN nodes, hence providing an easily deployable stand-alone infrastructure featuring a high degree of scalability and reconfigurability, with minimal intrusiveness or obtrusiveness. Environmental and process data are forwarded to a remote server and made available to the authenticated users through a rich user interface that provides data rendering in various formats and worldwide access to data. Experimental results show an excellent efficiency of the WSN system in terms of communication, making it a very flexible and cost-effective tool for environmental monitoring issues.**

Keywords–distributed VOC monitoring; wireless sensor networks; photoionisation detectors.

I. INTRODUCTION

Volatile Organic Compounds (VOCs) are widely used in industries as solvents or chemical intermediates. Unfortunately, they include components that, if present in the atmosphere, may represent a risk factor for human health. VOCs are also found as contaminants or by-products in many processes, i.e., in combustion gas stacks and groundwater clean-up systems. Detection of VOCs at sub-ppm levels is, thus, of paramount importance for human safety, and, consequently, critical for industrial hygiene in hazardous environments [1][2]. The most commonly used portable field instruments for VOC detection are the hand-held Photo-Ionisation Detectors (PIDs), which may be fitted with pre-filter tubes for specific gas detection. Wireless hand-held PIDs have recently become available on the market, thus providing ubiquitous operation, but they have a limited battery life, in addition to being relatively costly. This paper describes the implementation and on-field results of an end-to-end distributed monitoring system using VOC detectors, capable of performing real-time detection of gas emissions in potentially hazardous sites at minute data rate [3][4]. This paper describes the implementation of a distributed network for precise VOC monitoring installed in a potentially hazardous environment. The system consists of a WSN infrastructure with nodes equipped with weather-climatic sensors, as well as VOC detectors and fitted with TCP/IP over GPRS gateways to forward the sensor data via Internet to a remote server. The continuous monitoring of benzene emissions from a benzene storage tank is also demonstrated, using a unique wired/wireless configuration installed in ATEX Zone 0.

A user interface then provides access to the data, while offering various formats of data rendering. This prototype was installed in the eni Polimeri Europa (PEM) chemical plant in Mantova, Italy, where it has been in continuous and unattended operation since April 2011. The pilot site is currently testing and assessing both communications and VOC detection technologies.

II. SYSTEM OVERVIEW

A. The distributed VOC monitoring infrastructure

The distributed point sources method was chosen for this application as it provides reasonable installation and maintenance cost, a high scalability/reconfigurability and real-time data acquisition. A general overview of the deployed system is represented in Fig. 1.

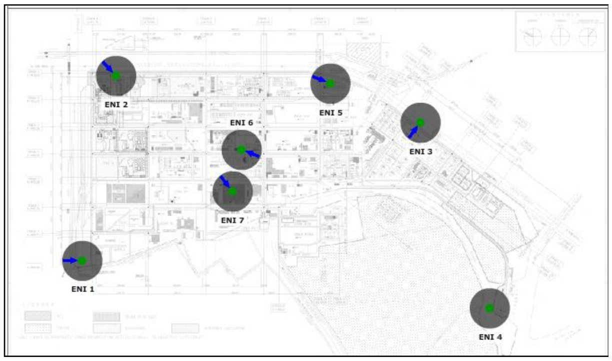


Figure 1. Installation overview; the grey circles indicate the position of each SNU; the blue arrows show wind direction.

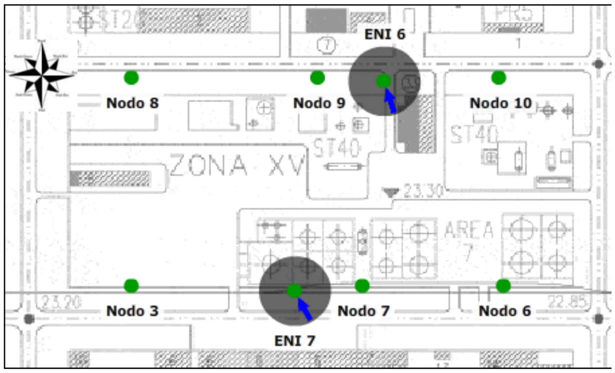


Figure 2. Close-up of SNU and ENU deployment around one of the chemical plants (left) and the pipeline (right); maps are oriented according to plant's axes rather than cardinal directions.

First off, representative locations were identified along the perimeter of the industrial area, along with several specific internal sites where hazardous emissions might potentially occur. Owing to the extension and complexity of the Mantova plant, covering some 300 acres and featuring complex metallic infrastructures, it was decided to subdivide the area involved into 7 different sub-areas. Each sub-area is covered by a sub-network consisting of a Sink Node Unit (SNU) equipped with weather-climatic sensors, such as wind speed/direction and relative air humidity/temperature (ENI 1 to ENI 7 in Fig. 1), and End Node Units (ENUs) equipped with VOC detectors. In addition, the ENI 2 unit is equipped with a rain gauge and a solar radiation sensor.

Each SNU is connected to one or more ENUs (see Fig. 2, for an example of configuration), appropriately distributed across the plant area. This modular approach allows the system to be expanded and/or reconfigured according to the specific monitoring requirements, while providing redundancy in case of failure of one or more SNUs.

Since the potential sources of VOC emissions in the plant are located in well-identified areas, such as the chemical plant and the benzene tanks, the deployment strategy includes a number (6) of VOC sensors surrounding the chemical plant's infrastructure, see Fig. 2, thus resulting in a virtual fence capable of effectively detecting VOC emissions on the basis of the concentration pattern around the plant itself. The SNUs forward meteorological data, as well as VOC concentration data, to a remote server; as noted above, Internet connectivity is provided via TCP/IP over GPRS using the GSM mobile network. Wireless connectivity uses the UHF-ISM unlicensed band. Electrical power is provided by both primary sources (batteries) and secondary sources (photovoltaic cells), as mentioned above.

VOC concentration and weather-climatic data are updated every minute. This intensive sampling interval allows the evolution of gas concentrations to be accurately assessed. Furthermore when all the weather-climatic measurements are collected, they provide a map of the relative air humidity/temperature (RHT) and wind speed/direction (WSD) in the area, that are crucial for providing accurate VOC-sensor read-out compensation [5]. The need for so many wind stations across the plant property is justified by the turbulent wind distribution in the area, as it can be observed by the different orientations of the blue arrows representing wind direction in Fig. 1.

Three of the ENUs, ENI 1, ENI 2 and ENI 3 were deployed along the perimeter of the plant to locally monitor VOC concentration while correlating it with wind speed and direction; the other seven were placed around the chemical plant and in close proximity of the pipeline, that are possible sources of VOC emissions.

In Fig. 2, the layout of the two sub-networks deployed around the chemical plant is represented. The sub-networks consist of two SNUs, ENI 6 and ENI 7, equipped with weather sensors (air/wind), each connected with three ENUs spaced at tens of meters from each other. The third sub-network located in the pipeline area consists of two ENUs located in close proximity of the end of the pipeline and connected to ENI 5. Sampling the VOC concentration at intervals of tens of meters allows the dispersion of VOC emissions to be evaluated; in addition, information about wind speed/direction allows the emission's source to be identified.

B. *Storage Tank monitoring infrastructure*

Storage tanks represent a potential source of VOC emissions and, thus, need to be appropriately monitored. The emissions from tanks can vary significantly depending on the size and design, liquid properties, tank maintenance, tank level, wind direction, wind speed, and whether the tank is filling, stable, or emptying. Benzene storage tanks in this settlement are of floating roof type, and are located in highly hazardous areas; the electrical equipments to be operated in those areas need a special safety certification for use in potentially explosive atmospheres. Certification ensures that the equipment or protective system meets the safety requirements and that adequate information is supplied with it to ensure that it can be used safely.

The lay-out of the Storage Tank Monitoring Network (STMN) is displayed in Fig. 3. The STMN consists of three Volatile Organic Compound (VOC) units, each equipped with a Photo-Ionisation Detector (PID) and a computational unit, serially interconnected by wire and connected to the Wireless Unit (WU), providing both power and wireless connectivity.

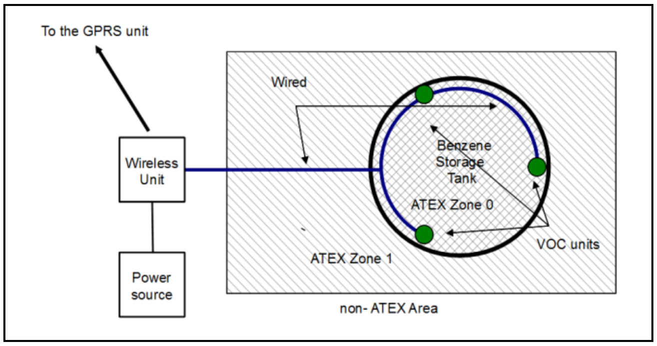


Figure 3. Schematic of the Storage Tank Monitoring Tank.

The WU is connected to the GPRS unit (ENI 3 in Fig. 1), which provides Internet connectivity via TCP/IP over GPRS.

The reason for choosing such a hybrid wired/wireless configuration is related to the VOC detector energy budget.

The PID, in fact, when operated at very low VOC concentration levels and in diffusion mode, requires to be continuously powered-on. Discontinuously operating the PID results in stabilisation times of the order of several tens of minutes for achieving a stable and reliable read-out.

That feature conflicts with the requirement of sampling the VOC concentration at minute data rates, a mandatory constrain for this application. Accordingly, it turns-out that the most suitable operation mode for the PID is the continuous power-on mode. In that mode, the current consumption for the VOC unit rises up to 50 mA, thus resulting in the need of a primary energy source with at least 80Ah capacity for ensuring a 60 days battery life; the requirement of bimonthly replacing 3 batteries on the top of the storage tank with skilled personnel was considered unpractical and too costly. On the other hand, the option of equipping the unit with a secondary energy source like a photovoltaic panel, thus prolonging the battery life, was discharged as it is hardly compatible with safety requirements of operation in ATEX Zone 0 and installation/maintenance above the benzene thank.

As a result, the hybrid configuration of Fig. 3 was envisaged, where the VOC units and communication/power supply units are split.

As it can be observed in Fig. 3, the three VOC detectors are located in Zone 0, very high level of protection, while the WU, along with the power unit consisting of the battery and the photovoltaic panel, is located in the non-ATEX area.

This allows for using a secondary energy source and for easily replacing the primary energy source as required by the maintenance programme.

III. THE COMMUNICATION PLATFORM

The distributed communications platform, already described [3][4], is able to support a scattered system of units collecting VOC emission data in real-time, while offering a high degree of flexibility and scalability, allowing for other monitoring stations to be added, as needed.

Furthermore, it provides reconfigurability, in terms of data acquisition strategies, while being more economically advantageous than traditional fixed monitoring stations.

A GSM mobile network solution featuring a proprietary TCP/IP protocol with DHCP, provides Internet connectivity. Dynamic re-connectivity strategies provide efficient and reliable communication with the GSM base station. All the main communication parameters, such as IP address, IP port (server's and client's), APN, PIN code and logic ID can be remotely controlled. Wireless connectivity between SNUs and ENUs is preformed in the unlicensed ISM UHF band (868 MHz).

A. The SN and WI units

The block diagram of the SNU is represented in Fig. 4. It consists of a GPRS antenna, a GPRS/EDGE quadriband modem, a sensor board, an I/O interface unit, and an ARM-9 micro-controller operating at 96 MHz.

The system is based on an embedded architecture with a high degree of integration among the different subsystems.

The unit is equipped with various interfaces, including LAN/Ethernet (IEEE 802.1) with TCP/IP protocols, USB ports and RS485/RS422 standard interfaces. The sensor board is equipped with eight analogue inputs and two digital inputs. The SNU is also equipped with a Wireless Interface (WI), shown in Fig. 4 right, which provides wireless connectivity with the ENUs.

B. The EN unit

The block diagram of the EN is shown in Fig. 5; it consists of a VOC sensor board and a VOC detector. The acquisition/communication subsystem of the ENU is based on an ARM Cortex-M3 32-bit micro-controller, operating at 72 MHz, which provides the necessary computational capability on the limited power budget available.

The block diagram of the ENU is shown in Fig. 5; it consists of a WI, similar to that previously described, and includes a VOC sensor board and a VOC detector. The acquisition/communication subsystem of the ENU is based on an ARM Cortex-M3 32-bit micro-controller, operating at 72 MHz, which provides the necessary computational capability on the limited power budget available.

To reduce the power requirement of the overall ENU subsystem, two different power supplies have been implemented, one for the micro-controller and one for the peripheral units.

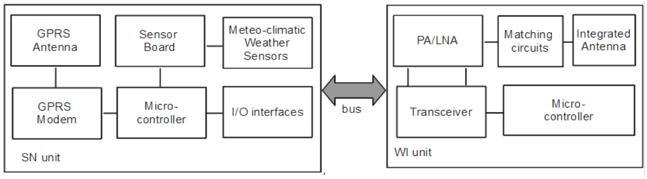


Figure 4. Block diagram of an SNU (left) and a WI unit (right).

The microcontroller is able to connect/disconnect the peripheral units, thus preserving the local energy resources.

The VOC detector subsystem is powered by a dedicated switching voltage regulator; this provides a very stable and spike-free energy source, as required for proper operation of the VOC detector itself.

The communication between the ENU and the VOC detector board is based on an RS485 serial interface, providing high-level immunity to interference as well as bidirectional communication capability, which is needed for remote configuration/re-configuration of the unit.

C. Network structure and routing schemes

A multiple GPRS gateway approach overcomes those limitations; even in the case of failure of one or more gateway units, Internet connectivity would be provided by the others still in operation, while the issue of the obstacles is circumvented.

Concerning the lower tier comprised of ENUs, it has been adopted the IEEE 802.15.4 standard to implement the communications features. This is motivated by the interoperability properties and flexibility when operating under different conditions [10]. In particular, the adopted MAC layer scheme follows the beacon enabled approach, in which a *coordinator* (SNU) periodically broadcasts a `beacon` packet for synchronizing the other nodes (ENUs) and arbitrating the access to the wireless shared medium through CSMA/CA protocol.

In designing the routing protocol, the IPv6 Routing Protocol for Low Power and Lossy Networks (RPL) has been adopted [11]. This recently standardized approach has been proposed to meet the forwarding requirements for Low Power and Lossy Networks (LNNs). In particular, RPL is a Distance Vector IPv6 routing protocol for LLNs that specifies how to build a Destination Oriented Directed Acyclic Graph (DODAG, sometimes referred to as a graph in the rest of this document) using an objective function and a set of metrics/constraints to prevent loops creation. The objective function operates on a combination of metrics and constraints to compute the 'best' path. There could be several objective functions in operation on the same node and mesh network because deployments vary greatly with different objectives and a single mesh network may need to carry traffic with very different requirements of path quality, involving, for example, latency, throughput, battery consumption or load balancing issues.

As for the wireless connectivity, a star configuration was preferred to a mesh configuration, given the limited number of nodes and the need to keep latency at a minimum.

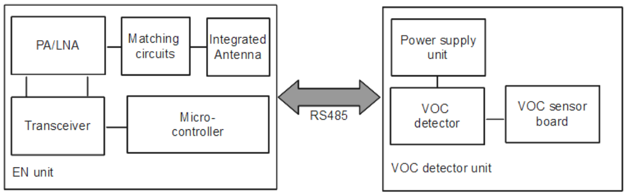


Figure 5. Block diagram of the ENU (left) and the VOC detector unit (right).

D. Protocols and WSN services

Two levels of communication protocols, in a mesh network topology, were implemented. The upper level handles communications between the SNUs and the server; it uses a custom binary protocol on top of a TCP layer. This level was designed and calibrated for real-time bidirectional data exchange, where periodic signalling messages are sent from both sides. Since our sensor network necessitates a stable link, quick reconnection procedures, for whenever broken links should occur, were especially important. To ensure minimal data loss, the SNUs have non-volatile data storage, as well as automatic data packet retransmission (with timestamps) after temporary downlink events. Furthermore, this design is well suited for low-power embedded platforms like ours, where limited memory and power resources are available. In fact, our protocol stack currently requires about 24 KB of flash memory (firmware) and 8 KB RAM.

The lower level, in contrast to the upper one, concerns the local data exchange between the network nodes. Here a cluster tree topology was employed; each node, which both transmits and receives data packets, is able to forward packets from the surrounding nodes when needed. In this specific application, the topology and routing schemes are based on an ID assigned to each EN unit, where the ID can be easily adjusted using selectors on the hardware board. This choice allows for easy support and maintenance, even when non-specialized operators have to install, re-install or service one or more units.

E. Energy budget issues

Energy budget plays a key role in the maintainability of the WSN [6]. In our case, this is made even more critical by the necessity of providing stand-alone operation with periodic maintenance intervals exceeding four months.

Since electrical energy from the plant could not be used, secondary sources had to be locally available; photovoltaic panels (PVP) fit the bill. The SNUs are all equipped with PVPs, as they have to support a number of functions, including connectivity and data collection from sensors. The ENUs, when equipped with low-energy demanding sensors, have 3 to 5 years of battery life using primary sources. However, in this installation the ENUs have to support the power-hungry VOC sensors. For this reason, the ENUs are also equipped with PVPs.

The ENUs have been fully deployed since May 2011; since that time we have noticed that the VOC sensor energy budget is predominant compared to that of the computational/communication unit. This is a critical issue for the ENUs, as the PIDs used for reading the VOC concentration need to be continuously powered-on to operate efficiently. The actual current drawn by the PIDs resulted in some 30 mA, corresponding to 720 mAh a day, almost twice the amount required by the communication/computational units, ranging to some 360 mW a day. The ENU's primary source capacity is 60 Ah, which provides more than 2 full months of continuous operation.

To rely on autonomous energy resources, while providing continuous operation, a secondary energy source was integrated into the ENU in order to supply the 360 mW+ 720 mW average power required. A 5 W photovoltaic panel can fulfil the task only under ideal sunlight conditions, i.e., in summer, but hardly at all in winter. The photovoltaic power supply unit includes a charge regulator, which was specifically designed to provide maximum energy transfer efficiency from the panel to the battery under any operative condition. Great attention was paid to the design of the voltage regulator, as the secondary energy source plays a key role in ensuring the stand-alone and unattended operation of the communication platform.

Fig. 6 displays the battery voltage plots of the ENUs connected to SNU 1, 5 and 6. As it can be observed, the ENUs exhibit quite satisfactory charge conditions. ENU 10 (eni 6 nodo 10) exhibits a lower voltage level, probably due to a deployment in a partially shadowed area. For the other two ENUs, the battery voltage remains above a 11.6 V value, with a slight decreasing trend, in the period December 2011-January 2012. due to the lower solar energy because of the onset of wintertime.

F. The VOC detector

The VOC detector is a key element for the monitoring system's functionality. For this application two criteria were considered mandatory. The first is that the VOC detector should be operated in diffusion mode, thereby avoiding pumps or microfluidic devices, which would increase the energy requirements and make the maintainability issues more critical. The second criteria was that the system should be able to operate in the very low part per billion (ppb) range, with a Minimum Detectable Level (MDL) of some 2,5 ppb with a ± 5% accuracy in the 2,5 to 1000 ppb range, which represents the range of expected VOC concentration. The Photo-Ionisation Detector (PID) fulfils most of the above requirements [7][8].

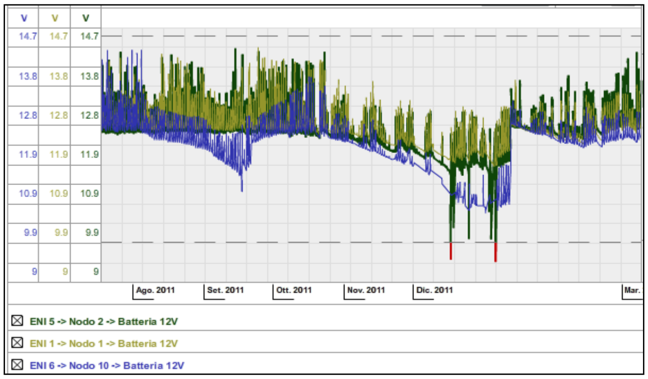


Figure 6. Battery voltage of the ENUs of the ENI 1, 5 and 6 sub-networks from July 18th 2011 to March 9th 2012.

A high sensitivity PID was chosen, featuring the specifications listed above [9]. Two major issues were identified; however, that could potentially affect the efficient use of the PID in our system. The first issue was that in the low ppb range the calibration curve of the PID shows a marked non-linearity; this would require an individualized, meticulous multipoint calibration involving high cost and complexity. The second issue was that, when operated in diffusion mode at low ppb and after a certain time in power-off, the detector requires a stabilisation time of several minutes, hence it wouldn't be able to operate at our required one-minute intervals.

Since both of the above-mentioned limitations are intrinsically related to the PID's physical behaviour, this was carefully investigated and a behavioural model of the PID was developed to explain these phenomena. As a result, a mathematical expression of the PID calibration curve was derived. Accordingly, the PID calibration procedure could be performed by measuring only two parameters i.e., the zero gas voltage and the detector sensitivity in mV/ppm.

IV. EXPERIMENTAL RESULTS

A. Long term operation

Data gathered from the field are forwarded to a central database for data storage and data rendering. For this purpose, the system has a web based interface for retrieving and displaying data and for post-processing.

The interface features different formats to display the gathered data. It is possible both access to raw data, and generate summary reports relating to specific periods and specific network areas. All monitored parameters can be geo-referenced.

Data from the individual sensors deployed on the field, either micro-climatic or VOC, can be directly accessed and presented in various formats. Fig. 7 shows the VOC concentration read-out of one of the VOC detector deployed in the PEM settlement. The graph shows undiscontinued operation over a period of ten months (May 2011-March 2012); the background concentration value is below 150 parts per billion (ppb), as expected.

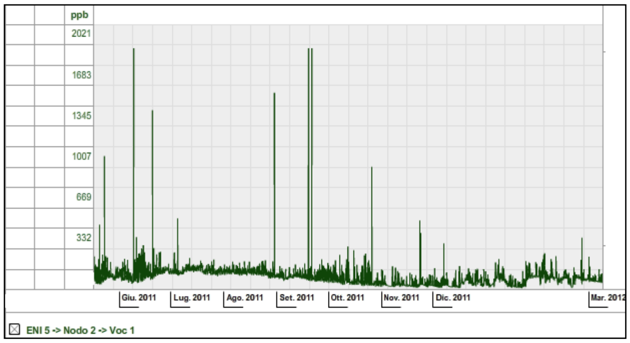


Figure 7. VOC concentration read-out during the period May 2011, March 2012.

The spikes observed in the graph are probably related to some operation and/or maintenance running in the plant.

Fig. 8 shows the trend of VOC concentrations values (detected by the six PIDs deployed around the chemical plant) in a monthly period of almost two months (January 15th 2012-March 10th 2012). Background measured values are coherently comparable to each other, demonstrating the effectiveness of the calibration procedure.

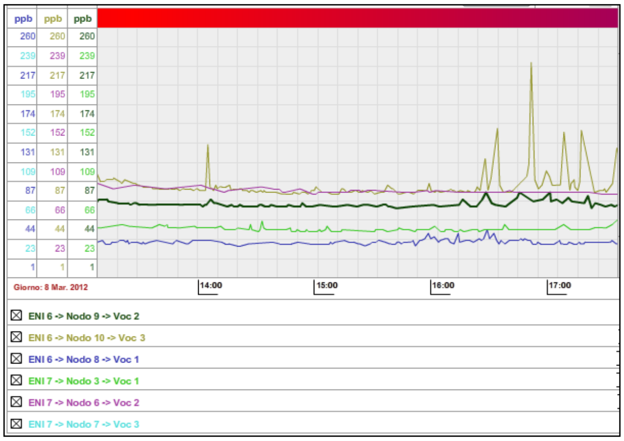


Figure 8. Graph of VOC concentration measured by six detectors deployed around the chemical plant (two months).

In Fig. 9, the VOC shows the concentration background, that is around 50 ppb; thanks to the very intensive sample-interval, 1 minute, the evolution of the concentration in time, along with other relevant meteo-climatic parameters can be very accurately displayed; it should be noted that the spikes, which can be observed in the blue trace, Fig. 13 left, have a duration of some 3 minutes. The multi-trace graphic feature is very useful to perform correlation between different parameters.

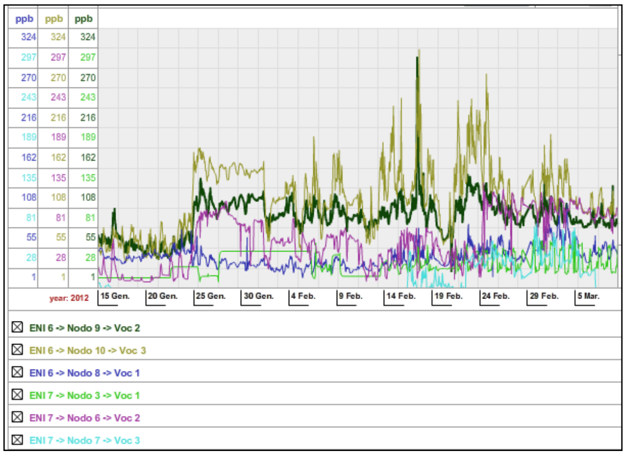


Figure 9. Graph of VOC concentration measured by six detectors deployed around the chemical plant (5 hours)

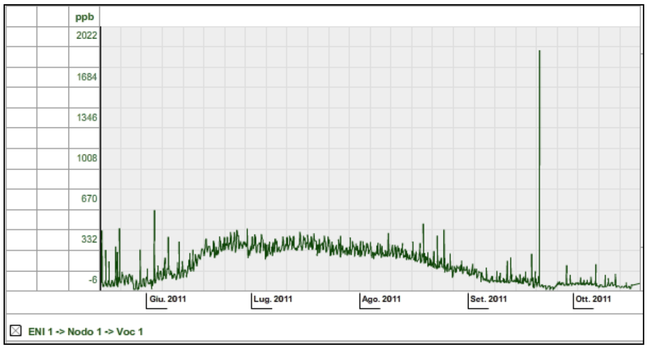


Figure 10. Graph of VOC concentration measured by a peripheral VOC detector over a period of four months.

Fig. 10 shows VOC concentrations in the long term (5 months) for a sensor positioned along the perimeter of the industrial area.

As it can be observed, data show an increase of the background value during the summer, due probably to higher temperatures.

Values tend to decrease from September. Peak value shown around 25th of September (concentration greater than 500 ppb) is due to meteorological conditions that may affect the dispersion of pollutants.

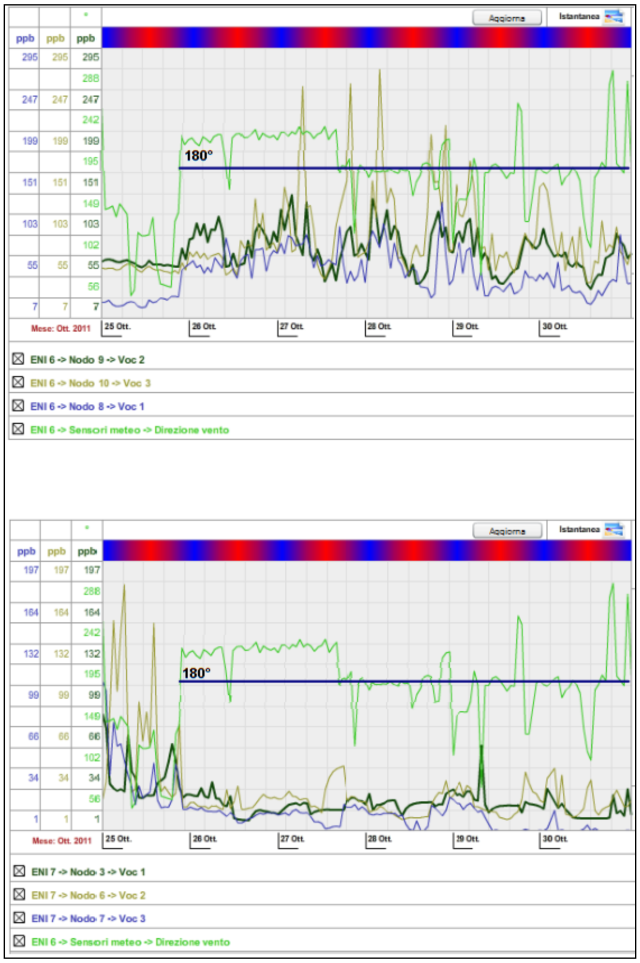
B. VOC concentration and weather-climatic variations

Correlating the microclimatic and wind parameters (air temperature/humidity and wind speed/direction) and VOC concentration proved to be very effective for increasing VOC read-out accuracy and, moreover, to map the VOC concentration with respect to wind direction, in order to identify possible VOC source.

When VOC sources need to be identified, in fact, the correlation between wind/speed direction and VOC concentration is vital. For this reason, a graphic representation that relates these two parameters can be very useful for interpretation of results.

An example of that possibility is given by the plots of Fig. 11. The graph represents the trend of VOC concentrations values (detected by the six PIDs deployed around the chemical plant) in a five days period.

With reference to the lay-out of Fig. 2, the VOC read-outs from the array deployed in the northern side of the plant are represented in Fig. 11 up, while the VOC read-outs from the array deployed in the southern side of the plant are represented in Fig. 11 down. The wind direction is also plotted in both the graphs of Fig. 11; it turns-out that in the most of the period the wind blows from south to north; By comparing the VOC read-out of the two arrays it is observed that the values of the former exhibits a much higher mean value than the latter; this is consistent with the direction north-south of the wind and demonstrates the effectiveness of the correlation between wind direction and speed to identify possible VOC sources.



☒ ENI 6 -> Nodo 9 -> Voc 2

☒ ENI 6 -> Nodo 10 -> Voc 3

☒ ENI 6 -> Nodo 8 -> Voc 1

☒ ENI 6 -> Sensori meteo -> Direzione vento

☒ ENI 7 -> Nodo 3 -> Voc 1

☒ ENI 7 -> Nodo 6 -> Voc 2

☒ ENI 7 -> Nodo 7 -> Voc 3

☒ ENI 6 -> Sensori meteo -> Direzione vento

Figure 11. Source identification by correlation between VOC concentration read-outs and wind direction.

In Fig. 12 different representations of VOC concentrations combined with the wind direction data is shown for one of the detectors located on the left side of the plant perimeter, namely ENI 1.

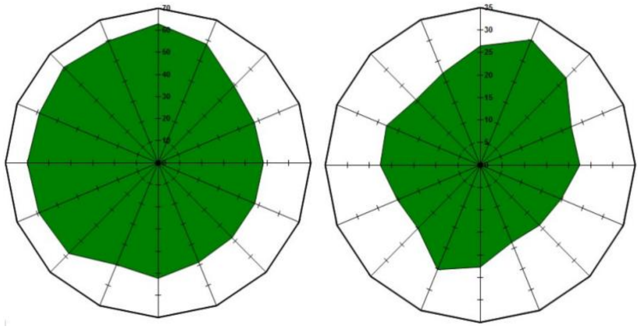


Figure 12. Plot in polar coordinates of the relationship between wind direction (angle) and concentration (radius).

The plot, in polar coordinates, represents the wind directions referenced to the North and the VOC concentration in ppb. This type of representation allows to quickly identify the wind directions corresponding to the higher levels of concentration, giving an overview of the predominant orientation of the VOC flux during the day.

The plot in Fig. 12 left represents the concentration along 24 hours in a working day, while the plot in the right represents the same on a Sunday. As the detector eni 1 is located at the western side of the plant, see Fig. 1, it turns out that the concentration in the III and IV quadrants of the plot represents the contribution of the VOC sources outside the plant, while the concentration in the I and II quadrants represents the contribution of sources inside the plant.

Sources outside the plant are very likely the benzene emission by vehicles running on the motorway in direction North-South, along the western side of plant, or possible emissions from other industrial sites

Thermohygrometric parameters are useful to compensate for the air temperature/ relative humidity variations. Fig. 13 represents the typical relative response of PID as a function of temperature and relative humidity. It turns-out that, particularly in summertime, the climatic conditions in the plant can significantly affect the accuracy of the detected VOC concentration and need to be compensated for. This seems to be confirmed by the observation that the emissions in the III for IV decrease on Sunday, with respect to the working day, and by the highest value of the emission, 60 ppb in the left against 30 ppb in the right.

Fig. 14 shows the effect of thermohygrometric parameters on VOC detector read-out on three different VOC detectors located in the southern side of the chemical plant; see Fig. 2.

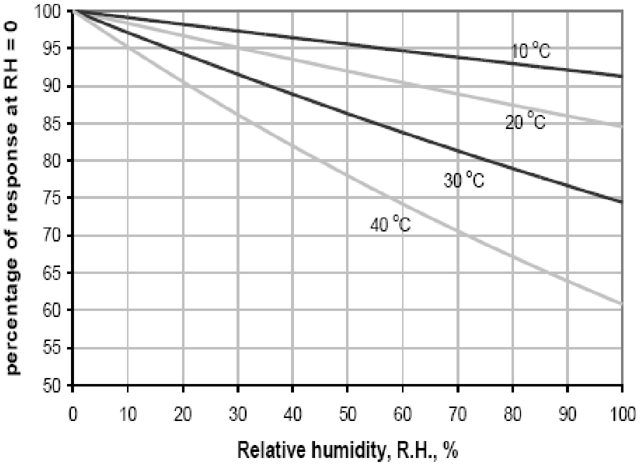


Fig. 13. VOC detector relative response as a function of thermohygrometric parameters (courtesy of Alphasense Ltd).

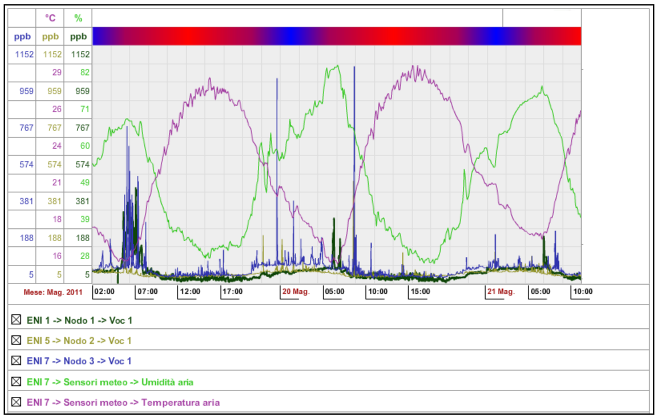


Figure 14. Correlation between air temperature/relative humidity and VOC detector read-out.

Maximum air temperature is as high as 30°C in the day, while air humidity reaches almost the 100% in the night. The predominant factor is the relative humidity; it is clearly observed that the VOC concentration values of all the three detectors follows the behaviour of the relative humidity, in accordance to what predicted by the plots of Fig. 13.

A post processing algorithm was implemented at the server side to compensate for the previously described effect. Fig. 15 shows the effect of the compensation, which results in smoothing the previously illustrated day/night effect.

C. Monitoring the benzene Tank

The network infrastructure displayed in Fig. 3 and previously described was operated starting from December 2011.

The VOC concentration recorded by the three detectors located on top the tank are plotted in Fig. 16, in the period December 11th 2011-March 15th 2012. The monitoring network installed in the proximity of the benzene storage tank roof proved to be very useful to identify and keep under control the process steps more significant in terms of emissions.

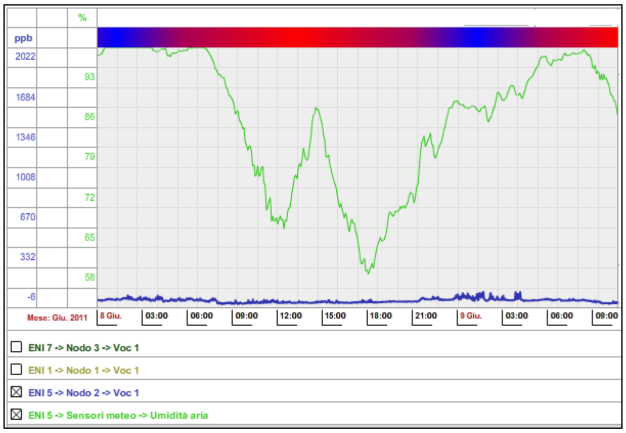


Figure 15. Compensation of the effect of thermohygrometric conditions.

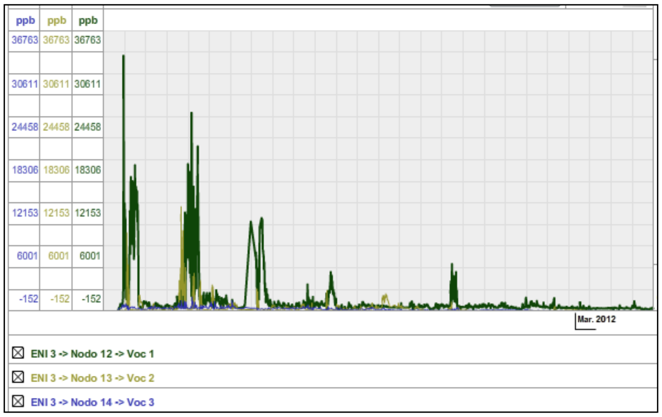


Figure 16. Graph of VOC concentration measured by three detectors deployed on top the benzene storage tank (December 11th 2011- March 15th 2012).

Periodic concentration peaks are observed, see Fig. 15, after complete filling of the tank; in such condition, the floating roof is located closer to the sensors and this allows to verify, along the perimeter of the roof, if there are sealing problems due to wear or deformation of the seals.

In the summer period the high volatility of the compounds could lead to variations in significant concentrations even during the emptying of the tank. The values found are in any case widely acceptable, since it is a direct emissions source.

D. Communications performance analysis

Before evaluating the performance of the proposed communications architecture, it has been addressed the available band the WSN is able to provide. In Fig. 17, a single-sink scenario where IEEE 802.15.4 nodes transmit data to the sink through one link (star topology), or possibly two-hops (3-level tree, rooted at the sink), is considered. A network composed of 30 nodes, working in beacon-enabled mode is accounted for. The throughput as a function of the size of the packets transmitted by nodes.

The throughput here represents the number of bits (of the payload) per second correctly received by the sink when all the 30 nodes try to access the channel and transmit their packets, assuming that nodes transmit packets of the same size.

It is possible to notice that, for low values of the packet size, star topology outperforms tree, whereas trees perform better when the packet size increases, since less nodes compete to access the channel at the same time (nodes are split in two levels). The optimal performance can be achieved for tree topology with super-frame order (SO) equal to zero and beacon order (BO) equal to two [10], which represents the best compromise between the duration of the active part of the sink super-frame (where level one nodes transmit) and that of the inactive part (where level 1 super-frames are located).

With the aim of characterizing the performance of the communications architecture proposed is presented, with regard to the fault tolerance. As a matter of fact, the RPL protocol is able to update the routing tables, thus facing frequent disconnections and node failures.

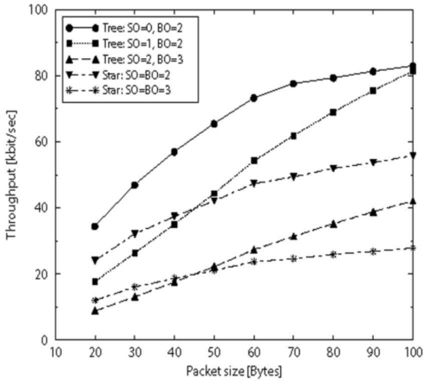


Figure 17. Throughput as a function of packet size for an IEEE 802.15.4 network arranged in star and tree-based topologies.

In deriving the key figures, three different scenarios have been investigated, as to the mobility, that has been assumed as the reason why the nodes get disconnected.. The reference play-ground, that is a 800m×800 m^2 square area, with 4 SNUs deployed in fixed position, i.e., at the center of each of the four sub-squares, and 200 ENUs uniformly distributed. The differences consist in the mobility patterns, as it assumed for ENUs, respectively:

- No mobility;
- Random Way Point (RWP) mobility model with speed uniformly distributed in the interval [1-2] m/s (slow mobility);
- RWP mobility model with speed uniformly distributed in the interval [3-6] m/s (fast mobility).

For each scenarios the following figures of merit have been evaluated for the setup phase:

- Association efficiency, i.e., the number of ENUs that are connected with one SNU with respect to the total number of SNs;
- End-to-end (E2E) data delivery latency, i.e., the time interval needed to a data message to be received by correct ER;
- The length of established e2e paths (between an ENU and its SNU).

The most relevant simulation parameters, characterizing the communication links, are summarized in Tab. I.

TABLE I. PARAMETERS VALUES USED IN THE SIMULATIONS

Parameter	***Value***
Carrier Frequency [GHz]	2.4
Transmitted power [dBm]	0
Receiver sensitivity [dBm]	-110
Coverage radius [m]	≈ 80
Transmission Bit-rate [kb/s]	250
Packet error-rate [%]	5
Simulated time interval [s]	2000
Setup period [s]	3
Data period [s]	1

In Fig. 17, the association efficiency as a function of simulated time is depicted for each scenario. The initial transient time needed to reach a full network connectivity is very short due to the presence, in RPL of an explicit solicitation of DODAG information from the neighbors. As the network connectivity is achieved, data message can be delivered from ENUs to their respective SNUs. To evaluate the effectiveness of this phase, the E2E path length experimental cumulative distribution function (ECDF) has been outlined in Fig. 18. It is evident that RPL protocol is able to establish, maintain and exploit stable minimum hop paths. The proposed scheme is also able to face low-to-medium speed without significantly performance degradation, while for fast mobility pattern it no longer holds. Specifically, short E2E paths become more frequent while longer ones are more rare; this is due to the fact that a longer multi-hop path is not likely to be always guaranteed, because each ENU involved in can randomly move toward another area.

In Fig. 19, the E2E delivery time for a data message from an ENU to the reference SNU is presented again for three different mobility patterns. The robustness of RPL approach allows to keep latency below a reasonable value for low-to-medium speed range. Nevertheless, if SNs move more quickly, network get often disconnected and, as pointed out in Fig. 19, only shorter paths are stable during all the message delivering; thus, the E2E latency decreases.

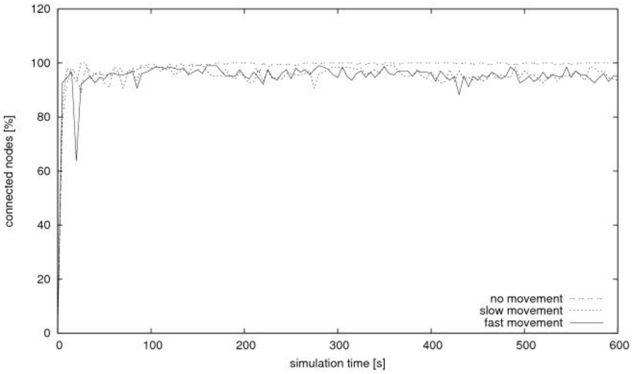


Figure 17. RPL: Association efficiency as a function of time.

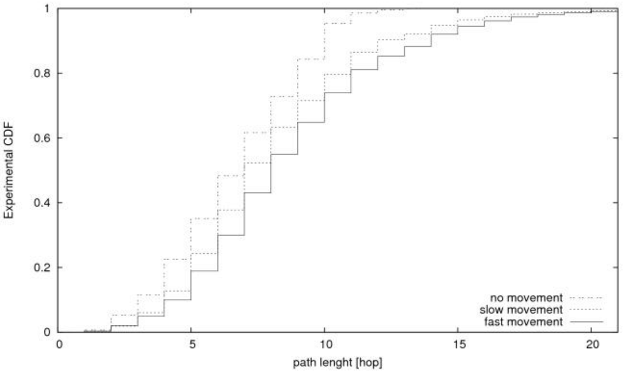


Figure 18. RPL: E2E path length CDF for different mobility pattern.

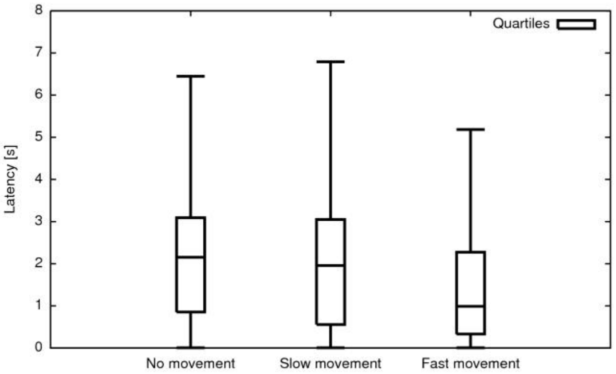


Figure 19. RPL: E2E data delivery latency for different mobility pattern with the indication of mean value and confidence interval.

V. Conclusion and future work

An end-to-end distributed monitoring system of integrated VOC detectors, capable of performing real-time analysis of gas concentration in hazardous sites at an unprecedented time/space scale, has been implemented and successfully tested in an industrial site. The aim was to provide the industrial site with a flexible and cost-effective monitoring tool in order to achieve a better management of abnormal situations, to identify emission sources in real time, and to collect continuous VOC concentration data using easily re-deployable and rationally distributed monitoring stations.

The piloting of the system allowed us to pinpoint key traits. Collecting data at 1-minute time intervals meets several needs: identifying short-term significant events, quantifying the emission impacts as a function of weather conditions as well as of operational process, in addition to identifying potentially VOC sources in the plant area. Moreover, the choice of a WSN communication platform gave excellent results, above all in allowing for redeploying and re-scaling the network's configuration according to specific needs as they arise, while, at the same time, greatly reducing installation costs. Real-time data through a web-based interface allowed both adequate levels of control and quick data interpretation in order to manage specific situations.

Further developments will concern with the development of a standard application to allow the deployment of WSN in other plant (e.g., refineries), in addition to an assessment of potential applications for WSN infrastructure monitoring of other environmental indicators.

Acknowledgment

This work was supported by eni SpA under contract No. 3500007596. The authors wish to thank W. O. Ho and A. Burnley, at Alphasense Ltd., for their helpful comments and clarifications concerning PID operation, as well as E. Benvenuti, at Netsens Srl, for his valuable technical support. Assistance and support from the management and technical staff of Polimeri Europa Mantova is gratefully acknowledged.

References

[1] Tsujita W., Ishida H., and Moriizumi T. (2004), “Dynamic gas sensor network for air pollution monitoring and its auto-calibration,” in Proceedings of the IEEE Sensors, Vol.1, pp. 56–59, October 2004.

[2] Tsow F., Forzani E., Rai A., Rui W., Tsui R., Mastroianni S., Knobbe C., Gandolfi A.J., and Tao N.J. (2009), “A wearable and wireless sensor system for real-time monitoring of toxic environmental volatile organic compounds,” in Sensors Journal, IEEE 9, Vol. 9 Issue 12, pp. 1734–1740, January 2009.

[3] G. Manes, R. Fusco, L. Gelpi, A. Manes D. Di Palma and G. Collodi, “Real-time monitoring of volatile organic compounds in hazardous sites,” in Intech Book, Environmental Monitoring, Chapter 14, pp. 219-244. ISBN 978-953-307-724-6.

[4] G. Manes, R. Fusco, L. Gelpi, A. Manes D. Di Palma and G. Collodi, “A Wireless Sensor Network for Precise Volatile Organic Compound Monitoring,” in International Journal of Distributed Sensor Networks, Special Issue on Ubiquitous Sensor Networks and Its Application. To appear.

[5] MiniPID User Manual V1.8. (2000). IonScience Ltd.

[6] G. Manes, R. Fantacci, F. Chiti, M. Ciabatti, G. Collodi, D. Di Palma, and A. Manes (2009), “Energy efficient MAC protocols for wireless sensor networks endowed with directive antennas: A cross-layer solution,” in Proceedings of IEEE Radio and Wireless Conference 2009, pp. 239-242, Orlando, FL, January 2009.

[7] Price JGW, Fenimore DC, Simmonds PG, and Zlatkis A. (1968), “Design and operation of a photoionization detector for gas chromatography,” in Analytical Chemistry, Vol. 40, No. 3, March 1968, pp. 541-547.

[8] Locke DC. and Meloan CE. (1965), “Study of the photoionization detector for gas chromatography,” in Analytical Chemistry, Vol. 37, No. 3, March 1965, pp. 389-395.

[9] Alphasense Ltd. Technical specifications; Doc. Ref. PID-AH/MAR11.

[10] “IEEE Standard for Information Technology - Telecommunications and Information Exchange Between Systems - Local and Metropolitan Area Networks Specific Requirements Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs),” IEEE Std 802.15.4-2003, pp. 1–670, 2003.

[11] L. Bartolozzi, F. Chiti, R. Fantacci, T. Pecorella, F. Sgrilli, “Supporting Monitoring Applications with Mobile Wireless Sensor Networks: the 'eN Route" Forwarding Approach”, *in Proc.* IEEE ICC 2012.

MAD Science: Increasing Engagement in STEM Education through Participatory Sensing

Scott Heggen, Osarieme Omokaro, Jamie Payton
Department of Computer Science
The University of North Carolina at Charlotte
9201 University City Blvd., Charlotte, NC, USA 28223
{*sheggen, oomokaro, payton*}*@uncc.edu*

***Abstract*—In this paper, we introduce the Mobile Application Development for Science (MAD Science) curriculum, which utilizes *participatory sensing* as a central theme to increase middle school students' engagement and interest in science and technology. Participatory sensing involves the general public in collecting and sharing information about the surrounding environment through the use of sensing (e.g., camera, GPS, accelerometer) and input capabilities on handheld mobile devices, such as smartphones. We present the results of a pilot offering of the MAD Science curriculum as part of a 10-week after-school program for middle school children. Our results indicate the potential for participatory sensing as a tool for increasing engagement in technology; after participating in the MAD Science program, students viewed technology more favorably, indicated increased enjoyment of technology, and indicated increased interest in pursuing education and careers in science and computing.**

***Keywords*-participatory sensing; public participation in scientific research; broadening participation; education**

I. INTRODUCTION

In order to support innovation and competitiveness in a global economy, governments are making significant investments to encourage interest in and improve educational methods for science, technology, engineering, and mathematics (STEM) fields. Providing early interventions is essential to increasing engagement in STEM education and interest in STEM careers; the majority of youth have formed "life aspirations" that impact educational and career choices even before the age of fourteen [1].

In response to this need, we have developed the Mobile Application Development for Science (MAD Science) curriculum as an intervention for middle school children to increase engagement with technology, increase engagement with and knowledge of science, and increase the desire to pursue an education and career in science and technology. The MAD Science curriculum is centered around *participatory sensing*, in which participants collect data samples for a focused data collection campaign using the sensors embedded in their mobile phones, such as the camera, microphone, and GPS sensors. Participatory sensing has its roots in public participation in scientific research (PPSR) projects, which have been shown to be effective in their use of open inquiry and investigation to increase engagement in science [1][2] and to improve knowledge about science [2][3]. The use of participatory sensing as a tool for investigating scientific questions promotes inquiry-based learning, which has been advocated for many years as a strategy to increase engagement in STEM education. Given the focus on modern sensing, computing, and communication technology as a tool for data collection, we believe that participatory sensing is well-suited to build upon the impact of PPSR to increase engagement, interest, and knowledge in technology as well as science.

In this paper, we present the MAD Science curriculum and the results of a pilot implementation as part of a 10-week after-school program for middle school students. In the MAD Science program, students engage as practitioners of STEM by creating a participatory sensing application, collecting data samples using mobile phones to address a civic problem that they have identified in their community, analyzing the results, and presenting them to the community. In the pilot offering of MAD Science, twenty-one middle school students created two software applications in order support their socially relevant participatory sensing data collection campaigns, and used the applications during local field trips to collect scientific data. Students analyzed the data in the MAD Science program, and presented findings to over 50 visitors at a community event. Our results indicate that engagement with technology was affected positively by the MAD Science program. Students' desire to pursue an education and career in science and technology also increased as a result of the MAD Science program.

The remainder of the paper is organized as follows. Section II describes the previous works to motivate our research. Section III describes the MAD Science curriculum and its implementation. Section IV provides insight into the impact the curriculum had in our pilot program. Finally, Section V provides our concluding remarks and future direction for the project.

II. RELATED WORK

Public participation in science has proven to be an extremely valuable tool for increasing knowledge and engagement in science. Raddick et al. [4] identify four ways in which PPSR can be used to increase scientific literacy: increasing content knowledge, providing an experience of the process of science, creating opportunities for changes in attitude toward science, and providing an opportunity for direct communication with scientists. The Center for the Advancement of Informal Science Education (CAISE) released a report on the value and potential of PPSR projects as a form of informal science education [2], using metrics that assessed the engagement or interest in science; skills in using technology; and awareness, knowledge and understanding of scientific concepts and processes. The CAISE report concludes that "…enlisting people into PPSR projects is probably one of the most expedient methods for informal science educators to engage people in science in a fun and meaningful way."

Additional studies have shown that directly engaging participants in the process of inquisitive thinking in PPSR projects improves knowledge about science [3]. For example, ReClam the Bay (RCTB) [5] promotes education in water quality, bay ecosystems, and the environmental benefits of shell fish by involving students in the growth and study of a shellfish in a science classroom. The project reported significant gains in content knowledge for middle, high school and even college students, and participant commitment to continuous education and habitat management goals. Similarly, the Salal Harvest Sustainability Study [6], a PPSR project promoting responsible harvesting of Salal shrubs, revealed that training in research design, data collection and data interpretation methods improved the harvesters' knowledge of scientific concepts and processes. The study showed increased skill set on the job and empowered community involvement, which informed better resource management and harvesting practices for the harvesters. Lastly, the Alliance for Aquatic Resource Monitoring (ALLARM) Acid Rain Monitoring Project [7] was able to empower participants, promote stewardship, increase knowledge and awareness of aquatic systems and health via various training and mentoring programs in Pennsylvania. The participants were trained in data management, data interpretation and statistical analysis skills, as well as effective presentation of data to address their issues. An evaluation of the ALLARM project revealed limited gains in project knowledge, but showed increased engagement, deep project commitment and solid data collection skills.

Advances in mobile computing, sensing, and communications technology has made it possible to utilize smartphones as a platform to support remote data collection for scientific purposes. Modern smartphones are typically equipped with several standard sensing modalities (e.g., camera, GPS, accelerometer) that can be used to capture and report observations of phenomena. This idea, in which volunteers use a software application deployed on a mobile phone to collect and share data for a purpose, is referred to as *participatory sensing* [8]. With over 5.8 billion users of mobile phones [9] worldwide, participatory sensing has the potential to reach a large number of volunteers, and can be used to collect data across large geographic areas with differing habitats.

Participatory sensing can be viewed as an extension of PPSR that incorporates the use of digital sensing technology and software applications to capture, report, and analyze data samples. The positive impact of PPSR projects for engaging the public in science can be viewed as an indicator of the potential for participatory sensing as a mechanism for increasing engagement in science. In addition, the use of computing, communication, and sensing technologies, including purposed sensors [10] and mobile phones [11], introduces the students to modern technology that is familiar and exciting, and has the potential to increase engagement in both science and technology. For these reasons, our MAD Science curriculum uses participatory sensing to engage middle school students in science and technology, to increase interest in education and careers in science and technology, and to increase knowledge of STEM concepts. This approach is closely related to the Mobilize program at UCLA [12], in which high school students use traditional programming languages to develop software applications for participatory sensing. While both programs utilize participatory sensing to increase engagement and knowledge of STEM, our MAD Science program is designed as an early intervention, targeting middle school students in the years before they form their life aspirations.

III. The MAD Science Curriculum and Implementation

The MAD Science curriculum was implemented at a local middle school through a national after-school program, which aims at expanding the learning day beyond the classroom in low-income communities. The goal of the after-school program is to provide extended learning time for students in an effort to close the "achievement gap"; the majority of students involved in the after-school program are Latino or African American, and over 75% qualify for free or reduced lunch. The after-school program recruits volunteers from the community to teach about their areas of expertise in an "apprenticeship," where students gain access to professionals in the community who volunteer as teachers. Apprenticeships emphasize hands-on learning activities to keep kids engaged and promote information retention. Each apprenticeship runs for 1.5 hours, one day a week, for ten weeks. The national after-school program in which we implement the MAD Science apprenticeship is supported by the National Science Foundation and its impact has an expanding reach; with programs in 18 cities and 31 middle schools across the United States, the after-school program serves approximately 4,500 middle school students each year

ISBN: 978-1-61208-236-3

and aims to grow to serve over 10,000 students in the next four years.

The first MAD Science apprenticeship was held in spring 2012 with twenty-one students from grades six to eight. Throughout the apprenticeship, the students applied the scientific method within the context of a participatory sensing data collection campaign. Students identified issues within the local community and put forth a hypothesis about the cause and a possible solution. Students then identified what data would be needed to evaluate the hypothesis, and created a participatory sensing campaign to collect the needed data. In doing so, students formulated the requirements for a participatory sensing application to support the data collection campaign, which was then implemented by our research team and deployed on mobile phones to enable data collection by the students. Once data was collected using the participatory sensing application, students analyzed how the data supported or refuted the hypothesis. At the end of the apprenticeship, the students demonstrated their acquired skills and knowledge to their friends and family. We describe the details of the apprenticeship below, and the results of its initial implementation are presented in Section IV.

A. MAD Science Lesson Structure

Each apprenticeship lesson is designed as an active learning experience, with a focus on hands-on activities to engage students and reinforce learning. Lessons follow the structure presented in Table I. First, every lesson began with a modified version of the classic Taboo© game that was used to teach vocabulary related to participatory sensing and STEM concepts, particularly focusing on the scientific method, sensing and communication technology, data collection, and data analysis. This activity encouraged the students to actively participate in learning the vocabulary of STEM by describing the terms in their own words and connecting terms to their own personal knowledge bases. As a member of the after-school program's staff noted, "…this activity got students to really think about what these words mean, which enabled them to better understand and articulate the material throughout the lesson."

Table I: The MAD Science Lesson Structure

Activity	Approx. Time	Purpose
Ritual	10m	A fun, short activity or game to engage students and is relevant to the topic of the day.
Introduction	5m	Share the lesson objectives and agenda.
Activity 1	20m	An activity centered around the objectives for the lesson
Activity 2	20m	An activity centered around the objectives for the lesson
Activity 3	20m	An activity centered around the objectives for the lesson
Teachback	10m	Guided questions to ensure the students understood the day's content

Following the ritual Taboo© game, an overview of the lesson was presented to the students, connecting the ideas of the planned activities to the broader theme of participatory sensing. The remaining one-hour lesson was broken into three hands-on activities, a time frame that seems effective for keeping the middle school students focused and engaged. Each activity was used to introduce and reinforce ideas related to the scientific method, standards and procedures for data collection, validity of data samples and data sets, sensing technology, and mobile phone communication. Each lesson ended with a "teachback," where the students would describe the knowledge they'd gained throughout the lesson. To encourage participation, rewards in the form of "mascot bucks" were commonly distributed for correct responses; these mascot bucks can be redeemed at the middle school's "reward store" for school-related items and apparel with the school logo.

B. MAD Science Lessons and Objectives

The MAD Science apprenticeship consisted of 10 sessions, each lasting 90 minutes. Table II provides an overview of the topics introduced in each session and the associated lesson objectives. Below, we describe lesson activities and their connection to the MAD Science goals of increasing engagement and interest in science and technology.

Weeks 1 through 4: (Re)Shaping Students' Opinions of Science and Technology

Our first objective was to challenge negative perceptions of science and stereotypes of scientists, and to present positive role models. In one activity, students were asked to view photos of people performing various activities and to comment on whether or not the person was acting as a scientist (Week 1 in Table II). The photos were selected to include a racially diverse group of scientists who do not fit the "nerd" or "geek" stereotype, and often depicted people using technology in pursuit of science. The primary purpose was to dispel student's notions about scientific work; we intended to show that scientists do not spend all of their time in a sterile lab environment, that scientists engage in work that helps to better society, and that scientists were accessible people that could be from their own community. The second activity (Week 4 in Table II) introduced the students to a scientist, who talked about career path in becoming a professor of computer science, answered students' questions about her education and the daily activities of a scientist, and interacted with students as they addressed their own scientific activities in the MAD Science lesson. The students seemed to gain new knowledge about professions in science and valued this opportunity to hear from an actual scientist; one student responded, "I learned that as a researcher you can learn about things that really interest you."

The second objective was to get the students to feel more comfortable with the idea of doing science. During week

Table II: The MAD Science Curriculum

Schedule	Topic		Activity Objectives
Week 1:		Introduction to Science, the Scientific Method, and MAD Science	Dispel misconceptions about the role of a scientist, and the procedures they use to conduct science
Week 2:	Participatory Sensing Taboo	Let's get excited about Science!	Show the students how they can be scientists through citizen science, and introduce them to the two campaigns they will be conducting
Week 3:		Sensors, sensors, sensors!	Describe each sensor in the phone and demonstrate how they can be used to collect samples for a scientist
Week 4:		How to Collect Data... The right way! + Guest Speaker	Connect the students to a real scientist; prepare the students to collect data by introducing them to standard data collection procedures
Week 5:		Data collection Field Trip 1	Take the students to the local stream to collect images of pollution, pipes, wildlife, and interesting phenomena
Week 6:		Data collection Field Trip 2	Take the students to the gymnasium to collect accelerometer readings while the students perform different physical activities
Week 7:		What does it all mean?!	Categorize, analyze, and interpret the results of their collected data; draw a conclusion about what was found in both campaigns
Week 8:		Student Presentation Preparation	Plan the presentations and demonstrations the students will be giving; reinforce the subjects taught in the prior weeks
Week 9:		Student Presentation Preparation	Build and practice the presentations and demonstrations the students will be giving
Week 10:		Student Presentations	The students present their work to family, friends, and guests

2 of the apprenticeship, the students were reintroduced to the scientific method, which had been previously covered in their traditional middle school science courses. The students then formed small groups to identify and solve a problem in their local community. The majority of students immediately identified pollution as a problem. The students were then led through a series of group activities to help them solve the problem of pollution in their community. Unknowingly, they were following the steps of the scientific method, which we discussed at the end of the lesson during a "teachback" activity. Tying the students personal interest in a community problem to the technical aspects associated with the scientific method showed the students the value of conducting science in a methodical manner. A student said of this activity, "MAD Science taught me that we can be citizen scientists and that we can use science to make a difference in our community." To further engage students in science by doing science, the students were introduced to an important concept in the scientific method: the validity of experimental data (Week 4 in Table II). The lesson discussed poor data collection practices and errors in data samples. Students were then shown a series of images of people as they collected data for a particular purpose, and were asked to identify if the methods being used were "good" or "bad." The students were adamant about justifying their reasoning for selecting "good" or "bad" without prompting from the session leader. The students were not simply answering the question, but they were reasoning through each data collection method and applying their knowledge of the scientific method to evaluate the appropriateness of the method for collecting data.

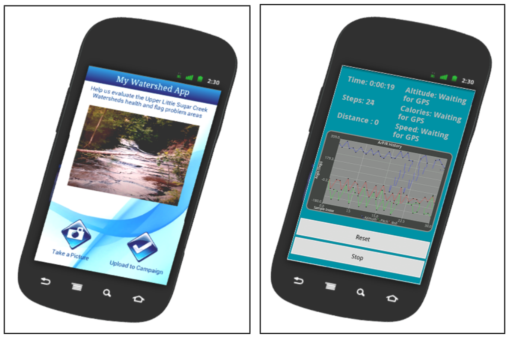
The third objective was to engage the students with technology through participatory sensing. To begin, the students needed some basic knowledge of mobile phones and embedded sensors. Six sensors common in mobile phones were introduced during week 3 (accelerometer, camera, camcorder, microphone, ambient light sensor, and GPS) using a hands-on activity. For example, in an activity used to explain GPS and localization, four students volunteered to represent GPS satellites, and a fifth student to represent the GPS receiver. Using strings to represent the distance between the receiver (whose position is unknown) and each satellite (whose positions are known), the students learned why GPS requires four satellites to accurately determine the receivers location. Reflecting on this activity, an after-school program staff member that supervised MAD Science stated, "Instead of sitting in a desk and listening to a lecture about the science behind mobile sensors, this activity had students learning this science by getting them out of their seats and demonstrating these processes. As a result of these types of hands-on activities, students were not only engaged but possessed a good understanding on the concepts covered as evidenced through frequent "teachbacks" during and at the end of lessons."

Weeks 5 and 6: Running a Participatory Sensing Campaign

Midway through the apprenticeship, the students are ready to begin their participatory sensing campaign. The students were split into two groups, with each group responsible for conducting a specific participatory sensing campaign. To do so, each group followed the scientific method, defining the problem, forming a hypothesis, and identifying data collection procedures for their campaign. All students participated in data collection for both campaigns. The group responsible for the campaign analyzed the collected data.

The first group focused on pollution in the local watershed. In designing their data collection procedures, students identified the need for a participatory sensing application that uses the camera and GPS sensors to identify stretches of water with pollution, pipe run-offs, construction near the stream, and other factors that would affect the stream's health. The MAD Science research team implemented a participatory sensing application to meet these specifications, and students took a field trip in the next session to a local park adjacent to a small stream that is part of the Upper Little Sugar Creek Watershed, which is in the students' local community. Using the mobile application in Figure 1a deployed on five mobile phones, the students worked in teams to gather 52 images of the stream where they identified pollution, pipes feeding into the stream, and other unhealthy activities. One student remarked, "My favorite activity in the MAD Science apprenticeship was the field trip. We were able to go outside and interact with the environment. I really like taking pictures with the phone and then looking at them afterward to see what they told us about the creek."

The second group wanted to show that certain physical activities exert more energy than others, and therefore may have more impact on personal health. They chose to use the accelerometer to capture data about a person's motion while performing these activities, and use this data to determine which activity is best for your health. Again, the MAD Science research team implemented a participatory sensing application that met these specifications (Figure 1b). Mobile phones were attached to the students' arms, legs, or placed in their pockets, and students took turns using the participatory sensing application to collect data while performing a variety of activities, including playing basketball, running, and jumping rope. A student said of this activity, "I especially liked the sports that we played to collect data about the types of movements we did. This was fun, and afterward I could explain the types of movements that the phone collected."


My Watershed App
2:30
Help us evaluate the Upper Little Sugar Creek Watersheds health and flag problem areas
Take a Picture
Upload to Campaign
Time: 0:00:19
Steps: 24
Distance : 0
Altitude: Waiting for GPS
Calories: Waiting for GPS
Speed: Waiting for GPS
A/P/R History
Reset
Stop

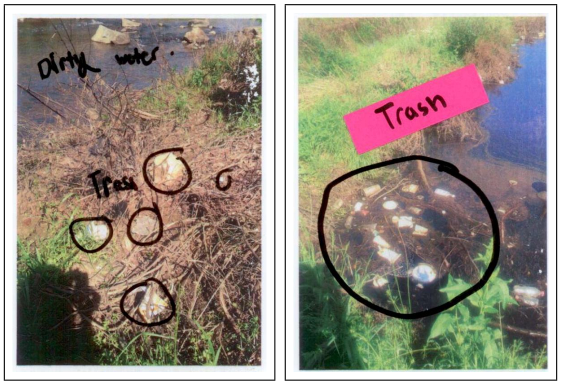
(a) (b)

Figure 1: Participatory sensing applications for (a) watershed pollution campaign and (b) physical activity campaign

Both of these activities marked a significant point in the program. Students were now able to see the connection between a technological novelty (the mobile phone) and the value they can provide as a tool that can be used to answer questions of interest to their peer group and the broader community. Even without having seen any data that they had collected, the activities provided the students with a sense of accomplishment, and they understood they were acting as scientists, providing meaningful data in a systematic way for a purpose. The after-school program staff that supervised MAD Science said of these activities, "The students not only did a great job working in groups to collect data at [the park], they really enjoyed themselves." and "Every student I spoke to about gathering [the physical activity] data was fluent in the terminology and what the data represented. I was very impressed!"

Weeks 7 through 10: Analyzing and Presenting the Results

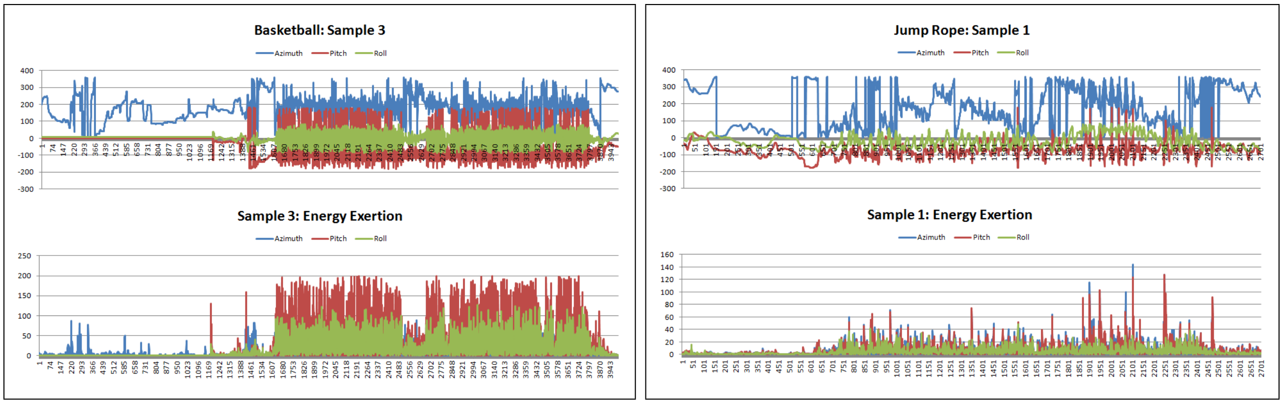
Starting in week 7, the students focused on interpreting the meaning of their data and presenting the results. The watershed pollution team focused on tagging and categorizing the images based on their content. Figure 2 shows a sample of the images and the tags produced by the students. The physical activity team compared the data from different activities to determine which activity exerted the most energy. Figure 3 shows the accelerometer readings that students used to analyze two activities, playing basketball and jumping rope. The after-school program staff that supervised our apprenticeship said that "...the data analysis activities were a success because they had students analyzing the data without even knowing it. I think this was possible because our data (pictures and interactive graphs) and the way students collected it (taking pictures and performing fun physical activities) attracted students attention."



(a) (b)

Figure 2: Sample data collected by the students for the watershed pollution campaign

While all of the activities were important, the final three weeks played a significant role in ensuring comprehension. At the end of the apprenticeship, the after-school program asks the students to show off their hard work to their parents, friends, and invited guests. The MAD Science students chose to do two activities: a presentation of their application of participatory sensing for watershed monitoring and physical activity, and a demonstration of the two mobile applications they used to collect data. The students developed their slides,



(a) (b)

Figure 3: Sample data collected by the students for (a) basketball and (b) jump rope

practiced their speeches, and worked together in teams to prepare for the big event. During this time, the students were reflecting on their prior weeks to remember what they had learned. For example, in week 3, the students were introduced to the terms “azimuth,” “pitch” and “roll” while learning about the accelerometer sensor. In order to create a clear, concise presentation for the physical activity data they collected, the students had to revisit and fully understand these terms. In week 8, the students would ask what these terms meant. By week 9 and 10, however, the students would approach one of the teachers and demonstrate the concepts to the teacher out of sheer pride. The students were excited to show what they had learned, and were enthusiastic to explain the concepts to their friends and family.

IV. MAD Science Impact

Our pilot offering of the MAD Science apprenticeship included 21 students (5 female, 16 male) between the ages of 10 and 14 that were participants in the after-school program at a large urban middle school. This group consisted primarily of students that are underrepresented in STEM: 12 students were African American, 8 were Latino, and 1 was Native American. Eighteen of the 21 students in our MAD Science apprenticeship qualified for free/reduced lunch. Three of the students were identified as having special needs.

The students were issued a pre- and post-survey (Table III) to assess the impact of the MAD Science apprenticeship on engagement in science, engagement in technology, and attitudes towards education and careers in both science and computing. To protect the privacy of students, these surveys were administered without any identifying information. To evaluate the impact of MAD Science on knowledge acquisition in science and technology, student grades were also collected by the middle school and provided to the research team in aggregate form, and interviews were conducted with the students and the after-school program staff after the MAD Science apprenticeship concluded.

A. Engagement with Technology

Our first objective in the MAD Science apprenticeship was to increase the students’ engagement with technology. In the pre- and post-surveys given to students (Table III), questions 1, 2, 3, 11, and 18-21 are intended to assess this objective. Figure 4 summarizes the results of each of these items from the pre-survey and post-survey. Since the number of respondents to the pre-survey (16) was different than for the post-survey (19), responses were normalized and are displayed as a percentage. For questions with a positive implication (e.g., I like computers), we expect to see an increase in size for bars at the top of each column (strongly agree and agree), i.e., a decrease in size of the blue and orange bars in Figure 4 that correspond to disagree and strongly disagree responses, respectively. The reverse is true for questions with a negative implication (e.g., I think computers are boring), i.e., an increase in size of the blue and orange bars that correspond to strongly disagree and disagree (respectively) and a decrease in purple and green bars that correspond to strongly agree and agree (respectively) in Figure 4 are expected. Questions with a negative implication are indicated in the chart labels with asterisks.

The results in Figure 4 indicate that the MAD Science apprenticeship had a small but positive effect on the students’ engagement with computers. While only question 20 (I will use mobile phones in many ways in my life) resulted in a statistically significant change (P-value = 0.0379), the responses to all questions from pre-survey to post-survey shifted in a positive way, indicating that the students’ engagement with technology increased throughout the apprenticeship, and the students viewed technology more favorably by the end of the apprenticeship. Questions 1, 3, and 11 (I know a lot about computers; I am good at using computers; It is fun to use computers) all showed positive

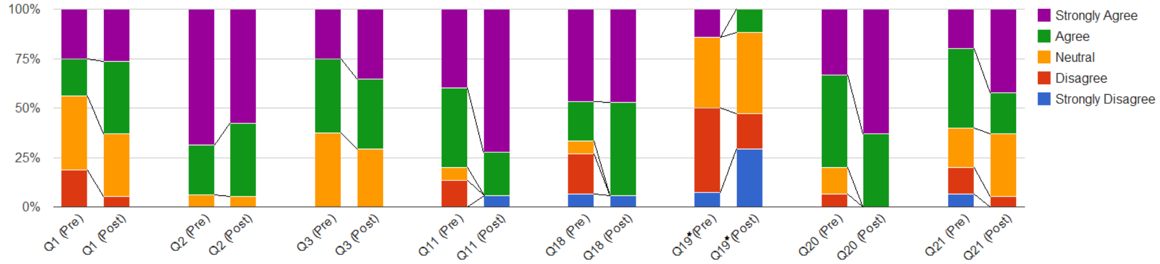


Figure 4: Pre- and post-survey results regarding engagement with technology.

Table III: The pre-survey and post-survey taken by the MAD Science students. Questions 1 to 21 used a Likert scale with strongly agree, agree, neutral, disagree, and strongly disagree options.

Rate each item by how much you agree or disagree with the statement:

1. I know a lot about computers.
2. I like using computers.
3. I am good at using computers.
4. People like me are interested in computers.
5. I am interested in learning more about what I can do with computers.
6. I might be interested in a career in the field of computing.
7. Someday, I might like to major in computing in college.
8. People who like computers are often weird.
9. Studying computing in high school would be a good idea.
10. I like to figure things out for myself.
11. It is fun to use computers.
12. I don't think I would like working with computers in my job.
13. I am not smart enough to be good at computing as a major or career.
14. Learning about science to solve problems is interesting.
15. I am not smart enough to be good at science as a major or career.
16. Learning about science is boring.
17. I am good at science
18. Mobile phones can be used to help people.
19. Mobile phones are only for fun
20. I will use mobile phones in many ways in my life.
21. Knowing how to work with mobile phones will help me get a good job someday.
22. Please check beside the ways you use computers:
    a. Word processing
    b. Computer Games
    c. Web search for school
    d. Chatting online
    e. Sending email
    f. Web search for personal interests
    g. Solving math and science problems
    h. Myspace/Facebook
23. Please check beside the ways you use mobile phones:
    a. Texting
    b. Games
    c. Search the web
    d. Chatting online
    e. Sending email
    f. Myspace/Facebook

gains, indicating the engagement with computers were having a positive impact on the students. The results for question 2 (I like using computers) indicates a decrease in the average response value from pre to post, but this is because three new respondents that did not participate in the pre-survey selected "agree" in the post-survey; overall, the percentage of students that agreed or strongly agreed that they liked using computers remained the same at approximately 95%.

Table IV summarizes the results of questions 22 and 23, which measure the students' usage of computers and mobile phones. Computer usage saw an increase in 8 of the 11 categories, and mobile phones saw an increase in 5 of the 6 categories, suggesting an increase in the students' interactions with both technologies. The largest gain for both computer and mobile phone usage was seen in their usage of Myspace/Facebook (20% for computers, and 22.3% for mobile phones). The most interesting gain is in computer usage for solving math and science problems, which showed a 17.3% gain.

Table IV: Students' usage of technology from Pre-survey and Post-survey

How do you use computers?	Pre-survey	Post-survey
Word processing	73.3%	61.1%
Computer Games	100.0%	94.4%
Web search for school	80.0%	83.3%
Chatting online	73.3%	77.8%
Sending email	73.3%	61.1%
Web search for personal interests	80.0%	88.9%
Solving math and science problems	53.3%	70.6%
Myspace/Facebook	46.7%	66.7%

How do you use mobile phones?	Pre-survey	Post-survey
Texting	66.7%	83.3%
Games	80.0%	88.9%
Search the web	80.0%	88.9%
Chatting online	66.7%	66.7%
Sending email	46.7%	55.6%
Myspace/Facebook	33.3%	55.6%

B. Engagement with Science

Our second objective was to increase the students' engagement with science; assessment of this objective is addressed by survey questions 14, 16, and 17 (Table III). Figure 5 summarizes the results of these questions from the pre-survey and post-survey. The survey results were contradictory to our expected results; responses showed very little variation from pre-survey to post-survey. This result is surprising, given the previous success of PPSR projects for engaging students in science. However, if we look more closely at the pre-survey results, we see that approximately 60% of students agreed or strongly agreed that "learning about science to solve problems is interesting", over 50% disagreed or strongly disagreed that "learning about science is boring", and 75% of students agreed or strongly agreed that "I am good at science." We plan to investigate this result further by surveying a larger population of middle schools students that do not participate in the after-school program and those that do, in order to determine if this is an issue related to "self-selection" of students that are already interested in science that chose to attend our apprenticeship, which has the word "science" in the title. We must also address possible limitations of our survey, increasing the number of science-based questions. Finally, we plan to interview students and teachers to better understand what activities they found to be challenging and interesting activities that center around science, and to try to distill characteristics that will help us to create more engaging activities in the future.

Students who participated in the MAD Science apprenticeship performed better in science than their middle school peers; 95% of MAD Science students maintained an A/B grade or improved a C/D/F grade in science, compared to 70% in the middle school as a whole. Again, since the after-school program is voluntary, this result may be accountable to self-selection bias; unfortunately, aggregated science grades for the entire middle school, participants in other apprenticeships within the same middle school, and participants in other apprenticeships across the nation-wide network of the after-school program were not available at the time of this publication for comparison.

Post-program interviews with after-school program staff that supervised the MAD Science apprenticeship and with the MAD Science students provide some anecdotal evidence that the program did, however, have a positive impact on students' engagement with science. A staff member stated "Turning data into a hands-on activity made this type of science come alive to the students. I witnessed several "light bulb" moments in the kids as they understood the material through real-world examples." A student explained, "I think you make science better by making it fun with hands-on activities. Instead of writing and sitting in desks the whole time, we should be interacting with materials and doing experiments." Another says, "This apprenticeship sparked my interest in science more because it was fun to gather data and make conclusions about things that impact me and my community." Lastly, another student claims, "MAD Science helped me become more confident in my science abilities because it revisited topics that we learned in our science class, like the scientific method. I was able to recall what the scientific method is, which allowed me to understand what we were talking about and made me want to participate more in activities."

C. Aspirations to Pursue a STEM Education or Career

Our third objective of the apprenticeship was to increase the students' desire to pursue a STEM-based education or career. Questions 6, 7, 9, 12, 13, and 15 evaluate this objective. Figure 6 summarizes the results of these questions from the pre-survey and post-survey. All responses improved from pre-survey to post-survey, indicating the students viewed STEM-based learning more favorably after the apprenticeship. Questions 13 (I am not smart enough to be good at computing as a major or career) is of particular interest, as it indicates the students understand computing is an attainable long-term goals. Questions 6, 7, and 9 (I might be interested in a career in the field of computing; Someday, I might like to major in computing in college; Studying computing in high school would be a good idea) also showed small improvements, indicating the students are becoming more interested in studying computing. The after-school program staff state that "After the apprenticeship the kids seemed fired up about learning, and excited to learn more. I would definitely say that includes going to college" and "This apprenticeship exposed students to a number of possible careers in computer science that require a college degree. I believe that because students were exposed to these careers and skills through hands-on activities that were engaging and fun, they definitely became interested about these careers and going to college."

Based on the student responses to engagement in technology, we expected an interest in pursuing an education and career in computing. However, since we did not see a gain in engagement with science, we did not expect to see a gain in the students' interest in an education and career in science. Nonetheless, question 15 (I am not smart enough to be good at science as a major or career) speaks differently, as it was the only question resulting in a statistically significant change (P-value = 0.0259). One student validated these findings by stating, "MAD Science made me more interested in pursuing a career in science, specifically research." Our future work includes identifying the cause of this discrepancy and ensuring that our survey questions reflect the students' opinions about science more accurately.

D. Secondary Responses

While not a direct objective of this study, a small improvement was also noticed in students' math scores. In the MAD Science apprenticeship, 86% of students maintained an A/B grade or improved a C/D/F grade in math. That compares to 82% for students in the after-school program, 73% in

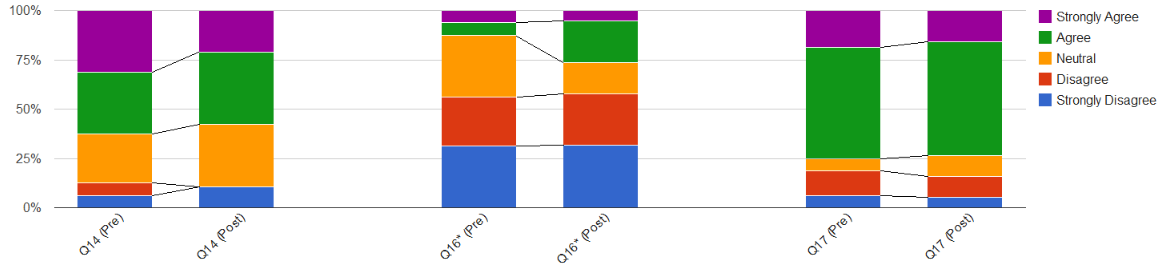


Figure 5: Survey results regarding engagement with science. Responses to questions marked with an * are expected to trend toward "Strongly Disagree" from pre-survey to post-survey.

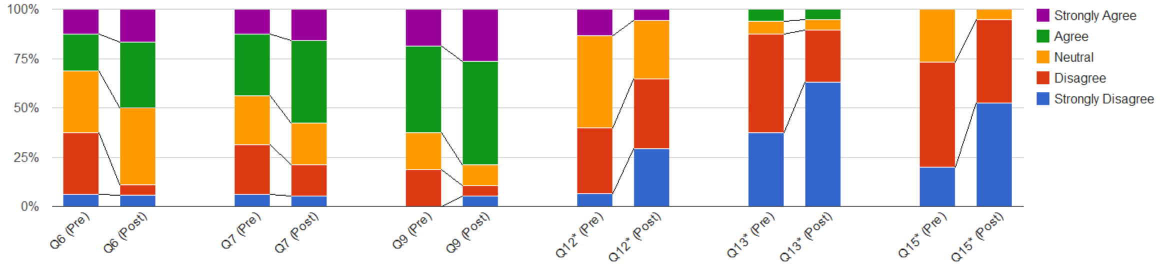


Figure 6: Survey results regarding aspirations to pursue an education in STEM. Responses to questions marked with an * are expected to trend towards "Strongly Disagree" from pre-survey to post-survey.

North Carolina, and 60% in the nation-wide network of the after-school program.

E. *Summary of Results*

The results of the MAD Science apprenticeship indicate the potential for participatory sensing as a tool for increasing engagement in science and technology. Students viewed technology more favorably, enjoyed interacting with technology, and aspired to pursue a career in computing because of the apprenticeship. While students did not indicate more interest in science, the students did indicate an interest in continuing an education or career in science.

V. Conclusion and Future Work

In this paper, we have presented the MAD Science curriculum, which aims to increase engagement and interest in science in middle school students through the use of participatory sensing as a central theme. The program addresses the issues of engagement and interest in science and technology by taking an inquiry-based learning approach to problems that are relevant to the students and their community, and exploiting "cool", accessible, and ubiquitous technology (i.e., mobile phones) to empower the students to understand and address the problem. The pilot offering of the MAD Science apprenticeship indicates that participatory sensing shows promise as a tool for increasing engagement in technology, and the desire to pursue an education in science and technology, for middle school children.

We plan to use the results of this initial pilot offering of MAD Science to extend and improve the curriculum and the study of its impact. While our pre- and post- surveys included items to assess engagement with mobile phones, the questions were repetitive and did not provide meaningful information. In addition, students that participated in the MAD Science program showed improvement in grades in mathematics, which we did not anticipate; we will include assessment items on future surveys to evaluate the impact of the program on engagement in and attitudes about mathematics. Finally, we plan to extend the curriculum in a way that increases knowledge about and skills related to computational thinking. While our current curriculum takes an instructionist approach (i.e., students gain knowledge by using participatory sensing applications), we plan to develop additional activities that take a constructionist approach (i.e., students gain knowledge by creating participatory sensing applications). We plan to develop a Scratch-like programming interface that allows students to write code for their own participatory sensing applications in a high-level, visual programming language. In addition to introducing computational concepts, we believe that an approach in which students build the technology that they will be using throughout the apprenticeship will result in a greater sense of ownership, and we suspect that this sense of ownership will ripple across all aspects of engagement.

Acknowledgments

The authors would like to thank Mr. Jeremy Gaden and the Citizen Schools staff for their feedback and support in the implementation of the MAD Science apprenticeship, the STARS Alliance for providing mobile phones to support this project, and most importantly, the Citizen Schools students for their participation in the MAD Science apprenticeship.

References

[1] R. Tytler, J. Osborne, G. Williams, K. Tytler, and J. Cripps Clark, "Opening up pathways: Engagement in STEM across the primary-secondary school transition," Australian Department of Education, Employment and Workplace Relations, Tech. Rep., 2008.

[2] R. Bonney, H. Ballard, R. Jordan, E. McCallie, T. Phillips, J. Shirk, and C. Wilderman, "Public participation in scientific research: Defining the field and assessing its potential for informal science education," Center for Advancement of Informal Science Education (CAISE), Tech. Rep., 2009.

[3] C. Cooper, J. Dickinson, T. Phillips, and R. Bonney, "Citizen science as a tool for conservation in residential ecosystems," *Ecology and Society, Journal of integrated Science and Residence Sustainability*, vol. 12, no. 2, p. 11, 2007.

[4] M. Raddick, G. Bracey, K. Carney, G. Gyuk, K. Borne, J. Wallin, S. Jacoby, and A. Planetarium, "Citizen science: Status and research directions for the coming decade," *AGB Stars and Related Phenomenastro 2010: The Astronomy and Astrophysics Decadal Survey*, vol. 2010, p. 46P, 2009.

[5] ReClam the Bay.org, "Reclam the bay," http://reclamthebay.org/edu.html/, 07.27.2012.

[6] H. L. Ballard and L. Fortmann, *Collaborating experts: integrating civil and conventional science to inform management of salal (Gaultheria shallon)*, K. Hanna and D. Slocombe, Eds. Oxford University Press, Oxford, UK., 2006.

[7] C. C. Wilderman, A. Barron, and L. Imgrund, "Top down or bottom up? ALLARMs experience with two operational models for community science," in *Proceedings of the Fourth National Monitoring Conference*, 2004.

[8] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. Srivastava, "Participatory sensing," in *World Sensor Web Workshop*, 2006, pp. 1–5.

[9] International Telecommunications Union, "Key global telecom indicators for the world telecommunications service sector," http://www.itu.int/ITU-D/ict/statistics/at_-glance/KeyTelecom.html, 07.27.2012.

[10] M. Silva, J. Lopes, P. da Silva, and M. Marcelino, "Sensing the schoolyard: using senses and sensors to assess georeferenced environmental dimensions," in *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*. ACM, 2010, p. 40.

[11] S. Vihavainen, T. Kuula, and M. Federley, "Cross-use of smart phones and printed books in primary school education," in *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. ACM, 2010, pp. 279–282.

[12] Mobilize, http://www.mobilizingcs.org, 07.27.2012.

Evaluation of a Cooperative Caching Scheme for Grid Ad Hoc Networks

Francisco J. González-Cañete, Eduardo Casilari
Department of Electronic Technology
University of Málaga
Málaga, Spain
{fgc, ecasilari}@uma.es

***Abstract*—In this paper, we evaluate the performance of the CLIR (Cross-Layer Interception and Redirection) cooperative caching scheme for ad hoc networks. Although this caching scheme was developed for Mobile Ad Hoc Networks (MANET) we study the application of this kind of algorithms in static grid ad hoc networks. By means of simulations, we evaluate the mean traffic generated in the wireless network, the delay perceived by the users and the percentage of failed searches as a function of the mean time between requests, the Time To Live (TTL) of the documents, the traffic pattern and the cache sizes. We compare the performance of our proposal with another five cooperative caching schemes as well as the option of no using a caching scheme. The simulation results show that our proposal outperforms the other caching schemes in terms of the studied parameters.**

Keywords-cooperative caching; grid; ad hoc network.

I. INTRODUCTION

The aim of a caching scheme is to reduce the traffic generated in the network, as well as the delay perceived by the users and the servers' load [1]. The reduction of the traffic in a wireless network also decreases the probability of collisions and interferences, and hence, the probability of packet loss. Reducing the delay perceived by the users when they request documents improves the user experience and makes the network more attractive to be used. Finally, as a consequence of the caching mechanism, the document requests can be served by other nodes in the wireless network instead of the servers. In a very loaded network, the servers could be a bottleneck as all the requests are sent to them. The caching mechanism mitigates this effect by moderating the overload of the servers so they can reply more requests.

Although many cooperative caching schemes have been proposed for MANETs (Mobile Ad Hoc NETworks) [2], they have not been evaluated for static ad hoc network, that is, wireless networks where the nodes do not move (which may be the typical case of many networking applications such as the sensor networks). The objective of this work is to evaluate the performance of different caching schemes proposed for MANETs in a static grid network.

The rest of this document is organized as follows. In Section II, the related work about cooperative caching schemes for MANETs is presented. In Section III, the proposed caching scheme is described. Section IV defines the system model and shows the performance evaluation of the caching schemes. Finally, Section IV enumerates the main conclusions of this work.

II. RELATED WORK

The cooperative caching schemes for ad hoc networks can be classified into four groups: broadcast-based, information-based, role-based and direct-request. The broadcast-based caching schemes employ broadcast messages as the first choice in order to find the documents in the network. These broadcast messages can be sent to the entire network, as in the case of MobEye [3]. Other schemes such as SimpleSearch [4], follow a more restrictive approach that limits the distance of the messages to four hops. ModifiedSS [5] is an evolution of SimpleSearch that employs GPS (Global Positioning System) in order to send the requests to the direction where the servers are located. Similarly, the caching scheme proposed by Moriya in [6] sends the broadcast messages to the neighbourhood so that, if the document is not found, the request is transmitted to the server.

The information-based cooperative caching schemes employ information of the location of the documents in the network. Nodes obtain this information by analysing the messages that they forward. As examples of this category of caching schemes we can mention: DGA (Distributed Greedy Algorithm) [7], Wang [8], Cho [9] and POACH (POware Aware Caching Heuristic) [10].

Under a role-based caching scheme, each node in the wireless network has a predefined role. That is, they can be caching nodes, requesting nodes, coordinator nodes, gateway nodes, etc. The role-based caching schemes are usually applied to cluster networks. CC (Cluster Cooperative) [11] and Denko [12] are examples of this kind of caching policy.

Finally, the direct-request caching schemes directly send the requests to the server with the hope of being served by an intermediate node in the route from the requester to the server. The proposal by Gianuzzy in [13] is an example of this kind of caching schemes.

However, the groups in this classification of caching schemes are not mutually exclusive. Thus, the caching schemes COOP [14], ORION (Optimized Routing Independent Overlay Network) [15], IXP/DPIP (IndeX

Push/Data Pull/Index Push) [16] and COCA (COoperative CAching) [17] are schemes that employ network information and broadcast requests. On the other hand, COACS (Cooperative and Adaptive Caching System) [18] and GROCOCA (GROup-based COoperative CAching) [19] are role-based caching schemes that also utilize information obtained from the network. In addition, CacheData, CachePath, HybridCache [20] and GroupCaching [21] are direct-request caching schemes that also employ the location information. Finally, ZC (Zone Cooperative caching) [22] and Sailhan [23] use direct requests and broadcast requests depending on some heuristic.

The CLIR cooperative caching scheme was proposed in [24]. It can be classified as a direct-request and information-based cooperative caching scheme. The main novelty of CLIR is the implementation of a cross-layer interception cache technique as well as the optimization of the redirection technique. Its performance was evaluated for MANETs and compared to other five cooperative caching schemes. The objective of this paper is to study the performance of CLIR in a static grid ad hoc network and compare this performance with other caching schemes.

III. PROPOSED CACHING SCHEME

CLIR implements a local cache in every node in the network. This local cache is managed using the LRU (Least Recently Used) replacement policy. Using this cache, every node stores the received documents. Therefore, further requests to the same document will be resolved by the local cache. This is called a local cache hit. As the requests must be forwarded hop by hop from the requester node to the server node, the intermediate nodes in the route from the source to the destination of the requests can reply directly if the requested document is stored in their local cache. This is called an interception cache hit.

When the route from the source node of the request to the destination node has not been created, CLIR utilizes the routing protocol to piggy-back the request in the routing protocol messages. By using this technique, the routing protocol is able to create the route to the destination node and search for the requested document at the same time. If any node that receives the route request message has a copy of the requested document in its local cache, it will reply using the route reply message informing that this node has a copy of the document. When the requester node receives the route reply message, the route between both nodes is created and the requester node will forward the request to the node that has the copy of the document. This is called a cross-layer interception hit. This mechanism allows finding the documents in the network even if the server is not temporarily available.

CLIR also implements a redirection cache that stores information about where the documents are located in the network. This information is obtained from the messages that are forwarded by the mobile node. The redirection cache manages information about the source of the requests and the corresponding replies. It also stores the number of hops and the TTL of the documents and it estimates the time that the documents are stored in the local caches. The redirection cache is managed by means of two LRU lists, one for the documents whose TTL is known and the other with the documents with an unknown TTL. When a node receives a request and the redirection cache contains information of a node that is closer to the original destination of the request, the request is forwarded to this closer node. When the redirected node receives the request, it replies with the document. This is called a redirection cache hit. In the case that the redirected node has evicted the document from its local cache, a redirection error message is sent to the redirection node in order to update the information of the redirection cache.

Finally, CLIR also implements the storage of the replied document in the node located in the middle of the route from the source and destination of the reply. So, the documents can be easily disseminated along the network. In order to avoid the excessive replication of documents, this mechanism is performed if the distance between both nodes is greater than four hops.

IV. PERFORMANCE EVALUATION

In order to evaluate the performance of the proposed cooperative caching scheme we have implemented CLIR using the NS-2.33 [25] network simulator. Additionally, for comparison purposes, the cooperative caching schemes MobEye, HybridCache, COOP, DPIP and SimpleSearch have also been implemented. Each point represented in the figures shown in this paper corresponds to the mean performance evaluation of five simulations using the same parameters but changing the seed. Depending on the simulation, the analysed variable is changed while the rest of the parameters are set to a default value. All figures include a confidence interval of 95% for each performance parameter.

A. *Simulation model*

Table 1 summarizes the main simulation parameters. We suppose that the nodes in the ad hoc network do not move. Depending on the evaluated configuration, nodes form a regular grid of 5x5, 7x7 or 9x9 nodes. Moreover, the nodes located in the corners of the simulation area, that is, in the positions (x,y)=(0,0) and (x,y)=(1000,1000), are considered to behave as Data Servers (*DS*). For simulation simplicity, we have considered a numeric identification for each document although the caching scheme can be extended to manage URLs. In order to distribute the traffic along the network, the documents with even identification are located in one server while the documents with odd identification are stored in the other *DS*.

Every node that is not a server is programmed to generate requests to the servers during the simulation time. When a request is served, another request is generated after a waiting time period. If the request is not served after a predefined timeout, the request is sent again. The document request pattern follows a Zipf-like distribution that has been demonstrated to properly characterize the popularity of the documents in the Internet [26]. The Zipf law asserts that the probability *P(i)* for the *i*-th most popular document to be requested is inversely proportional to its popularity ranking as shown in (1).

$$P(i) = \frac{\beta}{i^{\alpha}} \, . \qquad (1)$$

The parameter α is the slope of the log/log representation of the number of references to the documents as a function of its popularity rank (*i*).

TABLE 1. SIMULATION PARAMETERS

Parameter	Default values	Other utilized values
Simulation area (square meters)	1000x1000	
Number of nodes	49	25-49-81
Number of Servers	2	
Number of documents	1000	
Document size (bytes)	1000	
Timeout (s)	3	
TTL (s)	2000	250-500-1000-2000-∞
Mean time between requests (s)	25	5-10-25-50
Traffic pattern (Zipf slope)	0.8	0.4-0.6-0.8-1.0
Replacement policy	LRU	
Local Cache size (number of documents)	35	5-10-35-50
Redirection Cache size (number of registers)	35	
Simulation time (s)	20000	
Warm-up period (s)	4000	
Coverage radio (meters)	250	

As the coverage radio of the nodes is 250 meters and the simulation area is 1000x1000 m^2, the connectivity among neighbour nodes is different for each evaluated grid configuration. Figure 1 shows the connectivity for the 5x5, 7x7 and 9x9 grid configurations. As it can be observed, as the density of nodes increases the number of neighbour nodes grows.

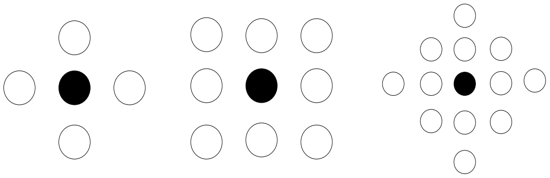


Figure 1. One hop connectivity of a node for 5x5, 7x7 and 9x9 grids.

As performance metrics we consider:

- Traffic load: It measures the mean amount of traffic generated or forwarded by each node during the simulation. As the wireless medium is limited, the greater the generated traffic the greater the probability of interferences and collisions.
- Delay: It is defined as the mean time that a request requires to be served, that is to say, the mean time that a user will have to wait to receive the requested document.
- Timeouts: This metric defines the percentage of requests that have failed and have been requested again because the document has not been received before the timeout.

The figures presented in this section correspond to the evaluation of a 7x7 grid network as the results obtained with the 5x5 and 9x9 networks are very similar. The performance evaluation will be studied as a function of the time between requests, the TTL of the documents, the Zipf slope and the local cache size.

B. *Time between requests*

Figure 2a represents the mean processed traffic by each node as a function of the time between requests. CLIR, DPIP and HybridCache are the caching schemes that generate the lowest traffic, followed by No Cache and SimpleSearch. MobEye generates more traffic because of the use of broadcast messages.

Figure 2b compares the mean delay of the requests and replies. CLIR is the caching scheme with the lowest delay. In fact, it is the only scheme that obtains a lower delay than the option of not using caches. SimpleSearch and MobEye employ a four request-reply messages method, and hence, they experience a greater delay and a greater traffic generation as previously observed. COOP has not been shown in this figure due to the high delay obtained. This behaviour is caused by the timeout needed to perform the direct request to the *DS* after the broadcast request has failed. DPIP also achieves a high delay due to the *DPIP_Timer* parameter that fixes a lower bound to the messages delay. Finally, HybridCache achieves a low performance for high loaded networks although this performance is improved as the traffic load is decreased. This fact is due to redirection loops caused by a wrong redirection management. When time between requests increases, the information stored in the redirection table is obsolete related to the documents stored in the local caches as they are evicted from the local caches before the information can be considered obsolete. As the number of evictions in the local caches decreases the redirection cache is able to obtain more redirection hits because it only takes into account the TTL of the documents to delete the information of the redirection cache.

Figure 2c shows the mean percentage of timeouts per node. HybridCache obtains a high percentage of timeouts due to the bad redirection management as previously explained. Similarly, COOP presents the same behaviour as HybridCache because of the same reasons. Finally, the rest of the caching schemes obtain a percentage of timeouts close to zero. In fact, this should be the normal behaviour of the caching schemes as the servers are always available and it is always possible to create a route to them.

C. *TTL of the documents*

Figure 3a represents the mean traffic processed by each node as a function of the mean TTL of the documents.

CLIR, DPIP and COOP generate less traffic than no Caching for all the studied TTLs. HybridCache is very sensitive to the TTL of the documents and, as the TTL is increased, the generated traffic also soars. This behaviour is due to the redirection cache, which only takes into account this parameter to delete the information in the redirection cache. Consequently, if a node evicts a document from its local cache, the nodes with information about the location of this document in their redirection caches will maintain incorrect data.

Figure 3b compares the mean delay as a function of the mean TTL of the documents. CLIR is the caching scheme that obtains the lowest delay. HybridCache, as shown in the previous study, is very sensitive to the TTL and the delay is highly increased as the TTL is incremented. The rest of the caching schemes obtain delays greater than the case of no Caching due to the four messages needed to obtain the document.

Figure 3c shows the evolution of the percentage of timeouts as a function of the TTL of the documents. COOP and HybridCache are the caching schemes with a percentage of timeouts greater than zero due to the previously commented reason. In fact, the percentage of timeouts is highly increased in HybridCache for TTLs greater than 2000 seconds.

D. *Zipf slope*

Figure 4a depicts the mean traffic processed by node as a function of the Zipf slope. CLIR is the caching scheme that obtains the lowest delay for all the slopes while MobEye and SimpleSearch generate more traffic than the No Caching option due to the broadcast requests. On the other hand, HybridCache also generates more traffic than the No Caching scheme for low slopes. This behavior is due to the replacement policy implemented by HybridCache, called SxO (Size x Order). This replacement policy is very sensitive to the popularity of the documents. Consequently, a low Zipf slope causes the reduction of the local cache hits, increasing the traffic generated in the network.

Figure 4b compares the mean delay as a function of the Zipf slope. The delay obtained by COOP is not shown because it is much greater than the rest of the caching schemes. Only CLIR and HybridCache (for a slope of 1.0) obtain a lower delay than the No Caching scheme. DPIP has a delay of even three times greater than CLIR although this difference is reduced as the Zipf slope increases. CLIR is the caching scheme with the lowest delay for all the considered Zipf slopes.

Figure 4c shows the mean percentage of timeouts per node as a function of the Zipf slope. As observed in previous studies, only HybridCache, COOP and MobEye present a percentage of timeouts different to zero. The behaviour of HybridCache and COOP is due to the incorrect implemented redirection technique. Nevertheless, the percentage of timeouts of these caching schemes is decremented as the Zipf slope increases because, as the Zipf slope increases, the percentage of local and remote cache hits increases and the documents can be served before the timeout. The rest of caching schemes obtain a percentage of timeouts close to zero.

E. *Cache size*

Figure 5a depicts the mean processed traffic by the nodes as a function of the local cache size. As the cache size rises the generated traffic is decreased because the probability of a local cache hit is increased. CLIR, DPIP and COOP are the caching schemes that generate a traffic lower than the No Caching scheme for all the studied cache sizes. MobEye is the caching scheme that generates more traffic due to the use of broadcast requests. On the other hand, HybridCache only performs better than No Caching when the cache size is greater than 20 documents. Hence, HybridCache does not work correctly when using small caches due to the implemented SxO replacement policy.

Figure 5b compares the mean delay as a function of the local cache size. CLIR is the caching scheme with the lowest delay and, in this case, is the one that performs better than the No Caching scheme for all the studied cache sizes. HybridCache presents a big delay for small caches, although it is drastically reduced as the cache size increases. In addition, SimpleSearch and MobEye always obtain a bigger delay than the No Caching scheme for all the studied cache sizes due to the four messages needed to obtain a document. Finally, DPIP shows a delay close to 150 milliseconds due to the limit imposed by the *DPIP_Timer*.

Figure 5c presents the mean percentage of timeouts as a function of the local cache size. As observed in previous studies, only HybridCache, MobEye and COOP show a percentage of timeouts different to zero. This percentage is reduced, especially in HybridCache, as the cache size increases because the probability of local and remote cache also augments.

V. CONCLUSIONS

In this paper, we have evaluated the performance of the CLIR caching scheme applied to static grid ad hoc networks. This evaluation has been performed using the metrics: mean traffic processed by the node, the delay perceived to obtain the requested documents and the percentage of mean timeouts. We have evaluated the influence of the traffic load in the network, the TTL of the documents, the traffic pattern (Zipf slope) and the local cache size. In addition, we have compared the performance of CLIR to the caching schemes HybridCache, COOP, DPIP, SimpleSearch and MobEye. Finally, the performance of CLIR has also been compared to the performance of an ad hoc network that does not implement any caching scheme.

From the set of developed simulations we can conclude that MobEye, COOP and HybridCache are not suitable for static ad hoc networks. We base this assumption in the fact that they obtain a mean percentage of timeouts different to zero. This behaviour is not acceptable in this kind of networks where the servers are always available because the wireless nodes do not move. Taking into account the rest of caching schemes (DPIP, SimpleSearch and CLIR), CLIR always obtains the lowest traffic generation as well as the lowest delay for all the studied situations. In addition, CLIR always presents a better performance than the No Caching Scheme for all the studied parameters and, hence, we can assert that it is suitable for this kind of networks.

ACKNOWLEDGEMENTS

We would like to thank Adela Isabel Fernandez-Anta for revising the syntax and grammar of this article. This study was partially supported by the National Project No. TEC2009-13763-C02-01.

REFERENCES

[1] D. Wessels, Web Caching: Reducing Network Traffic. O’Reilly, 2001.

[2] P. Kuppusamy, K. Thirunavukkarasu, B.Kalaavathi, “A Review of Cooperative Caching Strategies in Mobile Ad Hoc Networks”, International Journal of Computer Applications, vol. 29, no. 11, 2011, pp. 22-26

[3] G. Dodero and V. Gianuzzi, “Saving Energy and Reducing Latency in MANET File Access”, Proc. 26th International Conference on Distributed Computing Systems Workshops (ICDCSW'06), 2006, pp. 16-20.

[4] S. Lim, W.C. Lee, G. Cao and C.R. Das, “A novel caching scheme for improving Internet-based mobile ad hoc networks performance”, Ad Hoc Networks, vol. 4, no. 2, 2006, pp. 225-239.

[5] S. Lim, W.C. Lee, G. Cao and C.R. Das, “Cache invalidation strategies for Internet-based mobile ad hoc networks”, Computer Communications, vol. 30, no. 8, 2007, pp. 1854-1869.

[6] T. Moriya and H. Aida, “Cache Data Access System in Ad Hoc Networks”, Proc. 57th IEEE Semiannual Vehicular Technology Conference (VTC 2003), April 2006, vol. 2, pp. 1228-1232.

[7] B. Tang, H. Gupta and S.R. Das, “Benefit-Based Data Caching in Ad Hoc Networks”, IEEE Transactions on Mobile Computing, vol. 7, no. 3, 2008, pp. 289-304.

[8] Y.H. Wang, J. Chen, C.F. Chao and C.C. Chuang, “A Distributed Data Caching Framework for Mobile Ad Hoc Networks”, Proc. 2006 International conference on Wireless communications and mobile computing, 2006, pp. 1357-1362.

[9] J. Cho, S. Oh, J. Kim, K.H. Lee and J. Lee, “Neighbor Caching in Multi-Hop Wireless Ad Hoc Networks”, IEEE Communications Letters, vol. 7, no. 11, 2003, pp. 525-527.

[10] P. Nuggehalli, V. Srinivasan and C.F. Chiasserini, “Energy-Efficient Caching Strategies in Ad Hoc Wireless Networks”, Proc. 4th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc 2003), June 2003, pp. 25-34.

[11] N. Chand, R.C. Joshi and M. Misra, “Cooperative Caching in Mobile Ad Hoc Networks Based on Clusters”, International Journal on Wireless Personal Communications, no. 43, 2007, pp. 41-63.

[12] M.K. Denko, “Cooperative Data Caching and Prefetching in Wireless Ad Hoc Networks”, International Journal of Business Data Communications and Networking, vol. 3, no. 1, 2007, pp. 1-15.

[13] V. Gianuzzi, “File Distribution and Caching in MANET”, Technical Report DISI-TR-03-03, DISI Tech University of Genova (Italy), 2003.

[14] Y. Du and S. Gupta, “COOP – A cooperative caching service in MANETs”, Proc. Joint International Conference on Autonomic and Autonomous Systems and International Conference on Networking and Services (ICAS-ICNS 2005), October 2005, pp. 58-63.

[15] A. Klemm, C. Lindemann and P.D. Waldhorst, “A Special-Purpose Peer-to-Peer Sharing System for Mobile Ad Hoc Networks”, Proc. IEEE Semiannual Vehicular Technology Conference (VTC 2003), October 2003, pp. 2758-2763.

[16] G. Chiu and C. Young, “Exploiting In-Zone Broadcast for Cache Sharing in Mobile Ad Hoc Networks”, IEEE Transactions on Mobile Computing, vol. 8, no. 3, 2009, pp 384-397.

[17] C.Y. Chow, H.V. Leong and A. Chan, “Peer-to-Peer Cooperative Caching in Mobile Environments”, Proc. 24th International Conference on Distributed Computing Systems Workshops (ICDCSW’04), March 2004, pp. 528-533.

[18] H. Artail, H. Safa, K. Mershad, Z. Abou-Atme and N. Sulieman, “COACS: A Cooperative and Adaptive Caching Systems for MANETs”, IEEE Transactions on Mobile Computing, vol. 7, no. 8, 2008, pp. 961-977.

[19] C.Y. Chow, H.V. Leong, A. Chan, Group-based Cooperative Cache Management for Mobile Clients in a Mobile Environment, Proceedings of the 33rd International Conference on Parallel Processing (ICPP’04), 2004, pp. 83-90.

[20] L. Yin and G. Cao, “Supporting Cooperative Caching in Ad Hoc Networks”, IEEE Transaction on Mobile Computing, vol. 5, no. 1, 2006, pp. 77- 89.

[21] Y. Ting and Y. Chang, “A Novel Cooperative Caching Scheme for Wireless Ad Hoc Networks: GroupCaching”, Proc. International Conference on Networking, Architecture and Storage (NAS 2007), 2007, pp. 62-68.

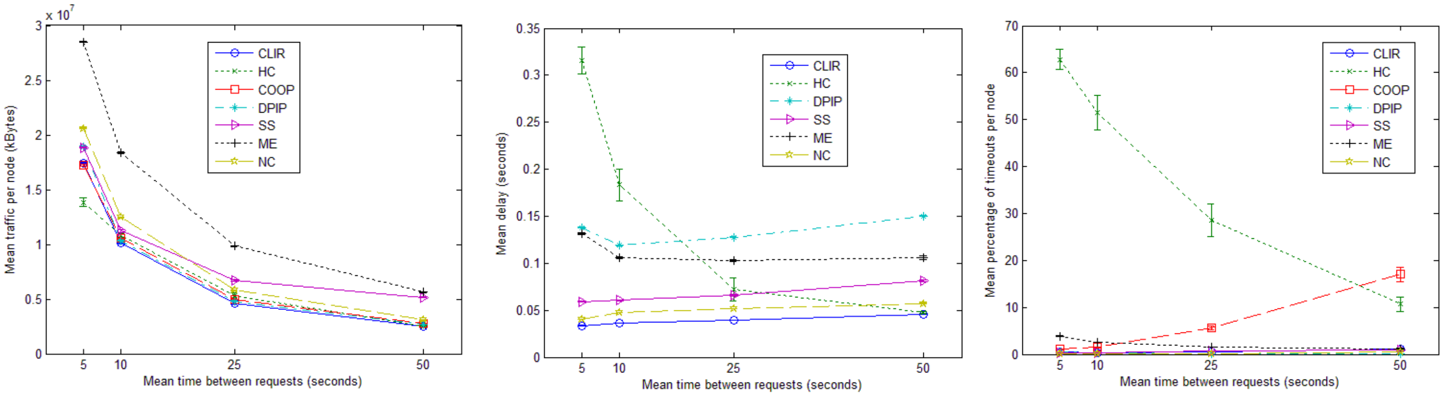
[22] N. Chand, R.C. Joshi and M. Misra, “Efficient Cooperative Caching in Ad Hoc Networks”, Proc. 1st International Conference on Communication System Software and Middleware (Comsware'06), January 2006.

[23] F. Sailhan and V. Issarny, “Cooperative Caching in ad hoc Networks”, Proc. 4th ACM International Conference on Mobile Data Management (MDM’2003), January 2003, pp. 13-28.

[24] F.J. González-Cañete, E. Casilari and A. Triviño-Cabrera, “A cross layer interception and redirection cooperative caching scheme for MANETs”, EURASIP Journal on Wireless Communications and Networking 2012, 2012:63, doi:10.1186/1687-1499-2012-63

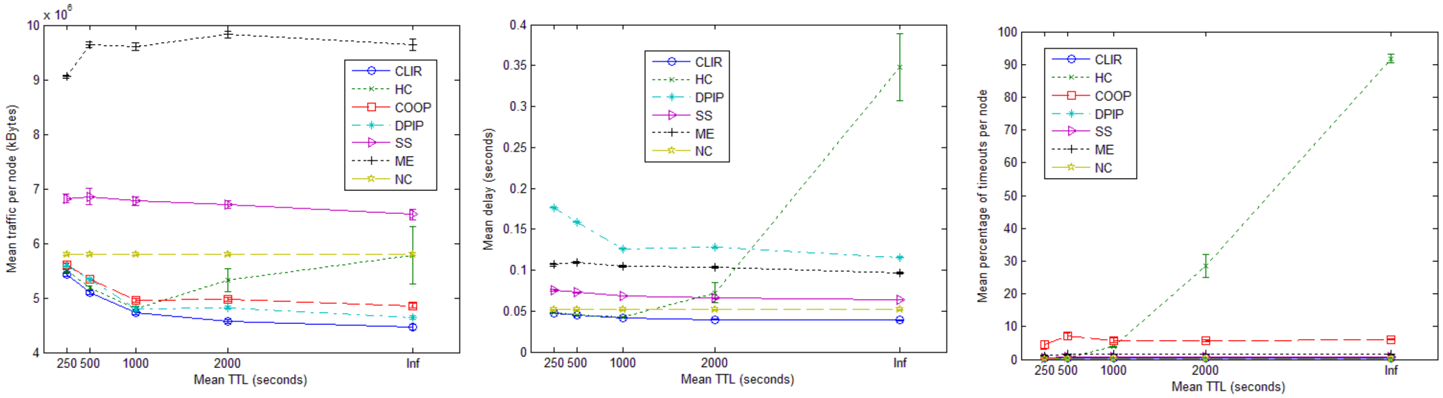
[25] http://www.isi.edu/nsnam/ns/ [retrieved: July, 2012]

[26] L.A. Adamic and B.A. Huberman, “Zipf’s law and the Internet”, Glottometrics, vol. 3, 2002, pp. 143-150.



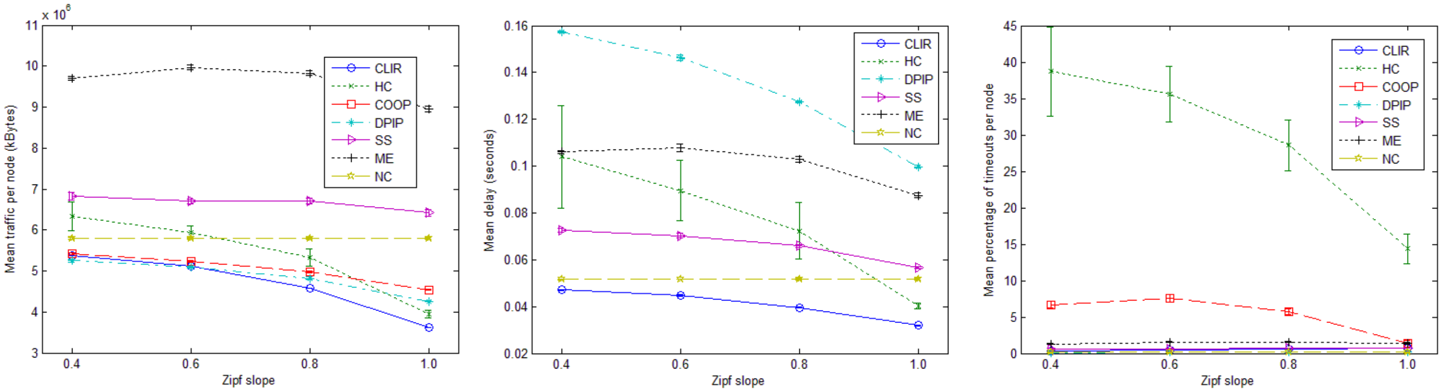
(a) (b) (c)

Figure 2. Mean traffic processed by node (a), delay (b) and percentage of timeouts (c) as a function of the mean time between requests.



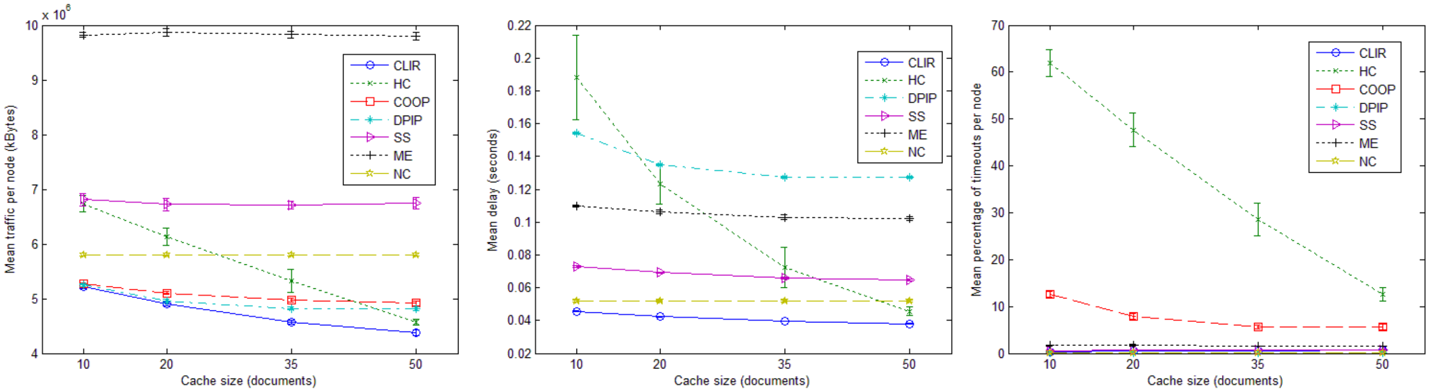
(a) (b) (c)

Figure 3. Mean traffic processed by node (a), delay (b) and percentage of timeouts (c) as a function of the mean TTL of the documents.



(a) (b) (c)

Figure 4. Mean traffic processed by node (a), delay (b) and percentage of timeouts (c) as a function of the Zipf slope.



(a) (b) (c)

Figure 5. Mean traffic processed by node (a), delay (b) and percentage of timeouts (c) as a function of the kcache size.

User-assisted Semantic Interoperability in Internet of Things

Visually-facilitated Ontology Alignment through Visually-enriched Ontology and Thing Descriptions

Oleksiy Khriyenko, Vagan Terziyan, Olena Kaikova
IOG group, MIT Department and Agora Center
University of Jyväskylä, P.O. Box 35 (Agora)
FIN-40014 Jyväskylä, Finland
e-mail: oleksiy.khriyenko@jyu.fi, vagan.terziyan@jyu.fi, olena.o.kaikova@jyu.fi

***Abstract* — Today, we make a separation between the real/physical world and the Internet. It is time for these two be blended and provide ubiquitous access and interoperability online. We are approaching Internet of Things - a forthcoming technological revolution that will radically change our environment and enable innovative applications and services. To make this happen, we have to eliminate the fragmentation in used technologies and have to make the devices be used across various applications and services. We need to find a way to actually carry out the necessary and massive deployment of ubiquitous devices. So we need to put more effort into the design of tools to automate deployment and configuration of devices. This paper tackled a problem of an effective way to support interoperability in Internet of Things. We propose visually-enriched approach for user-powered ontology alignment and semantic description of Things.**

Keywords-Mashup supported semantic visual mapping; visual ontology alignment; visual semantic human interface; semantic interoperability.

I. Introduction

With a purpose to better understand the need of proposed contribution, let us start with short samples of use case scenarios:

Scenario 1: Person is traveling by car. Suddenly something is happened with the car and it needs to be repaired. Instead of searching nearest car service station, booking a time and filling a request form describing current state of the car; the car itself searches for correspondent services in the web, collects necessary data from the correspondent modules of the car and books a time for maintenance service. During the maintenance, car gets new spare-parts and integrates them to the central diagnostic system of the car (regardless of the fact that parts are produced by different vendors). In the same manner, car might negotiate and book appropriate time for annual technical check-up taking into account timetable of the owner, been connected to his/her personal organizer. During the trip, cat might suggest optimized schedule of refueling taking into account fuel consumption, location of gasoline stations and their prices, discounts and bonuses available for the driver and other relevant contextual information.

Scenario 2: Person has bought “smart-home” system from some vendor. Vendor installs smart-home network with a set of smart-entities (sensors and actuators) and one control unit. So far, all the elements of the network belong to the same vendor and interoperate via the same ontology and communication protocol. A couple of month later, house owner buys a new smart-entity for good price from another vendor and connects it to the existing network. Later, friend of the house owner suggests some generic software application, which could be used as an upgrade of the smart-home network control unit and provides new useful features in comparison to the functionality of initial software of the control unit. This software application is produced by totally different vendor, and still can be installed to the control unit and communicate with all the connected to the smart-home network entities.

Scenario 3: Person has several measurement units (produced by different vendors) that can measure his/her heartbeat rate, arterial pressure, distance person walked or run, and some other parameters related to his/her health condition and physical activities. Person easily connects all these devices to a smart-phone to be able to log and observe them. Later, from an application store, person buys an application that suggests correspondent diet, taking into account all the measured personal data. Entering a supermarket and to be connected to the local infrastructure of it, application starts to navigate person to the correspondent location of the suitable products for his/her diet or alert the person when he/she puts to the basket a product which consists inappropriate ingredients.

All mentioned above stories are not fantasies. It is our tomorrow and, in some cases, even our today. Unfortunately, in case of our today, we have integration of systems produced by the same vendor. Supporting one interoperability modem in several products, vendor creates integrated environment for various applications and interaction scenarios to be run on it. All these applications should support correspondent predefined API and data model. But, it is not what we expect to be in our tomorrow. We need an open environment with possibility to integrate various systems and components (hardware, apps, communication channels, etc.) produced by different vendors (see Figure 1). With a goal to achieve such requirements, we are approaching Internet of Things (IoT) - a forthcoming technological revolution that will radically change our environment and enable innovative applications and services.

Above the personal level, the IoT will also have an important impact on enterprises and on society in general. IoT will enable a global connectivity between physical objects (connecting “things”, not only places or people), will bring real-time machine-published information to the Web, as well as will enable a better interaction of people with the physical environment by combining ubiquitous access with ubiquitous intelligence. IoT will consist of a heterogeneous set of devices and communication strategies between them. Such a heterogeneous system should evolve into a more structured set of solutions, where “things” are uniformly discoverable, enabled to communicate with other entities, and are closely integrated with Internet infrastructure and services, regardless of the particular way (RFIDs, sensors, embedded devices) in which they are connected to the IoT. In this context, one of the challenging bottlenecks is to support interoperability between “entities” on a semantic level [1][2]. Therefore, in this paper we propose an approach towards visually-enriched semantics as an infrastructure for user-powered semantic technology enhancement.

Paper consists of two main sections. Section 2 addresses semantic integration platform for IoT and a vision of user-powered consumption of semantic technologies. Section 3 presents a human-assisted ontology alignment approach.

II. Thing Integration Environment

A. Smart Gateway - semantic integration platform for IoT

We are already in the middle of era of automated machine communication. There is already a lot of machine-to-machine communication going on out there; parking meters are connected, and vending machines automatically report when new supplies are needed. Every minute huge amount of data are being exchanged between machines for various purposes within various sectors. However, there is a big challenge in moving beyond application-specific devices and establishing an information model that will create re-use of the data generated by devices for new applications in different domains. Finding the right horizontal points in the solutions is a key. There are already useful deployments within the transport, automotive, building, health and utility sectors, but everything is still very sector-specific. We need to create an infrastructure that will make information generated from a car or a building understandable not only within their own specific application/system, but across of various applications and domains.

The IoT will require interoperability at multiple levels. On the hardware side, such problems have to be addressed as handling a capability mismatch between traditional Internet hosts and small devices, as well as handling widely differing communication and processing capabilities in different devices. In the interface between the device and network domains, IoT gateways will provide a common interface towards many heterogeneous devices and networks [3]. We assume that all “things” (devices, sensors, actuators, etc.) are connected to the web. Digital “things” such as services usually are accessible through the web. Applications might be downloaded and installed to the integration platform - Smart Gateway. Thus, we have correspondent requirements for such a platform. Smart Gateway should allow installation of applications and further configuration of communication model with it, based on accompanied annotation of the application. In case of services, Smart Gateway should be able to access semantic annotation through service access point and configure communication model with it as well.

Talking about physical world objects (device, sensors, etc.), usually they are accessible through the gateway - a control unit of a network provided by the same vendor. The only requirement - gateway should be presented in the web as a service with a set of capabilities provided by “things”

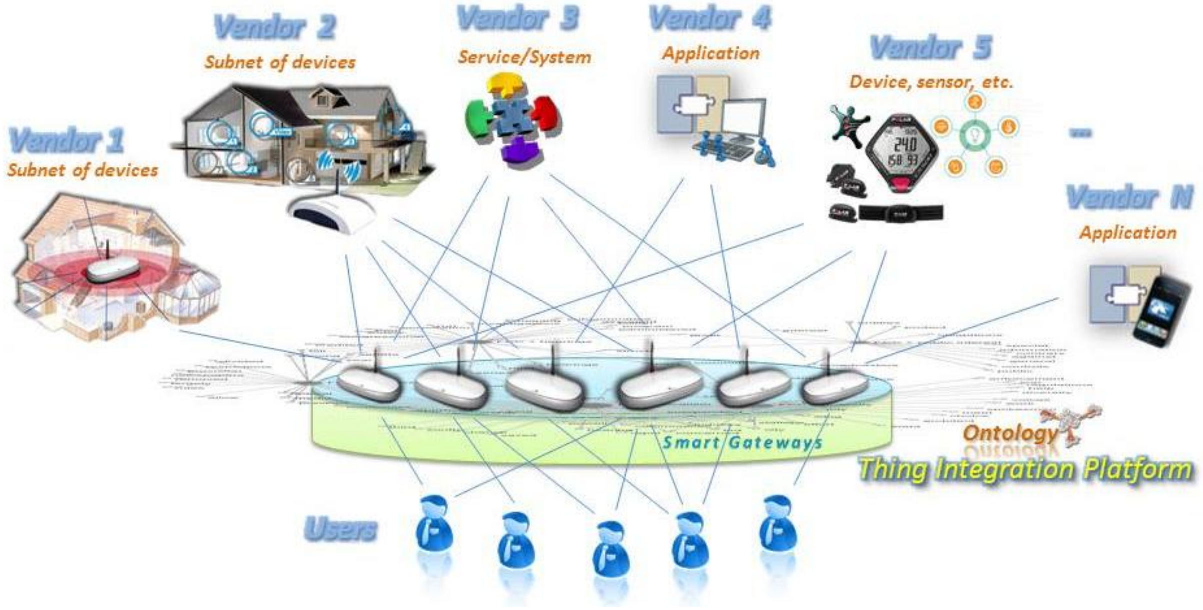


Figure 1. Thing Integration Environment.

connected to the gateway (providing data or doing some actions). In case we cannot have single "thing" personally connected to the web, we should deal with sub-network that consists of mentioned "thing" and correspondent gateway. Thus, we will have a set of gateways connected to the web and ready to become a part of the integration environment. Having all the gateways accessible as web-services, all connected to them real world "things" become digital entities and might be registered to the Smart Gateway.

Depending on a business model, Smart Gateway might be a part of a gateway, provided by certain vendor that would like to promote own network solution as en extendable open environment for entities produced by other vendors. Having Smart Gateway as a part of local network of connected "things", services and applications is a suitable model in case of time-limited and highly secured runtime systems. At the same time, Smart Gateway might be considered as a separate integration solution - services located in the Cloud and accessible through the web. Been easily accessible, such "thing" integration service might be very popular among ordinary people who would like to create and manage their own smart spaces, integrate various services with ubiquitous "things". Relevant research has been done in "Smart Resource" and "UBIWARE" projects with respect to Global Understanding Environment (GUN) [4].

B. *User-powered consumption of semantic technologies*

To achieve the vision of ubiquitous 'things', the next generation of integration systems will need different methods and techniques such as Semantic Web [5][6], Web Services [7][8], Mashups [9], Linked Data [10][11], etc.

Semantic based technologies are viewed today as key technologies to resolve the problems of interoperability and integration within the heterogeneous world of ubiquitously interconnected objects and systems. Semantic Web is a vision with an idea of having data on the Web defined and linked in a way that it can be used by machines not just for display purposes, but for automation, integration and reuse of data across various applications. Semantic Web is considered as a standardized approach to achieve automated interoperability of heterogeneous systems/applications. Heterogeneity of systems and various data sources become a bottleneck for automated service integration, data processing and reuse. To make data ready to be consumed and processed by external systems, data sources and data should pass through the semantic adaptation [4][12] and be accessible in common uniform way. Due to the huge amount of application areas that Semantic Web technology tried to cover, community started to elaborate different standards and techniques to solve interoperability problems. As a result, we have a big variety of separated islands of information and management systems. These information islands internally follow the Semantic Web vision, but are heterogeneous from the general (global) interoperability point of view. This leads to the fact, that society and especially its business-oriented part has started to doubt that such widely spread activity will be so much beneficial for them. Only some applications and systems in restricted domains became really useful. Most probably, the reason for this is the decentralization of uncontrolled activities, which creates new problems on the way towards ubiquitous Semantic Web. There are no doubts that Semantic Web is a very promising technology, but it definitely lacks more centralized management or at least an environment that plays coordinative and supportive role and directs users towards proper technology utilization.

Services providers, as well as producers of "things", are the end users of the service-oriented technologies. They need appropriate controlled support from the infrastructure that facilitates interoperability of services/devices, integration of heterogeneous data sources, and provides platform for new services/application development. Thus, we have to provide such a coordinative and supportive environment that will facilitate development and growth of service and smart-entity market. With respect to the current state of the art, we cannot expect that community of service providers and smart-entity vendors will build one global integration infrastructure with common ontology. We cannot expect that someone else (alone or in a consortium) will do the same. Current achievements in the area of interoperability of heterogeneous systems present technologies and tools for experts to build and manage adapters between heterogeneous systems or their components. Semantic Web is a technology for machines to better perform, providing services for human in automated or semi-automated way. In a case of unavailability of a common data model, we have to deal with semi-automated performance of the system when human become involved to the process not just as a consumer, but as an expert - necessary part in the chain to supervise and correct the process performed by machines [13].

With an increase in the development of ontologies, we need tools and techniques for solving heterogeneity problems between different ontologies. Therefore we need ontology alignment [14][15][16][17], which helps us to bring different knowledge representations into mutual agreement. With respect to the scenarios mentioned above, ontology definitions of all the smart-entities and applications/services should be (semi-)automatically aligned by control unit of the network to ensure interoperability of them in a unified way. Regarding to the mentioned ontology alignment techniques, we may expect automatic alignment for simple and similar ontologies, but in all other cases, we will definitely need a human be involved into the process. This is largely a human-mediated process. There are existing tools which can help with identifying differences among ontologies [18], but user interaction is still essential in order to control, approve, and optimize the alignment results.

Unfortunately, approaching the era of ubiquitous services and IoT, we cannot expect availability of huge amount of professional experts involved to the daily processes of "things" interoperability support. We have to find a solution to bring technology closer to the ordinary user and make him/her able to not only utilize services, but to setup, configure and supervise interoperability process. We expect a human to be not only an end-user/consumer of technology world, but also to become an integral part of it, providing own expertise and capabilities. In all mentioned scenarios, person (owner of the smart-network) should be able to help the system to perform a proper ontology alignment through correspondent human interface of the alignment system. Owner of interoperable system does not only consume a service delivered by smart machines, but also plays a valuable role as a supervisor of interoperability process. Therefore, among variety of other adapters between heterogeneous entities, bridge to the human (human-to-machine H2M and machine-to-human M2H interfaces) becomes one of the most important tools of next generation integration infrastructures.

Such adaptation of the human to the technology world might be provided by Personal Assistant (PA) - supportive agent assigned to every user [13]. From one side, it should deal with human personality and adapts to his/her personal ontology and personal perception of environment. From another side, it should support common semantic standards and approach to be interoperable with other surrounding digital world entities (applications, services and systems). The main features of PA (among others) are:

- Enabling personal user ontology creation and ontology driven resource annotation;
- Ability to adjust to the personal user ontology, to the way user perceives the environment, information and knowledge;
- Ability to build personalized semantic mind-map based on user behavior and preferences;
- Enabling personalized natural user-driven way of querying, filtering, browsing and presentation of information.

Personalized representation of information very much concerns a human supervised ontology alignment process. Ontologies very much differ from each other. The more specific, detailed and complex ontology we make, the more semantic value it has, but, it makes harder to integrate ontology with others. Taxonomies of different ontologies are not likely to be the same. Even developed by professionals, we still have different ontologies for the same problem domain. It would seem that experts, involved to the same domain, should operate with the same terms, use the same vocabulary and knowledge representation model. But, people are different, context and personal perception of surrounding world brings problems to interoperability process. As a part of the processes, human brings a certain level of uncertainty, and only human my help to solve the problem so far. Thus, to avoid heterogeneity in the resource annotations and simplify ontology alignment for automated interoperability between digital elements of the technology world, we may admit a necessity of personalized adaptation of every human (no matter whether it is an expert (knowledge provider) or user/customer) to the common information/data model.

III. HUMAN-ASSISTED ONTOLOGY ALIGNMENT

A. *Visually-facilitated semantic matching*

Let us consider a scenario of installation of a new floor-heating regulator to a “smart-home” system. Assuming that we have two different vendors (Vendor A - producer of the Control Unit for the smart-home system, and Vendor B - producer of the regulator for a floor-heating system), we have two different ontologies Ontology^{V}_{A} and Ontology^{V}_{B}. Vendor A logically defines all the floor-heating systems with respect to the room the system is associated with. Thus, Ontology^{V}_{A} might contain such concepts as: living room floor-heating system, bedroom #1 floor-heating system, bedroom #2 floor-heating system, kitchen floor-heating system, bathroom floor-heating system, etc. From the Vendor A point of view, all these concepts refer to absolutely different sub-systems in the “smart-home” network. On the other side, association of the floor-heating system with particular room/place does not matter for Vendor B. Therefor, “floor-heating system” concept in the Ontology^{V}_{B} is a more general and independent entity. Moreover, most probably “floor-heating system” concept will be named very much different in those two ontologies and automated alignment will be absolutely impossible.

Since the Control Unit of the smart-home is a more general device (in comparison to specific Floor-heating system) and deals with many other devices and systems in the installed network, it utilizes more wide ontology. Therefore, to allow interoperability between the Control Unit and Floor-heating system, we have to map Ontology^{V}_{B} to wider Ontology^{V}_{A}. At the same time, we have to pay attention to the user’s (“smart-home” owner’s) Ontology^{H}_{i}. In general case, every human has own personal ontology that will be supported by his/her PA for interaction with devices, services, applications and systems. But, for any system/application, to be a mediator between the human and some other system with its own ontology, personal human ontology itself should be mapped with ontology of mediator-system in advance. PA will collect correspondent alignments of personal human ontology with ontologies of various mediating systems that human will be interacted with.

Assuming that fully automated alignment is not possible, we do not consider the cases with very simple and self-descriptive ontologies, where automated alignment might be done based on matching of synonyms of the property manes. Correspondent example of the research at this direction is a work performed in the IoT project, where authors are trying to minimize human involvement to the process of establishing interoperability between heterogeneous systems [3]. They try to retrieve (to build) ontologies from examples of massages that systems operate in communication process (requests, response, etc.). Authors build simple plane ontologies based on names of parameters used in the messages. Later, ontologies are automatically aligned and correspondent alignments are used for automatic interoperability between heterogeneous systems in runtime. But, as was mentioned, it might work in case of self-descriptive messages, where parameters are named by words that make sense, without abbreviations and shortenings, and preferably in the same language. In all other cases (cases with complex hierarchy of sub-classes, cases of different domain description models, cases of multilingual and multicultural ontology definitions, etc.) this would not work automatically and will require human assistance. Thus, in cases of human-assisted alignment of personal human ontology or ontologies provided by different vendors, we need an innovative suitable for non-expert mechanism and correspondent user interface for ontology alignment.

With respect to the research [19][20][21][22], there are some available Ontology Alignment and visualization tools: Foam algorithm [23], multiple-view plug-in for Protégé [24] - AlViz [25], BLOOMB system [21] and Knowledge Modeler[26]. There graphical primitives such as point, line, area, or volume are currently utilized to encode information. These objects are characterized by position in space, size, connections & enclosures, shape, orientation, and visual cues like color and texture, with temporal changes, and viewpoint transformations. Unfortunately, all these tools were elaborated for domain experts who know what ontology is and what information models might be used. Such tools present a lot of statistical data and analytics that might be very useful for the ontology engineer, but not for the ordinary user of a service. Information visualization should aim at making complex data easy accessible and understood for interactive investigation by the user. In case of smart-home, we expect that user has a basic knowledge about a domain and functionality of the system. Therefore, we have to find more suitable approach for user-assisted ontology alignment. To be easily recognized by human, concepts and properties of different ontologies must be presented in the most understood form - in a form of image. An image (or other visual form) is the most common information representation model for human. It helps to understand the meaning and avoid verbal uncertainty presented in textual form. Therefore, user interface should be able to present semantics through interactive image mash-ups and user-friendly browsing mechanism.

Figure 2 shows us possible visual interpretations of the Vendor A, Vendor B and user ontologies with respect to the scenario of adding the living room floor heating system to the “smart-home” network. Since we are not consider “smart-home” owner as an expert in ontologies and complex control systems, we cannot expect that it would be possible for him/her to utilize currently available solutions for ontology alignment. Only we can expect is awareness of the user about purpose, capabilities and main functionality of the “smart-home” Control Unit and floor-heating system that he/she would like to add to the “smart-home” network. Having even such limited expertise of the problem domain, user is able to browse visual description of the Control Unit (structure of sub-systems, capabilities, inputs and outputs, properties, etc.) and description of the floor-heating system from another vendor to provide appropriate matching. User can intuitively map concepts and properties presented by images. Applying possible results, achieved on background from integrated modules of automated ontology alignment, Visual Ontology Alignment Tool assists user with suggestions and requests next necessary alignments caused by alignments made on the previous steps. Additional textual descriptions of visual annotations support user to make correct mapping. As a result, correspondent parts of ontologies Ontology^{V}_{A} and Ontology^{V}_{B}, which are related to the communication scenario between “smart-home” Control Unit (Vendor A) and floor-heating system (Vendor B), will be mapped and correspondent alignment will be used in runtime operation of the “smart-home” network.

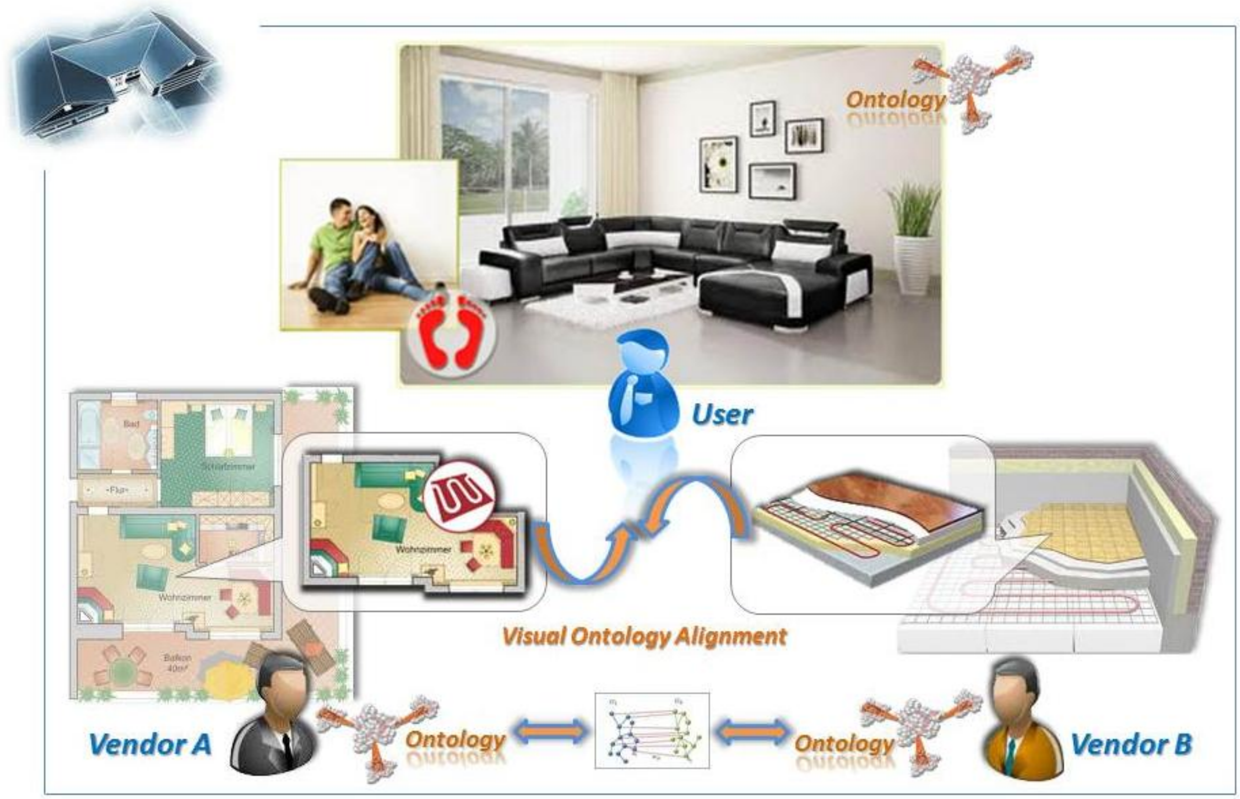


Figure 2. Visually-facilitated Ontology Alignment.

B. Visually-enriched ontology and resource description

To operate with visual representation model in a smart way, visualization tool should retrieve correspondent images together with ontologies. It means that ontologies should be extended with additional layer that contains visual definition of the concepts, classes and properties. Later in the text, we will call such visually-enriched ontology as Visual Ontology (VisOntology). We consider two scenarios of human-assisted visual enrichment (see Figure 4). In the first case, Ontology/Domain Expert creates VisOntology using Ontology Visual Enrichment Tool that adds correspondent image layer to the ontology. Later, Vendor provides annotation of the Service/System that was produced by Vendor. In the second case, Vendor itself provides visually-enriched annotation (VisAnnotation) of the produced Service/System using Visually-enriched Resource Description Tool based on regular domain ontology provided by Ontology/Domain Expert. In this case, visually-enriched ontology might be automatically created from the visually-enriched resource description during the annotation process. In case when it is difficult to associate any image with some of the concepts, tool will create an image with a correspondent text (word, character, sign, etc.) retrieved from the name of ontology element. One more case might have a place if we consider possibility for some third party to substitute Vendor in the Service/System annotation process and provide visual annotation in both previous cases. Both tools that were mentioned in the above use-cases have the same nature and similar functionality. Thus, let us consider them as a single tool for visual semantic enrichment.

The main purpose of the tool is to help user to brows ontology and assign "visSemantics" property to every class, property and instance of the ontology. Figure 3 presents possible extension of RDF Schema with "visSemantics" property used for VisOntology and VisDescription. Talking about resource annotation, tool creates an annotation template based on assigned ontology and provides possibility to add visual description. In such a way, tool extends the concepts of the ontology with "visSemantics" property and correspondent value in a form of image. Currently we consider the range of this property as a literal URL of an image. In more advance version of VisOntology and VisDescription of the resource the range might be extended to video, audio or any other multimedia content.

```
@prefix :    <http://www.example.org/sample.rdfs#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#>.

rdfs:visSemantics
  rdf:type rdf:Property;
  rdfs:domain rdfs:Resource;
  rdfs:range rdfs:Literal.

:FloorHeatingSystem
  rdfs:subClassOf :HeatingSystem;
  rdfs:visSemantics "www.example.org/FHSystem.jpeg".
```

Figure 3. "visSemantics" property.

Visual enrichment is individual, as long as a set of images, used by VisOntology and VisAnnotation providers, is individual. Tool allows user to make key-word based image annotation/tagging and create a personal pool of annotated visual content for further use. Later, such annotated content for visual enrichment will be easily retrieved based on user search request or automatically suggested to a user based on attributes of self-descriptive elements of ontology. In case we have old-fashioned service/system description based on ordinary ontology, enrichment of the description might be still automated on some extend. Based on names of ontology concepts, Visual Enrichment Tool may search among annotated visual content and build visual layer automatically. Quality of automated enrichment might be relatively low in comparison to human assisted enrichment. But, even in worst cases, when we do not have any human involvement at the stage of resource annotation, it might help to retrieve at least some visual content for further visual ontology alignment process.

Taking into account growing trend towards sharing and reuse of content, annotated visual content might be shared through various clouds and common spaces. Thus, tool can use not only own user's visual content, but also will allow user to manage and extend his/her virtual visual content space with external sources. As a continuation of the work, authors also consider a possibility to utilize Social Web to share visual annotation content and VisOntologies.

IV. CONCLUSION AND FUTURE WORK

With the aim to elaborate an environment that enables integration of heterogeneous "things" and intelligent distributed systems within the Internet of Things framework, authors address the mechanism of human-assisted simplification of semantic matching to allow interoperability of entities in the IoT. Assuming unavailability of a sufficient amount of professional experts to be involved to the daily "things" integration support process, we proposed the way to make user be not just a consumer of thing-based services, but also an expert capable to compose and establish interoperability among the things. Taking into account specifics of the potential user and unsuitability of current ontology alignment tools for it, this paper presents a human-driven approach towards visually-facilitated ontology alignment through visually-enriched ontologies and resource (thing) descriptions. Current implementation of correspondent toolset is concentrated on and consists of an interface for the final stage - Visual Ontology Alignment Tool that assumes existence of VisOntologies and VisDescriptions of Things. Implementation of the tool for visual enrichment of ontologies and resource descriptions is considered as a future continuation of presented work.

ACKNOWLEDGMENT

This research has been performed as part of IoT Tivit SHOK Program in MIT department (University of Jyvaskyla, Finland) funded by TEKES and consortium of industrial partners. Authors are grateful to the project team members from VTT (as well as other partners of the project) involved to the correspondent task for fruitful cooperation.

REFERENCES

[1] D.J. Cook and S.K. Das, “How smart are our environments? An updated look at the state of the art”. Pervasive and Mobile Computing. 3(2), pp. 53-73, 2007.

[2] J. Honkola, H. Laine, R. Brown, and O. Tyrkko, “Smart-M3 information sharing platform”. Proc. IEEE Symp. Computers and Communications (ISCC’10), pp. 1041-1046, 2010.

[3] K. Kotis and A. Katasonov, "Semantic Interoperability on the Web of Things: The Smart Gateway Framework", CISIS 2012, Palermo, Italy, 2012.

[4] O. Kaykova, O. Khriyenko, D. Kovtun, A. Naumenko, V. Terziyan, and A. Zharko, “Challenges of General Adaptation Framework for Industrial Semantic Web”, In: Amit Sheth and Miltiadis Lytras (eds.), Semantic Web-Based Information Systems: State-of-the-Art Applications, CyberTech Publishing, pp. 61-97, 2007.

[5] Semantic Web, 2001. URL: http://www.w3.org/2001/sw/

[6] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web”, Scientific American 284(5), 2001, pp. 34-43.

[7] A. Ankolekar, M. Burstein, J.R. Hobbs, O. Lassila, D.L. Martin, D: McDermott, S.A. McIlraith, S. Narayanan, M. Paolucci, T.R. Payne, and K. Sycara, “DAML-S: Web Service Description for the Semantic Web”, 2002. URL: http://www-2.cs.cmu.edu/~terryp/Pubs/ ISWC2002-DAMLS.pdf.

[8] M. Paolucci, T. Kawamura, T.R. Payne, and K. Sycara, “Importing the Semantic Web in UDDI”, 2002. URL:http://www-2.cs.cmu.edu/~softagents/papers /Essw.pdf

[9] EM. Maximilien, A. Ranabahu, and K. Gomadam, “An Online Platform for Web APIs and Service Mashups”. In IEEE INTERNET COMPUTING, IEEE Computer Society, 2008, pp. 32-43.

[10] T. Berners-Lee, “Linked Data - Design Issues”. 2006. URL: http://www.w3.org/DesignIssues/LinkedData.html

[11] T. Heath and C. Bizer, “Linked Data: Evolving the Web into a Global Data Space” (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. 2011.

[12] O. Khriyenko and M. Nagy, “ Semantic Web-driven Agent-based Ecosystem for Linked Data and Services”, In: Proceedings of the Third International Conferences on Advanced Service Computing, 25-30 September, 2011, Rome, Italy, 8 pp.

[13] O. Khriyenko, “Collaborative Service Ecosystem - Step Towards the World of Ubiquitous Services”. In: Proceedings of the IADIS International Conference Collaborative Technologies 2012, 19-21 July, 2012, Lisbon, Portugal.

[14] V. Spiliopoulos and G. A. Vouros, “Synthesizing Ontology Alignment Methods Using the Max-Sum Algorithm”, Knowledge and Data Engineering, IEEE Transactions on, vol.PP, no.99, pp.1-1.

[15] K. Kotis, A. Katasonov, and J. Leino, “Aligning Smart and Control Entities in the IoT”, In: Proceedings of the 5th Conference on Internet of Things and Smart Spaces, 27-28 August, 2012, St.-Petersburg, Russia.

[16] K. Kotis, A. Valarakos, and G. Vouros, "AUTOMS: Automating Ontology Mapping through Synthesis of Methods.", In: Proceedings of the International Semantic Web Conference (ISWC'06), Ontology Matching International Workshop, Atlanta USA, 00/2006.

[17] A. Valarakos, V. Spiliopoulos, K. Kotis, and G. Vouros, "AUTOMS-F: A Java Framework for Synthesizing Ontology Mapping Methods", i-Know,07, Graz, Austria, 00/2007.

[18] P. Shvaiko and J. Euzenat, “Ontology matching: state of the art and future challenges”. IEEE Transactions on Knowledge and Data Engineering, 2012.

[19] J. Pina, E. Cerezo, and F. Seron, “Semantic visualization of 3D urban environments”. Multimedia Tools and Applications, Volume 59, Number 2 (2012), 505-521, DOI: 10.1007/s11042-011-0776-3.

[20] M. Lanzenberger and J. Sampson, “Human-Mediated Visual Ontology Alignment”. HCI (9) 2007: 394-403.

[21] P. Jain, P. Hitzler, A.P. Sheth, K. Verma, and P.Z. Yeh, “Ontology Alignment for Linked Open Data”. In: Proceedings of the 9th International SemanticWeb Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Springer-Verlag (2010) 402–417.

[22] F. Kboubi, A.H. Chaibi, and M.B. Ahmed, 'Semantic Visualization and Navigation in Textual Corpus'. In: CoRR abs/1202.1841, 2012 .

[23] M. Ehrig and Y. Sure, “Ontology mapping - an integrated approach. In: Bussler, C., Davis, J., Fensel, D., Studer, R. (eds.) Proceedings of the First European Semantic Web Symposium, 10-12 May, 2004, Heraklion, Greece.

[24] Protégé-owl (Stanford Medical Informatics) - http://protege.stanford.edu/overview/protege-owl.html

[25] M. Lanzenberger and J. Sampson, “Alviz - a tool for visual ontology alignment”. In: Society, I.C.S. (ed.) Proceedings of the IV06, 10th International Conference on Information Visualization, London, UK, July, 2006.

[26] A. Sheth and D. Avant, "Semantic Visualization: Interfaces for exploring and exploiting ontology, knowledgebase, heterogeneous content and complex relationships," NASA Virtual Iron Bird Workshop, March 31 and April 2, 2004, CA.

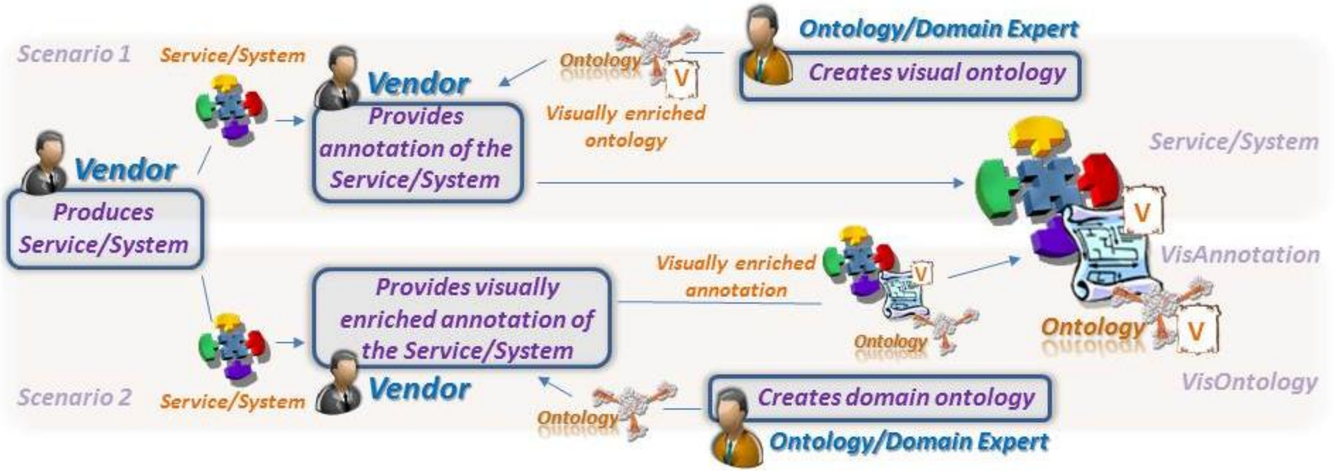


Figure 4. Human-assisted visual enrichment scenarious.

Fair Power Control in Cooperative Systems Based on Evolutionary Techniques

Konstantinos Chatzikokolakis [1], Roi Arapoglou [1], Andreas Merentitis [2], Nancy Alonistioti [1]

[1] National and Kapodistrian University of Athens, Greece, [2] AGT Group (R&D), Germany
{kchatzi, k.arapoglou, nancy}@di.uoa.gr, amerentitis@agtinternational.com

***Abstract*— Cognitive Radios have emerged as a promising paradigm for increasing spectrum utilization and alleviating the spectrum scarcity problem. However, the majority of works in the Cognitive Radio domain focus on the interaction between the primary and secondary users, while the efficiency and fairness of transmissions between secondary users are rarely explored. In this scope, we introduce an algorithm for fair transmissions in cooperative Cognitive Radio networks. The proposed scheme places particular emphasis on the QoS of underprivileged users, while maintaining a high overall network utility. Specifically, a Genetic Algorithm is designed and used to select transmission power values, under fairness constraints. The proposed algorithm is evaluated through extensive simulations. Results indicate significant improvement in the SINR of underprivileged users with minimal impact in the overall network utility.**

Keywords-component; cooperative power control; interference mitigation; fairness; genetic algorithm

I. INTRODUCTION

A plethora of novel communication technologies and wireless standards were developed during the last decade, in order to provide wireless users with enhanced Quality of Service (QoS). These novel telecommunication technologies coexist with legacy systems for extended periods of time. Furthermore, emerging wireless network environments are characterized by a growing need for spectrum, especially for high data rate applications. In this context, static frequency allocation schemes are considered too constrained for coping with the previous challenges. Actually, the paradox is that licensed spectrum use is not high, as mentioned by Defense Advanced Research Projects Agency (DARPA) in [1]. This leads to the observation that dynamic spectrum access techniques will play a catalytic role towards addressing the spectrum scarcity problem [2], [3]. A promising technology for efficient spectrum utilization is Opportunistic Spectrum Access (OSA) [4]. OSA introduces opportunistic reallocation of unused spectrum bands, also known as white spaces. Cognitive Radios (CRs) constitute a key enabler for OSA.

Cognitive Radios, first introduced by J. Mitola, are radio systems able to sense the unused spectrum and adapt their operating characteristics to the real-time environment [5]. In this direction, CRs should decide on the best spectrum band, over all available, in order to meet QoS requirements. A typical cognitive radio network comprises of secondary cognitive users trying to transmit in unused frequency bands. Their main objective is to communicate without causing interference to existing primary users. However, secondary users also compete with each other for resource allocation. In this scope, power control between secondary users is a particularly important aspect of the resource allocation problem, directly impacting the QoS, performance and energy efficiency of the wireless network.

In addition to high spectrum utilization, a key requirement for CR networks is that resource allocation should be fair and every cognitive user should have the opportunity to transmit. A comprehensive definition of "fairness" is difficult to be given, but it can be described intuitively as the ability to provide equal satisfaction to all users. Specifically for computer networks a formal performance parameter for fairness is given by the equation below [6]:

$$fairness = \frac{(throughput)^{\beta}}{\overline{delay}} \tag{1}$$

where β is a weight factor. The main obstacle in treating fairly each user is that fairness function is non-convex and may have several maxima. As a consequence, it is quite challenging to achieve the optimal throughput for every user in a network.

The throughput of each individual user, as well as various other aspects of the operation of a cognitive radio network [7] is largely dependent on the transmission power level (Tx). Therefore, a cognitive user is trying to choose an appropriately high transmission power value, targeting to keep the quality of the signal at the receiver at tolerable levels. However, if all cognitive users demonstrate selfish behavior and transmit using the maximum valid power, the outcome will be an increased interference among them and more importantly to the primary users. For these reasons, several cooperative algorithmic schemes were proposed for power control in cognitive wireless networks [8]. Such schemes mainly focus on optimizing the performance of the network as a whole, ignoring the characteristics and QoS requirements of each cognitive user. Under these assumptions, a typical phenomenon is that depending on their relative locations, a portion of cognitive users get high power values, in order to transmit, and the rest are assigned significantly lower ones, in order to mitigate interference and reach a steady state for the system. However, there is little point in maximizing overall network performance without taking into consideration the actual performance of each cognitive user. For this reason, there is a strong need for power control algorithms, which conform to the concept of fairness and provide increased opportunities for transmission to the underprivileged users.

A widely used resource allocation scheme in wireless networks is max-min fairness [9]. In this approach, wireless nodes try to achieve enhanced resource allocation starting from a minimum valid level, until all nodes are assigned resources fairly. An important drawback of max-min fairness scheme is the need for extensive message exchange among wireless nodes in order to be fully synchronized. Additionally, such schemes typically require a full knowledge model, which implies perfect message exchange, an assumption that is often not a realistic especially for cognitive radios operating under high uncertainty.

In this paper, a novel technique is proposed in order to enhance fairness properties in cooperative power control. The introduced approach is based on the distributed and cooperative power allocation scheme of [10] that is known to perform well under uncertainties. However, the original algorithm lacks fairness, as the power level of each cognitive user is not examined over time in order to reject consistently low level power values. The key contributions in the current paper are:

- The extension of the algorithm proposed in [10] with a fairness module that caters for underprivileged users. Specifically, a fairness check point is executed every time cognitive users calculate their power values to transmit. In this case, each cognitive user is examined i.e., if he was treated in an unfair way for a certain chronicle window in the past. If so, enhanced power values are generated by the evolutionary execution of Genetic Algorithm.
- The evaluation of the algorithm's behavior in cases of an incomplete knowledge model (i.e., some of the users may not know all the information). This is particularly important for real systems, since a full knowledge model is typically an unrealistic assumption.

The rest of the paper is organized as follows: Section II describes the baseline algorithm for cooperative power control. In Section III, fairness issues are discussed and a brief description of Genetic Algorithms is provided. Additionally, assumptions are formulated for the proposed fairness scheme. Furthermore, Section IV evaluates the performance of the proposed fairness scheme, comparing the Genetic Algorithm execution with a simplified fairness policy. Finally, in Section V, the key points of the proposed technique are summarized.

II. Cooperative Power Control Algorithm

In this section, a description of the algorithm in [10] is provided, in order to set the basis for the proposed fairness scheme. The main scope of the algorithm is to mitigate interference among cognitive users in licensed exempt spectrum bands. For this reason, each transmitter computes its power by taking into consideration both its Signal to Interference plus Noise Ratio (SINR) and the interference it causes to the other users. This formula prevents users from always setting their power to the maximum valid power level.

Initially, a set of L pair nodes is considered operating at the same frequency band, where K channels are available. The SINR of the *i*-th transmitter ($i \in \{1, 2,.., L\}$) in *k*-th channel ($k \in \{1, 2, ..., K\}$) is calculated by the equation given below:

$$\gamma_i(p_i^k) = \frac{p_i^k \cdot h_{ii}}{n_o + \sum_{j \neq i} p_j^k \cdot h_{ji}} \tag{2}$$

where,

- p_i^k is the power of *i*-th transmitter on channel *k*
- h_{ii} is the link gain between *i*-th receiver and *i*-th transmitter
- n_0 is the ambient noise level (equals 10^{-2}) [11]
- p_j^k is the power for all other users on channel *k*, assuming that $j \in \{1,2,\ldots,L\}$ and $j \neq i$
- h_{ji} is the link gain between *i*-th transmitter and *j*-th receiver

A flat faded channel without shadowing effects is considered (this assumption is only required for proving that the algorithm will converge in a limited number of steps [10], [11]). Since the channel is static, the only identified attenuation is the path loss *h* (channel attenuation or channel gain). Given that indoor urban environments are considered, the channel gain is $h_{ji} = d_{ji}^{-3}$, where *d* is the distance between the *j*-th transmitter and the *i*-th receiver.

We adopt from the literature [11] the notion of interference price, which expresses the marginal loss of utility for receiver *i* if all the other users marginally increase their transmission power. The equation below computes the interference price for user *i*:

$$\pi_i^k = \frac{\partial u_i(\gamma_i(p_i^k))}{\partial(\sum_{i \neq i} p_j^k \cdot h_{ji})} \tag{3}$$

where,

- $u_i(\gamma_i(p_i^k)) = \theta_i \log(\gamma_i(p_i^k))$ is a logarithmic utility function
- θ_i is a user dependent parameter

As mentioned before, cognitive users select their transmission power value by taking into consideration their own utility and the degradation in utility of the other users. They compute the appropriate power value to transmit by maximizing the formula given below:

$$u_i(\gamma_i(p_i^k)) - \alpha \cdot p_i^k \sum_{j \neq i} \pi_j^k \cdot h_{ji} \tag{4}$$

The first part of the equation is closely related to the Shannon capacity of the channel, while the second part expresses the utility loss to other users if user i increases its power level. It should be noted that factor α is included as a weight in order to prevent underestimation of interference that user i will cause to the others. Underestimation is caused due to uncertainties in message exchange (i.e., message loss), large delays in the message exchange between users and users' mobility. The value of α ranges from 1 to 2. As a consequence, factor α compensates for the underestimation of interference, as the second part of the equation is increased.

The algorithm consists of three steps. The first is the initialization, where each user sets its power to a valid value (usually a minimum one) and calculates its interference price. The second step is the power update, where each user computes the appropriate power value in order to maximize the equation (4). The third step is the interference price update, where each user computes its interference price based on the updated power value from the second step. Finally, it announces its interference price to the other users. The second and the third step take place asynchronously for all users until a final steady state is reached. As a steady state, we define a state where no further enhancement in utility of any node pair can be achieved without negatively affecting the total network utility.

III. Fairness Policy Based on Genetic Algorithms

It is clear from the previous section and particularly from equation (4) that the original algorithm does not impose any lower bounds for the minimum power value a user chooses to transmit. Each user only attempts to balance the trade-off between utility optimization and interference mitigation. Furthermore, recalculating appropriate power values for L users is a multi-dimensional problem. In order to find the optimal power values, a search space of L-dimensions needs to be investigated. A simplified scheme would be to assign higher power values to underprivileged cognitive users, (for example the maximum allowed). However, this would lead to increased interference to other users and sharp degradation to the utility of the network. As a consequence, a trade-off between enhanced power value assignment and increased interference exists and the system can be tuned towards the desirable behavior using appropriate policies. Specifically, fairness policies are introduced to the cooperative power control algorithm, targeting to benefit cognitive users that are considered as underprivileged, without disregarding the needs of the other cognitive users. For example, choosing the maximum permitted power value for the underprivileged users is not an attractive option, as the rest of the users will face a significant increment of interference that will lead to QoS degradation.

In general, policies can be used to formalize the concept of decision making, especially when closed loop optimization is concerned. Typically, policies are comprised by constraint rules, which represent the set of limitations (i.e., memory size or battery level) and action rules, which specify procedures to be executed when certain conditions are met (e.g., [12], [13]). Such rules can be incorporated in machine learning schemes (such as Genetic Algorithms, Neural Networks, etc.) in order to enhance the flexibility and performance of the system. In this work, Genetic Algorithms (GAs) are utilized as a function optimization technique since they are known to perform well in problems with multidimensional and large search space.

Genetic Algorithms (GAs), first introduced by John Holland in [14], belong to the overall category of Evolutionary Computing techniques (EC). Typically, a candidate solution is structured as a string and is referred to as chromosome. A chromosome consists of a series of genes, in accordance to the dimensions of the search problem. It is usual to represent chromosomes as binary strings, but other encodings are also permissible. GAs use the principles of evolution and natural selection to optimize an initial set of chromosomes in order to reach a final optimal solution.

The execution of GAs starts from a set of chromosomes, constituting the initial population. A series of crossovers and mutations on the initial population produces offsprings that are incorporated to the population. Afterwards, based on a fitness function each chromosome of the population is being evaluated. Finally, a subset of the population will proceed to next generation based on a selection scheme. The procedures of mutation, crossover and selection are repeated iteratively until a termination criterion is satisfied. Terminal condition of a GA could be a fixed number of generations, an optimal threshold value for the fitness function, or a minimum deviation between the best chromosomes of two consecutive generations. Figure 1 depicts the steps of a GA.

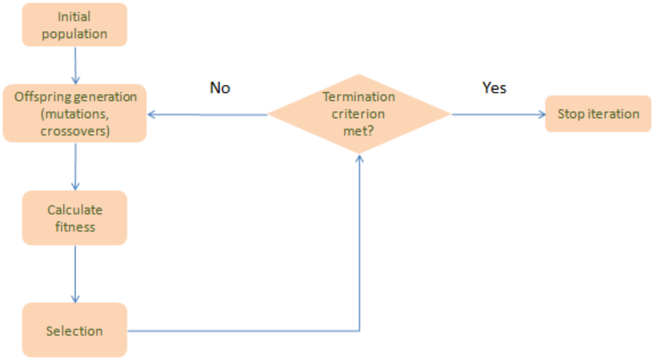

Initial population
Offspring generation (mutations, crossovers)
No
Termination criterion met?
Yes
Stop iteration
Calculate fitness
Selection

Figure 1. Flowchart of genetic algorithm

The main advantage of a GA is its capability to perform global search and, thus, converge efficiently to a near optimal solution [15]. This is due to the deviant nature of the candidate solutions that start from different points in the search space, in contrast to other heuristic methods that follow single candidate solution approach. Mutations and crossovers ensure production of different chromosomes (i.e., different candidate solutions to the search problem) during the generation process. Also, the ability of manipulating different chromosomes simultaneously makes GAs quick and robust. The main disadvantage of GAs is that for a high dimensional search space, it is complex to model the problem; however, this is not a major concern for the considered case, because the number of unprivileged users is a small percentage of the total number of users and therefore, exploring the search space is computationally feasible in an acceptable timeframe.

In our approach, a gene is a power value of a secondary cognitive user. Thus, a chromosome includes the power values of all the secondary cognitive users. The key point in GA execution is the evolutionary modification of the power values of the underprivileged users to more fair ones. Thus, appropriate assumptions and modifications were conducted for the phases of mutation, crossover and selection. In case of mutations, only genes, which correspond to underprivileged power values are mutated (i.e., increased). This modification is inline with the requirement for keeping cognitive users, which are not considered underprivileged, unaffected. Furthermore, the crossover procedure is designed to be simple. Thus, a single crossover point is selected randomly (the selection scheme followed in our approach is roulette wheel mechanism). Based on this scheme, the chromosome with the best fitness value passes to the next generation and following that, fitness values of the remaining chromosomes correspond to bounds between [0,1]. A random number on the same bounds determines which chromosome will follow the best chromosome to the next generation. The fitness function captures the trade-off between the increment of underprivileged users' Tx power and the increment of interference that will cause to the rest users and is computed for every chromosome. The proposed fitness function is given by the following equation:

$$\frac{\dfrac{\overline{Power_u}}{\overline{Interference}} - \dfrac{\overline{Power_u^0}}{\overline{Interference^0}}}{P_{\max} - P_{\min}} \qquad (5)$$

where,

- $\overline{Power_u}$ is the mean power of underprivileged users
- $\overline{Power_u^0}$ is the initial mean power of underprivileged users
- $\overline{Interference}$ is the current mean interference price
- $\overline{Interference^0}$ is the initial mean interference price
- $P_{\max}$ and $P_{\min}$ are the boundaries of users' Tx power

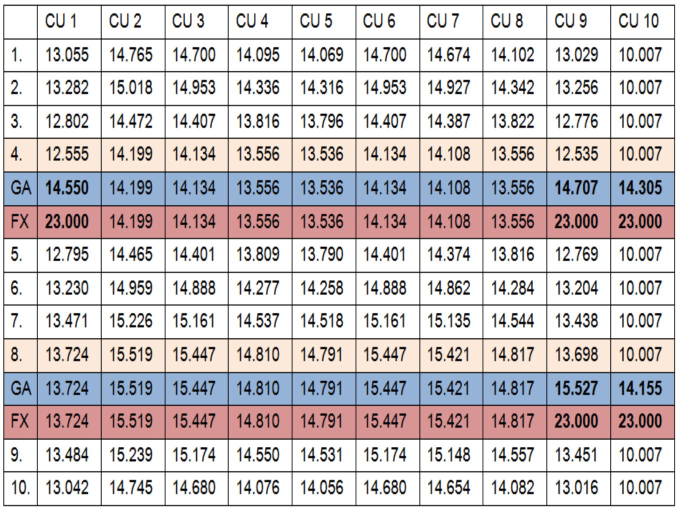
Finally, as mentioned before, GA is iterative and stops when a terminal criterion is met. In our approach termination criterion is considered to be the state where no significant enhancement is achieved between two consecutive generations.

IV. PERFORMANCE EVALUTATION

The performance of the proposed algorithm is evaluated through extensive MATLAB simulations. Towards this direction, our GA approach is compared to a scheme of fixed power value assignment (maximum valid power level). The main objective is to give "fairer" power values to the underprivileged cognitive users. This concludes to a more "fair" treatment, but incurs loss in system performance, as principles of the power control algorithm are violated. The major difference between the two proposed techniques is that in case of GA, underprivileged users get better power values, but not the maximum ones due to the negative impact of interference in the fitness function.

The proposed implementation examines a commonly used environment of 10 LTE mobile cognitive users (CUs) (e.g., [16], [17]) cooperating in order to transmit with an acceptable power value. The power range is between 10 and 23 dBm and the distances between the cognitive users is a random number in the [50, 550] meters range [18]. The users set their transmission power levels to maximize equation (4) until the algorithm converges to a steady state for a given topology. The whole procedure lasts for 10 topologies (i.e., steps) that reflect the mobility of the users in consecutive time frames. For every successive step, the fairness policy mechanism is called, in order to examine if underprivileged users exist. If so the GA algorithm is activated, so as to enforce fairness. In order to identify if a cognitive user is underprivileged, previous Tx powers are examined for a certain time window in the past. The size of the window is considered to be 3. Consequently, our fairness policy examines the current step and the previous 3 to detect underprivileged users.

Figure 2 illustrates 10 steps where each cognitive user (CU) chooses to transmit with a certain power value (in dBm) based on the original algorithm in [10]. On the first step, the average power value is 13.719 dBm, which is also the general upper bound for the "unfair" power values. For the simplified fixed power value schema (FX), maximum power values will be assigned to the underprivileged users. In such cases, an arbitrary increase in Tx power value of a CU usually results to a non cooperative state, where all CUs are negatively affected. Alternatively, fairness policy is called in every step and is enforced only in the fourth and eighth steps for the CUs 1, 9, 10 and CUs 9, 10 respectively. The initial power values for the fourth step will be re-calculated in case of GA. The same situation occurs in the eighth step as well. Both in case of GA or in case of fixed power values, the privileged users are not affected directly (i.e., by decreasing their transmission power).



	CU 1	CU 2	CU 3	CU 4	CU 5	CU 6	CU 7	CU 8	CU 9	CU 10
1.	13.055	14.765	14.700	14.095	14.069	14.700	14.674	14.102	13.029	10.007
2.	13.282	15.018	14.953	14.336	14.316	14.953	14.927	14.342	13.256	10.007
3.	12.802	14.472	14.407	13.816	13.796	14.407	14.387	13.822	12.776	10.007
4.	12.555	14.199	14.134	13.556	13.536	14.134	14.108	13.556	12.535	10.007
GA	**14.550**	14.199	14.134	13.556	13.536	14.134	14.108	13.556	**14.707**	**14.305**
FX	**23.000**	14.199	14.134	13.556	13.536	14.134	14.108	13.556	**23.000**	**23.000**
5.	12.795	14.465	14.401	13.809	13.790	14.401	14.374	13.816	12.769	10.007
6.	13.230	14.959	14.888	14.277	14.258	14.888	14.862	14.284	13.204	10.007
7.	13.471	15.226	15.161	14.537	14.518	15.161	15.135	14.544	13.438	10.007
8.	13.724	15.519	15.447	14.810	14.791	15.447	15.421	14.817	13.698	10.007
GA	13.724	15.519	15.447	14.810	14.791	15.447	15.421	14.817	**15.527**	**14.155**
FX	13.724	15.519	15.447	14.810	14.791	15.447	15.421	14.817	**23.000**	**23.000**
9.	13.484	15.239	15.174	14.550	14.531	15.174	15.148	14.557	13.451	10.007
10.	13.042	14.745	14.680	14.076	14.056	14.680	14.654	14.082	13.016	10.007

Figure 2. Converged power values for 10 cognitive users

Figure 3 illustrates the average Tx power values of the CUs for each of the 10 topologies. As can be seen again, the fairness policy is enforced in the fourth and eighth topology.

The purpose of a fairness scheme is to support the underprivileged users and minimize the negative impact to the network. Indeed, in the proposed scheme the underprivileged users get enhanced power values; however, this is done in a planned way, so that the impact in the overall performance of the network is limited (marginal reduction of the average network SINR by approximately 0.3 dB). This is a reasonable trade-off for enhancing the overall fairness, especially considering that the SINR of the underprivileged users and the related QoS is increased.

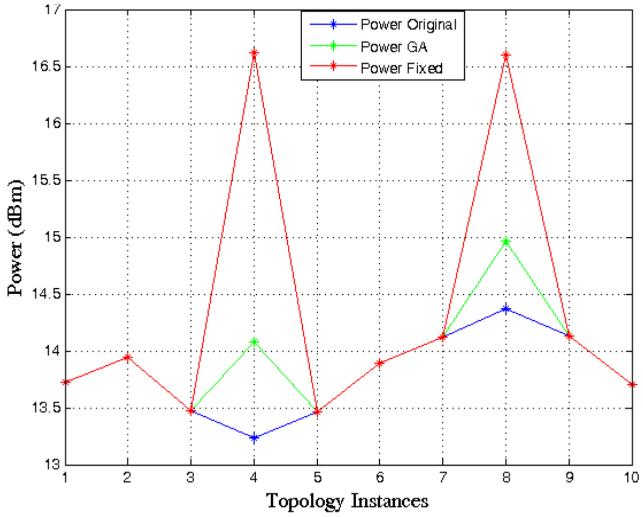


Figure 3. Average power values for 10 topologies

Figure 4 illustrates SINR values for the 10th CU in topologies where fairness was enforced. Specifically, concentrating on the underprivileged CUs 9 and 10, enhanced Tx power levels are calculated. This increase leads also to enhanced SINR at the receiver.

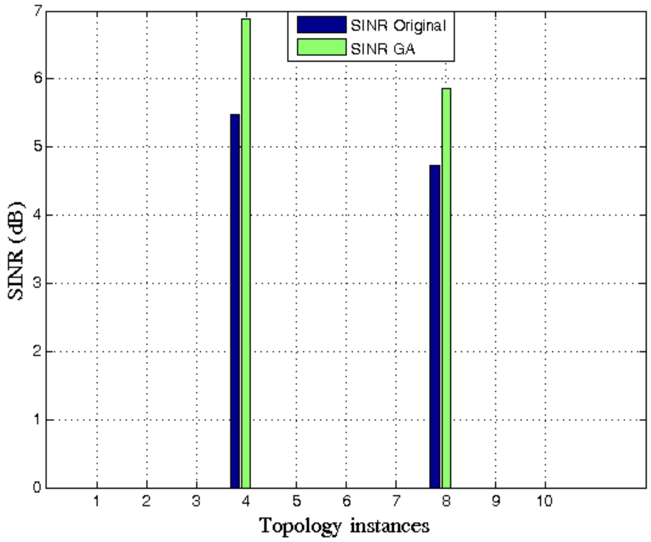


Figure 4. SINR improvement for underprivileged users

Since equation (4) strikes the optimal balance from a system utilization perspective between the selfish need for transmission at the highest level and the social conformance of reducing the interference to other neighboring users, altering the Tx Power to the constantly underprivileged users will also have a negative impact to the rest of the users in the environment. Figure 5 shows a comparative analysis of the average SINR gains of the underprivileged users against the average SINR degradation that the other users will experience.

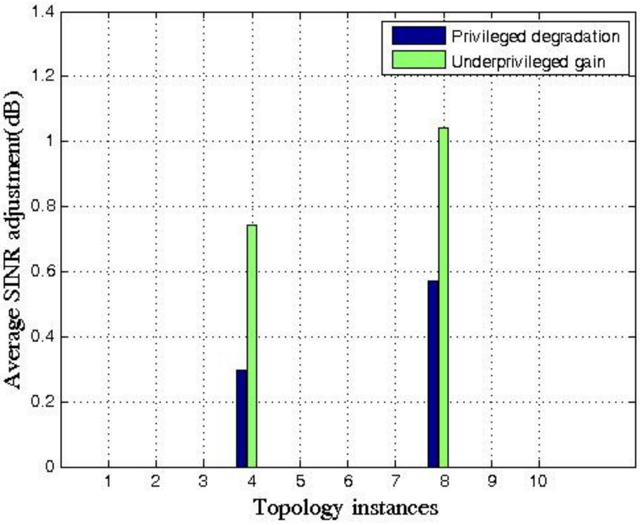


Figure 5. Fairness SINR gains against SINR degradation

As mentioned previously, many fairness schemes are challenging in their application to real world systems due to the full knowledge requirement and the stringent synchronization constraints among the wireless nodes that this requirement imposes. In our case the genetic algorithm can operate efficiently with a significantly relaxed knowledge model and synchronization scheme. For our evaluation of this highly desirable property we have conducted 1000 experiments assuming the same environment as before; the fundamental difference is that the system suffers a 10-20% message loss, thus leading to undesired effects for the nodes, as they will not have a complete knowledge of the environment. Figure 6 shows that in cases of an incomplete knowledge model the GA is triggered again exactly 2 times (as in the case with full knowledge) with probability equal to 42%. The results also show that cases of not triggering the GA when needed (false negatives) are not possible, but there are some false positive cases where the algorithm is triggered more times than actually needed.

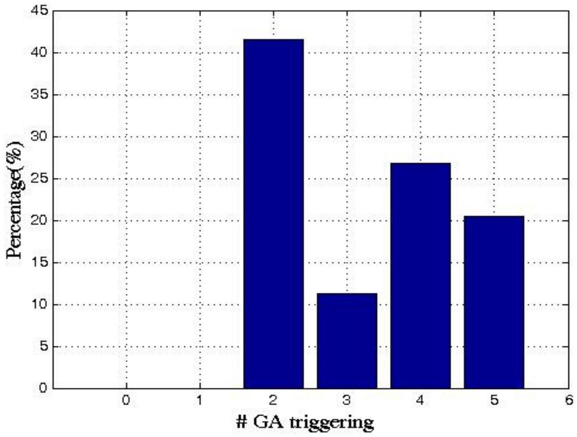


Figure 6. GA behavior in cases of an incomplete knowledge model

However, these false positives do not influence the efficiency of the algorithm as even in that cases the SINR of the users is only marginally affected. Figure 7 shows a characteristic example where the GA was triggered four times.

As it is shown, only on topologies 4 and 8 the SINR of the underprivileged users was adjusted while on other cases the algorithm did not change the transmission power of the users.

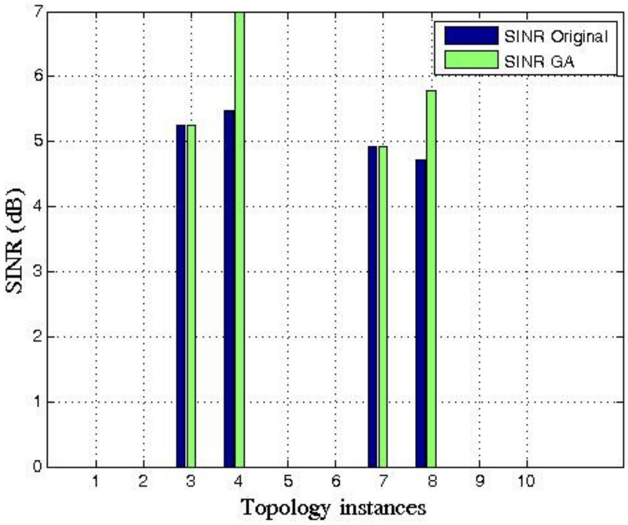


Figure 7. SINR improvement for underprivileged users in a false positive case

V. Conclusion

A novel technique for enforcing a fairness policy in cooperative power control for cognitive radio networks was presented. The proposed scheme extends the cooperative power control algorithm of [10] with a fairness check module. The power level values, which are assigned to the underprivileged cognitive users, are calculated through the evolutionary execution of a Genetic Algorithm. GAs were selected as a heuristic able to search multidimensional search spaces. The outcome of the GA algorithm was compared both with the original cooperative power control scheme and with a simplified fairness scheme. The results indicate that increased power values were assigned to the underprivileged users, considering also the negative impact in power gain of the network. Specifically, simulations show significantly improved SINR for the underprivileged users compared to the original algorithm with minimal impact in the SINR of the privileged users. Furthermore, in comparison to the case of a simplified fairness policy, which assigns underprivileged cognitive users with the maximum valid power level, the proposed scheme offers considerable power gains to the network. Finally, we have shown that the proposed algorithm can operate efficiently even in cases of partial knowledge models and imperfect message exchange/synchronization between the nodes, a property that is highly desirable for application in real world systems.

Acknowledgment

The research leading to these results has been partially funded from the EC FP7 SACRA project (no. 249060).

References

[1] P. Kolodzy, “Communications Policy and Spectrum Management,” in Cognitive Radio technology, B. A. Fette, Ed. London: Elsevier, 2006, pp. 29–71.

[2] W. Lehr and N. Jesuale, “Spectrum Pooling for Next Generation Public Safety Radio Systems,” Proc. IEEE International Symp. Dynamic Spectrum Access Networks (DySPAN 08), IEEE Press, Oct. 2008, pp. 1-23, doi:10.1109/DYSPAN.2008.72.

[3] A. Ghasemi and E. S. Sousa, “Optimization of Spectrum Sensing for Opportunistic Spectrum Access in Cognitive Radio Networks,” Proc. IEEE Conf. Consumer Communications and Networking (CCNC 07), IEEE Press, Jan. 2007, pp.1022-1026, doi:10.1109/CCNC.2007.206.

[4] S. Qijun, H. Xin, Z. Weixia, and Z. Xiangyu, “The Overview of Dynamic Frequency Spectrum Access Based on Some Advanced Techniques,” Proc. IEEE International Symp. Antennas, Propagation & EM Theory (ISAPE 08), IEEE press, Nov. 2008, pp. 964-967, doi:10.1109/ISAPE.2008.4735381.

[5] Y. C. Liang, K. C. Chen, G. Y. Li, and P. Mahonen, “Cognitive Radio Networking and Communications: An Overview,” Proc. IEEE Trans. Vehicular Technology, vol. 60, Sep. 2011, pp. 3386-3407, doi:10.1109/TVT.2011.2158673.

[6] L. Akter and B. Natarajan, “Modeling fairness in resource allocation for secondary users in a competitive cognitive radio network,” Proc. Wireless Telecommunications Symp. (WTS 10), IEEE press, Apr. 2010, pp. 1-6, doi:10.1109/WTS.2010.5479652.

[7] V. Kawadia and P. R. Kumar, “Principles and protocols for power control in wireless ad hoc networks,” IEEE Selected Areas in Communications, vol. 23, Jan. 2005, pp.76-88, doi:10.1109/JSAC.2004.837354(410) 23.

[8] Bingxiao Chen, Jun Sun, Shixiang Shao, Longxiang Yang, and Hongbo Zhu, “An improved power control method based on cooperative game for cognitive radio networks,” Proc. IEEE International Conf. on Communication Technology (ICCT 2010), IEEE Press, Nov. 2010, pp: 709-712 , doi:10.1109/ICCT.2010.5688588.

[9] A. Malla, M. El-Kadi, and P. Todorova, “A fair resource allocation protocol for multimedia wireless networks,” Proc. International Conf. on Parallel Processing, IEEE Press, Sept. 2001, pp. 437- 443, doi:10.1109/ICPP.2001.952090.

[10] A. Merentitis and D. Triantafyllopoulou, “Transmission Power Regulation in Cooperative Cognitive Radio Systems Under Uncertainties,” Proc. IEEE International Symp. on Wireless Pervasive Computing (ISWPC 10), IEEE Press, May 2010, pp. 134-139, doi:10.1109/ISWPC.2010.5483742

[11] J. Huang, R. A. Berry, and M. L. Honig, “Spectrum sharing with distributed interference compensation,” Proc. IEEE International Symp. New Frontiers in Dynamic Spectrum Access Networks (DySPAN 05), IEEE press, Nov. 2005, pp. 88-93, doi:10.1109/DYSPAN.2005.1542621.

[12] P. Magdalinos, S. Polymeneas, P. Gliatis, X. Fafoutis, A. Merentitis, and C. Polychronopoulos, “A Proof of Concept Architecture for Self-Configuring Autonomic Systems”, Proc. ICT Mobile and Wireless Communications Summit, IIMC International Information Management Corporation, Jun. 2008.

[13] E. Patouni et. al., “Protocol Reconfiguration Schemes for Policybased Equipment Management,” Proc. IEEE Vehicular Technology Conf. (VTC 06), IEEE Press, Sept. 2006, pp. 1-6, doi: doi: 10.1109/VTCF.2006.592.

[14] J. Holland, Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University Michigan Press, 1975.

[15] M. Gen, R. Cheng, and L. Lin, Network Models and Optimization: Multiobjective Genetic Algorithm Approach. Springer – Verlag, 2008.

[16] S. Guo, C. Dang, and X. Liao, “Distributed resource allocation with fairness for cognitive radios in wireless mobile ad hoc networks,” Journal on Wireless Networks, Springer Netherlands, vol. 17, Aug. 2011, pp. 1493-1512, doi:10.1007/s11276-011-0360-9.

[17] G. Yu, Z. Zhang, Y. Chen, P. Cheng, and P. Qiu, “Subcarrier and bit allocation for OFDMA systems with proportional fairness”, Proc. IEEE Wireless Communications and Networking Conf. (WCNC 2006), IEEE Press, Apr. 2006, pp.1717-1722, doi:10.1109/WCNC.2006.1696547.

[18] A. Agrawal, “Heterogeneous Networks. A new paradigm for increasing cellular capacity”, Jan. 2009 [retrieved May 2012].

Single-Handed Typing with Minimal Eye Commitment: A Text-Entry Study

Adrian Tarniceriu, Pierre Dillenbourg, Bixio Rimoldi
School of Computer and Communication Sciences
Ecole Polytechnique Federale de Lausanne
Lausanne, Switzerland
adrian.tarniceriu@epfl.ch, pierre.dillenbourg@epfl.ch, bixio.rimoldi@epfl.ch

***Abstract*—For most people, typing text on a mobile device requires visual commitment to the input mechanism. As a consequence, there are many situations in our daily life when we have to refrain from using these devices as our vision is already committed. An example is trying to text while walking in a crowded place. Chording devices allow us to type without looking at the input device but using them requires some training. We present the results of a study that evaluates the performance of a key-to-character mapping for a 5-key chording device. The mapping is designed to minimize the learning phase. After 45 minutes of training it was completely learned, and after approximately 250 minutes the average entry speed was 15.2 words per minute. A prototype that implements this mapping was mounted on a bike and tested by the authors who could comfortably ride and type while being focused on the road.**

Keywords-chording keyboard; text entry; key mapping; mobile device

I. Introduction

Mobile computing devices such as smart phones or personal digital assistants play an ever increasing role in our daily lives. More and more people appreciate their services and want to have access to them at all times, but their input methods are not suitable during activities for which vision is partially or entirely required. For example, in a crowded place most people need to stop walking in order to text via today's keypads or touchscreen keyboards.

Chording is a text-input method that utilizes a small number of keys. A character is formed by pressing a key combination, similarly to playing a chord on a musical instrument. With five keys, one for each finger, we have 31 different combinations in which at least one key is pressed. This is enough for the 26 letters of the English alphabet and five punctuation signs. If we can keep our fingers on the keys, then we can type with one hand and without looking at the keys. Our vision (or auditive feedback) is still needed occasionally to verify the output, but this requires considerably less commitment than continuously looking at the input device. Visual commitment can be further reduced by displaying the output in the natural field of vision, for instance on a windshield or on goggles.

The likely reason chording devices are not popular is that users require some training before being able to type. Compared to a QWERTY keyboard, where users can hunt and peck from the beginning, the mapping between keys and characters has to be learned for chording devices. The overhead needed to do so depends on the keyboard type and mapping and can vary by several hours. The main objective of this paper is to present the results of a study on a mapping designed to facilitate the learning task.

The paper is organized as follows. In Section II, we present a brief overview of related work. In Section III, we describe a key-to-character mapping for a 5-key keyboard designed to reduce the learning time. We denote this mapping in the following as 5keys. In Sections IV and V, we present two experiments that evaluate the learnability, text-entry rates, typing accuracy and common error patterns. In Section VI, we conclude the paper and discuss future directions.

II. Related Work

The first chording applications were used in stenotype machines (1830's), telegraph communications, and the Braille system that allows blind people to read and write.

Subsequent studies were performed by IBM [1], but the results were considered inconclusive and research was stopped in 1978. Douglas Engelbart, the inventor of the computer mouse also proposed a 5-key keyset, but this was not incorporated in any system [2], probably due to the popularity of the standard QWERTY keyboard.

As mobile devices and wearable technologies evolve, classic keyboards are no longer able to fulfill users' needs for mobility and ubiquitous access to computational resources. Therefore, chording keyboards have re-emerged as a popular research topic, leading to the appearance of several devices, like DataEgg [3], GKOS [4], Twiddler [5], EkaPad [6] and the chording glove [7]. Depending on the envisaged application, these keyboards can have different number of keys, mappings or shapes, each of these being a research topic. In the following, we will focus on a 5-key keyboard and a mapping designed to minimize the learning process.

III. Character Mapping

An important aspect of designing a chording keyboard is the mapping between key combinations and desired characters. One possibility is to assign easier combinations for more frequent letters, as in the Morse code, leading

ISBN: 978-1-61208-236-3

to higher typing speeds. Even if these mappings are easy to determine, the user must learn by heart the key-to-letter correspondence as there is no intuitive link between them. Another possibility is to use a semantically richer mapping, which would be easier to learn.

The key-to-character mapping studied in this work was designed with the primary goal of making it easy to remember. It was designed for a 5-key keyboard, where each character is represented by a different key combination. In this paper we will focus only on lowercase letters plus the period, space and backspace, as they are the most used; also because most typing studies consider only this set of characters.

To create enough possibilities for assigning an intuitive key combination to each character, we conceived five mnemonic categories.

1) Single-key category: remembering the map for the characters in this category should be totally trivial. Characters are produced by pressing a single finger and the letter is the initial of the finger. So by pressing the key under **t**humb, **i**ndex, **r**ing and **p**inky, we obtain "**t**", "**i**", "**r**", and "**p**", respectively. There is an exception to the rule: since "**m**" fits well in another category (see below), we have reserved the middle finger for the period.
2) Fingers-down category: the most natural way to produce the shape of an "**m**" with the hand is to stretch down the index, middle, and ring fingers (see Figure 1b). In a similar fashion we can produce the other letters in this category, namely "**n**", "**u**", "**y**", and "**c**".
3) Fingers-up category: similarly, a natural way to produce the shape of a "**w**" is to stretch up the index, middle, and ring fingers (Figure 1c). The associated character is obtained by pressing the key(s) under the remaining fingers. "**v**", "**l**", "**e**", and "**j**", follow the same idea. We have included **space** and **backspace** in this category as backspace can be associated with the thumb pointing to the left and space with the pinky pointing to the right.
4) Finger footprint category: usually, a character is produced faster with a 5-key device if the users first think of the fingers pressing the keys and not of the fingers that remain "up". Considering this, for "**h**" we can identify three landmark spots on the shape of the letter and associate them to fingers according to the following rule: the thumb is for spots that are left and low, the index for left high, the ring for right high, and the pinky for right low. The resulting mapping is given in Figure 1d. With a little bit of imagination in this category we can also fit "**a**", "**f**", "**k**", "**o**", "**s**", "**x**", and "**z**". For "**o**", we imagine five dots spread around a circle, and we obtain it by pressing all buttons.
5) Associative category: the letters "**b**" and "**d**" may be seen as an "**o**" with a vertical bar on the left and right, respectively. We use the index and the ring fingers to represent these bars. "**g**" was inspired from "**y**" (they look alike in handwriting) and in turn "**g**" inspires "**q**" (the tail ends left and right respectively, so for "**g**" we use the thumb and for "**q**" the pinky).

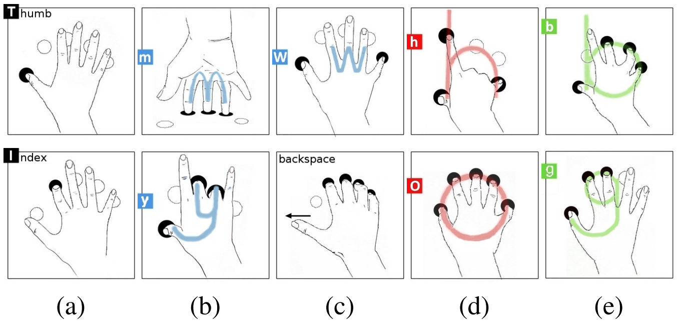


Figure 1. Examples for letter mappings. (a) Single-key mnemonics for "**t**"and "**i**". (b) Fingers-down mnemonics for "**m**" and "**y**". (c) Fingers-up mnemonics for "**w**" and "**backspace**". (d) Finger footprint mnemonics for "**h**" and "**o**". (e) Associative mnemonics for "**b**" and "**g**"

Two examples from each category are given in Figure 1.

Some of the above mnemonics are easier to remember than others. With five keys, however, there are only 31 usable combinations and we use them all to map the 26 characters plus the space, backspace, period, enter and comma. Hence any change aimed at improving one mnemonic implies at least one other change.

The effectiveness of the proposed mapping is assessed through two experiments described in the next sections. The first compares this mapping to two others from a learnability point of view, or, in other words, how easy users can remember the key-to-character correspondences. The second experiment estimates the usability of the mapping (typing speed, accuracy [8] and most common mistakes).

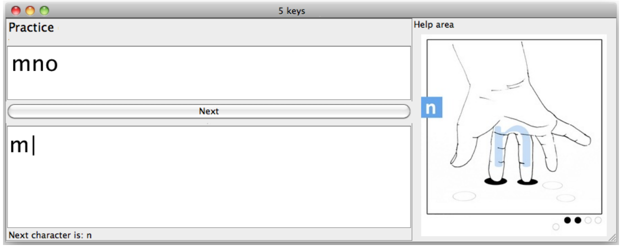
IV. Learnability Study

A. Experimental Setup

This first experiment compares, from a learnability point of view, the proposed mapping (5keys) to two others. The references are the Microwriter mapping [9], also based on intuitive mnemonics, and the Baudot code [10] that is based on letter frequency and assigns easier key combinations to most common characters. All three mappings are designed for 5-key keyboards.

A Java application was designed to simulate the chording keyboard on a regular QWERTY desktop keyboard. It only allows the use of five keys, each representing a key of the chording keyboard. Each of these keys correspond to a finger of the right hand. A typical choice was the spacebar for the thumb, and the keys for f, t, y and u for the index, middle finger, ring and pinky, respectively.

The experiment consisted of three sessions of three rounds each. For each round, the subjects had 5 minutes to look



mno

m|

Next character is: n

Figure 2. Typing application screenshot

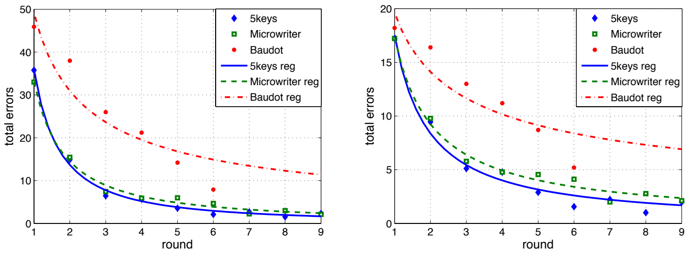
at a printed version of the mappings and try to remember them. Afterwards, they used the Java application to warm up by typing each letter of the alphabet. A help image showing the key combination for the letter to be typed was shown to the participants. A screenshot of the application is visible in Figure 2. The top-left window contains the target characters to be typed. The bottom-left window represents the typing area and the help image is displayed on the right. In the next step, the help image was not available any more and the participants had to type the alphabet three times. The order of the letters was random, but the same for all participants. The subjects had five seconds and only one attempt to type each target character. The correct key combination was displayed when the user typed a character (right or wrong) or when the five seconds expired.

30 participants, 10 for each of the three mappings, were recruited from the students of our university (undergraduate, master and PhD programs). They were between 19 and 30 years old, and four were female. For each session of the experiment (approximately 30 minutes) they received a fixed monetary compensation. None of the subjects had used a chording keyboard before. As the participants who know how to play a musical instrument could have had an advantage, they were equally distributed among the three experiment groups. We also tried to equally distribute them based on gender and study level. Two participants abandoned the experiment after the first session, one testing the 5keys and one the Microwriter mapping.

B. *Experiment Results*

To determine which of the mappings is easier to learn we compared the number of errors (wrongly typed or not typed characters) for each round. Exponential regressions [11] were derived to fit these error values. The average values for each mapping and for each round and the exponential regressions are presented in Figure 3a.

After two sessions (six rounds of approximately five minutes of typing each), the total number of errors was considerably lower for the mnemonic based mappings (5keys and Microwriter) compared to the mapping based on letter frequency. Therefore, we conclude that mnemonic based mappings are learned faster. The goal of the study was to evaluate which mapping is easier to learn and the Baudot



(a) (b)

Figure 3. (a) Average number of errors (for each mapping and for each round) and regression curves. (b) Average number of character errors (for each mapping and for each round) and regression curves

mapping is clearly more difficult. Hence, in the third session we only analyzed the 5keys and Microwriter mappings. By checking only the average values, no significant difference between these two was noticed. An advantage of 5keys can be observed from the analysis of the regression curves as the curve for 5keys is slightly below the curve for Microwriter.

Besides the total number of errors, we also compared the number of wrong characters per round. For example, if “**a**” was typed incorrectly two times, this counts only as one character error. The results are shown in Figure 3b. In this case the difference between 5keys and Microwriter mappings is more visible and, as expected, both lead to considerably less errors than the Baudot mapping.

At the end of the third typing session (after nine rounds or approximately 45 minutes of actual typing), the participants were asked how confident they feel about their knowledge of the mappings and if they could use the presented method as a text input mechanism. All of them answered affirmatively and most also mentioned that they completely learned the mappings. This is confirmed by a low error rate (3.16% after 6 rounds and 2.14% after 9 rounds for the proposed mapping).

From this experiment, we draw the conclusion that a mnemonic-based mapping facilitates the process of learning the code. We also conclude that the proposed 5keys mapping outperforms the Microwriter mapping, also mnemonic-based, in terms of average error rate.

The mnemonic set was designed based on the finger positions of the right hand. Two of the participants (one for the 5keys and one for the Microwriter) were left-handed. Yet they also typed with their right hand and, interestingly, their error rates were actually lower than the average.

V. Usability Study

The first experiment, aimed at evaluating the learning process, was followed by an independent experiment aimed at determining achievable typing rates, accuracy, and common error patterns for the 5keys mapping.

A. *Experimental Setup*

The second experiment was based on a similar Java application. This time, the subjects were asked to type full sentences, not just isolated letters. The experiment consisted of 11 sessions. For each of them, the subjects started by typing each letter from each category and continued with real sentences chosen from a set considered representative for the English language [12]. The first four sessions, of 25 minutes each, were for the participants to learn the mapping and to familiarize themselves with the keyboard. During these sessions the help image was always displayed. During the following sessions (5 to 10), each lasting 20 minutes, the help image was no longer available. Session 11 was similar, but only lasted for 10 minutes. For the first 10 sessions, the participants received a fixed monetary compensation. As an incentive, for the last one the reward was proportional to the subject's performance, measured by the number of correctly typed words. We recruited a new set of six students for this study. The number of participants is lower than for the first experiment due to the significant time commitment required.

In order to monitor the evolution of the experiment and to gather statistics about the subjects' activities and performances, for each session and for each participant the application generated several log files. These files recorded the typed text, the number of occurrences for each character, the corresponding key combination, the total number of errors, the number of corrected errors and the total time spent writing each character. When a typing error occurred, we checked what character was typed in lieu of the correct one.

B. *Text-Entry Speed*

We used the wpm (words-per-minute) measure to describe the text entry speed. This is defined as

$$wpm = \frac{60L}{t}\frac{1}{5} \tag{1}$$

where L is the total number of typed characters and t is the typing time in seconds. The scaling factor of 1/5 is based on the fact that the average English word length is approximately 5 characters. As the average word length for the typed text differed from one session to another, the use of the above formula provides a more reliable estimate than actually counting the words.

The average entry rate for the first session was 4.2 wpm and reached 15.2 by the end of the experiment (after approximately 250 minutes of typing). Even though these values were obtained using five keys from a classic keyboard, they do give an estimate of what can be achieved using a real implementation of the device (the shape of the hand when the fingers are placed on the buttons is almost the same for a flat surface, bike handlebar, or around a mobile phone case).

Figure 4 presents the entry rates (for each subject, average and exponential regression) for each session. We observe that the typing rates significantly improved from the 10th to the 11th session. This could be explained by the fact that for the last session the subjects were stimulated by a reward proportional to the number of correctly typed words.

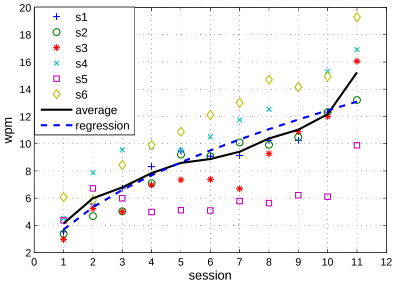


Figure 4. Typing rates per session for each user, average and regression

As a reference, the typing rates achieved after 250 minutes of practice are 12.4 wpm for multi-tap mobile phones [13] and 20.6 wpm for Twiddler [5]. Rates of 20.36 wpm were reached by expert T9 users [14]. We should point out that the experimental conditions were not the same for all devices. Hence, the above typing rates are only of indicative nature. For both multi-tap and T9 techniques visual attention is essential for most users. For the 5keys device it makes essentially no difference if the user has visual contact with the keys or not. It should also be taken into consideration that Twiddler uses 12 keys, whereas our mapping only requires 5 keys, thus providing a clear space advantage and more design flexibility. If placed in a position which is naturally under the fingertips (for example on the handlebar of a bike), the users will have continuous access to the keys.

C. *Error Analysis*

Starting with session 5 (when the help image was no longer displayed) we evaluated the accuracy based on the corrected and uncorrected errors. The error percentages are defined as

$$corrected\% = \frac{\#backspaces}{\#characters}100\,, \tag{2}$$

$$uncorrected\% = \frac{\#incorrect\_characters}{\#characters}100\,. \tag{3}$$

The errors could have two main causes: the subject does not recall the correct key combination or, alternatively, a coordination mistake is produced during execution. We call these error types cognitive and sensorimotor errors, respectively. We expect the cognitive errors to decrease faster, as a function of training, because it is easier to learn the code than to improve motor skills. This is confirmed by the statements of the participants in both experiments: they said that they had learned the mapping by the end of the training, and errors were due to lack of attention or finger combinations that seemed difficult.

Table I
ERROR RATES PER SESSION

Session number	5	6	7	8	9	10
Corrected errors %	8.04	6.07	7.35	7.51	7.59	6.21
Uncorrected errors %	0.52	0.47	0.19	0.36	0.21	0.22

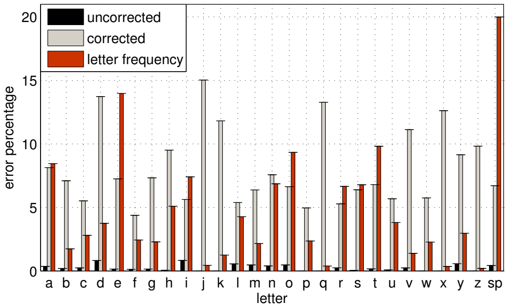


Figure 5. Error rates per character

Table I presents the error rates for sessions 5 to 10. For session 10 the uncorrected error rate is 0.2% as the subjects tended to correct most errors. The low uncorrected-error rate shows that the mapping was learned already by the end of session 4, as expected given the results of the first experiment.

Figure 5 presents the percentage of corrected and uncorrected errors per character for all users for the whole experiment. The figure also shows the scaled occurrence ratio for each character. Error rates are higher for less frequent letters (for example "**j**" and "**q**"), probably because the subjects had fewer opportunities to practice on them. Non-negligible error rates can also be observed for high-frequency characters, as users probably try to type faster as they gain more experience.

As in Section V-B, we compare the uncorrected error rates with those for multi-tap (5%) and Twiddler (3%) after 250 minutes of practice, and also with expert T9 users (0.52%). Even if the T9 entry method and Twiddler allow for higher typing rates, the error rates are also higher (by one order of magnitude for Twiddler). Again, these values are only indicative, due to different experimental conditions.

We also determined the dependency of the average error rate on the character category and on the number of keys that need to be pressed to compose a character. We observed that the error rates increase for characters that involve a larger number of keys, but the differences are not statistically significant (anova test p-values higher than 0.05). Letters from the single-key category have the lowest error rates and those from the associative category have the highest, but again, these results are not statistically significant.

D. Common Errors

To understand the error patterns that appear most frequently, we computed the confusion matrix [15] corresponding to the typed text. This is a square matrix with rows and columns labeled with all possible characters. 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*.

It is useful to represent a key combination by a 5-bit codeword in which the first digit represents the key under the thumb, the second digit the key under the index, etc. The value of a position is 1 if the corresponding key is pressed. So, for instance, 11011 is the codeword for "**x**", for which all fingers except the middle one press the keys. By analyzing the 5-bit code for the 10 most common substitutions, we notice that in 9 of the 10 cases the errors appear between characters that differ only by one bit (for example "**x**", code 11011 and "**h**", code 11001). In the other case, two single-key characters are substituted ("**t**" and "**i**").

If we check word by word and consider only substitution errors, i.e., errors that arise from substituting individual characters, 84% of the erroneous words contain one substitution, 13% contain two substitutions and 2% three substitutions. From a bit-error point of view, 51% of the erroneous words contain a one-bit error, 33% a two-bit error and 14% a three-bit error. One-bit errors occur when the user does not press one of the required keys (31% of the total errors) or presses an extra key (20% of the total errors). Most two-bit errors are substitutions, when a wrong key is pressed and a correct key not pressed. These values can be used to implement an error correcting mechanism that relies both on a dictionary and on the probability that a character be substituted for another.

E. Character Typing Duration

As the coordination effort is not the same for all key combinations, we expect that different characters require more time than others to be typed. Figure 6 shows the average time per key combination for sessions 5 and 10. The time needed to form a key combination, called composition time, is measured from the moment the first key of a combination is pressed until a key is released. It is when a key of the combination is released that the corresponding character is produced. From that moment on, the pressing of a key indicates the start of a new character. Instead of ordering the letters alphabetically, in Figure 6 we order them in increasing number of pressed keys (single-key letters are first and "**o**", for which all keys are pressed, is last). The letters containing the same number of keys are ordered in ascending composition time for session 10.

As expected, the composition time increases with the number of pressed keys, the dependence being statistically significant (anova test p-values lower than 0.05). We also notice that letters requiring key combinations perceived as more difficult (for example "**q**", code 01101 or "**d**", code 11101 for which the middle finger and the pinky are down while the ring finger is up) require more time than others.

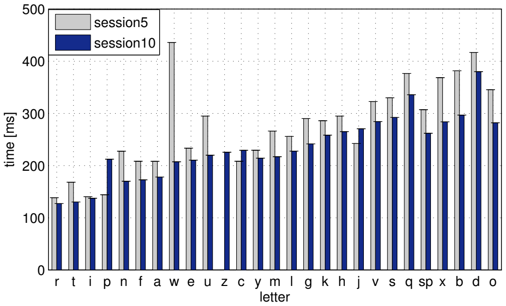


Figure 6. Average composition time per character for sessions 5 and 10

As subjects gained more experience, they were able to type faster and the average letter duration decreased from 273.9 milliseconds in session 5 to 234.5 in session 10, or by 14%. During the same period, text entry rates increased from 8.6 to 12.2 wpm, or by 41%. The difference is explained by the fact that the idle time between the end of one character and the beginning of the next also decreased.

VI. Conclusions and Future Work

In this paper, we have presented the results of a study evaluating the mapping for a chording input device. The overhead needed to learn the mapping was reduced by choosing easy-to-remember key combinations. A first experiment showed that the mapping was learned after less than 45 minutes of actual typing. Moreover, the total number of errors was considerably smaller than for a letter-frequency based mapping and slightly smaller than for another mnemonic based mapping.

A second experiment showed that after approximately 250 minutes of typing, the average text entry rate was 15.2 wpm and the uncorrected error rate 0.2%. We also analyzed which characters are perceived as more difficult to type and the most common errors. This data will be used to develop an error correction mechanism specifically designed for a chording keyboard using the proposed mapping. It will take into account a language model, as well as the probability that one character is typed for another.

During the experiments, the subjects sat at a desk. To go one step further, we designed, built and tested a prototype for a bike. We easily fit the five keys under the natural position of the fingers on the handlebar. Two of the authors tested the device and found that they could effortlessly ride and type while staying focused on the road. The position of the keys allowed them to control the bike with both hands while typing. Moreover, as the keys were directly under the fingers, they could also type accurately on a bumpy road. Though encouraging, these results are exploratory and a more accurate study should be performed.

There are numerous potential applications for a 5-key input device. For instance, with the buttons around a phone one can input a text message while walking (a stop to proofread before sending should suffice). By means of a wrapper application that captures the text, a user can control the operation of a smartphone: in the test on the bike, the authors could control the music, write a short note, interact with the map, etc. We can easily envision the potential benefit of a 5-key input device on the handlebar of a shopping cart. It could be used to browse or edit a shopping list on a PDA placed in the middle of the handlebar or to interact with the store's web site to check the availability and the location of an item. Another interesting application could be to have the keys on the side of a TV remote control. Although modern TV sets allow for Internet navigation, typing a URL and doing searches with a standard remote control can still be quite clumsy.

References

[1] F. C. Bequaert and N. Rochester , "Teaching typing on a chord keyboard," tech. rep., IBM Technical Report, 1977.

[2] D. C. Engelbart, "Design considerations for knowledge workshop terminals," in *Proceedings of the June 4-8, 1973, national computer conference and exposition*, AFIPS '73, pp. 221–227, ACM, 1973.

[3] http://www.xaphoon.com/dataegg/, July, 2012.

[4] http://gkos.com/, July, 2012.

[5] K. Lyons, T. Starner, D. Plaisted, J. Fusia, A. Lyons, A. Drew, and E. W. Looney, "Twiddler typing: one-handed chording text entry for mobile phones," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, (Vienna, Austria), pp. 671–678, ACM, 2004.

[6] http://www.ekatetra.com/products/ekapad.html, July, 2012.

[7] R. Rosenberg and M. Slater, "The chording glove: a glove-based text input device," *IEEE Trans Syst, Man and Cybernetics, Part C: Applic and Rev*, pp. 186–191, 1999.

[8] R. W. Soukoreff, "Text entry for mobile systems: Models, measures, and analyses for text entry research," Master's thesis, York University, 2002.

[9] http://www.ericlindsay.com/palmtop/mwrite.htm, July, 2012.

[10] A. Ralston and E. D. Reilly, eds., *Encyclopedia of computer science (3rd ed.).* Van Nostrand Reinhold Co., 1993.

[11] W. K. Estes, "A statistical theory of learning," *Psychological Review, vol 57*, pp. 94–107, 1950.

[12] I. S. Mackenzie and nd R. W. Soukoreff, "Phrase sets for evaluating text entry techniques," in *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems CHI '03*, (Fort Lauderdale, Florida, United States), pp. 766–767, ACM, 2003.

[13] I. S. MacKenzie, H. Kober, D. Smith, T. Jones, and E. Skepner, "Letterwise: prefix-based disambiguation for mobile text input," in *Proceedings of the 14th annual ACM symposium on User interface software and technology*, UIST '01, (Orlando, Florida, United States), pp. 111–120, ACM, 2001.

[14] C. L. James and K. M. Reischel, "Text input for mobile devices: comparing model prediction to actual performance," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '01, (Seattle, Washington, United States), pp. 365–371, ACM, 2001.

[15] K. Kukich, "Techniques for automatically correcting words in text," *ACM Comput. Surv.*, vol. 24, pp. 377–439, December 1992.

Ameliorating Compound Logistics Processes using Virtual Geo-Sensors

Wolfgang Narzt
Department of Business Informatics – Software Engineering
Johannes Kepler University
Linz, Austria
wolfgang.narzt@jku.at

***Abstract*—A virtual Geo-Sensor is a software module with associated geo coordinates which recognizes the physical presence of traceable mobile devices within a distinct interaction radius. At spatial proximity of an authorized device the Geo-Sensor automatically triggers electronically controlled actions (e.g., it opens gates, starts or stops engines, etc.) not distracting its users from their focused task by making them press buttons or glimpse at a display. This basic paradigm can be adopted for improving logistics processes inside premises, e.g., for supporting the unloading task of trucks from various suppliers. At every unloading site within a closed compound a virtual Geo-Sensor recognizes the arrival of trucks, detects congestion and automatically initiates re-routing procedures which are relayed to the drivers via a network of mobile devices carried within the trucks. This paper sketches the principles and the architecture of virtual Geo-Sensors and demonstrates their potentials in logistics fields in the frame of an experimental survey at the MAN truck manufacturing site in Steyr, Austria.**

Keywords: Geo-Sensors; Compound Logistics

I. INTRODUCTION

The improvement of supply chain event management in logistics domains utilizing new tracking and tracing technologies has been recognized as a key aspect for modern location- and context-based services. The benefit of knowing the live-position of delivered consignment is apparent for supplier and customer enabling them to carry out precise and concerted (pre-)calculations on the event chain with a supposed positive impact in terms of processing time and costs [20][24].

The technical prerequisites for implementing such a service are no longer considered an obstacle, nowadays. It is more an issue in terms of organization, law and work councils when numerous involved suppliers should participate in a multilateral transparent delivery process outside company borders. However, the potentials in utilizing location-based information for logistics fields cannot only be found in time extrapolations for arriving freight from different suppliers. The potentials can also be exploited *within* compound borders where a passed transitional liability grants more legal freedom of action to the compound carrier.

In the course of the nationally funded research program AGTIL [19] focusing on adaptive value creation by the means of the integration of technological, sociological and logistical issues the project consortium consisting of the University of Linz, the Upper Austrian University of Applied Sciences (Logistikum) and MAN Nutzfahrzeuge (truck manufacturer) in Steyr, Austria, envision a mobile location-based compound logistics system for controlling and accelerating the unloading process of delivered consignment within yard borders.

The technical approach for the vision is based on the use of an existing location-based service for mobile devices developed in the course of a research-cooperation between the University of Linz, Siemens Corporate Technology and the Ars Electronica Futurelab. The service is called “Digital Graffiti” [16][18] and is considered a social communication-, collaboration-, and interaction platform where arbitrary users are capable of observing their “friends”’ residences in near real-time (their revocable permission provided) and of consuming and placing location-bound information using their mobile phones. The novelty in this service is included in the type of information that can be placed: information does not only contain static data like text or pictures, it also encloses code fragments or triggers to external services which will automatically be executed when the appropriate privileged user reaches regional closeness to this information (virtual Geo-Sensor [17]).

Upon the technological core of Digital Graffiti the AGTIL project consortium has developed a service infrastructure for a mobile compound logistics system. The moving entities (trucks and their drivers) are identified and tracked by mobile devices. The cargo is modeled as a dynamic part of the driver’s personal user profile enabling a clear mapping of freight, driver and truck. Individually adjustable access- and visibility privileges guarantee privacy protection on a technical basis leaving legal privacy concerns up to negotiations among the participating parties. The control mechanism in this system is built upon the Geo-Sensors placed at all unloading sites and at neuralgic positions along the compound road network automatically announcing e.g., the arrival of trucks and re-calculating site stopover sequences on delays.

II. RELATED WORK

The combination of methods, technology and available information of location-based services and the dynamics of logistics domains has been a focal point of investigation since the emergence of (mobile) tracking devices capable of wirelessly transmitting data [29][30][31]. Various publications present design issues for logistics systems considering location-, context-, and situation-awareness [21][22] for op-

ISBN: 978-1-61208-236-3

timizing logistics processes through real-time vehicle routing and mobile technologies [23]. Many of them address the problem of cooperating different carriers within the supply chain and offer mathematical solutions based on location-bound information. They hardly address closed logistics optimizations based on mobile platforms within compound barriers.

Considering the technical basis of our proposed service (the Digital Graffiti system) we realize location-based services as an emerging focal point of investigation for an increasing number of research labs and industry [5][6][14][15] [27][28]. LocatioNet [9], Mobiloco [10], Plazes [12] or Socialight [13] are services for mobile phones that enable users to get in touch with friends and/or mark real physical locations with simple electronic tags. A comparable application is Google Latitude [4] connecting users to their friends and their current place of residence and providing location-based information within a virtual global public information space.

The concept of the Geo-Sensor handles the issue of seamless transitions between the real and the digital world [11] without distracting the user's attention [7]. Modern solution approaches use Near Field Communication (NFC) [25][26] for contactless initiated actions following the same objectives of dismissing the conventional display and key-controlled interaction paradigm in order to claim a minimum of attention for performing an action at a place of event (e.g. SkiData – contactless access control in skiing areas through RFID). However, the disadvantage in this solution lies within the fact that every location which is supposed to trigger an electronic action has to consider mandatory structural measures for engaging the NFC principle. Beyond, a remaining part of attention is still required as users are supposed to know the position of the NFC system and bring up the RFID tag or reader (depending on which part of the components carries the reading unit) close to the system for proper detection. Regarding the structural measures for implementing NFC this technology is only marginally applicable causing financial and environmental impairments.

Spontaneous interaction triggered upon physical proximity was further studied in numerous works [1][2][8]. These approaches share the aspect that radio sensors are used to determine mutual proximity between smart artifacts and humans. The simplest form of smart artifacts are Smart-Its [3], small computing devices that can be attached unobtrusively to arbitrary physical objects in order to empower these with processing, context-awareness and communication. Smart-Its are designed for ad hoc data exchange among themselves in spatial proximity. Gellersen et al. [3] underlined the importance of awareness of the environment and of the situation for inferring behavior of mobile entities.

III. ARCHITECTURE

Digital Graffiti as the technological basis is conceived as a platform to manage and visualize location-based information within the context of a mobile user. It is built upon a flexible network of mobile GPS-enabled devices (i.e., mobile phones, PDAs, netbooks, etc.) wirelessly obtaining and storing location-based information from and to a central server system (see Figure 1).

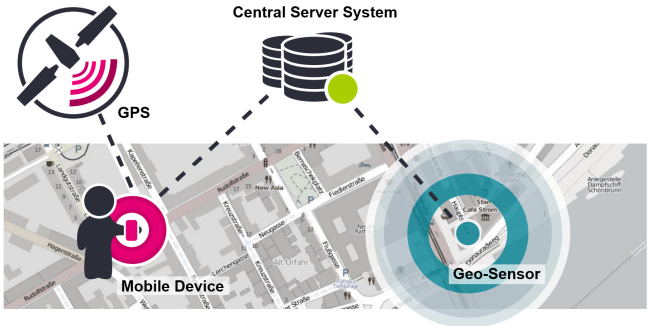


Figure 1. Digital Graffiti System Components.

It has been enhanced with functionality to fulfill the demands for a social network, comprises a map server (e.g., for custom floor plans or industry areas), provides an elaborated user and privileges management concept and additionally handles chat messaging and communication encryption for secure data transfer.

The clients are supposed to be executed on any mobile platform either as a native application particularly designed for the device (currently available for iOS, Android, Symbian and Windows) or as a web application (utilizing the novel W3C standard and HTML5 for accessing GPS out of a browser and complying with the requirements of a bare device without the needs of installing client software).

Once registered and logged in, the user is visualized as an avatar at his exact residing position in front of a map (see Figure 2) and his geographical position is textually resolved into a human readable address (e.g., building names, floor descriptions or office numbers). Alongside user's own position the system also offers to track the position of the user's friends, provided that the respective friend has granted permission. To sustain privacy this permission can be revoked by one click in the user interface.

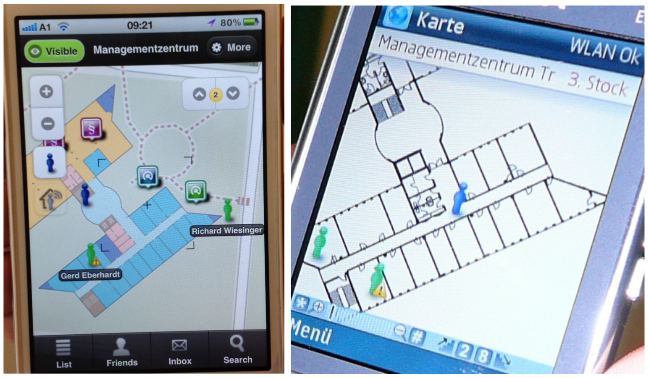


Figure 2. Digital Graffiti User Interface.

Similar to conventional cellular telephony the system uses a distributed provider model for the server-side component where users all over the world can join the provider of their choice in order to take part in the mobile location-based information service. This proven model distributes the load from (asynchronously) communicating users and guarantees scalability of the service all over the world as each provider only handles a limited number of clients.

Every provider stores a set of geographically linked information in appropriate fast traversable geo-data structures (e.g., r-trees) containing hierarchically combinable content modules (which we call gadgets) for text, pictures videos, sound, etc. The name gadget already refers to a possible activity within a module and is the key for a generic approach of integrating arbitrary system connections or electronic actions to be triggered automatically on arriving users (Geo-Sensors). They provide the basis for extensibility to third-party systems for which the number and variety of electronic connections is unforeseeable and simultaneously enriches the potentials of such a service [18]. Figure 3 illustrates the common principles of the Geo-Sensor architecture which enables fast connections to third party systems:

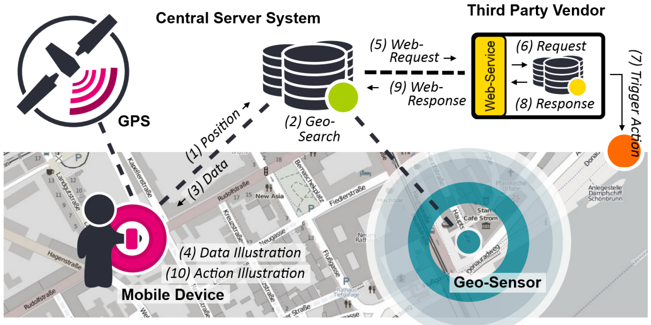


Figure 3. Geo-Sensor Architecture.

Clients repetitively transmit their own (commonly GPS-based) position to a server (1), which evaluates the geo-data considering visibility radiuses and access constraints (2) and transmits the corresponding results back to the clients (3). Generally, when the transmitted information contains conventional gadgets as text and pictures, it is immediately displayed on the output device of the client (4). The basic idea for executing code is to use the gadget metaphor and store executable code inside instead of text or binary picture data (smart gadgets). Therefore, we propose a web-service-based mechanism which is both effective and simple to extend: Smart gadgets contain a simple URL or XML-based web-request to a remote web-service which is the actual component to execute the code. When a client receives information containing a smart gadget, its URL is resolved (5) which is handled internally (6) and finally triggers the desired action at the third-party vendor (7). A response back to the client (8, 9) can additionally be illustrated as a visual confirmation whether the action could have been executed or not (10).

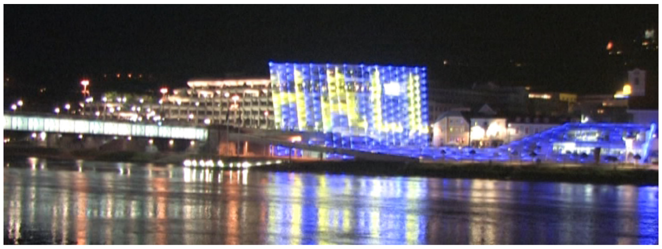


Figure 4. Geo-Sensor Example using Digital Graffiti.

This approach is simple because the clients just have to handle standardized web-requests. A majority of currently utilized mobile platforms support these mechanisms. Important for third-party vendors: Their internal data representations, servers and control units are hidden from the publically accessible location-based service guaranteeing a maximum degree of data security for the vendors.

Figure 4 gives an impression on this innovative interaction paradigm: We have put a Geo-Sensor containing executable code near the Ars Electronica Center building in Linz, the LED-facade of which is capable of displaying marquee text running around the walls of the building. An authorized person approaching the Geo-Sensor automatically triggers the execution of the contained code which causes the facade to welcome the user personally. Of course, this application is more of a playful approach rather than a business scenario, however, it demonstrates the potentials of the service enabling its users to initiate any electronically controllable action just by their physical presence.

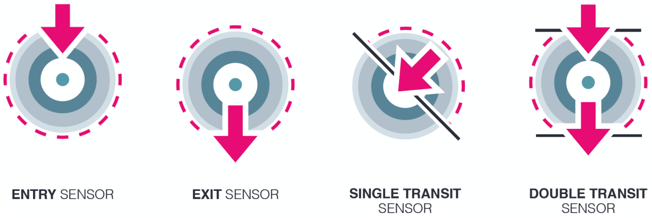


Figure 5. Geo-Sensor Types.

For even more flexibility the Digital Graffiti framework provides a series of differently triggering Geo-Sensors (see Figure 5) an application can select from in order to meet its particular requirements best possible: The simplest forms are Entry- and Exit-Sensors firing when a device either comes into or leaves the interaction radius of the sensor. A Single-Transit-Sensor defines a virtual line within its radius which must be passed from one direction in order to trigger it. Sensors of this type may be used in traffic scenarios where just the flow of a distinct direction is of interest. An extension of this sensor is the Double-Transit-Sensor firing twice at the entry and exit of a device from a given direction. Sensors of this type may e.g., detect congestion.

IV. Experimental Survey

A Double-Transit-Sensor also recognizes the stopping times of every truck in the course of an experimental survey at the MAN truck manufacturing site in Steyr, Austria, where the unloading process of consignment should be ameliorated within compound borders using the location-based Digital Graffiti service. At every unloading site such a sensor both records arriving as well as the departure times and reports potential congestion to a control center where re-routing procedures can be initiated for further trucks scheduled for a congested site. Re-routing can both be done manually due to a visual impression of capacity utilization on the compound or automatically considering dynamically adjustable constraints like unloading sequences.

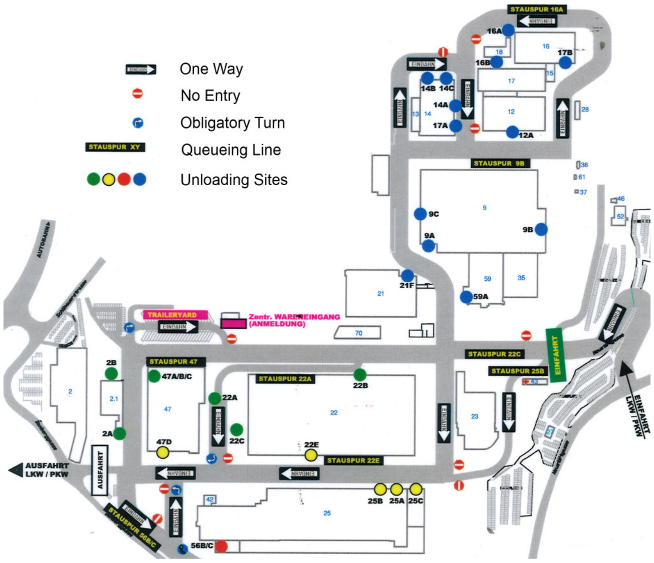


Figure 6. Geo-Sensors at 25 Unloading Sites at MAN Steyr.

Figure 6 gives an impression on the setup of the survey with in total 25 Double-Transit-Sensors at unloading sites (marked by green, yellow, red and blue dots) and several other Geo-Sensors at strategic points in the compound (e.g., at the two main entries or at the trailer yard where arriving drivers have to register and deposit their papers.

The test scenario works as follows: When the driver arrives at the trailer yard he is handed out a mobile device clearly identifying the truck and its cargo. A preceding registration click has been carried out by the operator connecting cargo data and device. Now the device provides tracking information to the control center and informs the drivers about the succeeding unloading site. At changes the operator is able to address an alternative destination directly to the appropriate driver and is therefore given a powerful instrument to dynamically interfere into compound processes.

The system architecture for this experimental survey (see Figure 7) contains an original unchanged Digital Graffiti kernel managing mobile users and Geo-Sensors (i.e., LBS data). It is controlled by a wrapping web-based Control Center, the actual application core handling unloading sequences or re-routing procedures. Proprietary cargo data (hosted at a special server system at MAN) is transferred via EDI (Electronic Data Interchange) interfaces to a temporary OFTP-server from where it is pushed to the Control Center.

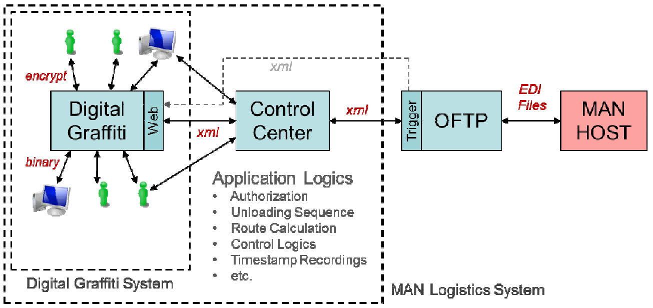


Figure 7. System Architecture of Experimental Survey at MAN Steyr.

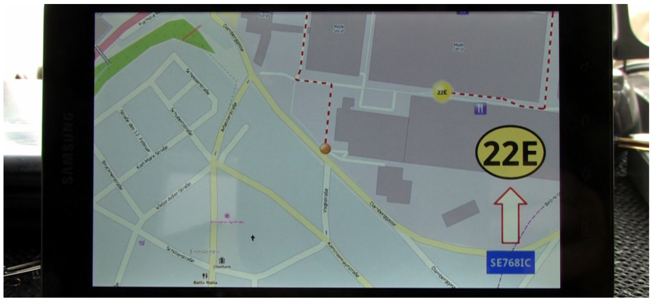


Figure 8. Compound System: Driver's Client.

Figure 8 shows the prototype mobile client application for the drivers (here: running on a Samsung 7" tablet with Android 2.2 mounted inside the windshield of a truck) which indicates the driver's own position on a detailed compound map, his next goal (here: unloading site "22E") and the route to it.

In the first test phase the technical requirements concerning feasibility, accuracy and real-time behavior have been evaluated: Is the compound system using Geo-Sensors capable and accurate enough to recognize waiting times in appropriate time intervals? Therefore, a number of selected trucks have carried mobile devices during their regular unloading process. The Geo-Sensors at the unloading sites have recorded all timestamps regarding entry- and exit times which have been cross-checked by manually noted timestamps of accompanying supervising persons.

As the assessment of this technical precondition succeeded (i.e., there is a clear correspondence between manually and automatically recorded timestamps) the second test phase could be initiated evaluating the economic potentials concerning time savings. In its final state this survey phase schedules for a compound-wide test with mobile devices in every truck and an automatic re-routing process due to detected delays. As such a test scenario is both expensive and organizationally elaborate (about 150 trucks arrive at the compound during the day with a maximum number of 40 trucks residing concurrently inside premises) only a light-weight version of this test has been carried out at this time of writing with an assortment of both manually and automatically recorded timestamps and a manual interference of an expert operator due to visually recognized impairments in the speed of the unloading processes. However, these data already reveal the economic potentials of this system in terms of reducing stopover times for individual trucks.

V. Preliminary Results

Figure 9 illustrates a glimpse on these data showing preliminary results of the system tests. The picture lists operating times for one specific test day on November 22nd 2011 at seven unloading sites (named "22A", "22B", "56", etc.) at the yard of MAN Steyr for 71 trucks and 112 unloading tasks performed by these trucks. So, the picture presents time measurements on site level, not for individual trucks (i.e., several trucks are listed more than once in this picture).

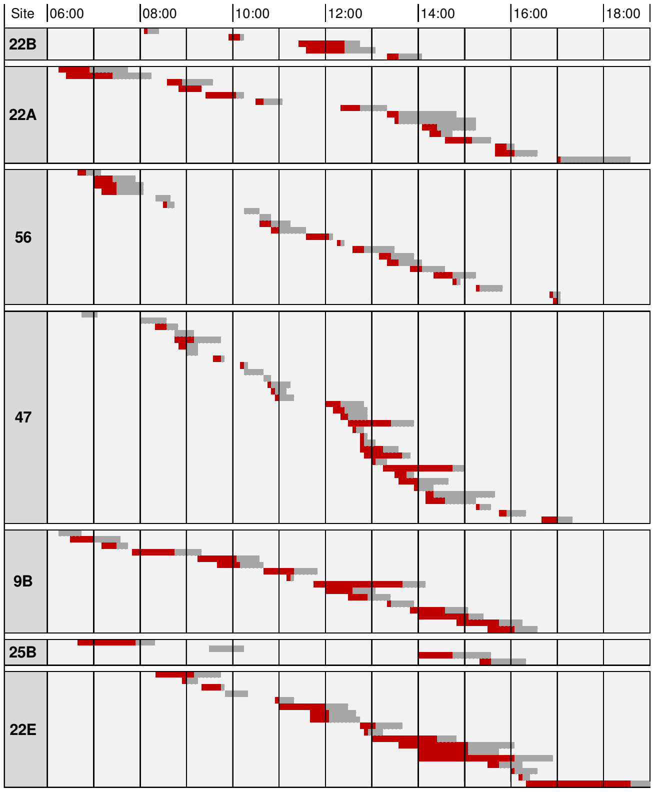


Figure 9. Recorded Timestamps at Unloading Sites.

The red lines indicate waiting times and the gray lines the time for the unloading process itself. In total figures this means that the drivers spend 45.6 hours waiting and 50.7 hours for the unloading tasks (or: approx. 25 minutes at every site for each driver waiting and half an hour unloading), which reveals nearly half of the time spent waiting. Thus, the theoretical potential considering these figures is an average reduction of unloading times by half.

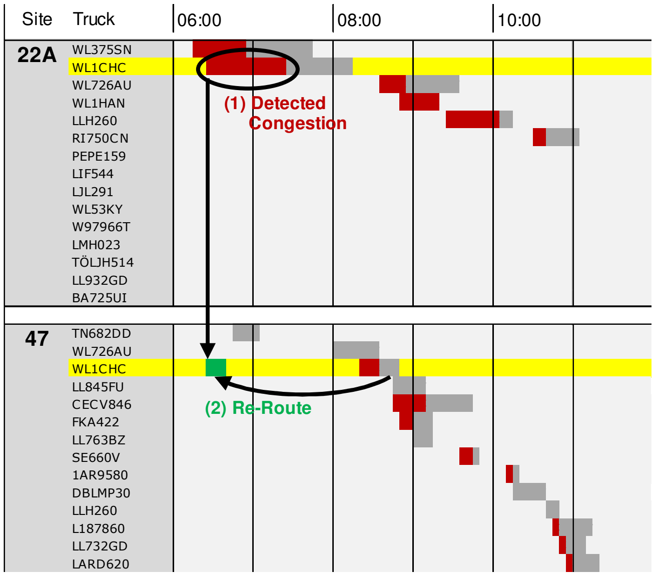


Figure 10. Re-Routing due to Congestion.

A closer look on the figures shows (see Figure 10) that the truck marked in yellow faces congestion at site 22A, whereas site 47 would be clear at the same time. It could be rerouted in order to switch the unloading sequence (assuming that the order is variable what might not always be the case) and in addition avoid a second recorded congestion later at site 47. This example results in a time improvement of 40% for this individual truck (before: 75 min waiting time, after: 45 min) and impressively demonstrates the potentials. However, it is still an excerpt for one individual entity and does not consider side effects which will likely occur on re-routing instructions, thus only an area-wide test which is still to be conducted will provide a clearer insight into the contingent average value of improvement.

The tests have also provided valuable information concerning social issues: Whereas drivers unaware of the regional conditions embrace a mobile guide escorting them through the compound there are more than half of the drivers who repetitively come along and refuse an additional gadget providing them with information they know anyway.

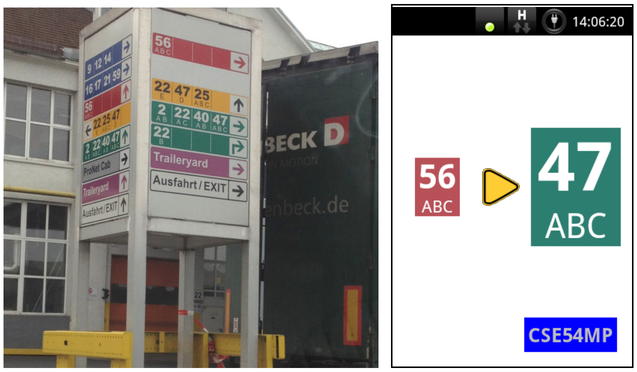


Figure 11. User-Interface Adaptations.

As a consequence, we have modified the user interface for the mobile device in a way that it does not show a map with one's own position on it, anymore. Instead, we utilize the local direction signs and appropriately display them on the screen (see Figure 11). Drivers unfamiliar with the place are still guided by the system whereas the others may keep their devices in their pockets (not perceiving well-known information) but are also notified on re-routings by an acoustic alarm and a firm depiction of the change.

VI. Conclusion and Future Work

Although, the system presented in this paper is still under development first prototypical implementations and tests confirm applicability of the virtual Geo-Sensor metaphor for being used for closed compound logistics operations. The project consortium is convinced that Geo-Sensors offer large potentials in terms of reducing congestion times while carrying out in-yard tasks. At every unloading site a virtual Geo-Sensor detects the presence of trucks and automatically notifies and re-routes on delays.

For verifying the supposed economic potentials further tests within the compound of MAN in Steyr are still necessary. Every truck driver will have to carry a mobile device and is requested to follow re-routing instructions relayed by the mobile device in order to create a quantifiable statement considering side effects of this dynamic interference. The final goal will be a downloadable mobile app to be installed by the participating parties in order to avoid registration routines at the entry gates with benefits for both sides regarding an improved use of their resources.

ACKNOWLEDGMENT

Special thanks goes to the project partners at the University of Linz, the University of Applied Sciences (Logistikum) and MAN Nutzfahrzeuge Steyr. We also want to express our gratitude to the Upper Austrian Government and the initiative AGTIL for funding this research project.

REFERENCES

[1] W. Brunette, C. Hartung, B. Nordstrom, G. Borriello, “Proximity interactions between wireless sensors and their application”, in: WSNA '03: Proceedings of the 2nd ACM international conference on Wireless sensor networks and applications, New York, NY, USA, ACM Press pp. 30-37, 2003.

[2] A. Ferscha, R. Mayrhofer, R. Oberhauser, M. dos Santos Rocha, M. Franz, M. Hechinger, „Digital aura”, in: Advances in Pervasive Computing. A Collection of Contributions Presented at the 2nd International Conference on Pervasive Computing (Pervasive 2004). Volume 176., Vienna, Austria, Austrian Computer Society (OCG) pp. 405-410, 2004.

[3] H. Gellersen, G. Kortuem, A. Schmidt, M. Beigl, “Physical prototyping with smart-its”, IEEE Pervasive Computing 3(3), 74-82, 2004.

[4] Google Latitude, www.google.com/latitude, verified May 2012.

[5] C. Gutwin, R. Penner, K. Schneider, „Group awareness in distributed software development”, ACM conference on Computer Supported Cooperative Work, 1999.

[6] L. E. Holmquist, J. Falk, J. Wigström, “Supporting Group Collaboration with Inter-Personal Awareness Devices”, Personal Technologies, Vol. 3, Nos. 1&2, 1999.

[7] H. Ishii, B. Ullmer, B, „Tangible bits: towards seamless interfaces between people, bits and atoms”, In: CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM, 234-241, 1997.

[8] G. Kortuem, J. Schneider, D. Preuitt, T. G. C. Thompson, S. Fickas, Z. Segall, „When peer-to-peer comes face-to-face: Collaborative peer-to-peer computing in mobile ad hoc networks”, In: Proceedings of the First International Conference on Peer-to-Peer Computing, 2001.

[9] Locationet, http://www.locationet.com, verified May 2012.

[10] Mobiloco, http://www.mobiloco.de, verified May 2012.

[11] W. Narzt, G. Pomberger, A. Ferscha, D. Kolb, R. Müller, J. Wieghardt, H. Hörtner, R. Haring, C. Lindinger, „Addressing concepts for mobile location-based information services”, Proceedings of the 12th International Conference on Human-Computer Interaction HCI, 2007.

[12] Plazes, http://plazes.com, verified May 2012.

[13] Socialight, http://www.socialight.com, verified May 2012.

[14] T. Sohn, K. Li, G. Lee, I. Smith, J. Scott, W. G. Griswold, “Place-Its: A Study of Location-Based Reminders on Mobile Phones”, UbiComp'05: Seventh International Conference on Ubiquitous Computing, pp. 232-250, 2005.

[15] F. Zambonelli, M. Mamei, „Spatial computing: an emerging paradigm for autonomic computing and communication”, 1st International Workshop on Autonomic Communication, Berlin (D), 2004.

[16] W. Narzt, W. Wasserburger, „Digital Graffiti – A Comprehensive Location-Based Travel Information System”, 16th International Conference on Urban Planning, Regional Development and Information Society (REAL CORP 2011), Editors: M. Schrenk, V. Popovich, P. Zeile, Essen, North Rhine-Westphalia, Germany, 2011.

[17] W. Narzt, H. Schmitzberger, “Location-triggered code execution — dismissing displays and keypads for mobile interaction”, UAHCI ’09: Proceedings of the 5th International Conference on Universal Access in Human-Computer Interaction. Part II. Berlin, Heidelberg: Springer-Verlag, pp. 374–383, 2009.

[18] G. Pomberger, “Digital Graffiti - A Framework for Implementing Location-Based Systems”, in International Journal of Software and Informatics (IJSI), Volume 5, Issue 1-2 (2011), Part II, ISSN 1673-7288, pp. 355-377, 2011.

[19] AGTIL, Adaptive Gestaltung der Wertschöpfung durch die Verknüpfung von Technologie, Industriesoziologie und Logistik, http://www.agtil.at, verified May 2012.

[20] H.-Ch. Graf N. Tellian N. “Smartphones as enabler of Supply Chain Event Management”, Management of Global and Regional Supply Chain – Research and Concepts, Posen, Poland, ISBN 978-83-7775-066-7, pp. 133–143, 2011.

[21] A. Langevin, D. Riopel, “Logistics Systems: Design And Optimization”, Springer New York, ISBN: 0-387-24971-0, 2005.

[22] Ul. Meissen, S. Pfennigschmidt, A. Voisard T. Wahnfried, “Context- and Situation-Awareness in Information Logistics”, Lecture Notes in Computer Science, Volume 3268/2005, 448-451, 2005.

[23] G.M. Giaglis, I. Minis, A. Tatarakis, V. Zeimpekis, “Minimizing logistics risk through real-time vehicle routing and mobile technologies: Research to date and future trends”, International Journal of Physical Distribution & Logistics Management, Vol. 34 Iss: 9, pp.749 – 764, 2004.

[24] H. Graf, W. Narzt: "Perspectives of Smartphone Technology as a Supply Chain Event Management Tool", 4th World Conference on Production and Operations Management, Amsterdam, July 2012, in press.

[25] N. Li, B. Becerik-Gerber: “Performance-based evaluation of RFID-based indoor location sensing solutions for the built environment”, Advanced Engineering Informatics, Volume 25, Issue 3, August 2011, pp. 535-546.

[26] W. Lu, G. Q. Huang, H. Li: “Scenarios for applying RFID technology in construction project management”, Automation in Construction, Vol. 20, Issue 2, March 2011, pp. 101-106.

[27] C. Liu, P. Rau, F. Gao: “Mobile information search for location-based information”, Computers in Industry, Volume 61, Issue 4, May 2010, pp. 364-371

[28] W. Ait-Cheik-Bihi, M. Bakhouya, A. Nait-Sidi-Moh, J. Gaber, M. Wack: “A Platform for Interactive Location-Based Services”, Procedia Computer Science, Volume 5, 2011, pp. 697-704.

[29] E. Kutanoglu, M. Mahajan: “An inventory sharing and allocation method for a multi-location service parts logistics network with time-based service levels”, European Journal of Operational Research, Volume 194, Issue 3, 1 May 2009, pp. 728-742.

[30] Y. Kayikci: “A conceptual model for intermodal freight logistics centre location decisions”, Procedia - Social and Behavioral Sciences, Volume 2, Issue 3, 2010, pp. 6297-6311.

[31] A. Javid, N. Azad: “Incorporating location, routing and inventory decisions in supply chain network design”, Transportation Research Part E: Logistics and Transportation Review, Volume 46, Issue 5, September 2010, pp. 582-597.

Techniques for Interacting With Small Devices

Javier Oliver
Department of Computer Engineering
University of Deusto
Bilbao, Spain
javier.oliver@deusto.es

Begoña García
DeustoTech
University of Deusto
Bilbao, Spain
mbgarciazapi@deusto.es

***Abstract*— Digital mobile devices keep reducing their size as time goes by. The limiting factor is no longer battery size or electronics miniaturization, but the dimensions of the input and output hardware devices that mediate the communication between the user and the machine. In this paper, the difficulties of interacting with small screen devices are underlined and some of the most promising techniques to address this issue are explained. These techniques include the use of magnetic field detectors, mobile phone cameras to track movement, tangible interfaces and voice controlled virtual joysticks.**

Keywords-interaction techniques; small device interaction

I. INTRODUCTION

Recent evolution of digital mobile devices has produced smaller and smaller products, to the point that some of them can be hidden inside the clothes or even implanted under the skin. The limiting factor of this miniaturization is no longer electronic design or even battery size. The main issue that conditions this constant reduction in the dimension of these mobile devices is the size of the input and output hardware that is required to implement the user interface [1]. The size of a screen should be big enough for a user with normal visual acuity to read the text, and buttons in a keyboard should be big enough for a normal sized finger to press them.

Although the size of the mobile devices keeps shrinking, the fingers of the users, or the size of the text they can read keep constant. If the use of smaller and smaller digital products wants to be fostered, existing technologies should be used creatively, and new technologies should be developed to implement a new generation of user interfaces that can overcome the difficulties of interacting with small mobile devices. Our main purpose in the rest of this paper is to describe some of the most promising innovative interaction techniques, giving a small sample of what interaction might look like in the near future.

In the following section we define what we understand by small mobile devices. In Section III we describe some of the techniques that can be used to interact with these small devices, and we present our conclusions in Section IV.

II. WHAT ARE SMALL MOBILE DEVICES?

Cellular phones are by far the most common mobile device available nowadays. It has been estimated that there were about 6.000 million of these devices in the world at the beginning of 2012 [2]. Most cellular phones fall under the category of small mobile devices because their design is not optimized for interaction and this produces a number of difficulties when the interaction tasks are carried out, namely low readability of small screens and hard to push small buttons.

But, even smaller mobile devices have been marketed, making the problem of fluid and effortless interaction even harder. For example, many devices designed for open air activities, such as GPS systems or training computers have screens of about 2.5’’ in diagonal. *Siftables* are small blocks that include wireless communication capabilities, sensing and a small screen of about two inches in diagonal. This design offers a whole new set of tangible user interface techniques to interact with digital information, making use of our high dexterity in manipulating digital objects [3]. Some of the smallest devices that have been designed are the *Telebeads*: electronic wearable jewelry objects that can be used as mnemonic aids and communication appliances to help in the managing of social network data. These systems have screens as small as a fraction of an inch [4].

III. ASSORTED INTERACTION TECHNIQUES FOR SMALL DEVICES

According to Fitt’s Law [5], target size and interaction time are inversely proportional. But in the small touch screens of portable devices, real state is very limited, so how big should targets be for a comfortable and fast interaction? Interaction guidelines offer different views on the subject [5]. In the *iPhone Human Interface Guidelines* [5] it is suggested that the minimum target size should be 44x44 pixels. *Windows Phone UI Design and Interaction Guide* says that the minimum should be 26x26 pixels, and *Nokia’s Developers Guidelines* propose 28x28 pixels. However, the average index finger is between 45 and 57 pixels wide, more than any of the above recommendations.

In the case of very small touch screens, the so called *fat finger problem* is exacerbated [6]. The finger occludes most of the screen real state and interaction is greatly hampered. In these situations, it has been proposed to move the touch sensitive hardware to the back of the device, so that the screen is visible and the position of the finger is shown by a small cursor [7]. Traditional touch screen pointing techniques try to alleviate screen occlusion by using offset cursors or a method called *Shift* where the user is shown a representation of the area occluded by the finger in a free region of the small screen. The exact position of the finger is shown by means of a cursor, and this helps performing the positioning task. When the *Shift* technique was compared with the interaction on the back of the device, it was shown that *Shift* did not work for screen sizes below one inch diagonally, whereas back-of-device interaction was successful almost independently of screen size.

Another approach to the interaction with a small screen is based on a magnetic field detector. Located behind the screen, this device is capable of very accurate positioning and leaves the screen completely visible [8]. In this study, a 1.5 inch screen was used, with a resolution of 280 x 220 pixels. The magnetometer used was capable of providing an angular accuracy of about two degrees at a cost of five US dollars. Users wore a small magnet in the index finger, which provided a useful range of about 10 cm, for a total active area of about 300 cm^2. This setup increased the operational area offered by the original 1.5 inch screen by a factor of more than 50.

TinyMotion is a software that uses the camera of a mobile phone as an input device. *TinyMotion* analyzes in real time a series of images taken by the mobile phone camera and extracts information about the movement of the phone. In order to evaluate the applicability of this approach, the *TinyMotion* team developed a number of applications and video games, all of which were controlled by moving the mobile phone is various ways [9]. In an application called *Mobile Gesture*, the user presses the *OK* button before writing a character, and then presses the *#* button to indicate the end of the writing process. Writing in this case means moving the mobile phone with the camera on to recognize the strokes of the character.

The recognizer code can detect western characters, punctuation symbols and more than 8,000 Chinese and Japanese characters. It takes about 20 ms to recognize a western character and about 40 ms to recognize one of the Chinese or Japanese characters (the hardware used is a Motorola v710 mobile phone, an unmodified model bought in 2005). *TinyMotion* has undergone several evaluations. To begin with, an informal usability test was carried out with 13 users. The results were very encouraging because the system was found to be very responsive. Many different backgrounds for the camera were used, and most of them worked very well. Even pointing the mobile phone camera to the blue sky gave good results.

The only failures to detect the movement were those with very extreme lighting conditions or very rapidly changing backgrounds such as a dark room, the surface of a computer screen switched off or pointing the camera through the window of a moving car. Some users even found it more convenient not to move the mobile phone and move the other hand in front of the camera instead. One of the most innovative interaction techniques for small screen is the combination of sensing technology and tangible user interfaces used in the *Siftables* project [3]. Tangible user interfaces are based on the notion of providing physical handles for digital objects, thus being able to access digital information by manipulating common objects. In some instances of tangible user interfaces, the system projects graphics onto the handles and in others, these handles are simply used to control more conventional graphical user interfaces. Some of the advantages of tangible user interfaces include [3]:

- Less significant cognitive requirements than an equivalent graphical user interface.
- Faster interaction
- Two handed input of data is supported, although multi touch screens offer this capability for more conventional interfaces.

The other technology that that *Siftables* project uses is the Sensor Network User Interfaces (SNUI). These are sets of elements capable of communication and sensing, that can have an organized behavior and be manipulated so that they conform a tangible user interface to access digital data. *Siftables* are small square tiles of about 36 mm per side and 10 mm thick. Each of them has a small color screen, an accelerometer, a set of infrared transceivers, a battery and an RF radio. With this hardware, the *Siftables* can sense their own motion, and also contacts with another objects. They can detect movements like elevation, tilting or vibration. They can also detect other tiles other tiles situated close by.

The communication capabilities of the system allows tiles to share information with other tiles or with a central computer located in the vicinity. The capabilities of the *Siftable* system are allowing the development of new interaction techniques in the domain of SNUI's, analogous to the more familiar metaphors of graphical user interfaces:

- Shaking or piling several of the elements at the same time could be interpreted as classifying them as belonging to the same group.
- Putting several tiles together could form a bigger screen to show large documents.
- Shaking vertically could mean *yes* and shaking horizontally could mean *no*.

A prototype photo sorting application has been developed by the *Siftable* team to illustrate the possibilities of SNUIs [3].

The hands can be avoided altogether in the interaction with very small screens, thus eliminating the *fat finger problem*. The *Vocal Joystick* is a system that allows the control of a pointing device by means of the voice. The technique can be used for onscreen selection, arbitrary point navigation and path following as required in drawing applications and videogames [10]. *Vocal Joystick* can recognize verbal and non verbal vocalizations, and other sound characteristics, such as loudness and pitch, and transform them into movements of the cursor. The process is continuous, and mouse movements are generated without delay. The *Vocal Joystick* can be implemented in an average personal computer and only requires a microphone and a sound card. The sound produced by the user is continuously monitored, and the movement of the pointer is immediately generated. Vowel quality depends on the articulation configuration of the mouth, and depending on the ability of the user two methods can be used: four direction or eight direction modes. In addition to the above, the system can also recognize a number of short sounds that can be used as trigger actions to perform functions such as a mouse click. There is a standard set of sounds, but adaptation to particular users is also possible, improving the performance of the interaction. As users become more experienced with *Vocal Joystick* they can reach interaction speeds comparable to hand operated joysticks.

A technology that completely eliminates the need of a screen is the use of passive magnetic tags. Magnetic tracking does not need a direct line of sight, but in the case of motion capture devices, they require active sensors, and this involves complex equipment and a wire connection between the detectors and the processing unit. On the other hand, passive magnetic tags can be powered by radiofrequency energy sent by the base station, and if attached to common everyday objects, these can be tracked, their orientation can be detected, and they can even respond to other actions such as pressure, finger position, etc. With this passive magnetic tag technology, and a few plastic objects, an interesting digital musical instrument with a tangible user interface has been developed.

A total of sixteen small plastic objects have been used to control a music producing application. Some of the objects have three orthogonal magnetic tags to monitor orientation in addition to distance to a reference point.

Each tagged object produces a different output when it is close to the receptor. An attached computer generates the corresponding MIDI messages that are sent to a number of music synthesizers. In addition to the sound output, the computer also generates background graphics. Although tags working in neighboring frequencies may show small interferences, all of the tagged objects can be used together.

The general public has had a very positive reaction to this system, and because of the simple interface, its use is very intuitive. The output of the system is somehow limited, so improvements are being made to turn this enjoyable demo into a full fledged musical instrument.

Another approach that can be taken is that of implanted interfaces. Just as pacemakers or hearing aids can be surgically placed underneath the skin, small input and output devices can also be permanently implanted under the skin of the users. In a recent work, it has been proposed that user interfaces could be implanted to allow users to perform simple interactions with their own bodies. In this study, a simulated implant was made to obtain an initial qualitative feedback on the use of implanted interfaces. A device with three inputs (button, tap sensor and pressure sensor) and three outputs (LED, vibration motor and piezo buzzer) was placed on the left arm of four volunteers and covered with silicon artificial skin. The volunteers had to perform some simple everyday tasks, such as taking a bus or asking for directions to go to the post office. As a secondary task, they had to pay attention to the output of the device and *answer* with the appropriate input control. In general, the participants considered that the device was easy to use, and all felt that the vibration motor was the easiest output channel to perceive. All users were able to see the blinking LED when looking at it, even in direct sunlight. Although the authors conclude that it is feasible to operate small and simple user interfaces implanted under the skin, they put forward some challenges associated with the use of these interfaces. Regarding input, it has to cross the skin, so the use of sound or light is somehow limited. Also, accidental operation of the controls has to be considered and avoided. Output is normally visual, auditory or tactile, and the bandwidth is small. For example, in the case of visual stimuli, typically a small LED flashing through the skin is used. Tactile feedback could be specially appropriate because it would not be perceived by anyone except the users.

IV. Conclusion and Future Work

The continuous reduction in size of digital mobile devices is presenting new challenges to interaction designers. When input and output hardware is reduced beyond a certain point, traditional interaction techniques have to be applied in creative ways or new interaction techniques have to be developed [11-15] to meet communication needs between the user and the system. In this paper, several innovative techniques have been described, including the *Shift* method, magnetic field detectors, the use of mobile phone cameras to track movement, tangible interfaces and voice controlled pointer management systems. Future work should widen the spectrum of the techniques reviewed here, and provide some kind of categorization.

Educators and practitioners should be familiar with these new trends in interaction with small devices to prepare future professionals for the interface design scenarios that they will meet in their careers.

References

[1] Ni T. and Baudisch P. Dissapearing mobile devices. 22nd Annual Symposium on User Interface Software and Technology , October 4-7, Victoria, British Columbia, Canada, 2009, pp. 101-110.

[2] Whitney L. 2011 ends with almost 6 billion mobile phone subscriptions. CNET News, January 4, 2012. Available at: http://news.cnet.com/8301-1023_3-57352095-93/2011-ends-with-almost -6 -billion-mobile-phone-subscriptions/, 2012 [retrieved: september, 2012].

[3] Merrill D., Kalanithi J., and Maes P. Siftables: Towards Sensor Network User Interfaces. In Proceedings TEI'07, 2007.

[4] Labrune J.B. and Mackay W. Telebeads: Social Network Mnemonics for Teenagers. In Proc IDC '06, 2006.

[5] Anthony T. Finger-Friendly Design: Ideal Mobile Touchscreen Target Sizes. Smashing Magazine. Available at: http://uxdesign. smashingmagazine.com/2012/02/21/finger-friendly-design-ideal-mobile-touchscreen-target-sizes/, 2012 [retrieved: september, 2012].

[6] Siek K.A., Rogers Y., and Connelly K.H. Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. In Proc. INTERACT’05, 2005.

[7] Baudisch P. and Chu G. Back-of-Device interaction Allows Creating Very Small Touch Devices. In Proc CHI 2009.

[8] Harrison C. and Hudson S.E. Abracadabra: Wireless, High-Precision, and Unpowered Finger Input for Very Small Mobile Devices. 22nd

Annual Symposium on User Interface Software and Technology , October 4-7, Victoria, British Columbia, Canada, 2009, pp. 121-124.

[9] Wang J., Zhai S., and Canny J. Camera phone based motion sensing: interaction techniques, applications and performance study. In *Proceedings of the 19th annual ACM symposium on User interface software and technology*, 2006.

[10] Harada S., Landay J.A., Malkin J., Li X., and Bilmes J.A. The Vocal Joystick: Evaluation of Voice-based Cursor Control Techniques. In Proc ASSETS, 2006, pp. 197-204.

[11] Holtz C., Grossman T., Fitzmaurice G., and Agur A. Implanted User Interfaces. CHI´12, Austin, Texas, 2012, pp. 503-512.

[12] Butler A., Izadi S., and Hodges S. SideSight: multi-"touch" interaction around small devices. Proceedings of the ACM Symposium on User Interface Software and Technology (UIST '08), 2008.

[13] Cho S.J., Murray-Smith R., and Kim Y.B. Multi-context photo browsing on mobile devices based on tilt dynamics. In *Proceedings of the 9th international conference on Human computer interaction with mobile devices and services*, 2007, pp. 190-197.

[14] Paradiso J.A., Hsiao K.Y., and Benbasat A. Tangible music interfaces using passive magnetic tags. In *Proceedings of the 2001 Conference on New interfaces for Musical Expression*, 2001.

[15] Yatani K. and Truong K.N. SemFeel: A User Interface with Semantic Tactile Feedback for Mobile Touch-screen Devices. 22nd Annual Symposium on User Interface Software and Technology , October 4-7, Victoria, British Columbia, Canada, 2009, pp. 111-120.

Development of m-Sahayak- the Innovative Android based Application for Real-time Assistance in Indian Agriculture and Health Sectors

Biswajit Saha, Kowsar Ali, Premankur Basak, Amit Chaudhuri
ICT & Services
Centre for Development of Advanced Computing (C-DAC), Kolkata
Kolkata, India
e-mails: { biswajit.saha, kowsar.ali, premankur.basak, amit.chaudhuri }@cdac.in

***Abstract*—Mobile or Smart phones (Android based) is becoming an essential device for all types of people irrespective of the age group and literacy (literate or illiterate). In India, mobile technology has unleashed a paradigm shift in the communication medium to reach out to the masses. Crops/plants need regular physical and scientific handling for proper growth. Extension of mobile-phone based any-time, any-where scientific expert advices to the farmers is a possibility in India, i.e., information was available earlier for electronic processing, and now, communication is merged with Information Technology to create ICT impacts as a whole. ICT enabled environment is becoming a day-to-day reality everywhere in India. Tele-health allows health care professionals to diagnose and treat patients in remote locations using ICT. In this paper, the development of a special Android-based application is reported. The application takes care of certain problems in agriculture and health care by concurrently capturing images, audio and video and sending them to a specified server. Agricultural Scientists or doctors can view or listen to this images/video/audio information and provide proper solutions, accordingly. The development was tested satisfactorily. The paper will report on system concepts, design details and results.**

Keywords-Android; Mobile Application; Remote area; Real-time Assistance; Telehealth

I. INTRODUCTION

The agriculture sector is changing the socio-economic environments of the population due to liberalization and globalization. About 75% people are living in rural India and still depend on agriculture. About 43% of India's geographical area is used for agricultural activity. The agriculture continues to play a major role in Indian economy, by contributing 1/6th of the export earnings. Agriculture is crucial to India's economy, as it provides 20% of GDP and employs 60% of the workforce. However, most of India's poorest people are subsistence farmers who have little or no access to technology for their proper solution of the damaged crops. Farmers lack knowledge on medicine or procedure to control the plant damages [1]. Every year, significant amounts of agricultural products are lost in India due to some critical diseases and improper maintenance. In remote areas, very often, the farmers do not get any suggestion regarding the correct scientific procedure to be followed for a particular cultivation. For this reason, they produce lesser amounts of crops by incurring more expenditure.

Sahayak, in some Indian languages, means 'Assistant'. Therefore, the application we present has been given a local name, Mobile Sahayak or m-Sahayak. In India, problems are faced by common people in rural areas in the health sector, too. In urban areas, people get better facility in health care. However, the qualified doctors are not always willing to go to serve in remote rural areas. Many patients in remote villages die without proper treatment [11]. The population is increasing day by day, but the number of physicians, doctors, nurses or government hospitals serving rural population is not increasing proportionately. Therefore, the health care needs of rural population are not being addressed properly [12]. In remote areas, there is a dearth of modern healthcare facilities. This situation demands introduction of tele-medicine support in order to provide fast and high quality medical consultancy covering broad areas. The significant increase of capabilities in modern telecommunication and data processing enables advanced tele-service solutions to assist medical treatment at remote locations [9]. As an affordable and accessible means of communication, rural communities are realizing the potential of mobile telephony to create economic opportunities and strengthen social networks. Mobile telephony effectively reduces the "distance" between individuals and expert scientist/doctors, making the sharing of information and knowledge easier and more effective. It is hoped that the development of this Android-based application can be widely implemented in the near future. This will benefit people in rural areas. Even rehabilitation centres, village schools, mobile health care units and industrial units like mines [5] of developing countries may use the application.

II. RELATED WORKS

In recent days, there has been an attempt to assist the farmer by telephony service but, this service is not 24X7 hours service. Sometimes, the farmers are not able to connect with experts due to communication failures [6]. Another important problem is that in a critical situation, if the farmers are not able to explain or if the disease is a new one, then farmers would not be able to identify the diseases of the crops [1][2][4][9]. Captured images from crop surfaces can provide a better solution where the remote agri-scientist can see instantly the image for disease diagnosis. Similarly,

ISBN: 978-1-61208-236-3

captured skin, face, or other images through the developed application may be sent to expert doctors to extend tele-health advice to remote areas.

III. System Architecture

The overall architecture of the developed system includes Google Application Engine(GAE) and Google Web Toolkit(GWT) functioning as server .

The real-time assistance ensures information flow to remote areas using internet (GPRS) or via SMS. Mobile application can go beyond restricted information flow adding real values to the information transferred.

Figure 1 shows overall flow chart of the system.

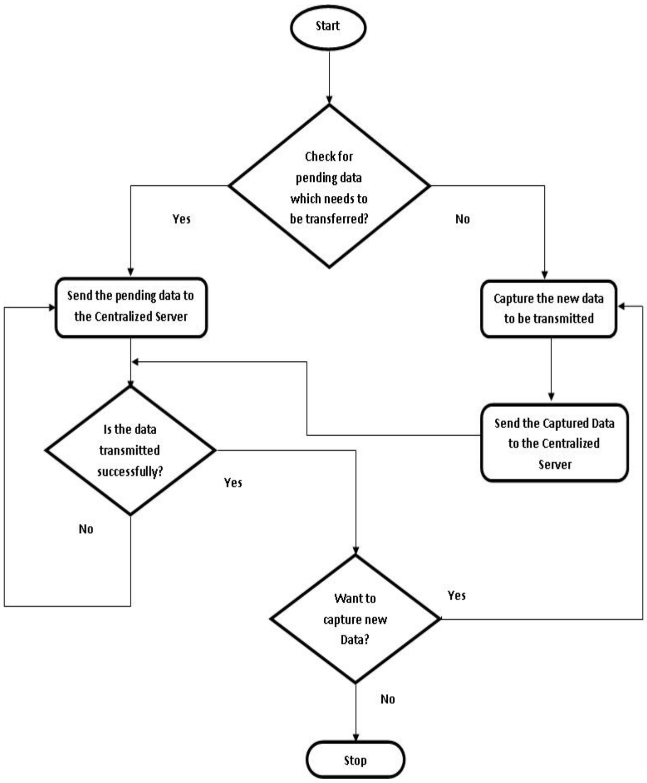

Start
Check for pending data which needs to be transferred?
Yes
No
Send the pending data to the Centralized Server
Capture the new data to be transmitted
Is the data transmitted successfully?
Send the Captured Data to the Centralized Server
Yes
No
Want to capture new Data?
Yes
No
Stop

Figure 1. Overall Flow Chart for Real-time Assistance in Remote

Figure 1 shows the overall flow of the process. Depending on the user's selection, the process will execute.

This flow chart clearly shows the whole system flow of our developed application. As soon as the application starts, it will check for the images, which are needed to be transmitted to the centralized Server. Depending on the user action, the application will capture new data or will transfer the captured data on the centralized Server. It allows the user to capture data repeatedly and stores the data locally even after transmitting the data to the centralized server. After the particular file has been transmitted, the details of the file will be deleted from the spinner so that the user can have the actual list of images, which needs to be transmitted to the centralized server.

A. m-Sahayak System overview for Farmers

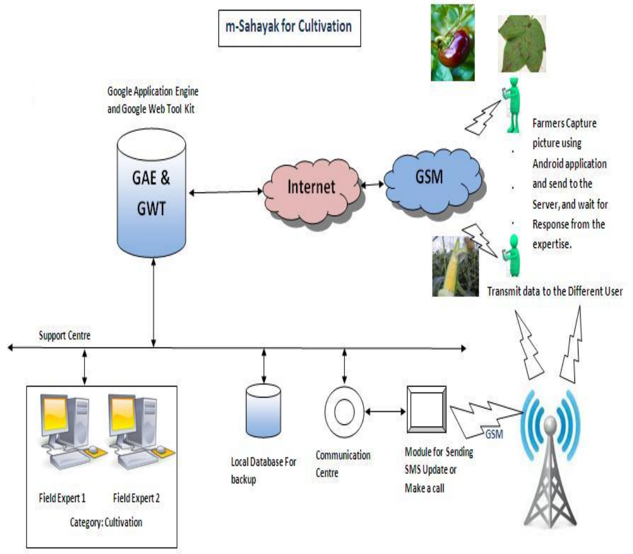

m-Sahayak for Cultivation
Google Application Engine and Google Web Tool Kit
GAE & GWT
Internet
GSM
Farmers Capture picture using Android application and send to the Server, and wait for Response from the expertise.
Transmit data to the Different User
Support Centre
Field Expert 1
Field Expert 2
Category: Cultivation
Local Database For backup
Communication Centre
Module for Sending SMS Update or Make a call
GSM

Figure 2. m-Sahayak System Architecture for Farmers

Figure 2 shows the architecture of m-Sahayak for Indian Farmers. The farmers can capture the picture with Android-based mobile handset with the developed Android application. Within the application, there is a provision to customize the camera settings like pixel resolution, flash, etc. When the application executes, it shows all the supported properties and the user can set them as required [7]. If the GPRS connection is slow, low resolution should be used. All the data (video, audio, image) are sent to the specified Server via HTTP connection using GPRS. On the server side the agricultural scientist is able to see all incoming data and images and provides proper solutions accordingly by call or sms.

B. m-Sahayak System overview for Tele-Health Care

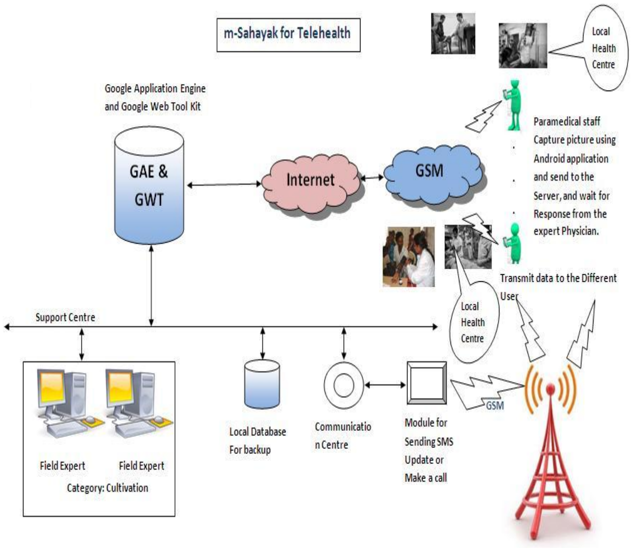

m-Sahayak for Telehealth
Local Health Centre
Google Application Engine and Google Web Tool Kit
GAE & GWT
Internet
GSM
Paramedical staff Capture picture using Android application and send to the Server, and wait for Response from the expert Physician.
Transmit data to the Different User
Local Health Centre
Support Centre
Field Expert
Field Expert
Category: Cultivation
Local Database For backup
Communicatio n Centre
Module for Sending SMS Update or Make a call
GSM

Figure 3. m-Sahayak System Architecture for Telehealth Care

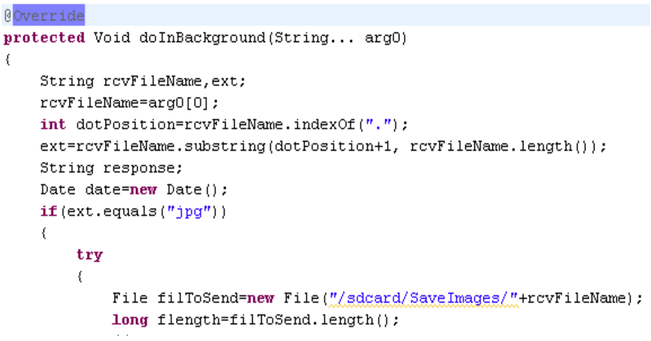
Figure 3 shows the system architecture of m-Sahayak for tele-health service. Even today, in many places in India, there is a local health center but not any physician on 24x7 basis. The paramedical personnel who are normally available in the local health centers can use the developed android-based application [3]. The paramedical staff may capture the picture (photo) and send to the server. On the server side, the corresponding expert doctor could then view the patient's specific zone of the body and give expert guidance to those patients in the remote area.

IV. DESIGN CONCEPT AND SOFTWARE DEVELOPMENT

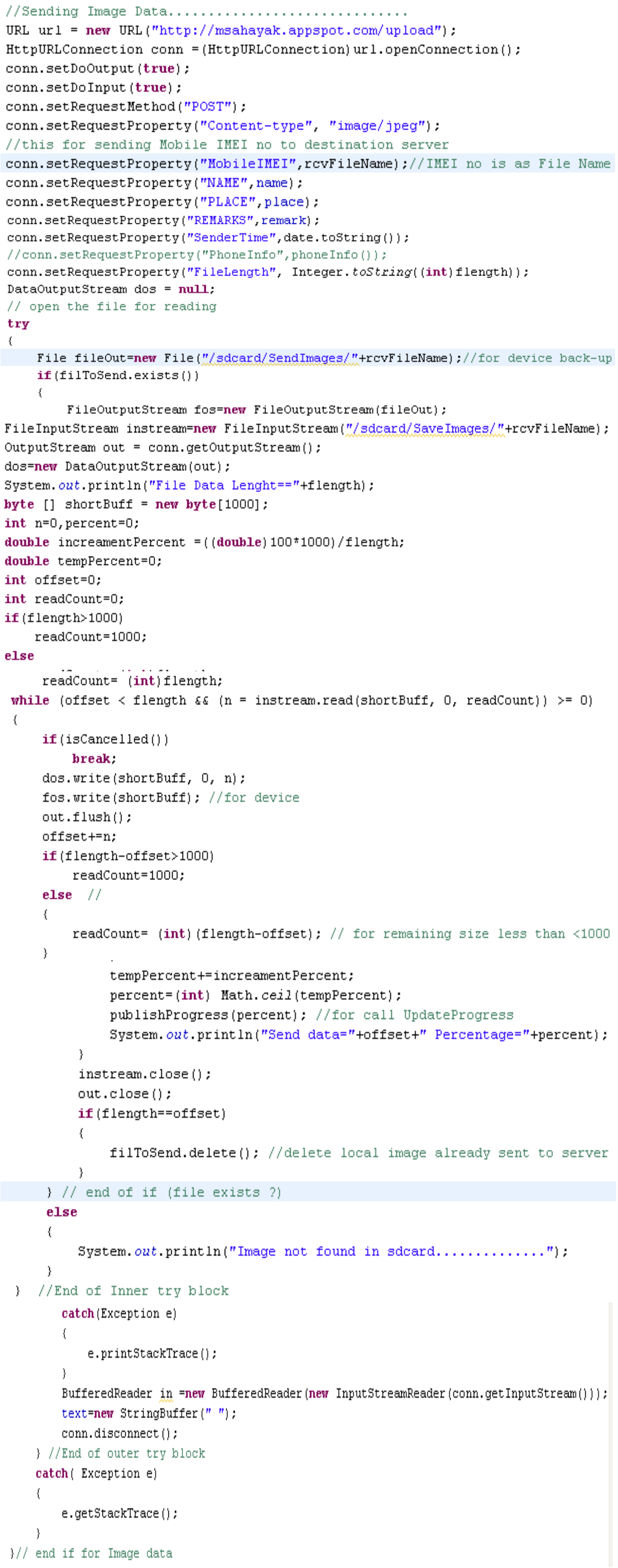
m-Sahayak is based on client-server architecture. At the server side related doctors or scientists can retrieve all the incoming problem data from server. The application has the following technologies and user features [5].

- The application has easy to use Graphical User Interface (GUI) with the capability of providing the information about the image when it will be captured and stored in the mobile device.
- It's Graphical User Interface (GUI) is very simple to use, even the illiterate person can use it very easily. The picture can be captured by pressing a single button and then can be sent by pressing another .
- The data (video, audio and image) are transmitted through GPRS. The connection cost in this case is reduced to a minimum since only those few bytes requested by the user will be downloaded to the mobile phone.
- The data response comes through call or sms. All the information or advices are provided to the user by Call/SMS at 24X7 hours from the experts.
- Security: A secured connection using HTTP protocol would be there to prevent information fraud.
- Durability: In this application, the image will be saved in a directory after it is captured. Till the image is completely sent, the application will not begin to send any other image from any other directory. So there would be no ambiguity so as to which image has been sent and which one has not been sent.

The vital part of the application is the data transfer part. In this mechanism at first, it will connect to the central server through HTTP call and connect to [13]. Some source codes for transferring image data to the server are exemplified as below.



```
@Override
protected Void doInBackground(String... arg0)
{
    String rcvFileName,ext;
    rcvFileName=arg0[0];
    int dotPosition=rcvFileName.indexOf(".");
    ext=rcvFileName.substring(dotPosition+1, rcvFileName.length());
    String response;
    Date date=new Date();
    if(ext.equals("jpg"))
    {
        try
        {
            File filToSend=new File("/sdcard/SaveImages/"+rcvFileName);
            long flength=filToSend.length();
```

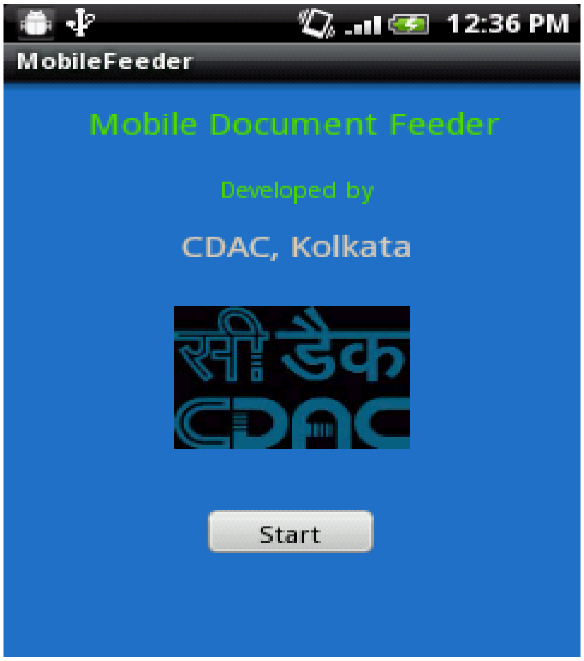


```
//Sending Image Data.............................
URL url = new URL("http://msahayak.appspot.com/upload");
HttpURLConnection conn =(HttpURLConnection)url.openConnection();
conn.setDoOutput(true);
conn.setDoInput(true);
conn.setRequestMethod("POST");
conn.setRequestProperty("Content-type", "image/jpeg");
//this for sending Mobile IMEI no to destination server
conn.setRequestProperty("MobileIMEI",rcvFileName);//IMEI no is as File Name
conn.setRequestProperty("NAME",name);
conn.setRequestProperty("PLACE",place);
conn.setRequestProperty("REMARKS",remark);
conn.setRequestProperty("SenderTime",date.toString());
//conn.setRequestProperty("PhoneInfo",phoneInfo());
conn.setRequestProperty("FileLength", Integer.toString((int)flength));
DataOutputStream dos = null;
// open the file for reading
try
{
    File fileOut=new File("/sdcard/SendImages/"+rcvFileName);//for device back-up
    if(filToSend.exists())
    {
        FileOutputStream fos=new FileOutputStream(fileOut);
FileInputStream instream=new FileInputStream("/sdcard/SaveImages/"+rcvFileName);
OutputStream out = conn.getOutputStream();
dos=new DataOutputStream(out);
System.out.println("File Data Lenght=="+flength);
byte [] shortBuff = new byte[1000];
int n=0,percent=0;
double increamentPercent =((double)100*1000)/flength;
double tempPercent=0;
int offset=0;
int readCount=0;
if(flength>1000)
    readCount=1000;
else
     readCount= (int)flength;
 while (offset < flength && (n = instream.read(shortBuff, 0, readCount)) >= 0)
 {
     if(isCancelled())
         break;
     dos.write(shortBuff, 0, n);
     fos.write(shortBuff); //for device
     out.flush();
     offset+=n;
     if(flength-offset>1000)
         readCount=1000;
     else  //
     {
         readCount= (int)(flength-offset); // for remaining size less than <1000
     }
             tempPercent+=increamentPercent;
             percent=(int) Math.ceil(tempPercent);
             publishProgress(percent); //for call UpdateProgress
             System.out.println("Send data="+offset+" Percentage="+percent);
         }
         instream.close();
         out.close();
         if(flength==offset)
         {
             filToSend.delete(); //delete local image already sent to server
         }
     } // end of if (file exists ?)
     else
     {
         System.out.println("Image not found in sdcard..............");
     }
 }  //End of Inner try block
       catch(Exception e)
       {
           e.printStackTrace();
       }
       BufferedReader in =new BufferedReader(new InputStreamReader(conn.getInputStream()));
       text=new StringBuffer(" ");
       conn.disconnect();
    } //End of outer try block
    catch( Exception e)
    {
       e.getStackTrace();
    }
}// end if for Image data
```

ISBN: 978-1-61208-236-3

V. TESTING , RESULTS AND DEPLOYMENT

The application is tested on HTC Wild Fire Handset and the application is developed for Android 1.5 to 2.2 Version. Some of the screen shots of the application are given in Figure 4, Figure 5 and Figure 6.



Mobile Document Feeder

Developed by

CDAC, Kolkata

Figure 4. Screen Shot for Application start

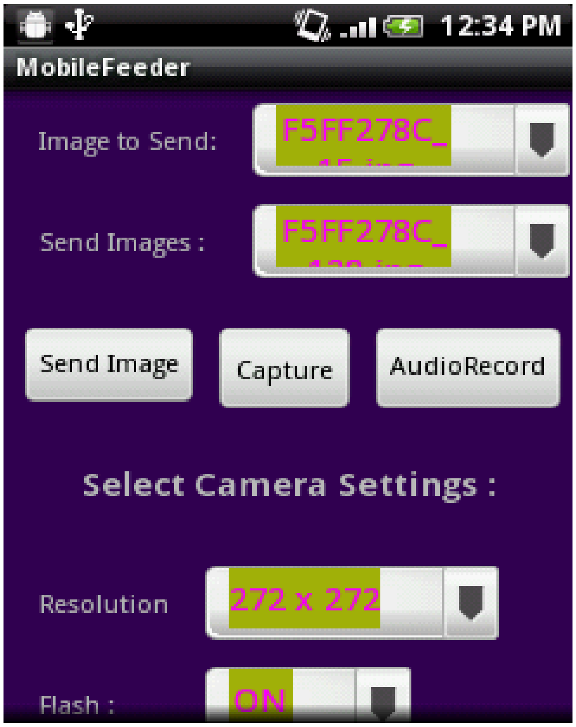


Image to Send: F5FF278C_

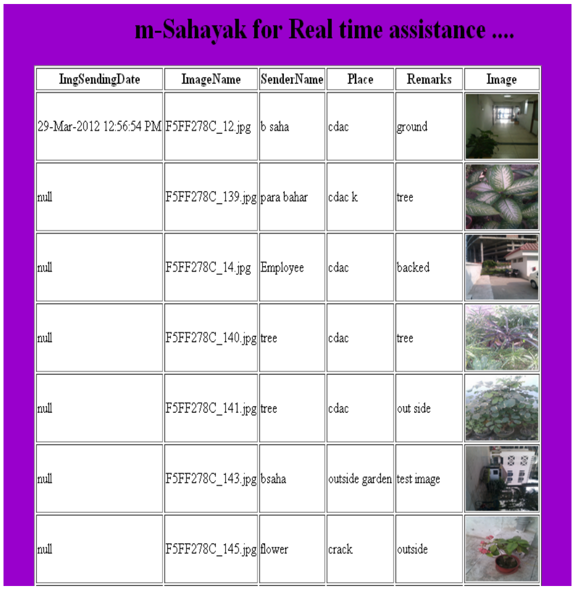
Send Images : F5FF278C_

Select Camera Settings :

Resolution 272 x 272

Flash : ON

Figure 5. Screen shot of main Application layout



m-Sahayak for Real time assistance ....

ImgSendingDate	ImageName	SenderName	Place	Remarks	Image
29-Mar-2012 12:56:54 PM	F5FF278C_12.jpg	b saha	cdac	ground	
null	F5FF278C_139.jpg	para bahar	cdac k	tree	
null	F5FF278C_14.jpg	Employee	cdac	backed	
null	F5FF278C_140.jpg	tree	cdac	tree	
null	F5FF278C_141.jpg	tree	cdac	out side	
null	F5FF278C_143.jpg	bsaha	outside garden	test image	
null	F5FF278C_145.jpg	flower	crack	outside	

Figure 6. Demo application Web Screen using google application engine (GAE) at [13]

Figure 4 shows the initial screen of the Application when it is initiated by the user in the Android mobile set. When the “Start” button (as shown in Figure 4) is pressed the next screen which is provided to the users is about the parameters which needed to be configured (as shown in Figure 5). These parameters are for basic camera settings like resolution, Flash etc. The spinner titled “Image to Send”(as shown in Figure 5) shows the file name which will be transferred when the “Send Image” button (as shown in Figure 5) is pressed by the user. On successful transmission of all the images to the Central Server this spinner list will be empty. This helps the user to keep a track of pending data which needs to be transferred to the Central Server. This happens when no GRPS/Wifi/3G i.e., internet service is not available on the mobile handset and the transmission of data fails. The mobile handset stores the data in its internal memory and transmits it when the service is available. In this way it helps the user from repeatedly capturing the data. The spinner titled “Send Images” (as shown in Figure 5) holds the list of filenames which have been successfully sent to the central server. A button titled “Capture” (as shown in Figure 5) is provided to capture the real time scenario by the user. Another button titled “AudioRecord” (as shown in Figure 5) is provided to record the voice or query of the user along with the captured image. This button causes to open a new screen along with user setting to record the voice and also having a button to send it to the central server. A copy of the data which has been sent is available on the mobile handset also . Figure 6 shows the administrative screen of the m-Sahayak application available on the field experts terminal. On the basis of the available data, a solution is provided to the corresponding users. The service is available to registered users only. The current centralized server is based on google application engine (GAE) and the data which needs to be viewed by the field experts is available over HTTP [13].

ISBN: 978-1-61208-236-3

Testing of the developed application is done on Android 2.2 Emulator. The test result on the emulator comes out with satisfactory outputs. The whole application is broken into 3 main units i.e., the capturing of the image, recording of the voice and sending of the data. When tested individually these units show satisfactory results on an emulator. After the successful testing on an emulator, we have installed the application on the HTC Wildfire handset having Android 2.2. The behavior of the application is satisfactory in the initial stage that has been mentioned. This application is also compatible with higher versions of Android (2.2 and above).

The m-Sahayak prototype has been deployed in the experimental site i.e., a garden near the design lab. Field conditions in this garden were similar to the real farming site. The natural capturing of images from at least 5 types of plants were examined from a distance by agricultural experts. Final deployment in a district of the state of West Bengal, India is under consideration through m-Governance scheme initiated by the Government of West Bengal, India.

VI. Future Works

The proposed architecture is at the initial stage of development whose results have been mentioned. The deployment of the application in the experimental site is successful. Using the current version of the application, the architecture of the systems stands firm.

VII. Conclusion and Remarks

The application m-Sahayak would be a boon to Indian farmers as well as common people in the remote areas. Those who are already using an Android phone, can register with their phone number and get an account. The data sent by them will be stored in the corresponding account. The experts may provide their advice to that particular phone number. By this application, the health care units, may be able to provide better treatment using limited resources to the Indian common people. Using this application, farmers nay control the crop damage and prevent the food problem.

Special features which would popularize the application are

- Simple Graphical User Interface (GUI) that can be used by everybody.
- “One Stop Solution” to all kinds of disease problem in crops and tele health care for people.
- User cost (GPRS) is minimal.

The most common benefit of mobile devices, as found by the survey is its penetration in rural India as the largest basic medium of basic communication. The mobile phone is the only convenient mode of communication to which farmers have access. So it would help the farmers and the rural people if used properly and would be beneficial to most of them.

As far as infrastructure is concerned in India, the Mobile communications services reach to each and every remote place. We have surveyed the current market which shows the basic requirement for running the application is available easily which Indian rural people can afford.

References

[1] Manav Singhal, Kshitij Verma, and Anupam Shukla “Krishi Ville – Android based Solution for Indian agriculture,” Advanced Networks and Telecommunication Systems (ANTS), 2011 IEEE 5th International Conference, pp. 1-5, Dec 2011, Bengaluru, India

[2] Daniel Schuster, Thomas Springer, and Alexander Schill “Service based Development of Mobile Real-time Collaboration Applications for Social Networks”, Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 232-237, 2010 8th IEEE International Conference, March-April 2010.

[3] Octavian Postolache, Pedro S. Girão, Mário Ribeiro, Fernando Santiago, and António Pena “Enabling telecare assessment with pervasive sensing and Android OS smartphone” Medical Measurements and Applications Proceedings (MeMeA), 2011 IEEE International Workshop, pp. 288-293, May 2011.

[4] Ralph Morelli, Emmet Murphy, and Trishan de Lanerolle “An Open Source Mobile App for Assisting Health andAgricu ltural Aid in Haiti” Global Humanitarian Technology Conference (GHTC), 2011 IEEE , pp. 102-107, Oct-Nov 2011

[5] Eko Supriyanto and Head of Advanced Diagnostics and E-Health Research Group ”A Suitable Telehealth Model for Developing Countries” 2011 International Conference on Instrumentation, Communication, Information Technology and Biomedical Engineering, pp. 414, 8-9 November 2011, Bandung, Indonesia

[6] Ivo M. Lopes, Bruno M. Silva, Joel J.P.C. Rodrigues, Jaime Lloret, and Mario L. Proença Jr “A Mobile Health Monitoring Solution for Weight Control” Wireless Communications and Signal Processing (WCSP), 2011 International Conference, pp. 1-5, Nov 2011.

[7] Yu-Hsien Chu, Yen-Chou Hsieh, Chia-Hui Wang, Yu-Chun Pan, and Ray-I Chang “UPHSM: Ubiquitous personal health surveillance andmanag ement system via WSN agent on open sourcesmartphone” e-Health Networking Applications and Services (Healthcom), 13th IEEE International Conference, pp. 60-63, June 2011.

[8] Charalampos Doukas, Thomas Pliakas, and Ilias Maglogiannis “Mobile Healthcare Information Management utilizing Cloud Computing and Android OS ” 32nd Annual International Conference of the IEEE EMBS Buenos Aires, Argentina, pp. 1037-1040, August 31 - September 4, 2010.

[9] Matthias Görs, Roland Marx, Michael Albert, Markus Schäfer, and Klaus Schilling “Tele-Medicine Techniques for Remote Support of Patients in Dialysis and COPD” 2011 International Conference on Instrumentation, Communication, Information Technology and Biomedical Engineering, pp. 23-28, 8-9 November 2011, Bandung, Indonesia

[10] Ben Falchuk “Visual and Interaction Design Themes in Mobile Healthcare”, Mobile and Ubiquitous Systems: Networking & Services, MobiQuitous, 2009. MobiQuitous '09. 6th Annual International, pp. 1-10, July 2009, Telcordia Technol., Piscataway, NJ, USA.

[11] Mei-Ying Wang, John K. Zao, P.H. Tsai, and J.W.S. Liu “Wedjat: A Mobile Phone Based Medicine In-take Reminder and Monitor” 2009 Ninth IEEE International Conference on Bioinformatics and Bioengineering, pp. 423-430, June 2009 Comput. Sci. Dept., Nat. Chiao Tung Univ., Taiwan .

[12] Feming Pang, Linying Jiang, Liu Yang, and Kun Yue “Research of Android Smart Phone Surveillance System” 2010 International Conference On Computer Design And Applications (ICCDA2010), pp. V2-373 - V2-376, June 2010, Software Coll., Northeastern Univ., Shenyang, China.

[13] http://msahayak.appspot.com/upload [retrieved: Aug, 2012].

ISBN: 978-1-61208-236-3

Mobile Gesture Recognition using Hierarchical Recurrent Neural Network with Bidirectional Long Short-Term Memory

Myeong-Chun Lee
Dept. of Computer Science
Yonsei University
Seoul, Korea
lmspring@sclab.yonsei.ac.kr

Sung-Bae Cho
Dept. of Computer Science
Yonsei University
Seoul, Korea
sbcho@cs.yonsei.ac.kr

***Abstract*—As the sensors embedded to a smartphone are proliferating, many application systems for context-aware services are actively investigated. This paper proposes a gesture recognition system with smartphones for better interface. It is important to maintain high accuracy even with the large number of gestures. To improve the accuracy, we adopt the recurrent neural network based on hierarchical BLSTM (Bidirectional Long Short-Term Memory). The first level BLSTMs are used to discriminate the gestures and non-gestures, and the second level BLSTMs classify the input into one of twenty gestures. Experiments with 24,850 sequence data consisting of 11,885 gesture sequences and 12,965 non-gesture sequences confirm the high performance of the proposed method over the competitive alternatives.**

Keywords-mobile interface; gesture recognition; hierarchical neural network; bidirectional recurrent neural network; long short-term memory

I. INTRODUCTION

A variety of sensors such as accelerometer, ambient light, proximity, dual cameras, GPS, dual microphones, compass, and gyroscope are embedded to a smartphone. They are not only sophisticated, but also show good performance [1]. Especially, accelerometer is one of the most commonly used sensors for the physical movements of the user carrying the phone. For this reason, several user interfaces with gesture and activity recognitions have been developed by using the accelerometer [2].

However, there are two crucial problems to develop user recognition systems with the smartphone sensors. One is to identify the non-gesture or non-activity data. The gesture or activity data includes meaningful and non-meaningful parts. Sometimes the amount of non-meaningful data can be even more than gesture data. In this case, it is time-consuming to recognize both meaningful and non-meaningful data. The other is to maintain high-accuracy even with the large number of classes. Many pattern recognition systems depend on machine learning methods to learn the complex patterns. However, the performance degrades when classifying a large number of classes. It is an important problem when providing the various services to users.

In this paper, we propose a mobile gesture recognition system where data is collected from the accelerometer sensors embedded to a smartphone. Figure 1 shows example of a gesture data set. To alleviate the aforementioned problems, a recurrent neural network based on BLSTM (Bidirectional Long Short-Term Memory) is used to classify twenty classes. To discriminate the gestures and non-gestures, hierarchical structure is exploited. In experiments, the proposed method outperforms the standard BLSTM.

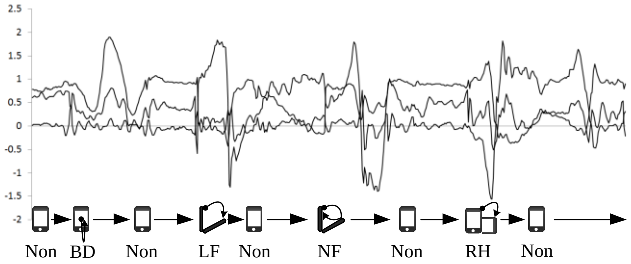


Figure 1. Example of a gesture data set

The paper is organized as follows. Section II presents the related works for gesture recognition. Section III describes the overall architecture, BRNN (Bidirectional Recurrent Neural Network), and LSTM (Long Short-Term Memory) respectively. Section IV shows the experimental results and we conclude with some remarks in Section V.

II. RELATED WORKS

The related works using accelerometer is shown in Table I.

TABLE I. GESTURE RECOGNITION WORKS USING ACCELEROMETER

Author	Classifier	No. of gestures	Collector	Overview
J. Liu *et al.* (2009) [3]	DTW	8	Wiimote	Personalized gesture using uWave
G. Niezen *et al.* (2009) [4]	HMM, ANN, DTW	8	Smartphone	comparison of classifier algorithms
J.–K. Min *et al.* (2010) [5]	DTW, NB, K-means	20	Smartphone	DTW model selection through NB
T. Marasovic *et al.* (2011) [6]	K-NN	7	Smartphone	Combination of the PCA and K-NN
A. Akl *et al.* (2011) [7]	DTW, AP	18	Wiimote	Dimensional reduction through RP

The success of the gesture recognition is subject to which classification method is used, how many gestures are used, and what kind of collector is used for collecting the data. As can be seen in Table I, most of the research use static algorithms such as MLP (Multi-Layer Perceptron), k-means clustering and combination of DTW (Dynamic Time Warping) and other methods. However it is more suitable to use the dynamic classification algorithms directly for the time-series data. For this reason, we adopt a dynamic classification algorithm which is a kind of recurrent neural network. BLSTM shows better performance than other algorithms for recognizing the time-series patterns in several domains. F. Eyben *et al.* proposed an audiovisual approach and LSTM for recognizing conversational speech. From the experiments, they showed that the LSTM outperformed SVM (Support Vector Machine) [8]. T. Thireou *et al.* applied BLSTM to the sequence-based prediction of protein localization, and showed that the proposed method is better than FFNN (Feed Forward Neural Network) and BRNN[9].

III. Architecture and Method

A. *The overall system architecture*

This paper aims at enhancing the accuracy by using hierarchical structure. The entire system configuration is shown in Figure 2. The accelerometer data collected from a smartphone are segmented by using sliding window and average variation. The preprocessed data is hierarchically classified after training. We adopt the recurrent neural network based on BLSTM[10], which is a hybridization of BRNN[11] and LSTM[12]. First, training data are used to classify the gestures and non-gestures. Second, classified gesture data are used for classifying the twenty gesture classes.

B. *Bidirectional recurrent neural network*

The basic idea of BRNN is to present each training sequence forwards and backwards to the two separate recurrent hidden layers, both of which are connected to the same output layer. This provides the network with past and future context for every point in the input sequence. Figure 3 shows a structure of unfolded bidirectional recurrent neural network over two time steps. BRNN shows better performance than other approaches such as regression and classification experiments [11].

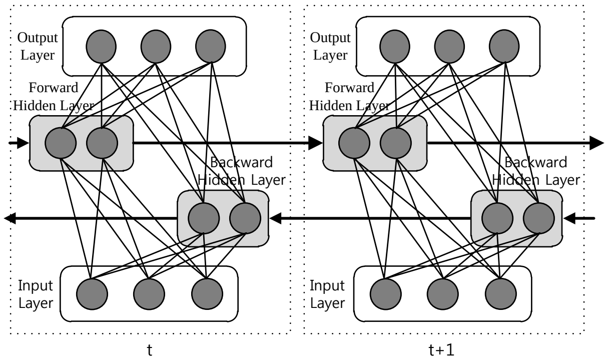


Figure 3. The structure of bidirectional recurrent neural network

C. *Long Short-Term Memory*

LSTM is an extension of the recurrent neural network. It uses the three gates that can store and access the data collected from rest of the network. Gates are activated from logistic sigmoid activation function. Figure 4 shows a memory block of LSTM. The Hyperbolic tangent activation function is used for squashing functions. The basic calculation of each gate is the same as the standard artificial neural network [12].

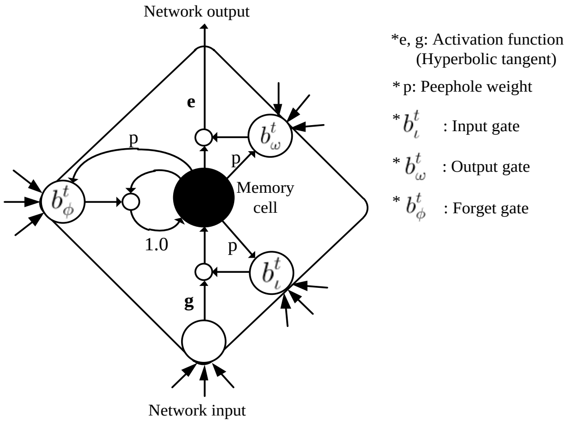

Network output
e
p
p
Memory cell
1.0
p
g
Network input
b_ϕ^t
b_ω^t
b_ι^t
*e, g: Activation function (Hyperbolic tangent)
* p: Peephole weight
* b_ι^t : Input gate
* b_ω^t : Output gate
* b_ϕ^t : Forget gate

Figure 4. A LSTM memory cell

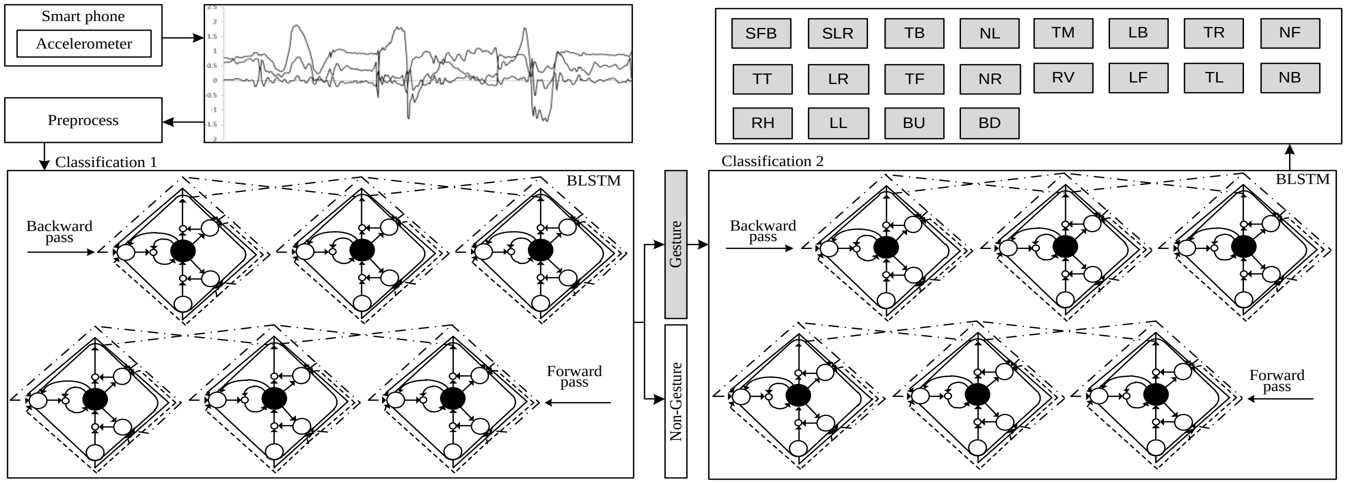


Figure 2. The overall system architecture

The input gate determines whether the input values put the memory cell or not.

$$\alpha_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota}\beta_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1} \qquad (1)$$

$$\beta_\iota^t = f(\alpha_\iota^t)$$

where, α_ι^t is a state of the input gate at time *t*. It is calculated from input values, the output of other networks, and state of the memory cell. *I*, *H*, and *C* mean the number of input node, hidden node, and cell, respectively. *w* is the weight of connected nodes. f is the logistic sigmoid function to activate the input gate.

The output gate determines whether the information is output or not. The calculation of the output gate is similar with the input gate.

$$\alpha_\Phi^t = \sum_{i=1}^{I} w_{i\Phi} x_i^t + \sum_{h=1}^{H} w_{h\Phi}\beta_h^{t-1} + \sum_{c=1}^{C} w_{c\Phi} s_c^{t-1} \qquad (2)$$

$$\beta_\Phi^t = f(\alpha_\Phi^t)$$

where, α_Φ^t is a state of the forget gate at time *t* and β_Φ^t is a state after applying the activation function.

Equation (3) is a calculation that is generated by forget gate, the state of cell, the state of input gate, and state of α_c^t after applying hyperbolic tangent activation function. Note that the fixed weight value 1.0 is used for preserving the information in the memory cell.

$$\alpha_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc}\beta_h^{t-1} \qquad (3)$$

$$S_c^t = \beta_\Phi^t S_c^{t-1} + \beta_\iota^t g(\alpha_c^t)$$

Forget gate provides the information to reset the memory cell.

$$\alpha_\omega^t = \sum_{i=1}^{I} w_{i\omega} \alpha x_i^t + \sum_{h=1}^{H} w_{h\omega}\beta_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^{t-1} \qquad (4)$$

$$\beta_\omega^t = f(\alpha_\omega^t)$$

where, ω is the output gate and β_ω^t is the state of an output gate after applying the activation function at time *t*.

Equation (5) is a definition of the cell output. To activate the cells, hyperbolic tangent is used and multiply the state of output gate.

$$\beta_c^t = \beta_\omega^t e(s_c^t) \qquad (5)$$

For training the LSTM recurrent neural network, we use the Back Propagation Through Time (BPTT).

IV. EXPERIMENT

A. *Data preparation*

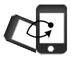
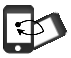
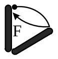
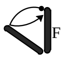
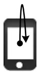
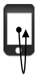
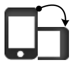
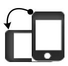
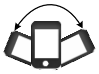
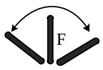
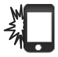
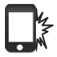
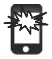
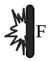
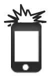
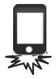
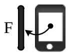
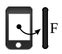
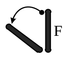
For the experiment, Samsung Omnia smartphone with MS Windows Mobile 6.1 was used as the platform. The acceleration data are sampled at 50Hz. 30 people of 10~60 years old participate in the experiment. The collected data is divided into generations and date. Total amount of the data consists of 11,885 gesture sequences and 12,965 non-gesture sequences. The number of files used in the experiment is 1075.

Table II shows the detailed description of twenty gestures. Rotating and tilting hold their physical states after the movement. Tapping represents the hand or finger stroke on a smartphone surface. In the case of shaking, subjects shake the devices two or more times in a specific direction. Snapping has an angular acceleration while bouncing moves straightly to a direction and reflected back where both are the kind of pendulum movement. We set learning rate and momentum as 0.0001 and 0.9 respectively. BPTT which is one of the dynamic learning algorithms is used for training the gesture data.

B. *The results*

For the first experiment, the data are divided into a ratio of seven to three as training and test data, respectively. 17,470 sequences are used for training, and 7,380 sequences are used for test. Each sequence is distributed randomly. The results through all experiments in this work are compared with the standard BLSTM. The accuracy rate for the first experiment is shown in Figure 5. The average accuracy of Hierarchical BLSTM is 91.15%, whereas the standard BLSTM is 89.20%. As can be seen in Figure 5, the hierarchical BLSTM generally outperforms the standard BLSTM.

TABLE II. DESCRIPTION OF THE GESTURE DATA

Symbol	**NL**	**NR**	**NF**	**NB**	**BU**	**BD**	**RH**	**RV**	**SLR**	**SLF**
Meaning	Snapping				Bouncing		Rotating		Shaking	
Direction	Left	Right	Forward	Backward	Up	Down	Horizontal	Vertical	Left-Right	F-Backward
Movement										
Symbol	**TL**	**TR**	**TF**	**TB**	**TT**	**TM**	**LL**	**LR**	**LF**	**LB**
Meaning	Tapping						Tilting			
Direction	Left	Right	Forward	Backward	Top	Bottom	Left	Right	Forward	Backward
Movement										

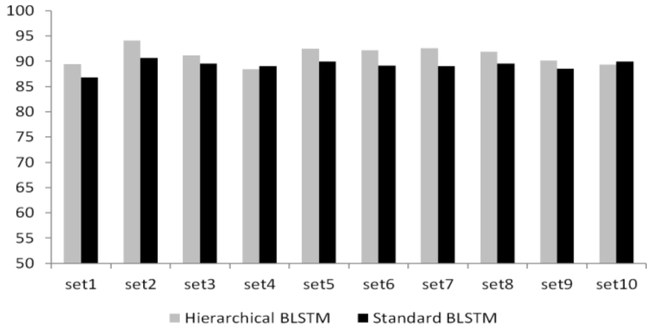


Figure 5. The Results for randomly distributed data

For the second experiment, we group the data as generations. 18,490 sequences are used for training and about 2,100 sequences are used for testing at the each generation. Table III also shows the hierarchical BLSTM outperforms the standard BLSTM in most cases.

TABLE III. GENERATION RESULTS

		Set1	Set2	Set3	Set4	Set5	Avg.
10~20	Hierarchical BLSTM	90.13	90.5	88.33	90.1	86.1	89.032
	Standard BLSTM	88.5	89.9	85.8	86.3	84.9	87.08
20~40	Hierarchical BLSTM	94.81	97.15	94.4	96.9	96.23	95.898
	Standard BLSTM	95.32	96.69	93.73	96.9	95.79	95.686
40~60	Hierarchical BLSTM	82.54	86.5	88.43	89.21	85.8	86.496
	Standard BLSTM	79.1	86.7	85.2	87.6	85.8	84.88

To get the fair comparison, we conducted ten-fold cross validation test. The raw data are randomly partitioned into ten subsamples. Of the ten subsamples, a single subsample is retained for the validation data for testing and remaining nine subsamples are used for training data. This process is repeated ten times. The results of the hierarchical BLSTM except the set3 and set8 are more accurate than the standard BLSTM.

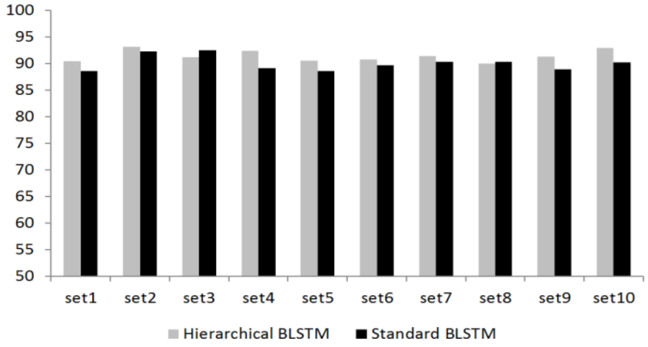


Figure 6. 10-fold cross validation result

V. CONCLUSION AND FUTURE WORK

In this paper, we collect the accelerometer data from a smartphone and classify the data by using hierarchical BLSTM. Since the gesture data contain a lot of non-gesture data, we classify the non-gesture sequences before classifying the meaningful sequences. More than 20,000 sequences were used for reliable experiment and total classes of the data were twenty one including non-gesture data. The performance of the standard BLSTM was compared with the hierarchical BLSTM and our approach outperformed the standard BLSTM. For the future work, it can be possible to achieve higher accuracy if the data are grouped with the similar meaning because some gestures have similar characteristics.

ACKNOWLEDGMENT

This research was supported by the Original Technology Research Program for Brain Science through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2010-0018948).

REFERENCES

[1] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Cambell, “A survey of mobile phone sensing,” *IEEE Communications Magazine*, vol. 48, no. 9 pp. 140-150, 2010.

[2] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles. T. Choudhury, and A. T. Campbell“A. survey of mobile phone sensing,” *Communications Magazine*, vol. 48, no. 9, pp. 140-150, 2010.

[3] J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, and V. Vasudevan, “uWave: accelerometer-based personalized gesture recognition and its applications,” *IEEE Int. Conf. on Pervasive Computing and Communications*, pp. 1-9. 2009.

[4] G. Niezen and G. P. Hancke, “ Evaluating and optimising accelerometer-based gesture recognition techniques for mobile devices,” *AFRICON*, pp.1-6, 2009.

[5] J.-K. Min, B.-W. Choe, and S.-B. Cho, “ A selective template matching algorithm for short and intuitive gesture UI of accelerometer-builtin mobile phones, ” *Cong. on Nature and Biologically Inspired Computing*, pp. 660-665, 2010.

[6] T. Marasovic and V. Papic, “ Accelerometer-Based Gesture Classification Using Principal Component Analysis,” *Int. Conf. on Software, Telecommunications and Computer Networks*, pp. 1-5, 2011.

[7] A. Akl, C. Feng, and S. Valaee, “ A novel accelerometer-based gesture recognition system,” *IEEE Trans. on Signal Processing*, vol. 59, no. 12, pp. 6197-6205, 2011.

[8] F. Eyben, S. Petridis, B. Schuller, G. Tzimiropilos, S. Zafeiriou, and M. Pantic. “Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks,” *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 5844-5847, 2011.

[9] T. Thireou and M. Reczko, “Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins,” *IEEE/ACM Trans. on Computational Biology and Bioformatics*, vol. 4, no. 3, pp. 441-446, 2007.

[10] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, “A novel connectionist system for unconstrained handwriting recognition,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[11] M. Schuster, “Bidirectional recurrent neural network,” *IEEE Trans. On Signal Processing*, vol. 45, no. 11, pp.2673-2681, 1997.

[12] S. Hochreiter and J. Schmidhuer, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

Business Impact Assessment for Autonomic Network Management

Identifying Business Impact Indicators From Use Case Requirements

Sander Spek, Vânia Gonçalves, Simon Delaere
IBBT-SMIT, Vrije Universiteit Brussel
Pleinlaan 2, 1050 Brussels, Belgium
{alexander.spek, vania.goncalves, simon.delaere}@vub.ac.be

***Abstract*—In this paper, a methodology is presented to identify business impact indicators from the use case requirements in an early phase of the development stage. It is based on the Quality Function Deployment method and the business model design framework. This methodology is applied to the design for a unified management framework to facilitate autonomics in complex, ubiquitous and large-scale networks. The analysis concludes that this innovation will be particularly disruptive in the functional architecture and value proposition aspects of the business model.**

Keywords—autonomics; self-x; business impact; business model

I. INTRODUCTION

Determining business impacts of systems in pre-commercial development is a challenge that requires both a clear view of technical results and limitations of the system in development as well as a grasp of business insights. Technical objectives are often explicated in an early stage of the development by means of functional, non-functional and business requirements. While business requirements might already give some indication on the eventual business impacts of the system once implemented they are often developed by technical partners and thus limited in scope. Considering business requirements in an early stage is important to anticipate on the market impact of the innovation and possible conflicts between actors. However, it is far from straightforward to relate the functional and non-functional requirements to business impacts.

This paper explores an endeavor to link the identified requirements of autonomic management systems to business impacts by creating a matrix inspired by the Quality Function Deployment (QFD) method (see e.g., [1–5]), which allows for the mapping of individual requirements to business model design parameters. This is applied in the early design phase of the UniverSelf project [6][7], a collaborative project with over 15 partners, aiming to develop a management framework for autonomics in existing and emerging network architectures.

In the next section, the details of the UniverSelf framework to be developed will be explained in more detail, including the six use cases that form the basis for the services' requirements. Section III describes the methodology used, including an elaboration on its two building blocks: the Quality Function Deployment and the business model design framework. The results will be presented in Section IV, while a discussion on the findings takes place in Section V. Finally, the conclusion and future work is presented in Section VI.

II. UNIFIED MANAGEMENT FRAMEWORK

The goal of the UniverSelf project is to overcome "the growing management complexity of future networking systems, and to reduce the barriers that complexity and ossification pose to further growth" [6][7]. It designs a Unified Management Framework (UMF) in order to enable autonomic principles in current and emerging networking architectures. This framework should constitute a cross-technology, common substrate for both systems and services, and include the necessary functions to achieve self-management in the autonomic network, its systems and its network equipment. The project partners have formulated six use cases, which are considered to be representative and complementary, reflecting the network operator's desires to reduce its costs caused by complexity and reducing the dependency on human operation for operational tasks. This desire includes both the reduction of operational costs (OPEX) as well as network equipment and infrastructures (CAPEX). These use cases are labeled as follows:[1]

1. *Self-Diagnosis/Healing for IMS VoIP & VPN:* self-diagnosis and healing features with applications for IP networks and IMS services as well as VPN networks.
2. *Networks' Stability and Performance:* simulation and emulation results about stability and performance of a network (with a great number of nodes and real impairments) with cross-layer and cross-domain self-configuration mechanisms.
3. *Dynamic Virtualization and Migration of Contents and Servers:* the dynamic virtualization and migration of data/content and network entities (gateways and servers) nearer to users.
4. *SON and SON collaboration according to operator policies:* to design novel SON to improve network operation and performance, and to demonstrate the operation of a mobile network empowered by SON entities within a general management framework.
5. *Operator-governed, end-to-end, autonomic, joint network and service management:* to enable

[1] For historic reasons, use case 5 and 6 are referred to as use case 6 and 7 respectively in the UniverSelf project. For optimal clarity, we use a consecutive numbering in this paper.

operators to describe their goals and objectives, through high-level means and govern their network, to achieve policy-based operation of Radio Access Network and backhaul/core network segments, and to achieve coherence between these segments through cooperation, negotiation and federation.

6. *Network and Service Governance:* facilitates network and service governance through the use of IPTV services running on top of both fixed and mobile networks.

These use cases have been formulated from a technical perspective, complete with functional, non-functional and business requirements, about 200 in total. While these requirements might represent valid technical goals, one also needs to assess their feasibility and viability from a business perspective. Since the systems are complex, the networks are heterogeneous and the stakeholders are multiple, this is not a straightforward exercise. In this paper an endeavor is made, based on the requirements from the use cases. While the cost reduction is the main aim of the framework and its use cases, this analysis is an effort to discover secondary impacts that without this analysis might be overlooked.

III. METHODOLOGY

The methodology is aimed at deriving business-impact indicators from the requirements that have been developed by technical partners at an early stage of the design. Two analytical frameworks are taken as the basis. The Quality Function Deployment provides a method of analyzing requirements to derive characteristics and controls from them, although in the original QFD the results are not business impacts. The business model design framework, on the other hand, provides clear business design choices for development in telecommunications and other systems; but, so far has had no connection to the technical requirements outlined at an early stage in a development project. The last subsection contains our synthesized method that takes elements from both.

A. Quality Function Deployment

Quality Function Deployment is a method developed in the late 1960s by Mizuno and Akao. It can be defined as "an over-all concept that provides a means of translating customer requirements into the appropriate technical requirements for each stage of product development and production (i.e., marketing strategies, planning, product design and engineering, prototype evaluation, production process development, production, sales)" [8] (via [4]). The underlying idea is that (potential) customers have valuable input for the design of the product, but that cannot or will not express this in a technical terminology [5]. The total QFD process is described in what is called the *House of Quality*, see Figure 1, but the essence is its four-phase approach, consisting of the following steps: (1) *Product planning,* transforming customer demands into quality characteristics; (2) *Product design,* transforming quality characteristics in product characteristics; (3) *Process planning,* transforming product characteristics into a manufacturing process; and (4) *Process control,* transforming the manufacturing process into quality controls. These phases are presented visually in Figure 2.

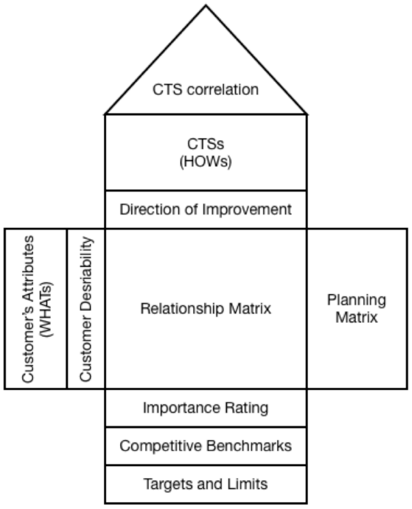


Figure 1. The House of Quality. (Taken from [4].)

Interesting about the four phases of QFD is that they all provide a transformation of certain requirements or characteristics of one kind into requirements or characteristics of a different kind by using a scorecard matrix. Typically this is done by placing both input and outcome on one of the axes, and to add scores to the intersections. These scores can be weights (e.g., to indicate importance) or categories (e.g., strong relationship (9), medium relationship (3), weak relationship (1); example taken from [1]).

B. Business model design framework

The business model design framework is a model developed by Ballon (see a.o., [9–10]). It "follows the multi-parameter approach by defining four levels on which business models operate, and by identifying three critical design parameters on each level" [9]. These refer to a whole ecosystem for a product or service rather than to a specific organization within that ecosystem. In Ballon's view, the essence of a business model is the (re)configuration of

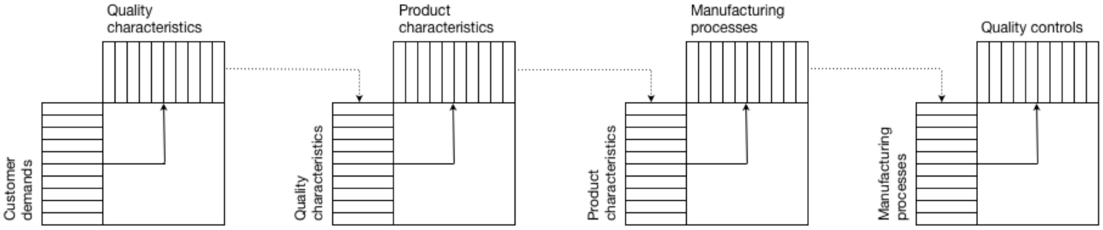


Figure 2. The QFD four-phase approach. Based on [1].

TABLE I. THE BUSINESS MODEL DESIGN PARAMETERS BY BALLON [9–11].

Control parameters		**Value parameters**	
A. Value network parameters	***B. Functional architecture parameters***	***C. Financial model parameters***	***D. Value proposition parameters***
A1. Combination of assets	B1. Modularity	C1. Cost(-sharing) model	D1. Positioning
A2. Vertical integration	B2. Distribution of intelligence	C2. Revenue model	D2. User involvement
A3. Customer ownership	B3. Interoperability	C3. Revenue-sharing model	D3. Intended value

control and value [9][10]. The value reflects the traditional view of a business model as an elaboration of a value proposition, while the control aspect raises attention to the questions on who controls the value network and the system design. As displayed in Table I, the framework reflects this by being divided into two parts: control parameters and value parameters. The control parameters consist of three value network parameters and three functional architecture parameters. The value parameters consist of three financial model parameters and three value proposition parameters. This business model approach based on the (re)configuration of two elements, each providing a more detailed set of parameters, dissolves the disadvantages of both two-parameter schemes and multi-parameter approaches as they existed during the development of this approach. [10]

The design framework consists of two sets of three configuration parameters each. For the control parameters, these are value network parameters and functional architecture parameters. For the value parameters, these are financial model parameters and value proposition parameters. The parameters will be used as indicators of business impacts in this analysis of the requirements in a system design.

C. Synthesized method

A matrix has been created for scoring the functional, non-functional and business requirements to the business model design parameters. Since the requirements per use case vary, this exercise is performed for every use case separately. The scoring entails that a point is given in case the requirement impacts the said design parameter. When counting the points per design parameter, one can gain insights into the importance of that parameter given the requirements of a specific use case. It might feel as a self-fulfilling prophecy to also score the business requirements on the business parameters to determine their business impact. This, however, is justified because the same persons who developed the use cases and also extracted the other requirements have formulated them, and did so from a technical perspective. In other words, they only concern direct business effects of the technology, and do not yet make an analysis of the business consequences, hence the use of them as inputs in the matrix.

In the QFD-methodology, one can give a score to a field in the matrix to quantify the impact it has. In this exercise, only a binary assessment is used; as 197 requirements were scored on 12 parameters, thus creating a 2364 field matrix, this has provided sufficient granularity for the purpose of this method.

A comparable method has been developed in [11], which analyzes the business impacts for mobile self-organizing networks. Technical parameters have been identified on the basis of technical functionalities and key performance indicators (KPIs). The business model design framework was operationalized in the form of business parameters, which were then linked to the technical parameters to create a scorecard of business impacts. Other efforts in extracting business impacts from technical documentation also exist. For instance, Raju *et al.* [12] utilize a framework based on seven business model parameters extracted from the business model design framework, specifically adapted to energy aware self-growing business ecosystems. The method in this paper distinguishes itself on several aspects. First of all, it starts from the requirements, a document that is a part of many development tracks already. Also, it arrives at the business model design framework, which is a suitable and objective general framework for further analysis. As a final argument, the scoring of the use cases is a semi-formalized process. Several persons can perform it in parallel, after which the results can be easily compared and discrepancies can be discussed. In the present case, this also happened.

IV. RESULTS

After scoring the requirements of the six use cases on the business parameters, a matrix is created that can be interpreted generally, as well as aggregated in two directions.

TABLE II. RESULTS OF THE SCORING

	A. Value Network			**B. Functional Architecture**			**C. Financial Model**			**D. Value Proposition**		
	A1	***A2***	***A3***	***B1***	***B2***	***B3***	***C1***	***C2***	***C3***	***D1***	***D2***	***D3***
Use case 1	1	1	1	5	17	20	4	1	1	0	9	11
Use case 2	0	0	0	3	7	2	2	1	0	1	7	7
Use case 3	2	1	0	2	9	8	4	2	0	2	1	8
Use case 4	0	0	0	1	4	6	3	2	0	2	4	4
Use case 5	5	2	0	14	11	11	9	3	0	2	3	13
Use case 6	1	0	0	6	12	2	4	4	0	1	8	8
Total	**9**	**4**	**1**	**31**	**60**	**49**	**26**	**13**	**1**	**6**	**28**	**47**

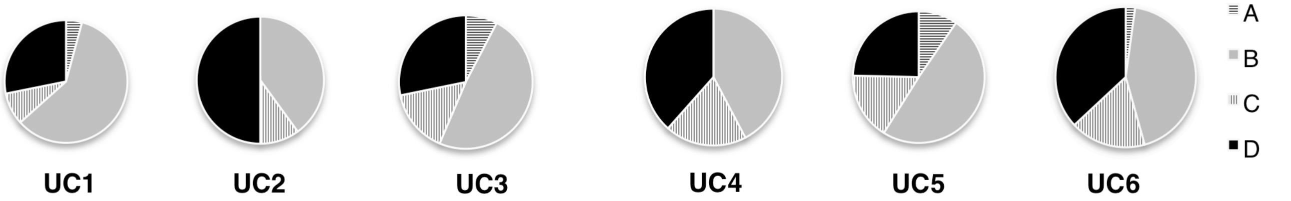


Figure 3. Results of the scoring per use case.

First, the results per use case will be considered, to see if the method gives significantly different outcomes for these different cases. Second, a reverse approach is taken, where the scores are aggregated per business parameter, in order to analyze for single business parameters which of the use cases have a large impact on it.

A. *General results*

Many requirements have scored on multiple business parameters, while some did not have an impact on any of the parameters. For the 197 requirements, in total 275 scores have been given. The results are displayed in Table II. The individual requirements have been grouped per use case. The business parameters that the requirements have been scored on can be found in the top. One can observe that the system, as designed in this stage, will especially have an impact in sets B and D of the business parameters, namely the functional architecture and the value proposition. It seems less obtrusive in terms of value network and financial model. The biggest emphasis will be on the distribution of intelligence (B2), followed by interoperability (B3) and intended value (D3). Also the modularity (B1) and the user involvement (D2) are stressed. These results can be interpreted as an indication that the service as currently designed puts great emphasis on its functional architecture. Even though distribution of intelligence receives the greatest attention, it scores high on all three parameters. The emphasis on distribution of intelligence results from the aim to build contextual knowledge and monitoring information from both control and data planes gathered from elements distributed over core, distribution and access network levels, which are essential for the orchestration of autonomic functionalities. Since the management framework targets the integration between current and future networks and interfaces with NMSs and OSSs, interoperability also stands out. In terms of value proposition, the service will have a high impact on the user involvement and the intended value. This emphasis on user involvement is a reflection of the efforts to reduce human (operator) involvement, as tasks will be taken over by the autonomic network mechanisms and elements. The system is thus clearly designed to facilitate the reduction of OPEX, although the limited scoring on the cost model parameter indicates that the requirements do not point directly to lowering OPEX; it is often implicit. The emphasis on intended value is a reflection of the requirements focused on improving performance and stability through self-x functionalities, thereby enhancing Quality of Service (QoS) and Quality of Experience (QoE) for end-users.

B. *Results per use case*

The results per use case are displayed in Figure 3. What becomes immediately clear is that there are significant differences between the different use cases. Especially use case 2 seems to stand out, as it is the only use case that has the biggest emphasis in the value proposition parameters. Both use case 2 and 4 seem to have no impact on the value network parameters, which means that the innovations in these use cases are not likely to alter or disrupt the current value network configuration. Looking at the goals of the use cases, this seems right. Use case 1, *Self-Diagnosis/Healing for IMS VoIP & VPN*, is a technical use case, so it should not come as a surprise that the requirements reflect this by putting most emphasis on the functional-architecture parameters. Similarly, use case 2, *Networks' Stability and Performance*, focuses on the quality of the service to the end-user, and as expected it scores high on the value proposition parameters. Use case 6, *Network and Service Governance*, relies on both the functional architecture and value proposition parameters, and hardly has any impact on the value network and financial model parameters. This can be explained by the fact that this use case is of a supportive nature, improving the functionality and the value proposition of the service by setting the orientations for a governance tool that will contribute to the translation of human high-level service business goals into network policies.

To study this example in more depth, Figure 4 contains the breakdown of use case 6 again, but this time up to the level of the individual parameters. For clarity, the value network and financial model parameters have been blanked out, to focus on the functional architecture (B) and value proposition parameters (D). What becomes clear is that the impact in this section should be mostly attributed to B1 and B2, modularity and distribution of intelligence, as they make up the major share of the solid-line bordered slice of the pie. The impact of the third functional architecture parameter, interoperability, is limited in this use case. In the value proposition parameter set, one can see a similar effect, with the impact being attributed to D2 and D3, user involvement and intended value. There is little impact related to the positioning of the product or service in this use case. One can make such a detailed description for each of the use cases, but space restrictions prevent doing so in this paper.

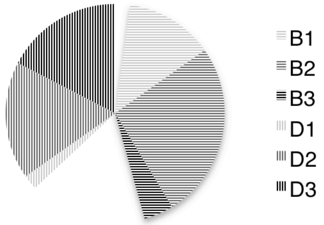


Figure 4. Results of use case 6, specified to individual parameters. Parameter sets A and C have been blanked out in the chart.

C. *Results per business parameter*

The previous analysis can be reversed: taking the design parameters as a basis and study the matrix to determine which of the use cases contribute the most or the least to a specific parameter. These results can be derived from Table II by looking at the columns rather than the rows. Examples are given in Figure 5, with parameters broken down into the use-case contributions. The main contributor for D2, *User Involvement*, is use case 1, with use case 6 and 2 following closely. The interpretation is that the decreasing human operator involvement will be largely determined by the result of the first use case, which deals with self-diagnosis and self-healing. The same reasoning also applies to use case 2 and 6, albeit with a slightly smaller impact. From the use case descriptions, one could also have expected use case 4 to score high here, but judging on the requirements the user involvement impact is present but limited. The parameter *Modularity* shows a different picture. Here it becomes clear that especially use case 1, 5 and 6 contribute to the weight of modularity—an indication that issues concerning integration, interfacing and standardization might occur in these use cases. This is an indication that issues concerning the syntax and semantics of external manifestation can become of high importance in this use case. The results do not provide any information on the specificities of these issues, but they do stress that there is a significant impact of these use cases on the business configuration concerning modularity — something that can serve as valuable input for further analysis. The last two use cases deal mostly with the management and governance of the system, so the stress on modularity issues is a plausible result. The large impact of use case 1 is more surprising, but is largely justified by this use case's aim to integrate several sources of information (service and network alarms, configurations, messages, performance indicators, etc.) spread across different network elements, service components and NMS which will be the primary source for self-healing and diagnosis decisions.

V. DISCUSSION

In this section, a discussion on the methodology will be given as well as a discussion on the results.

A. *Discussion on the methodology*

This paper provides work in progress on a novel method to identify business impacts in an early stage of the product or service design process by the use of analyzing the requirements drawn out by technical partners. The analysis of the results in this paper shows that in many cases the results of this exercise are plausible, as one can explain them from a business analysis. However, since this is a new type of analysis one should be careful for a bias, e.g., a bias of the technical partners creating the requirements, or a bias of the analyst(s) scoring the requirements on the business model design parameters. These biases cannot be excluded entirely, but can be contained by involving multiple parties in the exercise. Moreover, the value of the exercise is not so much in the exact numbers it produces, but rather in the indications that it gives. The results should not be interpreted rigorously, but should rather serve as an input for business analysis. It provides early indicators on where the design puts stress on certain business impacts that might turn into bottlenecks or require special attention later on during development. It could raise awareness on a multitude of issues that need to be tackled before implementation or commercial deployment: issues concerning backward compatibility, standardization, trust, the value network, and etcetera.

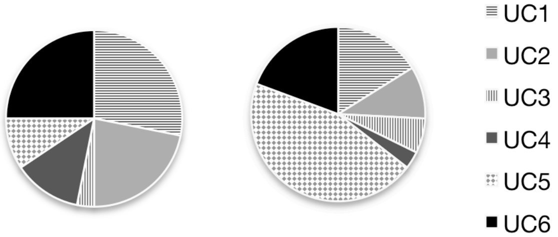


Figure 5. D2, User involvement (left), and B1, Modularity (right), broken down to the contributions of the different use cases.

Another bias might be in the method itself. The results indicate a strong emphasis on the functional architecture parameters. This could be a characteristic of this specific service, but it could also be a natural bias, since functional requirements make up a large part of the total set of requirements and they might favor the functional architecture more than the other sets of configuration parameters. By applying this method in more projects, such a bias could be discovered. A solution could then be to include an extra weighing factor to mute the discovered biases.

The weighing factors of the different types of requirements can also be considered. In this exercise, all types of requirements have been treated equal, despite there being 131 functional requirements and only 31 and 35 non-functional respective business requirements. Especially the business requirements already contain a first indication of the possible business impacts of the use cases, so one could consider weighing them in more than the other requirements.

B. *Discussion on the results*

The type of analysis as performed in this paper does not have a clear impact assessment as an outcome. It is rather an explorative exercise that provides valuable input for a qualitative analysis to follow. It explores the technical specificities from the requirements and indicates where they could have a business impact. It can be valuable input to discover impacts a straightforward analysis could have missed, and it provides some insights on which impacts are the heaviest, and which use cases specifically have impacts on a certain business parameter.

The main aim of the framework is a cost reduction, particularly in the form of OPEX. The use cases have been developed with this in mind, but this is not reflected in the results of the analysis. Most of the studied use cases are facilitating this higher goal, but do not refer directly to it. Therefore, the results should be interpreted as those business impacts that exist next to the cost reduction and that might facilitate this reduction.

The value network parameters are expected to have a relatively small impact. This means that the use cases, as currently formulated, will not be very disruptive in terms of the value network, although the requirements are very low level, and higher level disruptions could be discovered from other types of analysis. Therefore, there is not much impact on assets belonging to certain roles, on vertical integration and disintegration, and on customer ownership. Also the financial model is relatively stable. However, these results do not take into account the actual business scenarios, and the different value configurations that can arise. The matrix indicates some issues concerning the cost(-sharing) model and the revenue model, but these are less significant, or at least more straightforward, compared to other disruptions. Considering potential business scenarios (e.g., network virtualisation, network infrastructure sharing), the introduction of the management framework could however impact the value network and the split of costs and revenues.

The value proposition is subject to change. While positioning is stable, there is a high impact on user involvement and intended value. This is a reflection of the decreasing human involvement caused by autonomic configuration and interventions and the emphasis these use cases put on network performance and stability towards an improved QoS and QoE.

Finally, the functional architecture is the area with the highest business impact. All three parameters are impacted, which is partly a reflection of the design choice to create a modular, distributed system with autonomic mechanisms. The analysis recognizes this, and stresses the consequences this might have for the business model. It puts an emphasis on automatic collaboration and standardization of policy languages and interfaces. One consequence of this is that the roles responsible in the value network must have a good understanding between each other and must have a sufficient amount of trust. In networks where such an understanding and trust is not guaranteed, this design choice could cause severe problems when the system or service gets deployed. In such a case, a misalignment of business model configurations between the designed framework and the actual market, either one of the two has to be adapted.

VI. Conclusion and Future Work

A methodology has been proposed to use a QFD-inspired matrix to derive potential business impacts from use case requirements. The outcome is a set of impact indicators belonging to parameters of the business model design framework, which can serve as the basis of a qualitative impact analysis. This methodology is applied to the six use cases of a design for a unified management framework in order to enable autonomic principles in current and emerging networking architectures. Cost reduction in the form of OPEX and CAPEX was the main objective for these use cases, but with this methodology one can find other impacts — either facilitating the cost reduction or unrelated to it.

The model indicates the highest impacts in the parameters relating to the functional architecture and the value proposition. The first is caused by the distributed design of the framework, which has major business impact consequences. The value proposition is impacted by the changes in user involvement due to automation, and the changes in the intended value due to improvements in the QoS and QoE. As it turns out, the designed framework already imposes a certain business model configuration. This analysis provides a first insight in whether this configuration matches the actual market situation or that adaptations need to be made.

The methodology thus provided a promising starting point for further analysis. It highlights the most important parameters to be considered in a following qualitative study. However, the methodology needs to be applied to more designs in order to study and contain possible biases. Also, practical guidelines about validations of the matrix by multiple actors need to be developed.

Acknowledgment

The research leading to these results has been performed within the UniverSelf project (www.UniverSelf-project.eu) and received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 257513.

References

[1] A. T. Bahill and W. L. Chapman, “A tutorial on quality function deployment,” Engineering Management Journal, vol. 5, no. 3, 1993, pp. 24–35.

[2] C. P. M. Govers, “What and how about quality function deployment (QFD),” International Journal of Production Economics, vol. 46, 1996, pp. 575–585.

[3] Y. Akao, “QFD: Past, present, and future,” in International Symposium on QFD, 1997, vol. 97, pp. 1–12.

[4] L. K. Chan and M. L. Wu, “Quality function deployment: A literature review,” European Journal of Operational Research, vol. 143, no. 3, 2002, pp. 463–497.

[5] K. Yang and B. S. El-Haik, “Chapter 7: Quality Function Deployment (QFD),” in Design for Six Sigma: a roadmap for product development, Second Edition. Kindle ed., McGraw-Hill Professional, 2008.

[6] Univerself Project, “UniverSelf, realizing autonomics for Future Networks,” Oct-2010. [Online]. Available: http://cordis.europa.eu/fp7/ict/future-networks/documents/projects-univerself-presentation_en.pdf.

[7] Univerself Project, “About UniverSelf,” 2010. [Online]. Available: http://www.univerself-project.eu/about-univerself.

[8] L. P. Sullivan, “Quality Function Deployment,” Quality Progress, vol. 19, no. 6, 1986, pp. 39–50.

[9] P. Ballon, “Business modelling revisited: the configuration of control and value,” info, vol. 9, no. 5, 2007, pp. 6–19.

[10] P. Ballon, “Business Modelling as the Configuration of Control and Value,” in Proceedings of the 20th Bled eConference eMergence: Merging and Emerging Technologies, Processes, and Institutions, Bled, Slovenia, 2007.

[11] V. Gonçalves and S. Delaere, “Business Impact Assessment of Mobile Self-Organising Networks,” in New Frontiers in Dynamic Spectrum, 2010 IEEE Symposium on, 2010, pp. 1–12.

[12] A. Raju, S. Lindmark, O. Yaron, E. De Poorter, L. Tytgat, S. Delaere, and P. Ballon, “Business Model Assessment of Green Wireless Sensor Ecosystems,” in Proceedings of CTTE’12.

Permission Tracking in Android

Michael Kern
Micro Focus
Borland Software Corporation
Linz, Austria
michael.kern@microfocus.com

Johannes Sametinger
Dept. of Business Informatics – Software Engineering
Johannes Kepler University
Linz, Austria
johannes.sametinger@jku.at

***Abstract*—Mobile devices get smarter and increasingly provide access to sensitive data. Smart phones and tablet computers present detailed contact information, e-mail messages, appointments, and much more. Users often install apps on their devices to get additional functionality like games, or access to social networks. Too often, such apps access sensitive data and take privacy less serious than expected by users. In this paper, we will have a closer look at permissions that users grant to apps in Android, a wide-spread operating system for mobile devices like smart phones. As it turns out, Android does not provide sufficient control to their users about what apps are allowed to do. We demonstrate the feasibility of a permission tracking functionality, but conclude that thorough modifications in Android itself will be necessary to provide satisfying control of apps' permissions and users' privacy.**

Keywords-Android, mobile devices, privacy, permissions, tracking.

I. INTRODUCTION

Almost everything that users can access on their desktop can also be accessed on their mobile devices, in particular on smart phones. These devices have become powerful with capabilities that desktops did not even have several years ago. Increased capabilities have come hand in hand with increased security threats. Nowadays, all our private and business data is accessible on our phones. Thus, these devices increasingly have become the targets of malicious attacks. The most successful and most widespread operating system of today's mobile devices is Android, with Apple's iOS following second [1]. Needless to say, the operating system plays a key role for the security and for the privacy of these devices. Permissions play a major role for apps on Android. Giving permissions to apps can lead to data leaks because it may, for example, allow these apps to access contact data and to access the Internet. A malicious app may thus send all our contact data to a server on the Internet. Fine-grained permission setting is not (yet) possible in Android.

In this paper, we will introduce Android and its security features. We will then suggest a mechanism to track permissions in Android. For that purpose, we have developed an Android application that supports permission assignment, permission tracking and permission notifications. In Section II, we will provide an overview of the Android operating system. Section III follows with an overview of security mechanisms of Android. In Section IV, we will provide more details of Android permissions. The *PermissionTracker* tool will be presented in Section V. Implementation aspects follow in Section VI. A comparison with other approaches is given in Section VII. Performance issues are discussed in Section VIII. Eventually, a conclusion ends the paper in Section IX.

II. ANDROID

Android is a Linux-based operating system for mobile devices like smart phones and tablet computers. It is developed by the Open Handset Alliance led by Google. Android apps can be downloaded from online stores like Google's app store *Google Play* (formerly *Android Market*) and also from third-party sites.

Google leads the Android Open Source Project (AOSP) with the goal "to create a successful real-world product that improves the mobile experience for end users" [2]. Since its original release, there have been many Android updates, each of which fixed bugs and added new features. The main Android building blocks are device hardware, the Android operating system, and the Android application runtime [2]. Android supports a wide range of hardware configurations, e.g., smart phones, tablets, or set-top-boxes. Even though Android is processor-agnostic, it does take advantage of some hardware-specific features, e.g., security capabilities such as the no-execute page protection of the ARM architecture.

The core system is built on top of the Linux kernel. All device resources, like camera, GPS, Bluetooth, telephony, network connections, are accessed through the operating system. Most Android applications are written in the Java programming language. Core Android services and applications are native applications or include native libraries. The virtual machine and native applications run within the same security environment, contained within the application sandbox. Applications get a dedicated part of the file system in which they can write private data. Android applications are either pre-installed or user-installed. Pre-installed applications like phone, email, calendar, web browser, and contacts provide key device capabilities that can be accessed by other applications. Development of user-installed applications is supported by an open development environment.

Google also provides cloud-based services for Android devices. They include services to let users discover, install, and purchase applications, update services, and application services that, for example, easily let application developers backup user data in the cloud.

ISBN: 978-1-61208-236-3

III. Android Security

The Android platform security architecture provides several key security features in an effort to protect user data as well as system resources, and to isolate applications from each other. To achieve these goals, Android provides a sandbox for all applications, secure inter-process communication, and application signing, among others. User-granted permissions at the application level are also part of the security architecture. They are central for the content of this paper and, thus, will be described separately in Section IV. In the following subsections, we describe basic security risks and Android security mechanisms.

A. *Security Risks*

Security risks originate from various sources, e.g., from the operating system or from communication mechanisms. Android consists of a Linux kernel and various open source libraries implemented in C/C++. The effect of open source to security is being discussed contradictory. C/C++ is known to be more prone to vulnerabilities than other languages like Java or C#. A static security analysis of version 2.2 has revealed several hundred errors [3]. Almost 100 of these were rated critical, leading to crashes or constituting potential vulnerabilities.

Smart phones have numerous wireless interfaces like WLAN, Bluetooth, NFC, and of course GSM and UMTS or 3G. If confidential information is transmitted by the user or by an app, then attackers can easily listen in with a simple antenna. Attackers having direct access to a stolen or lost device are seen as one of the biggest threat. Data is not encrypted by default. Devices are used for critical services like online banking or virtual private networks (VPNs) to access sensible information.

Unlike Apple, Google does not impose control mechanisms for apps to be published on the market. If users report about malicious activities, then apps get removed from the market. Until that time, users are exposed to these malicious apps. Unaware users may also deactivate security functions countertrading enhanced usability. They may leave their device unattended or they may permanently turn on services like Bluetooth or WLAN. They may also install apps from un-trusted sources.

B. *Sandbox*

Android applications run in a sandbox, i.e., an isolated area of the operating system with no access to the rest of the system's resources. Access is granted only when users warrant explicit access permissions during application installation. Before installation, all the required permissions are displayed.

C. *Inter-process Communication*

Android processes can communicate with any of the traditional UNIX-type mechanisms, e.g., file system, sockets, signals. There are additional Android-specific mechanisms for inter-process communication. Google recommends that Android developers use these mechanisms, i.e., binders, services, intents, and content providers [3].

Binders are lightweight remote procedure call mechanisms that are designed for high performance for in-process and cross-process calls. Android services run in the background. They can run in their own process or in the context of another application's process. Services can provide interfaces that are directly accessible through binders. Intents are simple message objects that represent the intention to do something. For example, if an application wants to display a web page, it creates this intent to view a specific URL and hands it off to the system. The system locates the browser that knows how to handle that intent and runs it. Intents can also be used to broadcast system-wide events, e.g., notifications. A data storehouse provides access to data on Android devices, e.g., the user's list of contacts. Applications can access data that other applications have exposed via a content provider. Applications can also define their own content provider and expose data of their own.

Android developers are encouraged to use best practices to secure users' data and avoid the introduction of security vulnerabilities [4].

D. *Application Signing*

Android applications must be signed by their developers. Code signing allows users to identify the authors of applications and to update applications without having to deal with permissions again. Unsigned applications that attempt to install will be rejected by either the Google app market or by the package installer on the Android device. In addition to Android’s built-in security features, users may use software by various vendors. Solutions are available for tasks like access control, data encryption, traffic counting, anti-theft (device locking, device location, data wiping), and malware protection. Well-known vendors for PC security solutions like BitDefender or Kaspersky also offer products for mobile devices with the Android operating system.

IV. Android Permissions

A game may, for example, need to activate vibration but should not need to read messages or access contact information. After reviewing the permissions, users can decide whether to install an application [4]. Protected resources include camera, location data (GPS), Bluetooth, telephony, SMS/MMS, and network/data connections. Granted permissions are applied to applications as long as they are installed. Android’s permissions are some form of Mandatory Access Control, or MAC for short. In contrast to DAC which stands for Discretionary Access Control, access is not controlled by users or by user ids, but rather by permission labels that are assigned system functions. Accessing a resource requires the call of system functions. If an application wants access to a resource, it needs the permissions required by the appropriate system functions; see Figure 1 [4].

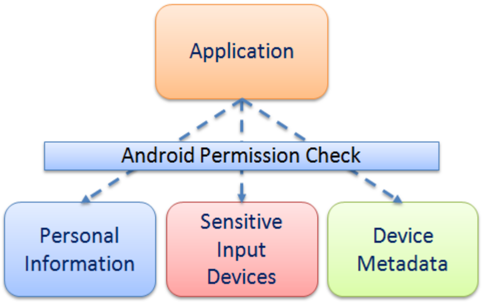


Figure 1. Access to sensitive data through protected APIs [4].

Protected APIs include [4]:

- Camera functions
- Location data (GPS)
- Bluetooth functions
- Telephony functions
- SMS/MMS functions
- Network/data connections

Required permissions of applications are stored in their manifest file; see Figure 2. The application described in the manifest of Figure 2 needs to send and receive SMS messages, as well as access to the user's contact list. During installation the user gets informed about the permissions that are requested by an application. A dialog will show that, for example, the application requests access to services that may cost money, access to the phone's location, to network communication, to account information, and to storage. We can trust the application and install it or we can cancel the installation process as a whole. There is no way in between like installing the application but denying access to account information. The package installer is the single point of interaction with the user; no further checks with the user are done while an application is running.

The Android operating system defines over one hundred different permissions. They are called standard permissions and distinguish various functions of the protected APIs mentioned above, for example, reading contacts, writing contacts, sending SMS, receiving SMS.

```
<manifest package="com.example.myapp">
    <application />
    ...
    <uses-permission android:name=
      "android.permission.SEND_SMS" />
    <uses-permission android:name=
      "android.permission.RECEIVE_SMS" />
    <uses-permission android:name=
      "android.permission.READ_CONTACTS" />
    …
</manifest>
```

Figure 2. Permission declaration in manifest.

A. *Application-defined permissions*

Permissions may also be defined by applications. Thus, developers can restrict access to their applications, i.e., to their activities, services, broadcast receivers, and content providers. The declaration of these application-defined permissions is again written in the application's manifest file. Figure 3 shows an example permission with the name "my.permission.MY_PERM". Other application may request this permission during installation and, thus, get access to activities and services of our sample application that are protected with this permission.

B. *Protection level*

The protection level specified in Figure 3 characterizes the potential risk that is implied in the permission [5]. This level indicates the procedure the system should follow when determining whether or not to grant the permission to an application requesting it. The value can be set to either *Normal*, *Dangerous*, *Signature*, or *SignatureOrSystem* [5]. *Normal* is the default value. Access is granted to isolated application-level features, with minimal risk to other applications, the system, or the user. Access is automatically granted to requesting applications. The protection level *Dangerous* provides access to private user data or control over device. It introduces a potential risk and, therefore, is not automatically granted to requesting applications. If *Signature* is specified, access is granted if a requesting application is signed with the same certificate as the application that declared the permission. The level *SignatureOrSystem* grants access to applications that are in the Android system image or that are signed with the same certificates as those in the system image. This level is used for certain special situations.

C. *Drawbacks*

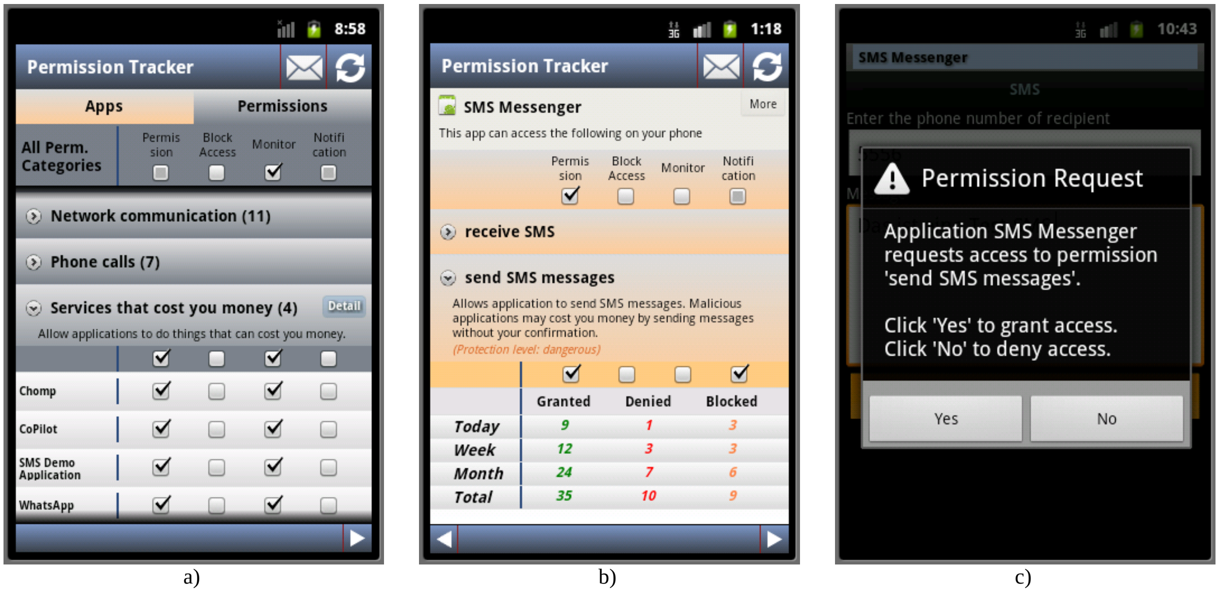
There are some drawbacks of the Android permission system.

1) *Static permissions*

Android's permission system is rather rigid and lacks flexibility. Users can only install applications by granting all permissions requested by that application. It is not possible to withdraw any permission, neither during installation nor after the installation process. The only option users have is to uninstall an application.

```
<manifest package="com.example.myapp">
    <application />
    ...
    <permission
      android:name="my.permission.MY_PERM"
      android:protectionLevel="normal"
      android:label="@string/myPerm_Label"
      android:description=
        "@string/myPerm_Description">
    </permission>
    …
</manifest>
```

Figure 3. Declaration of an application-defined permission.



All Perm. Categories	Permis sion	Block Access	Monitor	Notifi cation
	□	□	☑	□

Network communication (11)

Phone calls (7)

Services that cost you money (4) Detail

Allow applications to do things that can cost you money.

	☑	□	☑	□
Chomp	☑	□	☑	□
CoPilot	☑	□	☑	□
SMS Demo Application	☑	□	☑	□
WhatsApp	☑	□	☑	□

a)

SMS Messenger

This app can access the following on your phone

Permis sion	Block Access	Monitor	Notifi cation
☑	□	□	□

receive SMS

send SMS messages

Allows application to send SMS messages. Malicious applications may cost you money by sending messages without your confirmation.
(Protection level: dangerous)

	☑	□	□	☑

	Granted	Denied	Blocked
Today	9	1	3
Week	12	3	3
Month	24	7	6
Total	35	10	9

b)

Enter the phone number of recipient

Permission Request

Application SMS Messenger requests access to permission 'send SMS messages'.

Click 'Yes' to grant access.
Click 'No' to deny access.

c)

Figure 4. a) PermissionTracker showing permissions' applications. b) Granted, denied and blocked permissions. c) Permission request

2) Missing control

Users have no control over their resources. Once an application has been installed, it can access resources with the permissions that have been granted during installation. Users cannot neither watch which resources an application accesses, nor can they permit or deny any such access.

3) Over-privileged applications

Applications sometimes are over-privileged, which means they require access to resources they do not need to function. Over-privileged applications increase the impact of vulnerabilities.

4) Permission granularity

Some standard permissions are defined at a coarse granularity, e.g., INTERNET, WRITE_EXTERNAL_STORAGE. Applications with the permission INTERNET have arbitrary access to the Internet. There is no way to restrict access for example to specific domains or services.

5) Permissions across applications

Applications that have been signed with the same certificate and have the same user id can share their permissions. Shared user ids may grant permission to applications without explicitly declaring them in the application's manifest file. Applications may also combine their permissions. For example, application A may have access to the user's contact data but no access to the Internet. Application B may have access to the Internet but no access to the user's private data. This inoffensive situation may become threatening when application A hands over sensitive data to application B which in turn may send it to a server on the Internet.

V. Permission Tracker

We have developed an Android application that allows users to administer permissions of their applications. We have extended the existing permission concept and enable users to allow or deny permissions at any time. Additionally, we facilitate the observation of application's access to resources.

The *PermissionTracker* tool provides three consistent views with different levels of detail. Users can view application categories and inspect the permissions of single applications or groups of applications. Users can alternatively view permission categories and inspect applications with specific permissions or permissions in specific permission groups. At any time, users can modify the permissions of applications or groups of applications. Check boxes are available and can be selected individually for granting or denying permissions, for blocking access to resources, for monitoring access to resources, and for sending notifications when a resource is being accessed. If users block access to a resource, they will be asked for approval every time an application wants to access the resource. If access to a specific resource by a specific application or a group of applications is monitored, then statistics with information about resource access will be available.

The *PermissionTracker* also allows a detailed view with information about specific permissions, including the protection level, see Figure 4a. We can also see in Figure 4b the number of granted, denied and blocked access for the "send SMS messages" permission. The table in the bottom of the figure shows the numbers for today, for this week, this month, and the total number. These numbers get collected only if monitoring has been activated.

Notifications inform users immediately when an application requests access to a specific resource. Figure 4c shows the dialog that appears when the user opens a notification. We can see that the application SMS Messenger requests the permission to send SMS messages. We can either grant or deny this permission. If the user does not react to a notification, for example, because the phone has been left home, then the dialog will close automatically after five minutes. In this case, access will be denied.

PermissionTracker generates several reports to ease an analysis of applications and permissions. The reports are created in HTML format and either shown in the browser or sent to an email address. Reports include permissions of applications, the permissions' status at the time of access, as well as date and time of access, see Figure 5 for an example.

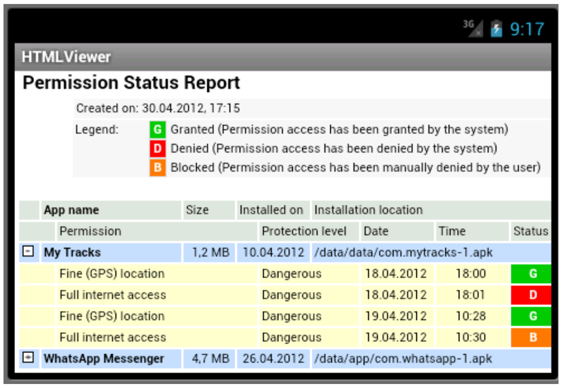
VI. Implementation Aspects

In Section IV, we have described Android's permission mechanism. The regular Android permissions do not support the implementation of an application like the *PermissionTracker*. Therefore, a few adaptations had to be made in the Android system. Subsequent sections describe these modifications.

A. Android

Android developers use an application framework that serves as a layer between applications and the mobile device. These Android APIs, i.e., short for application programming interfaces, support the creation of GUI elements, data repositories, data communication, etc. The manager classes get started during initialization of Android and run in separate threads. Figure 6 shows part of these manager classes. The first two boxes border the package manager service and the activity manager service. The activity manager interacts with the overall activities running in the system. It is responsible to consider permissions when components of applications interact among each other or access components of Android. The package manager retrieves various kinds of information related to packages of currently installed applications on the device. It stores the permissions that applications request during their installation and provides functions to application developers to retrieve information, for example, about these permissions. In order to implement the functionality of our *PermissionTracker* application, we had to add two functions to this manager, i.e., a block permission dialog and a permission notification, see the bottom box in Figure 6.

User settings about permission tracking are stored in a file app_configuration.xml in the directory of the *PermissionTracker*. As mentioned above, these user settings contain information about whether a user wants to individually grant or deny access to a resource, whether notifications should be sent upon access to a resource, or whether access should get logged.



3G 9:17

HTMLViewer

Permission Status Report

Created on: 30.04.2012, 17:15

Legend: G Granted (Permission access has been granted by the system)
D Denied (Permission access has been denied by the system)
B Blocked (Permission access has been manually denied by the user)

	App name	Size	Installed on	Installation location	
	Permission	Protection level	Date	Time	Status
⊟	My Tracks	1,2 MB	10.04.2012	/data/data/com.mytracks-1.apk	
	Fine (GPS) location	Dangerous	18.04.2012	18:00	G
	Full internet access	Dangerous	18.04.2012	18:01	D
	Fine (GPS) location	Dangerous	19.04.2012	10:28	G
	Full internet access	Dangerous	19.04.2012	10:30	B
⊞	WhatsApp Messenger	4,7 MB	26.04.2012	/data/app/com.whatsapp-1.apk	

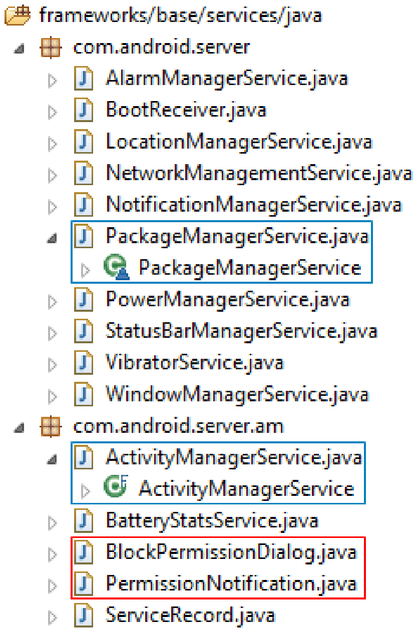
Figure 5. Permisson status report.

The Android Activity Manager interacts with the overall activities running in the system. For example, it returns attributes of the device configuration or information about the memory usage of running processes [6]. Our extension to the Activity Manager checks the permissions and performs monitoring and sending notifications if necessary. The Android Package Manager can be used to retrieve information about application packages that are installed on the device [7]. This information includes permissions that applications have assigned to. However, the package manager does not contain any methods that allow the modifications of permissions. This is not necessary as current Android implementations assign these permissions during the installation of an application. Later modifications have not been planned so far. The implementation of our *PermissionTracker* needs such methods. Therefore we have simply added them, i.e., methods to grant, to revoke and to log permissions.

B. Installation

Due to the extensions made to Android, the *PermissionTracker* cannot be installed on a system without these extensions. Care has to be taken to use a device where an original Android version is installed. If the installed Android contains extensions of the device's manufacturer, these extensions would be overwritten when installing the new Android version for the *PermissionTracker*. For details about how to install an Android build see [8].

A device's operating system that comes when buying it is called stock ROM. A custom ROM is a version of Android that includes the kernel, apps, and services, i.e., everything



```
frameworks/base/services/java
  com.android.server
    AlarmManagerService.java
    BootReceiver.java
    LocationManagerService.java
    NetworkManagementService.java
    NotificationManagerService.java
    PackageManagerService.java
      PackageManagerService
    PowerManagerService.java
    StatusBarManagerService.java
    VibratorService.java
    WindowManagerService.java
  com.android.server.am
    ActivityManagerService.java
      ActivityManagerService
    BatteryStatsService.java
    BlockPermissionDialog.java
    PermissionNotification.java
    ServiceRecord.java
```

Figure 6. Android's manager classes.

that is needed to operate the device. As the name suggests, this version has been customized by someone in some way. These custom ROMs can be installed via the recovery console. There are some downsides to custom ROMs. For example, when installing a new version with a custom ROM, then all existing data is lost. Thus, installation a new device may be easy, but doing such an install at a later point of time may be more sophisticated [9][10].

VII. Related Work

Weaknesses of the Android permission system have been known for a while. Concepts and tools have been suggested for improvements. We will start with extensions to Android, continue with applications and conclude with a comparison to our approach.

A. Android Extensions

APEX stands for Android Permission Extension. It is a framework that allows users to specify runtime constraints to restrict applications' access to resources. Users can specify constraints through a simple interface of the extended Android installer. The extensions are incorporated in Android with little changes of the code and the user interface [11]. Kirin is an alternate application installer and security framework to Android. It ensures that applications meet predefined security requirements. Applications are rejected, if any specified requirements are not met. Devices stay in a secure state without users having to make any security decisions. Challenges include the modeling of security mechanisms and the acquiring of appropriate policy primitives [12].

SAINT stands for Secure Application Interaction, a mechanism that allows application developers to define install-time and run-time constraints. Policies can be specified by application developers, but not by users [13]. MockDroid is also a modified version of Android. It allows users to 'mock' applications' access to resources. Mocking means that resources are reported as empty or unavailable whenever an application requests access. In contrast to Android's static permission system, users can revoke access to specific resources at run-time [14]. XMandroid stands for eXtended Monitoring on Android. It performs runtime monitoring and analysis of communication across applications in order to prevent potentially malicious control and data flow based on a defined policy [15].

TISSA stands for Taming Information-Stealing Smartphone Applications. It is a system that implements a privacy mode that empowers users to flexibly control application's access to personal information in a fine-grained manner. Granted access can be dynamically adjusted at run-time [16].

B. Applications

There are several applications available that help in getting an overview about the permissions that installed applications request. Some of these applications also allow to grant or to deny access to specific resources. aSpotCat is an ad-supported application that eases the process of finding out the permissions that specific applications have. It also provides lists of applications that use specific resources. For example, which applications use GPS or SMS? [20]

PermissionDog is an application that lists applications and checks the permissions they are using. Based on these permissions, PermissionDog rates the potential danger of applications. If wanted, every time an application is launched, a notification pops up and shows the number of used permissions and the potential danger of the application [21]. PermissionsDenied is an application that lists the amount of active and disabled permissions of each application on the device. Permissions can be enabled and disabled. Any changes require a reboot of the device in order to become effective. Brief descriptions of permissions can also be shown [17][22].

LBE Privacy Guard features a back-ground service that constantly monitors applications' activities. Users get alerted whenever an application attempts to access a sensitive resource like the location, the phone ID or the Internet. The requested access can then be permitted or denied by the user [17][23][24]. CyanogenMod offers a variety of features and enhancements to Android. Among others, permissions can dynamically be granted and denied [18]. WhisperCore is a security application for Android with firewall and encryption functionality. It also offers an extended permission mechanism. Similar to MockDroid, access to resources is not blocked, but results in empty or dummy data [19].

C. Comparison

Table 1 shows a comparison of the solutions introduced in this section. We use the following criteria for the comparison.

1) Availability

Not all the solutions that we have introduced are available for end-users. Some are available as applications (marked with an A in Table 1) while others can be installed as Custom ROM (marked with CR). *PermissionTracker* is a simple application but has to be installed as Custom ROM, because it depends on extensions in Android. The same holds for MockDroid.

2) Modified Android

Modifications or extensions in the Android source code are required by several solutions, because the original Android's permission mechanism is too simple to provide enhanced permission functionality. Some applications operate with the original Android system and, thus, can provide only limited functionality like listing permissions that had been requested and granted to applications during their installation process.

3) Policy

Control of resource access is based on policies by the system. The end-user does not have control over permissions other than to change policies of the system. This is in contrast to manual control, where the user has direct control over the permissions of single applications.

4) Conditions

Restriction of resource access can be controlled by the definition of specific conditions. This is only possible in APEX, where users have fine-grained control about permissions of applications.

TABLE I. COMPARISON

	APEX	Kirin	SAINT	MockDroid	XmanDroid	TISSA	aSpotCat	Permission Dog	Permissions Denied	LBE Privacy Guard	Cyanogen Mod	WhisperCore	PermissionTracker
1. Availability	-	-	-	CR	-	-	A	A	A	A	CR	CR	CR
2. Modified Android	X	X	X	X	X	X	-	-	-	-	X	X	X
3. Policy	-	X	X	-	X	-	-	-	-	-	-	-	-
4. Conditions	X	-	-	-	-	-	-	-	-	-	-	-	-
5. Blocking	X	-	X	-	X	-	-	-	X	X	X	-	X
6. Mocking	-	-	-	X	-	X	-	-	-	-	-	X	-
7. User confirmation	-	-	-	-	-	-	-	-	-	X	-	-	X
8. Logging	-	-	-	-	-	-	-	-	-	X	-	-	X
9. GUI	-	-	-	-	-	-	AP	AP	A	A	-	-	AP

5) Blocking

Permission to access a specific resource may be blocked by raising a security exception. It depends on an application how it handles such a situation. It may crash, close down or continue to run.

6) Mocking

An alternative to blocking access to a resource is to provide dummy or mock data. Thus, an application will not realize that it won't get access to real data and will continue to run as expected. We have not yet implemented this feature, but believe that this will be necessary. Tricking applications will prevent them from not functioning as expected.

7) User confirmation

In some situations users may prefer to explicitly confirm access to a resource by a specific application. Doing so for all applications and all permissions is too much trouble. But it is useful to have control over specific resources, especially when a new application has just been installed and is not (yet) completely trusted by the user.

8) Logging

Logging the access to resources allows for later inspection. Logs make it easy to find out which resources really had been accessed by specific applications. Logging is only provided by LBE Privacy Guard and *PermissionTracker*.

9) GUI

Android extensions typically provide some minor extensions to the Android permissions user interface. Permission applications provide extended user interfaces. For example, grouping applications and permissions into categories, which makes it easier to grant or deny resources. aSpotCat, PermissionDog and *PermissionTracker* support categories for applications and for permissions (AP in Table 1). PermissionDenial and LBE Privacy Guard provide categories for applications only (A in Table 1).

As can be seen in Table 1, *PermissionTracker* is an application that needs an extended Android version. Its permission control is based on user manipulation rather than the definition of policies and it provides blocking of permissions. It stands out by providing dynamic user confirmation and logging which is also available for the LBE privacy guard. However, the LBE privacy guard offers less functionality because it is based on an unmodified version of Android.

VIII. PERFORMANCE

Performance is a crucial issue that has to be taken into account for any changes to the existing permission control mechanism, as users are typically not willing to sacrifice performance for security. With *PermissionTracker*, users can define policies for apps to manage and monitor access to features. These policies are stored in an Xml file that is read into memory during each system startup. If an app wants to access certain permissions, the policy as specified in the *PermissionTracker* is evaluated.

We have measured the performance overhead of our code changes on the Android emulator with Android 2.3.6. 45 apps were installed on the system. For that purpose, we have exemplarily measured the time taken to resolve the permission checks for sending SMS (SEND_SMS) and for accessing GPS data (ACCESS_FINE_LOCATION). In particular, we have timed the intervals for checking permissions in the existing security mechanism of Android, in our modified permission checks and in our modified permission checks where we log permission access.

The difference between the time taken by the original permission-check mechanism of Android and that of our system enhancement is rather small, an average of 3.5 milliseconds for both permissions. When logging is activated, the additional time needed for policy evaluation increases to an average of 35 milliseconds. This is mainly caused by the fact that each permission access is written to a log file in the working directory of *PermissionTracker*. All in all the enhancements to the existing permission checking model of

Android are efficient and require only a little performance overhead that is not noticeable to users.

IX. Conclusion and Future Work

We have introduced Android with its security mechanisms. The permission system is rather rigid in Android and suffers from a few drawbacks. We have developed a more flexible permission system with a permission tracker tool. The system allows users to block and monitor access to resources by arbitrary applications. For implementation purposes, the Android system had to be slightly extended.

The use of an app like the *PermissionTracker* application is recommended for users who want to control access to their resources on a finer granularity than what is currently possible with Android. Android users who want to have advanced control over permissions granted to their apps have three options. First, they can download and use one of the permission apps with limited functionality. This helps in getting a better overview about permissions granted to apps, but does not prevent privacy breaches. Second, users can use a modified version of Android with a better handling of permissions. This provides more protection against privacy breaches. However, installing a modified version of Android is only practicable when activating a new device. The installation process would remove any settings and user data on a device that is already in use. In order to empower all end-users with flexible permission control, it will be necessary to include more flexible permission control into the regular Android system.

References

[1] IDC - Press Release. Android- and iOS-Powered Smartphones Expand Their Share of the Market in the First Quarter, According to IDC. 24 May 2012. www.idc.com/getdoc.jsp?containerId=prUS23503312 [retrieved: Aug., 2012]

[2] Android Open Source Project. Android. source.android.com/ [retrieved: Aug., 2012]

[3] Coverity Scan: 2010 Open Source Integrity Report. Featuring the Coverity Software Integrity Report for the Android Kernel. www.coverity.com/library/pdf/coverity-scan-2010-open-source-integrity-report.pdf [retrieved: Aug., 2012]

[4] Android Developers . Permissions. developer.android.com/guide/topics/security/security.html [retrieved: Aug., 2012]

[5] Android Developer. The AndroidManifest.xml File. developer.android.com/guide/topics/manifest/manifest-intro.html [retrieved: Aug., 2012]

[6] Android Developer. ActivityManager. developer.android.com/reference/android/app/ActivityManager.html [retrieved: Aug., 2012]

[7] Android Developer. PackageManager. developer.android.com/reference/android/content/pm/PackageManager.html [retrieved: Aug., 2012]

[8] Android Open Source Project. Initializing a Build Environment. source.android.com/source/initializing.html [retrieved: Aug., 2012]

[9] Artem Russakovskii. Custom ROMs For Android Explained - Here Is Why You Want Them. www.androidpolice.com/2010/05/01/custom-roms-for-android-explained-and-why-you-want-them/ [retrieved: Aug., 2012]

[10] Android Code. Installing the Latest Custom ROM. code.google.com/p/android-roms/wiki/Install_Custom_ROM [retrieved: Aug., 2012]

[11] Mohammad Nauman and Sohail Khan. Design and Implementation of a Fine-grained Resource Usage Model for the Android Platform. Department of Computer Science, University of Peshawar, 2010, csrdu.org/pub/nauman/pubs/apexext-iajit10.pdf [retrieved: Aug., 2012]

[12] William Enck, Machigar Ongtang and Patrick McDaniel. Mitigating Android Software Misuse Before It Happens. Department of Computer Science and Engineering, Pennsylvania State University, 2008, www.enck.org/pubs/NAS-TR-0094-2008.pdf [retrieved: Aug., 2012]

[13] William Enck, Patrick McDaniel, Stephen McLaughlin and Machigar Ongtang. Semantically Rich Application-Centric Security in Android. Department of Computer Science and Engineering, Pennsylvania State University, 2010, www.enck.org/pubs/acsac09.pdf [retrieved: Aug., 2012]

[14] Alastair Beresford, Andrew Rice, Nicholas Skehin and Ripduman Sohan. MockDroid: Trading privacy for application functionality on smartphones. Computer Laboratory, University of Cambridge, 2011, www.cl.cam.ac.uk/~acr31/pubs/beresford-mockdroid.pdf [retrieved: Aug., 2012]

[15] Sven Bugiel, Lucas Davi, Alexandra Dmitrienko, Thomas Fischer and Ahmad-Reza Sadeghi: Android Security XManDroid: A New Android Evolution to Mitigate Privilege Escalation Attacks: System Security Lab, Technische Universität Darmstadt, 2011, www.informatik.tu-darmstadt.de/fileadmin/ user_upload/Group_TRUST/PubsPDF/xmandroid.pdf [retrieved: Aug., 2012]

[16] Vincent Freeh, Xuxian Jiang, Xinwen Zhang and Yajin Zhou. Taming Information-Stealing Smartphone Applications (on Android). Department of Computer Science, NC State University, 2011, www.csc.ncsu.edu/faculty/jiang/pubs/ TRUST11.pdf [retrieved: Aug., 2012]

[17] CNET.de. Permissions Denied für Android: Alle App-Berechtigungen voll im Griff. (in German) www.cnet.de/blogs/mobile/android-app/41552649/ permissions_denied_fuer_android_alle_app_berechtigungen_ voll_im_griff.htm [retrieved: Aug., 2012]

[18] CyanogenMod. CyanogenMod Wesbsite. www.cyanogenmod.com/ [retrieved: Aug., 2012]

[19] Whisper Systems. Selective permissions for Android. whispersys.com/permissions.html [retrieved: Aug., 2012]

[20] Sam Lu. aSpotCat (app by permission). play.google.com/store/apps/details?id=com.a0soft.gphone.aSpotCat [retrieved: Aug., 2012]

[21] Android Freeware. PermissionDog. www.androidfreeware.net/download-permissiondog.html [retrieved: Aug., 2012]

[22] Google Play. Permissions Denied. play.google.com/store/apps/details?id=com.stericson.permissions [retrieved: Aug., 2012]

[23] Google Play. LBE Privacy Guard. play.google.com/store/ apps/details?id=com.lbe.security.lite [retrieved: Aug., 2012]

[24] Sameed Khan. LBE Privacy Guard For Android Monitors Access Requests, Guards Privacy. addictivetips. www.addictivetips.com/mobile/lbe-privacy-guard-for-android-monitors-access-requests-guards-privacy/ [retrieved: Aug., 2012]

MyVigi : An Android Application to Detect Fall and Wandering

Bruno Stanislas Beauvais
UFRIMAG
Joseph Fourrier University
Grenoble, France
e-mail: bruno.beauvais@e.ujf-grenoble.fr

Vincent Rialle and Juliette Sablier
AGIM laboratory
FRE 3405 CNRS-UJF-UPMF-EPHE
Grenoble, France
e-mail: juliette.sablier@agim.eu
vincent.rialle@agim.eu

Abstract—**According to the World Health Organization, nearly 35.6 million people live with dementia through out the world. These people, who often live at home, are exposed to the risk of wandering and falling. The use of a discreet monitoring device such as a smart-phone could assist them, increase their mobility and decrease the stress level of the caregivers. The objective is to implement a mobile wearable tool aimed at detecting the wanderings and falls of people with dementia or mild cognitive impairment living at home. The tool must be easy-to-use, cost effective and ethically acceptable by patients at risk and their caregivers. The selected supportive hardware is an Android based smart phone selected for its high performance and durability, worldwide availability, and low cost. The software design method is based on a participative design approach involving disabled persons, family caregivers, and health professionals. The client application uses a three-axial accelerometer embedded in the smart phone to detect falls. The smart phone is also able to detect wandering using a global positioning system. In case of alert, caregivers are automatically contacted by phone call, SMS and mail. A web site also provides them with a map of the localization in real-time. This work-in-progress paper presents the method and the technological implementation of the MyVigi application.**

Keywords-Alzheimer's disease; fall; wandering; telemedecine; cellular phone.

I. INTRODUCTION

Alzheimer's disease is the most common form of dementia. It is characterized by memory loss, slow disintegration of the personality and physical control, with manifestation of aggressiveness, wandering, incontinence, disinhibition, binge-eating, hallucination, delusion an depression[1][2]. This disease leads to an almost complete loss of independence.

Nearly 35.6 million people live with dementia through out the world[3]. According the results of the PAQUID cohort study, 61.5 % of this population is living at home in France[4]. Considering the increase of the life expectancy, the number of affected people will double in 2020 and triple in 2050.

According the France Alzheimer's Association, wandering concerns 11% of independent people and 28% of people who need occasional help[5]. They lose their way on well-known routes[6]. Falling is also particularly dangerous for them, especially because of the risk of femoral neck fracture[7]. In both risk situations, a rapid rescue is essential for a better prognostic. After disappearing, half die or get serious injuries if they are not rescued within 24 hours[8]. These facts show that the risks are important and arrive early in the disease, which is why they are a source of considerable anxiety for the caregiver.

The natural caregiver is harshly affected by this disease. In France, they spend 6 hours per day watching over their relative and the cost is 570€ per month[9]. They are threatened by exhaustion, half of them suffer from depression and spouses in particular have an excess death rate of 63 %[10].

Gerontechnology is more and more useful and appreciated for caregivers and the professionals[11]. Current tele-alarm systems provide assistance for the caregiver, in particular for wandering. The devices use a global position system (GPS) to locate the user and detect the way out of a specific area. However, these devices are costly and not much accepted because of their stigmatizing appearance[12]. Recently, work has been done to transfer this function to mobile phones. The applications Tweri[13], Ifall[14] and Iwander[15] are remarkable examples.

This technology breaches the privacy of the users. Its usage under medical control is recommended to avoid misuse[16]. The ethical dilemma is the following[17]:

- It is unacceptable to track somebody without his knowing about it.
- It is unacceptable to let somebody wander without aid.
- It is unacceptable to confine somebody to their home during months or years when he could go out with technical aid.

Although this technology is available, it is rarely used[16] [18]. The reasons are :

- Lack of knowledge of the technology by the potential user and prescribing doctor
- Ignorance of the true needs of the users
- Inadequacy between the financial cost for the users and their budget

Thanks to a user-centered approach, our goal is to provide truly beneficial results for the user with a tool that is easy-to-use, cost effective and ethically acceptable. Section II describes the 3 main phases of the project: Requirements Analysis, Implementation and Experiment. Section III describes the technology used and the methods chosen to detect falls and wandering. Section IV describes related benefits of our technological choices. During the time of writing, the prototype is still developing, and therefore, the paper is being submitted as a work-in-progress.

II. Method

We use a design framework[19] named TEMSED, which stand for 'Technology, Ergonomics, Medicine, Society, Economics, Deontology'. This framework is described as a base of health informatics[20] and models a holistic approach featuring a set of 6 consistent areas of human values to investigate:

- 'Technology' focuses on the purely technical values such as functioning and robustness.
- 'Ergonomics' puts in relation the user and the technical devices.
- 'Medicine' concerns the impact in terms of medical practice.
- 'Society' concerns the impact in terms of social practice.
- 'Economics' refers to the dissemination capacities of the device and relies essentially on its economical viability.
- 'Deontology' conforms to the respect of rights and duties of the stakeholders.

TEMSED is a general framework that provides the values we take into account during the design and the assessment. Concretely, human values described are investigated using interviews and literature analysis.

A. *Requirements Analysis*

The requirements analysis consists in two complementary activities: the scientific literature analysis and the interviews of stakeholders[21][22][23]. The people interviewed are persons with dementia and mild cognitive impairment, family caregivers, and health professionals. Interviews are semi directive and a comprehensive approach is used[24]. It divided into two parts: first, potential functionalities are described in order to be criticized by the interviewees. Then, prototypes of the user interface are presented and the interviewed people are asked to test it. The results of this phase are scenarios of use prioritized according their importance.

B. *Implementation*

The implementation consists in developing client applications for an embed device, as well as server applications to centralize data and make it permanently accessible to the family caregivers and rescue team. The development is iterative and scenario-based[21]. Fig. 1 describes architecture and details are given in the chapter III.

One of the major issues for the dissemination of the gerontechnology is their prohibitive rates for the users[18]. In order to reduce the production costs and to facilitate technology transfer, we choose an open source approach. The source produced is under CeCCIL-C license[25]. This license similar to the LesserGPL allows to include the source as component in proprietary software. Then, the share of production cost of these components is possible between industrial companies.

C. *Experiment*

The experiment consists in testing the system for 3 weeks. The tester population is composed of people affected by Alzheimer's disease and their natural or professional caregivers. After the experiment, people would be interviewed. The objective is to get feedback about the material to make a sociological analysis of its use. The interviews are semi directive and conducted thanks to the interview's guide used in ESTIMA[26].

III. Technological Implementation

A. *Hardware*

The prototyped application is designed for the Samsung Galaxy SII. The device has a dual-core Cortex-A9 processor running at 1.2 GHz and 1 Go of RAM. Its dimensions are 125.3 mm × 66.1 mm × 8.5 mm and it weighs 116 grams.

The device is equipped with the SiRF Star IV chipset for navigation with Global Positioning System (GPS). The device also is equipped with nine-axis MPU-9150 as motion tracking sensor. It associates compass, three-axis gyroscope and three-axis accelerometer. The range of the accelerometer is set on 4G, this resolution 0.1 m/s² and this rate of response 100Hz.

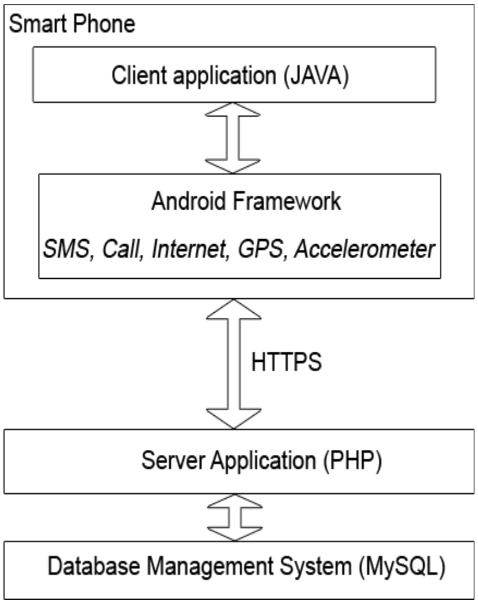

Smart Phone
Client application (JAVA)
Android Framework
SMS, Call, Internet, GPS, Accelerometer
HTTPS
Server Application (PHP)
Database Management System (MySQL)

Figure 1. Architecture

B. *Software*

This device is built for Android, a Linux-based operating system designed for mobile devices such as smart phones and tablet computers. It also packages middleware and key applications[27]. Applications are written in Java and run on the Dalvik virtual machine. The Android service development kit provides libraries to interface with hardware as communication capabilities and sensors. This operating system is also the most widespread[28] and is supported by a large community of developers.

Data concerning localisations and configuration of the client is synchronised with a server application. The server is written in PHP, run on Apache server and uses a MySQL database to store persistent data. The GSM or WIFI connection is used following their availability.

C. *Fall detection*

In case of a fall, a simple button to press is not enough, 30% of people don't use it and stay lying on the floor[29]. Further, a rapid response increases the functional prognosis[30][31]. Therefore, an automatic system of detection is essential, and in particular for people suffering from dementia. The fall differs from the Activities of Daily Living (ADL) by the acceleration applied on the body and by the change of the position. These differences can be detected by an accelerometer in condition that the sensor is hanged close to the trunk[32].

The implemented algorithm uses android API to get the orientation and the acceleration exerted on the device. The data is given according to three orthogonal axes (X,Y,Z). The absolute acceleration (A_T) is obtained from the computation of the root sum of the square of the accelerometer's three axes (A_x,, A_y, A_z).

$$A_T = \sqrt{\left(A_x^2 + A_y^2 + A_z^2\right)}$$

The algorithm uses thresholds and timers to detect three successive states: free fall, impact and lying position (Fig. 2):

- Free fall (T0 < t < T1): the fall starts with a free fall phase where the acceleration is under 1 G[33]. The laps of time is between 300ms and 500ms.
- Impact (T1 < t < T2): the body stroke the floor with an acceleration superior at 3 G[34].
- Lying position (T2 < t): the body stays lying with an orientation close to 90° when compared with the beginning of the free fall (T0).

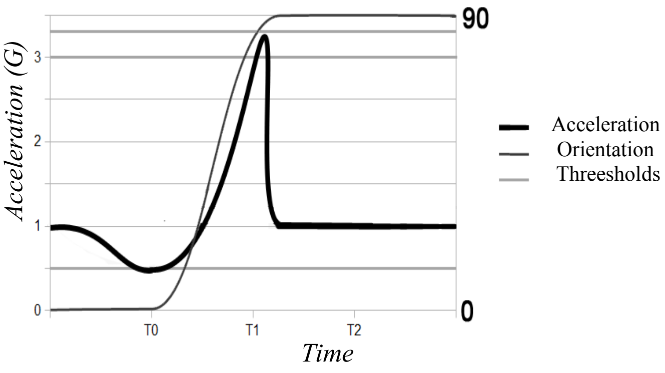


Figure 2. Fall modelling

In order to assess the sensibility of the algorithm we constitute a set of 10 scenarios summing up most of the fall[35].

The parameters taken under consideration are:

- The different sides: forward, backward and lateral.
- The end position: lying, sitting and on the knee.
- The possible quarter turning of the trunk.

The slow fall is not currently taken under consideration but will be included in the future using pressure sensors. The challenge of automatic detection is also to differentiate falls from ADL. The false positive detection is an important inconvenience that leads the user frequently to give up the system[26]. In order to decrease false positive, we set up scenarios that could provoke them. We take into account the nature of the sensors (gyroscope and accelerometer) to build these scenarios:

- Sitting down on a chair and get up.
- Lying on a bed and get up.
- Squat down and get up.
- Picking something on the floor and get up.
- Fall on the floor, lying and get up.

D. *Wander detection*

Wandering is a vital risk that is stressful for the caregiver. In a study of caregivers' wishes regarding technology[36], 53.3% of respondents reply “very much” to “would this technology be helpful to you ?”. According to a sociological survey[26], this technology is above all a tool to assist their vigilance. It's also essential that the system is easy to use and the detection algorithm is easy to understand by the caregiver.

The implemented algorithm to detect wandering. is based on the elapsed time when the user is outside its usual living areas. The algorithm handles :

- The living areas such home or shops.
- The time slot when these areas are possibly visited.
- The route duration from an area to an other.

This information is supplied by the caregiver from client or server applications. A timer is started when the user leaves a living zone. Wandering is detected when the timer goes past the longest duration to go to another area. If the areas take into account are the ones with a time slot matching the time of the day.

The life zones are bounded by a circle and located by latitude and longitude. The algorithm uses the Android API to track the device and check its inclusion in one of the living areas. The location with GPS consume highly the battery. Nevertheless, the challenge is to provide a system able to work as less over all day. Android offers the interesting possibility to localise the device using the identifiers of surrounding networks such cellular radio systems and wifi's. Android provides also a margin of error with Circular Error Probality (CEP) of 5%[37]. However, the accuracy is low and internet connection is required to associate network identifiers with locations.

The implemented algorithm uses preferentially networks. The GPS chipset is used only if the accuracy is too low. The required accuracy to establish inclusion in the area depends on (figure 3):

- X: distance between the location and the centre of the closest living area.
- φ: radius of the current living zone.
- ME: margin of error (CEP < 5%).

The tracking is stopped when the accuracy is sufficient to establish the device in the area (1) or out the area (2).

$$X < \varphi - ME \ (1) \qquad X > \varphi + ME \ (2)$$

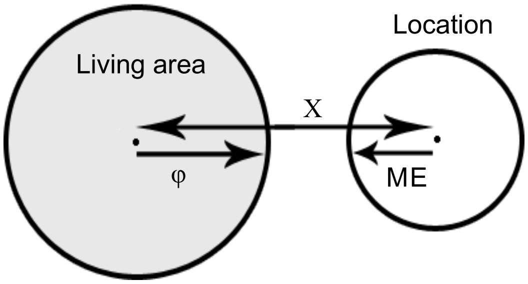


Figure 3. Detection of leaving or entering areas

E. Alert

The alert is activated automatically when a risky situation is detected (wandering or a fall). The user can also call for help using physical or tactile buttons. Another way to call for help manually, is to shake the phone. Once the alert is activated, the device tries to contact a list of caregivers in several ways until the alert is canceled. SMS and mails are sent every 5 minutes with the address of the last known position. At the same time, the smart-phone calls one caregiver after the other until someone responds. A map with the localization in real-time is also available on a website. The entire alert process is configurable online.

F. Privacy protection

The transport and the availability of personal information online such location involve risks for the privacy violation. We have followed European directives and tried to limit the risks as far as possible[38]:

- The access of personal information is password protected and can be deleted.
- The locations are encrypted with the advanced encryption standard algorithm[39] with the user password as the key.
- The locations are uploaded on the server only in case of alert. Then they are deleted.

IV. Discussion

The smart phone's different communication capabilities offer several interesting possibilities. First, the internet connection allows the client application to upload data such as localization and alerts onto a single centralized database. In this way, it is possible to know more about user behaviour, and also to check whether the system is running the way it is supposed to.

The use of open format-based data and standard protocol favorises interoperability and permits the client application to add other telemedecine application. The internet connection allows it to use web services to enrich the detection of risky situations with the local weather for example. The bluetooth and wifi allow it to integrate into the system additional sensors such as a cardio frequency meter.

The choice of the smart phone as device has the advantage to not be stigmatising. However, phones are often forgotten[26]. Watch-phones are an inadequate option, because the fall detector has to hang close the waist to be effective[32]. An alternative is to use an accessory to hang the device on a belt, decreasing the chance of it being forgotten. In this way, it is always accessible, less bulky and less likely to get lost or stolen.

V. Conclusion and future work

We ensure to match the user's need thanks to a participative design approach involving disabled persons, family caregivers, and health professionals. We also choose to support the TEMSED framework to make certain all the domain concerned by the use of the MyVigi application are investigated. The final result would be a mobile wearable tool aimed at detecting the wanderings and falls of people with dementia or mild cognitive impairment living at home. This tool would be easy-to-use, cost effective and ethically acceptable by patients at risk and their caregivers.

VI. *References*

[1] N. Nagaratnam, M. Lewis-Jones, D. Scott, and L. Palazzi, “Behavioral and psykjchiatric manifestations in dementia patients in a community: caregiver burden and outcome”, Alzheimer Dis Assoc Disord, 1998.

[2] Alzheimer’s Association, “Alzheimer’s disease facts and figures”, Alzheimer’s and Dementia, March 2010.

[3] Alzheimer’s Disease International, “World Alzheimer Report”, http://www.alz.co.uk/research/world-report, 2009.

[4] H. Ramaroson,, C. Helmer, P. Barberger-Gateau, L. Letenneur, and J. F. Dartigues, “Prevalence of dementia and Alzheimer's disease among subjects aged 75 years or over: updated results of the PAQUID cohort”, Rev Neurol (Paris), 2003.

[5] J. Prévot, “les résidents des établissements d’hébergement pour personnes âgées en 2007”, DRESS, Etudes et Résultats, n°699, August 2009.

[6] A. Nelson, and L. A. Donna, “Evidence-based protocols for managing wandering behaviors”, New York, NY: Springer Pub. Co, 2007.

[7] S. M. Friedman, I. B. Menzies, S. V. Bukata, D. A. Mendelson, and S. L. Kates, “Dementia and Hip Fractures Development of a Pathogenic Framework for Understanding and Studying Risk“,Geriatric Orthopaedic Surgery & Rehabilitation, vol. 1 no. 2 52-62, November 2010.

[8] R. J. Koester, and D. E. Stooksbury, “Behavioral profile of possible Alzheimer's disease patients in Virginia search and rescue incidents Wilderness and Environmental Medicine, 6,34-43 (1995)

[9] France Alzheimer, “Etude socio-économique - Prendre en soin les personnes atteintes de la maladie d’Alzheimer : le reste à charge - Principaux résultats”, 2010.

[10] R. Schulz, and S. R. Beach, “Caregiving as a risk factor for mortality : the Caregiver Health Effects Study”, JAMA, December 1999.

[11] V. Rialle, "Technology and Alzheimer's disease", Soins Gerontol, 2008.

[12] V. Faucounau, M. Riguet, G. Orvoen, A. Lacombe, V. Rialle, J. Extra, and A.-S. Rigaud, “Electronic tracking system and wandering in Alzheimer's disease: A case study Annals of Physical and Rehabilitation Medicine, Volume 52, Issue 7, Pages 579-587.

[13] “Tweri : Alzheimer cargiver”, www.tweri.com.

[14] F. Sposaro, and G. Tyson, "iFall: An android application for fall monitoring and response," in Conf Proc IEEE Eng Med Biol Soc, 2009.

[15] F. Sposaro, J. Danielson, and G. Tyson, “iWander: An Android Application for Dementia Patients” In: 32nd Annual International Conference of the IEEE EMBS Buenos Aires, Argentina, 2010.

[16] M. Laila, V. Rialle, C. Brissonneau, D. Princiaux, C. Sécheresse, D. Boukhalfa, O. Magnilliat, and A. Franco, “The utility and the feasibility of electronic tracking for the prevention of wandering in demented elderly patients living in an institution”, 6th conf International Society for Gerontechnology. Pisa, Italy, June 4-6, 2008.

[17] V. Rialle, “La géolocalisation de malades de type Alzheimer : entre urgence sociosanitaire et dilemme sociétal”, NPG Neurologie Psychiatrie-Gériatrie. 2009.

[18] A. Franco, "Living at home: autonomy, inclusion and life planning" (Vivre chez soi : autonomie, inclusion et projet de vie), report to the Secretary of State responsible for Senior Citizens, France, juin 2010.

[19] V. Rialle, N. Vuillerme, and A. Franco, "Outline of a general framework for assessing e-health and gerontechnology applications: Axiological and diachronic dimensions" Gerontechnology 9(2): 245, 2010.

[20] C. Quantin, F.A. Allaert, B.A. Auverlot, and V. Rialle, “Security, legal and ethical aspects of computerised health data” In: Venot A, editor, Medical Informatics - Foundations and applications: Springer-Verlag; 2012 (in press).

[21] B. Rosson, and C. M. John, “Usability Engineering: scenario-based development of human-computer interaction”, Academic Press, 2002.

[22] J. M. Carrol “Hci Models, Theories, and Frameworks: Toward a Multidisciplinary Science”, MIT Press, Cambridge, MA, USA, 2003.

[23] P. Topo, and B. Östlund, “Dementia, Design and Technology. Time to Get Involved”, Volume 24, Assistive Technology Research Series, IOS Press, 2009.

[24] J. C. Kaufmann, “L’entretien compréhensif”, Nathan, Paris, 1996.

[25] CEA, CNRS, INRIA, “CeCILL-C free software license agreement”, http://www.cecill.info.

[26] V. Rialle , C. Ollivet, C. Brissonneau, F. Leard, I. Barth, J. Extra, and J. Sablier, “Alzheimer's disease and geolocation: initial results of the Estima study”, Soins Gerontol, 2012.

[27] Google Inc., “What is Android ?”, developer.android.com.

[28] W3Techs: http://w3techs.com, retrieved May 2012.

[29] J. Flemin, and C. Brayne. “Inability to get up after falling, subsequent time on floor, and summoning help: prospective cohort study in people over 90”, BMJ, 2008.

[30] R.E. Roush, T.A. Teasdale, J.N. Murphy, and M.S. Kirk, “Impact of a personal emergency response system on hospital utilization by community-residing elders”, South Med J., 1995.

[31] R. J. Gurley, N. Lum, M. Sande, B. Lo, and M. H. Katz, “Persons found in their homes helpless or dead”, N Engl J Med., 1996.

[32] M. Kangas, “Development of accelerometry-based fall detection. From laboratory environment to life.”, Doctoral Thesis, University of Oulu, Oulu, Finland.

[33] A.K. Bourke, J.V. O’Brien, and G.M Lyons, ”Evaluation of a thresholdbased tri-axial accelerometer”, Gait & posture, September 2006.

[34] J. Chen, K. Kwong, D. Chang, J. Luk, and R. Bajcsy, “Wearable sensors for reliable fall detection”, Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, September 2005.

[35] N. Noury, P. Rumeau, A.K. Bourke, G. Ólaighi, and J.E. Lundy, ”A proposal for the classification and evaluation of fall detectors”, IRBM, 2008.

[36] V. Rialle, C. Ollivet, C. Guigui, and C. Herve, "What do family caregivers of Alzheimer's disease patients desire in smart home technologies? Contrasted results of a wide survey", Methods Inf Med 47(1): 63-69, 2008.

[37] Google, “Geolocation API Specification”, W3C Proposed Recommendation, 10 May 2012.

[38] The european parliament and the council of the european union, “Directive 2009/136/EC”, 25 November 2009.

[39] J. Nechvatal, E. Barker, L. Bassham, W. Burr, M. Dworkin, J. Foti, and E. Roback, “Report on the Development of the Advanced Encryption Standard (AES)”, surcsrc.nist.gov, National Institute of Standards and Technology, 2 Octobre 2000.

An Autonomous Traceability Mechanism for a Group of RFID Tags

Yann Glouche
INRIA, Unité de Recherche Rennes-Bretagne-Atlantique
Campus de Beaulieu,
35042 Rennes Cedex, France
email: yann.glouche@inria.fr

Paul Couderc
INRIA, Unité de Recherche Rennes-Bretagne-Atlantique
Campus de Beaulieu,
35042 Rennes Cedex, France
email: paul.couderc@inria.fr

***Abstract*—Coupled Objects are an innovative way to ensure integrity of group of objects, or complex objects made of parts. This principle can be used in various applications such as logistic or security. The main property of coupled objects is that integrity checking is autonomous and does not depend on external information systems: all the necessary data are self contained in radio-frequency identification tags associated to the objects. This avoids important issues such as scalability and privacy, but the self-contained approach makes error diagnostics difficult when an integrity check fails. In this paper, we propose a solution to this problem, with a resilient data structure supporting the identification of missing elements in a coupled object. When some elements among the coupled objects are missing, it is possible to detect if the group is corrupted. Moreover, our approach also allows to identify the missing elements.**

Keywords-RFID; checking; integrity; data structure; graph.

I. Introduction

In recent years, we have seen increasing adoption of the radio-frequency identification (RFID) technology in many application domains, such as logistic, inventory, public transportation, and security. In this paper, we focus on an innovative integrity checking approach, which uses a collection of RFID tags distributed over a set of object, discussed in [1], [2]. In the coupled objects approach, a set of physical entities are associated in a logical group by writing integrity data in an RFID tag on each object. The integrity of the group can then be checked when needed by reading the tag data and verifying integrity properties.

For example, package containing a group of tagged items could be transported in a secured way by different parties to the final recepient. Liability transfer and trust between parties would be ensured by checking the package integrity at each transfer step.

Unlike typical RFID systems, the tags store data, which are self sufficient for the integrity checking application, without needing access to an external data base or information system. Autonomous operation is an important feature as it avoids many issues associated with typical RFID systems: network access and scalability, database scalability, privacy and security.

However, the coupled objects architecture also has some limitations: when an integrity error is detected in a group of objects, characterizing the nature of the error is challenging. Typically, when some parts are missing in the group we would like to identify them, or to recover the application data associated with them. Unfortunately, as RFID tags provide very limited memory capacity, it is difficult to implement redundant storage and robust data structures distributed over a set of RFID tags. It is this problem that we address in this paper. We propose a resilient data structure for coupled objects, which can be used in particular for improving traceability when parts of an object group are lost.

The article is organized as follows. The next section presents the background on coupled objects and discusses some related works. The third section details our resilient data structure for coupled objects, while the fourth section describes the building and verification algorithms. Finally, Section V concludes the paper.

II. Background

The coupled objects concept finds an application in many RFID approaches. However, due to some memory limitations in the RFID tags, retrieving the data stored in missing elements is difficult whithout using an external information system.

A. Coupled Objects principles

Coupled objects consist in groups of physical objects that are logically associated together, meaning that they carry digital information referencing other objects of the set, or representing their membership to the group.

The physical objects are associated to an RFID tag. Each member of a group is represented by a tag. The group representation uses some data distributed over the set of tags. The structure is self-contained in data stored in the memory of a set of tags. Because the memory size in the tag is limited and the integrity check should be fast, the group will be represented by a digest, computed by a hash code function. The digest will be stores in the memory of each tag of the group. This approach enables full autonomous operation of both the association points and the checkpoints.

The group building phase works as follows. Let us consider a set of n tags with unique identifiers $t_1, t_2, ..., t_n$. The

ISBN: 978-1-61208-236-3

idenfiers are ordered in a determined sequence (using a chosen order relation). Then, a hash function is applied to this information to compute the digest: $d = hash(t_1, t_2, ..., t_n)$. This hash value is used as a group identifier gid, stored in each tag of the set.

This hash value makes possible the integrity checking phase. Once a group is formed, integrity checks can be performed. The principle is to read all the tag identifiers t_i of the objects of a given group (sharing the same group id gid), and verifying that the $hash(t_1, t_2, ..., t_n) = gid$. If the computed hash does not match the gid stored in the tags, the group is considered as invalid.

B. Ubicheck application

A simple application scenario of coupled objects is *Ubi-Check* [1]: consider someone at the airport who is going to cross the security gate. The person is required to remove his jacket, belt, and put in a tray mobile phone, music player, remove laptop from his bag, and may be other objects... All that in hurry, with other people in the queue doing the same. Obviously, personal objects are vulnerable to get lost in this situation: objects can get stuck inside the scanner, can stack up on each other at the exit of the scanner, and it is easy to forget something while being stressed to get a flight. A coupled objects solution would secure the process by associating all personal items into a group, and checking it at the exit gate of the screening area. The same application could also work when leaving an hotel, a train, a plane, etc. We will use the Ubi-Check scenario in the rest of the paper to illustrate the problem of integrity failure raised by lost objects.

C. Dealing with integrity errors

When an integrity error is detected, this mean that one or more objects are missing from the group. In UbiCheck, this would trigger an alarm or a visual signal to notify the user. However, the system is not able to characterize the nature of the error, by listing the missing parts. A simple solution would be to store in an external information system a detailed list of the object group, but this would ruin all the advantages of the coupled objects architecture: autonomous operation, independance of external information system, privacy, etc.

Our goal is to design a data structure for the tags that would help to charaterize the missing element from the group, with the two following properties :

1) Autonomy: the necessary data to characterize the integrity error should be self-contained in the tags,
2) Resilience: the data structure should survive losing one or more objects.

The exact nature of the data used to interpret the integrity error can depend on the application. In the context of this paper, we will consider that objects of a group have interpretable identifiers, such as names, that would be used to enumerate the missing objects when an integrity failure is detected. For example, in Ubi-Check, each personal item would be associated to a user defined shortname such as 'wallet', 'phone', 'passport', etc. These names constitute the data to be protected by the distributed data structure stored in the tags, and we should be able to recover them when some objects are lost.

D. Related works

The problem we are addressing shares some similarities with data resilience methods that have been widely studied for storage devices and file systems. Related to the FRD concept [3], the RAID-6 approach protects disk storage systems from any disk failures. Unlike RAID-1 through RAID-5, which detail exact techniques for storing and encoding data to survive single disk failures, RAID-6 is merely a specification. The exact technique for storage and encoding is up to the implementor. Various techniques for implementing RAID-6 protect disk storage systems from two disks failures, such as EVENODD coding [4], Row Diagonal Parity coding [5], Liberation Codes coding [6]. These techniques for implementing RAID-6 have been developed and are based on erasure codes. There is no one standard for RAID-6 encoding.

The RAID implementation proposed in [7], [8] can restore some data lost after at most two disks failures. For restoring the data of two disk, this approach is based on the resolution of an equation system. This equation system is build by the definition of two parity functions, with two unknown variables, which are the two pieces of data lost on each of the two defective disks. It can be generalized to restore the data of n disks by considering n parity functions, to build an equation system of n equations and n unknown variables. Then, n disks are used to store the parity data to restore the disk failures. Like the data disk, the RFID tags store some data.

The RAID-5 approach distributes parity data along with the data. It is possible to distribute the parity data over the memory of each tag. The impact on the memory consumption of each tag becomes less important. It is simple to use a RAID approach to restore the data of the tags. This approach can be easily adapted for restoring some lost data on the RFID memory banks. The RAID approach can be used to rescue the data of n tags lost. However, when more than n tags are lost, nothing can be said about the missing tags: nothing can be deduces about the n first lost tags.

The RAID approach is used to secure the data, minimizing the overhead on storage space. The goal of our approach is not exactly the same. We want to able to recover some data that characterize the physical objects that have been lost, or the parts of a group that have been lost.

In the domain of the communication network, some approaches are based on the graph theory [9] to enforce the connectivity between network stations. In [10], the graph optimally connected is used to enforce the links between a set of stations, which be can viewed as a set of vertices of a graph. This approach enforces the connectivity in the network, when the disconnection of some stations occurs. Close to our RFID problem, a link between stations can be seen like the capacity to describe each tag data by the other tags. Enforcing the link between stations is a similar problem to enforce the link between tags.

III. Resilient data structure for coupled objects

As explained previously, when parts of a coupled object are lost, integrity failure is detected, but we need a resilient data structure to improve the traceability of the lost parts. This structure will be distributed over all the parts of the group, stored in tags memory. We assume that each part is described by an application-specific data item. For the purpose of simplicity, in the examples we will consider the Ubi-Check case where objects are described by short names.

The structure is based on the following design principles. First, the traceability mechanism should be able to identify the missing parts, so the structure should implement data redundancy. Second, the same memory space is allocated on each tag of the structure: the data distribution is balanced over the set of tags. Third, the robustness of the traceability mechanism should be independent of particular tags that may be lost; this means that if we want the structure to be able to resist to k tags lost in a coupled object of n parts, any of the k out of the n tags could be lost.

Our approach is based on the notion of regular graph from graph theory [9].

A. A graph representation for the data of RFID tags

In this model, a group of tags is modeled by a graph. Each RFID tag represents a vertex of a graph. Each tag stores in its memory some data about the neighbors: for example a nickame, which can be considered, as a short description. Each tag knows the nickname of its neighbor in the graph representation. These data are used to store the edges of the graph representation in the memory of the tags of the group.

Definition 1 (Graph): A graph G is a pair (V, E) comprising a set V of vertices, and a set E of edges, which is a set of pairs of vertices belonging to V.

Definition 2 (Regular graph): A regular graph is a graph where each vertex has the same number of neighbors: every vertex has the same degree. Let $k \in \mathbb{N}$. A regular graph with vertices of degree k is called a k-regular graph or regular graph of degree k.

The regular graphs of degree at most 2 are easy to classify: A 0-regular graph (Figure 1) consists of disconnected vertices, a 1-regular (Figure 2) graph consists of disconnected edges, and a 2-regular graph consists of a connected cycle (Figure 5) or a set of disconnected cycles (Figure 3). A 3-regular graph (Figure 4) is also called a *cubic* graph.

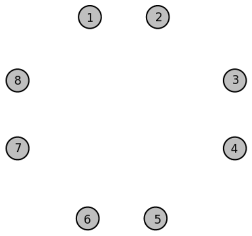


Figure 1. A 0-regular graph.

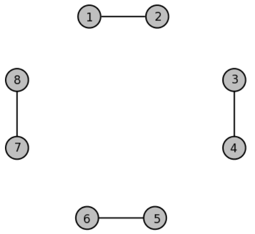


Figure 2. A 1-regular graph.

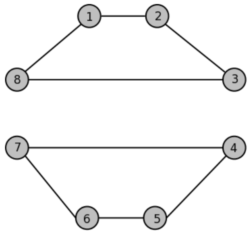


Figure 3. A 2-regular graph.

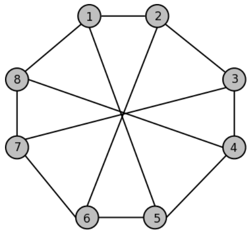


Figure 4. A 3-regular graph.

Let us consider the 2-regular graph G composed of 8 vertices presented on Figure 3. Let G_1 be the subgraph of G composed of the vertices 1, 2, 3, 8, and G_2 the subgraph of G composed of the vertices 4, 5, 6, 7. If the subgraph G_1 is deleted, then in the subgraph G_2, it does not exist a vertex for which one of its neighbors is missing. It is impossible to determine something about the missing vertices. In fact, it is impossible to deduce something by only considering the vertices 4, 5, 6, 7, because the subgraph G_2 is completly disconnected of the subgraph G_1. The graph structure should be connected to ensure a better tracability of the vertices by their neighbors in a graph representation.

Definition 3 (Path): In a graph, a *path* is a sequence of vertices such that from each of its vertices there is an edge to the next vertex in the sequence.

Definition 4 (Connectivity in graph): In a graph, the vertices of a path are said to be connected. Consider a graph $G = (V, E)$. If, for all vertices u and v of V, there exists a path from u to v, then G is connected.

Let us consider a graph $G = (V, E)$. This property ensures that, for all subgraph $G' = (V', E')$ such that $V' \subseteq V$ and $E' \subseteq E$, there exists two vertices $v \in V'$ and $u \in V \setminus V'$, such that $(u, v) \in E \setminus E'$ (there is an edge of G between a vertex of V' and a vertex of $V \setminus V'$). A 2-regular connected graph is shown on Figure 5. The regular graph presented on Figures 1 and 2 are also not connected. The 2-regular graph shown in Figure 3 is not connected. A $k-$regular graph can be connected if $k \geq 2$.

A connected graph ensures that, when some vertices are deleted, at least one edge from one missing vertex to a present vertex is deleted. When a vertex is missing, it becomes easy to ensure that it is missing by using a simple graph exploration algorithm. The $k-$vertex-connected graph enforces the connection property in the $k-$regular graph.

Definition 5 ($k-$vertex-connected graph): A graph $G = (V, E)$ is said to be $k-$vertex-connected if the graph remains connected when you delete fewer than k vertices from the graph.

A a graph is $k-$vertex-connected, if k is the size of the smallest subset of vertices such that the graph becomes disconnected if you delete them. So, it is sure that when less than k vertices are deleted in the graph, it is still connected. A $k-$vertex-connected graph ensures that: when some vertices are deleted, at least k edges from the set of missing vertices to the set of present vertices are deleted. By the following Property 1, a $k-$regular graph structure can at most ensure a $k-$vertex-connectivity.

Property 1: A $k-$regular graph is at most $k-$vertex-connected.

Proof (sketch): In the example presented on Figure 5, the $2-$regular graph is also $2-$vertex-connected.
In a $k-$regular graph, each vertex can be disconnected from the graph by deleting at least k neighbors. In this case, the graph becames disconnected, and one subgraph is defined by only one vertex. ■

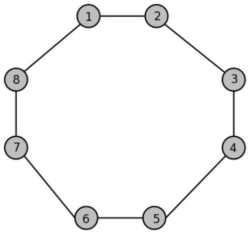


Figure 5. A 2-regular graph connected.

This structure of $k-$vertex-connected graphs induces the property of k-edge-connected.

Definition 6 (k-edge-connected graph): Let $G = (V, E)$ be an arbitrary graph. If $G' = (V, E \setminus X)$ is connected for all $X \subseteq E$, where $\mid X \mid< k$, then G is $k-$edge-connected. Trivially, a graph that is $k-$edge-connected is also $(k-1)-$edge-connected.

Property 2: Let's $G = (V, E)$ a k-vertex-connected graph. Then, G is a k-edge-connected graph.

Proof: Let's a graph $G = (V, E)$, such that G is not $k-$edge-connected.

- $\Rightarrow$ There exists a set of edges $E' \subseteq E$, such that $\mid E' \mid< k$, and the subgraph $(V, E \setminus E')$ is not connected.
- $\Rightarrow$ There exists a set $V' \subseteq V$ such that $\mid V' \mid=\mid E' \mid$ and $V' = \{s \in V \mid \exists t \in (V \setminus V'), (s, t) \in E'\}$ for which the subgraph $(V', \{(u, v) \in E \mid (u \in V \setminus V') \vee (v \in V \setminus V')\}))$ is not connected.
- $\Rightarrow$ There exists a set of vertices V', such that $\mid V' \mid< k$ and the subgraph $(V \setminus V', \{(u, v) \in E \mid (u \in V \setminus V') \vee (v \in V \setminus V')\})$ is not connected.
- $\Rightarrow$ G is not $k-$vertex connected.

Thus, if a graph G is not $k-$edge-connected, then it is not $k-$vertex-connected.
It is equivalent to say that a graph G is always $k-$edge-connected or not $k-$vertex-connected (by definition of the implication relation).
Thus, if a graph G is $k-$vertex-connected, then it is $k-$edge-connected.

■

Then, the structure of $k-$regular graph k-vertex-connected is also $k-$edge-connected.

Definition 7 (Optimally connected graph): Let G be a regular graph of degree k. If G is indeed $k-$vertex-connected and $k-$edge-connected, then it is called an *optimally connected graph*.

Thus, the structure of $k-$regular graph k-vertex-connected and $k-$edge-connected is also called an optimally connected graph of degree k. Then, the graph presented on Figure 5 is a optimally connected graph of degree 2, and the graph presented on Figure 6 is an optimally connected graph of degree 3.

By using this representation of graph $k-$vertex-connected and $k-$regular (with $k \geq 2$) to the RFID application presented in Section II, each tag is modeled by a vertex of the graph. Each association of two tags is modelized by an edge of the graph. Two tags are associated and they define an edge, when each of them contained the nickname describing the other. So, each vertex (or tag) knows who are its neighbors. With this representation of regular graph, the same quantity of memory is used in each tag, and the existence of each tag is equivalently stored in the others. By the $k-$vertex-connection property of this graph, this structure ensures that in a graph $k - regular$:

- when a set of n tags is lost, with $n \in \mathbb{N}, n \leq k$, the structure ensures that exactly n tags are missing, and the n nickames are known,
- when some tags (some vertices of the graph representation) are missing, at least k neighbors can give the nickname of some missing neighbors.

When a neighbor of a tag in the graph representation is missing, it is possible to determine explicitly that it is missing by using the knowledge of each vertex on the identity of its neighbors. This knowledge is stored in the RFID tags memory.

For example on Figure 5, it is possible to ensure that:

- if the tag 1 is missing, then the remaining tags 2 and 8 lost a neighbor.
- If the tags 1 and 2 are missing, then:
  - the remaining tag 8 looses the tag 1 as neighbor,
  - the remaining tag 3 looses the tag 2.
- If the tags 1, 2 and 3 are missing, then:
  - the remainingings tag 8 looses the tag 1 as neighbor,
  - the remaining tag 4 looses the tag 3,
  - Nothing can be deduces about the tag 2, because its neighbors are also missing.
- If the tags 1, 2 and 5 are missing, then:
  - the remainging tag 8 looses the tag 1 as neighbor,
  - the remaining tag 3 looses the tag 2,
  - the remaining tag 4, 6 looses the tag 5 as neighbor.

B. Robustness of the structure and memory cost

The greater the degree of a graph is the more robust is the structure. More robust is the structure, the easier it is to rescue some data when some tags are lost. It is also easier to determine the missing tags. The greater the degree of a graph is, in consequence, the more costly this representation is costly in tag memory. Let us consider the scenario application of Ubi-Check of the traveler in an airport presented in Section II-B. A traveler have 8 objects: Phone, Wallet, Bag, Belt, Jacket, Passport, Watch, Laptop, that he considers as very important. The traveler decides to group this set of objects with the Ubicheck application based on a RFID solution. Each object is associated to an RFID tag. In the example presented on Figure 6, the traceability mechanism links the tags by using an optimally connected graph of degree 3.

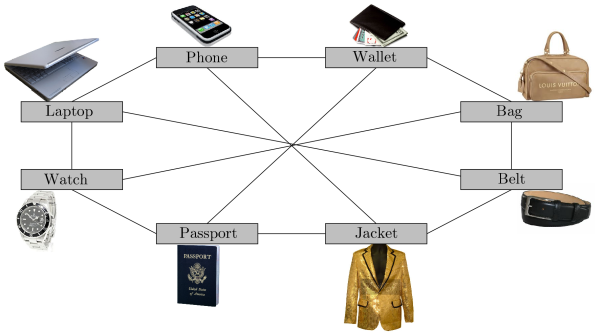
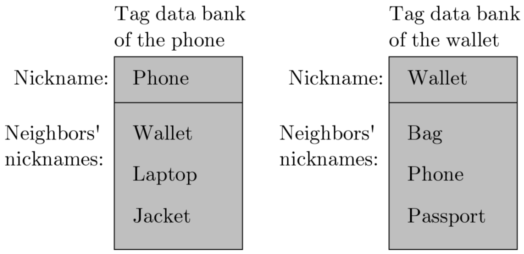


Figure 6. A 3-regular graph connected.

Each tag stores a nickname, which characterizes (for the owner) the object associated to it. To store the graph structure over the set of tags, each tag stores the nicknames of all its neighbors. Figure 7 represents the data stored in the memory bank of the tag associated to the phone (in the example of a $3-$regular graph presented on Figure 6).



Tag data bank of the phone

Nickname: Phone

Neighbors' nicknames: Wallet, Laptop, Jacket

Tag data bank of the wallet

Nickname: Wallet

Neighbors' nicknames: Bag, Phone, Passport

Figure 7. 3-regular graph representation in the tag data bank of the phone and wallet.

When the degree k of the graph representation increases, the robustness of the information also increases. But, increasing the degree also increases the memory space used in each tag for storing the graph structure. In fact, each tag stores an information about all of its k neighbors. So, the space used by the traceability mechanism is proportionnally increased when k increases.

With this optimal graph representation of degree 3, when at most 3 objects are lost, it is possible to explicitly list them. When more than 3 objects are lost, 3 of them can be listed. More generally, by using a graph representation of degree k, when at most k objects are lost, it is possible to explicitly list them. When more than k objects are lost, k of them can be listed.

For this example, the traceability mechanism can ensure that:

- if the phone is missing, then its absence is detected by the tag of laptop, wallet, and jacket.
- If the phone and wallet are missing, then:
  - the absence of the phone is detected by the tags of laptop, and jacket,

  - the absence of the wallet is detected by the passport, and the bag.
- If the phone, wallet and bag are missing, then:
  - the absence of the phone is detected by the tags of laptop and jacket,
  - the absence of the wallet is detected by the passport,
  - the absence of the bag is detected by the tags of watch and belt.
- If the phone, wallet, bag and passport are missing, then:
  - the absence of the phone is detected by the tags of laptop and jacket,
  - the absence of the bag is detected by the passport,
  - the absence of the passport is detected by the tags of watch and jacket,
  - nothing can be deduces about the missing wallet because in this 3-regular graph representation the three neighbors of the wallet are also lost.

As shown on Figure 8, when the traceability mechanism uses a regular graph of degree 3, four memory fields are used. For example, in the graph structure presented on Figure 6, the tag associated to the phone stores a nickname $phone$ that characterizes the object associated to it, and it also stores the nicknames of its neighbors: $Wallet$, $Laptop$, $Jacket$. In the same way, the tag associated to the wallet stores its nickname, and the nicknames of its neighbors.

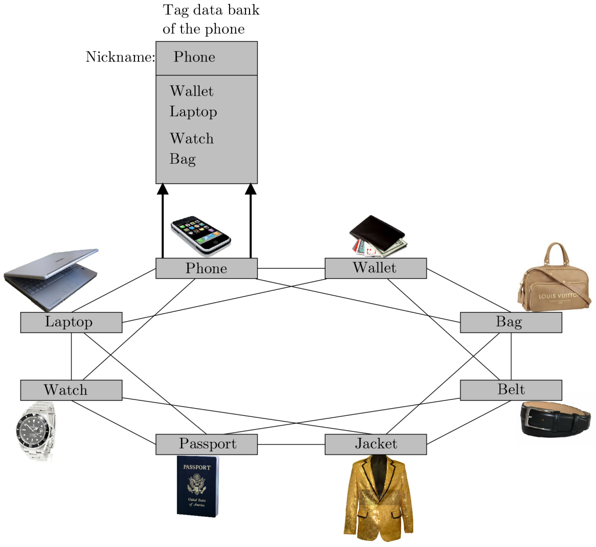


Figure 8. 4-regular graph representation in the tag data bank of the phone.

The characters can be translated in a hexadecimal value on two bytes. Let us consider, that the nicknames have at most 10 characters. In a 3-regular graph representation, each tag must store the three nicknames of these three neighbors. Then, it is necessary to use 60 bytes to store the nickname of the neighbors in 2-regular graph. More generally, for a $k-$regular graph representation, when the maximum length (in number of characters) of the nickname is equal to l, the memory space used to store the nickname of the neighbors of a tag is equal to: $2 * l * k$. Here, each tag is identified by a nickname, this a good way to extend the Ubicheck application. For other applications, it is also possible to save another information than a nickname.

IV. Building and verification algorithms

In this section, we present the algorithm of graph building, and the algorithm of graph structure checking.

A. Graph building

The graph building algorithm is used during the group creation phase of RFID tags, to store information about their neighbors in the graph structure.

It is essential to determine the necessary conditions to build a $k-$regular graph with a set of tags.

Property 3: Let $d \in \mathbb{N}$, and V a set a vertices. A regular graph $G = (V, E)$ of degree d exists, if and only if there exists $e \in \mathbb{N}$, such that $d <\mid V \mid$ and $d* \mid V \mid= 2 * e$.

$\Longrightarrow$:

Each vertex has at most $\mid V \mid -1$ potential neighbors. Then in a regular graph $G = (V, E)$ of degree d, $d <\mid V \mid$.

Let $d : V \mapsto \mathbb{N}$ the function that associates a vertex of V to its degree. In a simple graph each edge links two vertices. When the sum of the degrees is done, each vertices is counted twice: one time when the degree of the first vertex is counted, and a second time when the degree of the other vertex is counted. Then, in a simple graph: $\sum_{v \in V} d(v) = 2* \mid E \mid$.

In a regular graph of degree d, all the vertices have the same degree d, then $d* \mid V \mid= 2* \mid E \mid$. Thus, if a regular graph $G = (V, E)$ of degree d exists, then there exists $e \in \mathbb{N}$, such that $d* \mid V \mid= 2 * e$ and $d <\mid V \mid$. ■

$\Longleftarrow$:

Let a set of vertices V, and $d, e \in \mathbb{N}$ such that $d <\mid V \mid$ and $d* \mid V \mid= 2 * e$. Let us consider the set of vertices V as points regularly placed on a circle (as the vertices of a regular $n-$gon, with $n =\mid V \mid$). Let a set of edges E defined as follow:

- if d is even (the assumption: $d * n$ is even, is satisfied), then each vertex is connected to $d/2$ vertices that are after it as well as $d/2$ vertices that are before it,
- if d is odd and $\mid V \mid$ is pair (that satisfies the assumption: $d* \mid V \mid$ is pair), then each vertex is connected to $(d-1)/2$ vertices that are after it as well as $(d-1)/2$ vertices that are before it, and all the diagonals (those connecting two diametrically opposite points) of the $n-$gon are adding.

Then there exists a regular graph $G = (V, E)$ of degree d. ■

The algorithm 1 builds a regular graph of degree d from a set of vertices V, with $n =\mid V \mid$. To ensure the existence of the regular graph, it is assumed that $d <| V |$ and $d* | V |= 2 * e$ with $e \in \mathbb{N}$ (Property 3). The algorithm 1 works as follow. Let's us consider n vertices as points regularly placed on a circle (as the vertices of a regular $n-$gon). So, if d is even then each vertex is connected to $\lfloor d/2 \rfloor$ vertices that are after it as well as $\lfloor d/2 \rfloor$ vertices that are before it. And if d is odd, all the diagonals (those connecting two diametrically opposite points) of the $n-$gon are adding. In the case where d is odd, it is necessary that n is pair to satisfy the existence condition: $d * n$ is pair. In this case, each vertex is also connected to $\lfloor d/2 \rfloor$ neighbors before it, and to $\lfloor d/2 \rfloor$ neighbors before it.

Algorithm 1 Algorithm of building a graph optimally connected.

```
  $d$ : degree of the graph
  $V$ : a set of vertex
Ensure: ($d <| V |$) and (($d* | V |$) is pair)
  $set$ : Array of tags
  $nbTags$ : Integer
  $j$ : Integer
  $E$ : a set of edge
  $V \leftarrow \emptyset$
  $E \leftarrow \emptyset$
  $set \leftarrow HW\_read()$
  $nbTags \leftarrow set.length$
  for $i = 0$ to $nbTags$ do
    {//for each tag a vertex is added in $V$}
    add a new vertex $set[i]$ in $V$
    for all $j$ such that $j \neq i$ and $j \in [i - \lfloor \frac{d}{2} \rfloor ; i + \lfloor \frac{d}{2} \rfloor]$ do
      {//all the edges between the vertex $i$ and the $d/2$ vertices that are after him are added to $E$}
      {//all the edges between the vertex $i$ and the $d/2$ vertices that are before him are added to $E$}
      if the edge $(set[j \textbf{ modulo } nbTags], set[i]) \notin E$ then
        add the new edge $(set[i], set[j \textbf{ modulo } nbTags])$ in $E$
      end if
    end for
    if $d$ is odd then
      {//the edge between the vertex $i$ and its diametrically opposite points is added to $E$}
      add the new edge $(set[i], set[(i + \lfloor \frac{nbTags}{2} \rfloor) \textbf{ modulo } nbTags])$ in $E$
    end if
  end for
  return new Graph($V$,$E$)
```

$HW\_read()$ represents the inventory method provided by the RFID reader to return the set of RFID tags detected by the reader.

Let's a $k-$regular graph $G = (V, E)$. With the building principle of algorithm 1, it is impossible to delete less vertices for disconnecting a set of vertices of the $n-$gon, than for deleting only one vertex. The algorithm 1 builds a regular graph of degree d. Then, the d neighbors of a vertex must be deleted to disconnect one point. Thus, a regular graph of degree d built with the algorithm 1 is $d-$vertex-connected. Thus, the algorithm 1 builds some optimal graphs.

B. Checking graph

Algorithm 2 checks the integrity of an RFID structure modelized by a $k-$regular graph. For each vertex in a set $| V |$, the algoritm search all its neighbors in the same set $| V |$. The missing neighbors of all tags are stored in the set $| S |$. This set stores some information to depict the missing vertex. If S is empty, then no vertex is missing, the integrity of the graph is alright.

More formally, the algorithm is described as follows:

Algorithm 2 Algorithm of checking.

```
  $S$ : set of nicknames
  $V$ : set of tags or set of vertices in the graph representation
  $E$ : a set of edges
  $S \leftarrow \emptyset$
  for all $v \in V$ do
    {//The presence of all the neighbors of each vertex $v$ is tested}
    for all $n \in neighbors(v, E)$ do
      if $n \notin V$ then
        {//The nickname of the missing neighbor is added to $S$}
        add the nickname of $n$ in $S$
      end if
    end for
  end for
  {//If $S = \emptyset$ then no object is missing,}
  {//else $S$ contains the nicknames of the missing objects.}
  return $S$
```

V. Conclusion and future work

In this paper, we proposed a resilient data structure for coupled objects to help integrity error diagnosis: in particular, it can be used to identify the missing elements of a group of objects. The structure is stored in tags memory, in line with the idea of coupled objects to support autonomous operation. The robustness of the structure can be increased at the expense of memory overhead, making the structure configurable to application requirements.

Beside the Ubi-Check application described in the paper, we are considering other application scenarios where integrity checking of coupled objects should be diagnosed in case of errors. For example, it could be used to secure a medical prescription with a set of drugs. In case of an integrity error, it would be important to charaterize the nature of the error, such as identifying a missing drug. Some perspectives to this work include supporting dependency properties between elements inside a coupled object, in order to better charaterize integrity errors, or to characterize the global properties of a composite object when some elements are missing.

REFERENCES

[1] M. Banâtre, F. Allard, and P. Couderc, "Ubi-check: A pervasive integrity checking system," in *NEW2AN*, 2009, pp. 89–96.

[2] F. Allard, M. Banâtre, F. Ben Hamouda, P. Couderc, and J.-F. Verdonck, "Physical aggregated objects and dependability," Avalaible: http://hal.inria.fr/inria-00556951 [Last accessed in June 28, 2012], INRIA, Research Report RR-7512, Jan. 2011.

[3] J.-C. Fabre, Y. Deswarte, and L. Blain, "Tolérance aux fautes et sécurité par fragmentation-redondance-dissémination," *Technique et Science Informatiques (TSI)*, vol. 15, no. 4, pp. 405 –427, 1996.

[4] M. Blaum, J. Brady, J. Bruck, and J. Menon, "Evenodd: an optimal scheme for tolerating double disk failures in raid architectures," *SIGARCH Comput Archit News*, vol. 22, no. 2, pp. 245–254, 1994.

[5] P. Corbett, B. English, A. Goel, T. Grcanac, S. Kleiman, J. Leong, and S. Sankar, "Row-diagonal parity for double disk failure correction," in *Proceedings of the 3rd USENIX Symposium on File and Storage Technologies (FAST'04*, 2004, pp. 1–14.

[6] J. S. Plank, "The raid-6 liberation codes," in *FAST-2008: 6th Usenix Conference on File and Storage Technologies*, 2008, pp. 97–110.

[7] H. Anvin, "The mathematics of raid-6," 2011. [Online]. Available: http://kernel.org/pub/linux/kernel/people/hpa/raid6.pdf

[8] I. S. Reed and S. Solomon, "Polynomial Codes Over Certain Finite Fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960.

[9] F. Harary, *Graph Theory*. Reading, MA: Addison-Wesley, 1969.

[10] B. Myers, "Optimally connected communication networks with maximum diameters," *Electronic Circuits and Systems, IEE Proceedings G*, vol. 128, no. 6, pp. 289 –292, December 1981.

A Remote Control System to Configure Network Devices

Sandra Sendra, Emilio Granell, Ismael Climent, Jaime Lloret
Universidad Politécnica de Valencia
Camino Vera s/n, 46022, Valencia, Spain
sansenco@posgrado.upv.es; emgraro@posgrado.upv.es; ismael@cspteleco.net; jlloret@dcom.upv.es

Abstract— **Due to the emergence of new technologies and the proliferation of e-learning services and online courses, the need to provide remote access to real devices from any device, anywhere at any time is higher. There are several software applications that can simulate a laboratory virtually and others where you can configure the network devices remotely. But, but in order to provide more experience to the students, real devices should be used. In this paper, we present a remote control system that is able to manage the access to multiple console connections, in order to configure real devices. The advantage of our system is that it is an easy and low cost solution. This system provides ubiquitous access to the students, 24 hours a day, 7 days a week. Furthermore, it is also very easy to extend its system to other fields like Internet of Things.**

Keywords-Remote control System, Internet of Things, Remote laboratory.

I. INTRODUCTION

Nowadays, many people and enterprises that have Internet access are quite high, due to the growth of telecommunications infrastructure and the development of the next generation networks. It causes the need of well-trained professionals with good skills to configure and manage this type of networks.

The best way to train students and recycle professionals is to have a big infrastructure with many devices that can be accessible from anywhere at any time to let them practice. The courses about networks and new technologies based on the realization of practices on real devices, gives as a result, well-trained professionals capable of facing any challenge [1] [2].

The use of many network devices, allows students to do useful practices, getting reproduce any scenario that a student in his professional stage in a company could find [3].

More and more, it is tending to use collaborative teaching methods to improve learning, training and acquisition of skills required by any higher education [4]. However, in on-line courses where students can be in different cities, it is very difficult to carry out laboratory collaborative practices. But if instructors were in possession of a device management platform, where they can control the access to devices and even communicate in real time with their students, these problems would be solved.

The use of laboratory practices based on real devices improve the learning outcomes of training and experience for students [5], minimizing the difference between the lab practices and scenarios or problems that can happen in a company.

In this paper, we have developed a console server to access and manage network devices remotely. This is a software application for control of a teaching laboratory for courses about network management. The application is executed on a computer with a network operating system (NOS), which will be responsible for managing the access to several network devices. Network devices such as routers and switches have integrated a console port for their configuration. Our application is able to handle simultaneously up to 12 console ports, for the configuration of 12 devices. The advantage of our system compared to other existing systems is that it is a low-cost solution. This application has been developed initially for the management of network devices. However, its application scope is very wide and it may extend its use in Internet of things (IoT).

The rest of the paper is structured as follow. Section 2 shows some of the main proposals relating to the implementation of solutions for remote access to laboratories. Section 3 explains the composition of our console server and the hardware used. To manage our application and the access for all users, we need a NOS. Section 4 defines the NOS used and its settings. Section 5 explains the operation of the developed application and the operation of the protocol. It also shows how to connect to the console server. Finally, conclusions and future work are show in section 6.

II. RELATED WORK

It is easy to find several proposals regarding the possibility of implementing virtual labs. An example is the work presented by J. Lloret et al. [6], where they describe the experience carried out in the subject of computer networks at the Polytechnic University of Valencia. Authors implemented a laboratory that allows remote access to real devices for the students of the engineering of telecommunications.

We find a system similar to the one presented in this paper. The NDG NetLab developed by Cisco [7]. It is a computer whose physical format is a rack unit connected to other auxiliary network equipment, through which it controls the groups of network equipment such as routers and switches, enabling users to access to the system and the topologies laboratories. The architecture of this system is a server that controls the network laboratory equipment by certain actuators that turn on, shut down and restart the devices by orders of NetLab. With NetLab, a user may reserve time to use laboratory equipment and can perform exercises remotely through the Internet as if he was directly connected to them. However, the use of this system requires a license very expensive.

As Lloret et al. commented in their work [8], the combination of the new tools provided by e-Learning platforms and the new technologies, with a minimum combination of the traditional education instruments, shows satisfactory results, as much for the student as for the company.

In this paper, we show the development of a system that makes a real laboratory accessible from any place, at any time with any device (if the device has an application that allows Telnet connections).

III. System Description

In a teaching academy for networking courses, we can found a lot of network devices like routers and switches. These devices usually have a console port to be configured by the users. However, until now it is necessary to stay physically in the same place to connect the cable and configure it.

In this section we will see how we solved this problem and the equipment and hardware used for this.

Today it is relatively simple that a building used to teaching has a broadband connection to the Internet. We have used this fact to deploy our proposal. As we can see in Fig. 1, we have a preassembled network topology, which may be formed by several devices. All of them will be connected directly to the console server, which will be connected to Internet. The implementation of the console server will not preclude that instructors and local students could use the equipment. However, the possibility of getting them from the Internet through the server opens new possibilities for teaching. In this case, a local instructor can be teaching a class of students who are at home or in different place. It also raises the possibility that an instructor can teach a class on real devices without being physically at the academy.

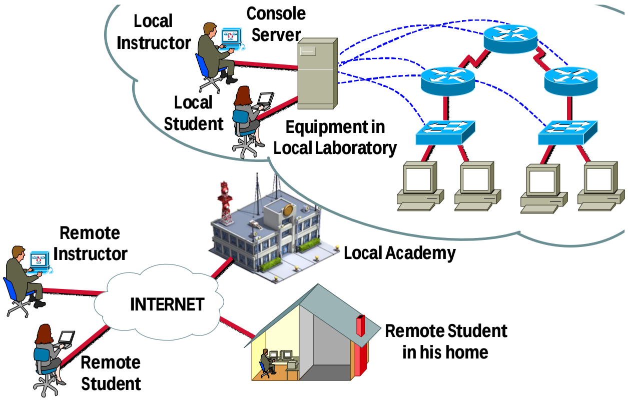


Figure 1. System Architecture.

The computer chosen is a computer-server, Fujitsu TS-C870I. This computer is formed by two processors Intel Pentium III of 550 MHz, with 512 MB of memory RAM. Because the processing capacity of the equipment that should be performed is not too high, these characteristics are sufficient. It is only need a 10 GB hard drive to store and run the OS. Moreover, this computer contains only two integrated serial ports that are insufficient for our proposal. Therefore, we need 2 expansion cards, which allow us to multiply the number of serial ports. To do this, we have used:

- Sunix's 4079T, serial and parallel combo expansion card: It is a serial and parallel combo expansion card [9]. It offers two ports of DB9 Male connection plus one internal parallel port. The serial ports have a data transfer rate up to 921.6 kb/s and the parallel ports offer speed up to 2.7 Mb/s. This board is supported for Windows CE, Windows 98SE / ME / NT/ 2000 / XP / 2003 / Vista, Linux, DOS (See Fig. 2).
- Quatech ESC-100D ISA Eight Port RS-232 Serial Board (DB-25) [10]: This board utilizes a single PCI slot to provide eight independent asynchronous serial ports that share a single interrupt. Serial port connections are made via standard DB-25 male connectors. This board is supported for Windows 95/98/Me/NT/2000/XP/VISTA, OS/2, DOS. It allows a data transfer rate up to 921.6 kb/s, full hardware and software flow control and full modem control (See Fig. 3).

With this hardware, we have 12 serial ports (COM ports), which can be used to configure 12 devices simultaneously. This has been our basic configuration, but it can be easily extended by adding more serial expansion cards (even using USB ports).

Table 1 shows the port distribution in function of available hardware. It is also shown the number of TCP port that is assigned to each serial port.

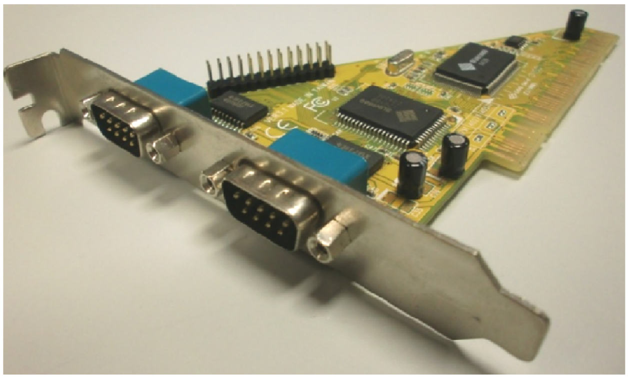


Figure 2. Sunix's 4079T, serial and parallel combo expansion card.

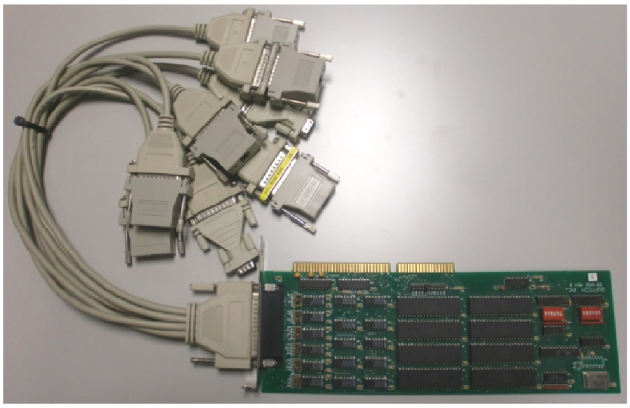


Figure 3. Quatech ESC-100D ISA board.

TABLE I. SERIAL PORT AND TCP ASSIGNMENT

Console Connection	**Compute Fujitsu TS-C870I**		
	Device	***Port connection***	***TCP Port Assignment***
COM 01	Motherboard Serial Port	1	1101
COM 02		2	1102
COM 03	Sunix's 4079T, expansion card	1	1103
COM 04		2	1104
COM 05	Quatech ESC-100D PCI board	1	1105
COM 06		2	1106
COM 07		3	1107
COM 08		4	1108
COM 09		5	1109
COM 10		6	1110
COM 11		7	1111
COM 12		8	1112

IV. SERVER CONFIGURATION

So far, we have a computer made up of 2 expansion cards, which allow you to configure up to 12 network devices. But the access management to each port must be done through an application that must run under a NOS. The choice of a NOS or another, depends mainly on the infrastructure that we need. In this section we will see the server configuration that has been selected as well as the NOS that has been used.

- **Level of network security.** This decision is based on the types of security that are considered more appropriate. Networks server-based allow more options regarding the security that workgroup could offer. On the other hand, when security is not a factor to be considered, it would be more appropriate, the utilization of a workgroup network.
- **Number of network users.** When the number of users is small, sometimes it is more practical and easy to manage a working group that a client/server network, where the maintenance, updating and management of resources will be small.
- **Number of computers in the network.** As in the previous case and for the same reasons, if we have a several computers on the network, it may be better to work in a client/server network.
- **Interoperability of the network.** After identifying the needs of security, users and computers on the network, the next step is to determine the types of interoperability required in the network to get it to behave as a unit.

Unlike the NetWare operating system, an OS not widespread, Windows combines a desktop computer OS and a network OS in a single OS. Microsoft provides client OS versions and server OS versions. On the other hand, the GNU/Linux OS's, and in particular the Ubuntu distributions, also offer different versions for client and server as Ubuntu Server.

GNU/Linux is a general-purpose operating system, multiuser and multitasking. The most popular OS based on GNU/Linux Debian, Ubuntu and Solaris. Typically, a GNU/Linux system is made up of a central computer and multiple terminals for users. This OS not only includes the networking features, designed specifically for large networks, but also some applications for personal computers. GNU/Linux OS's work well on an autonomous system, and as a result of its ability of multitask it also works well in a networked environment.

For our system, we chose the version of Windows 2003 Server from Microsoft. On the computer we installed the 2 expansion cards that we have presented in the previous section. We have also modified the allocation of IRQs for different resources that he OS performs by default, since some of the IRQs were shared by the hardware of the motherboard and the new hardware installed.

Finally, the setting of the serial ports in terms of transmission speed is 9600 bauds, 8 data bits, 1 stop bit and no hardware flow control.

V. PROGRAM CODE

In order to establish a secure communication between remote users and the equipment installed in the academy or learning center, we need an application that manages the access to the console server and devices connected to it. In this section, we will see the operation of the application developed.

The object **TRS232TCPXX** (where XX indicates the number of COM port), executes and loads the initial configuration of serial ports that will be used. These settings include the name of connection, the data transfer rate of the serial port in bauds, the port name that will be used and the TCP port number that is assigned to that COM port. This configuration will run every time that the service starts, and it is preconfigured for the 12 COM ports which can be used (See Fig. 4).

Program Code for TRS232TCPXX object

```
__fastcall      TRS232TCP10::TRS232TCP10(TComponent*      Owner):
TService(Owner) {

    const char * lpCommandLine = GetCommandLine();
    CommandLine=AnsiString(lpCommandLine);
    CommandLine.Delete(1,1);
    Path=ExtractFilePath(CommandLine);
    TIniFile *Config = new TIniFile(Path+"Serial2TCP.ini");
    ComPort1->Open();
    ServerSocket->Active=true;
    ServerSocket->Open();

}
```

Figure 4. Program Code for TRS232TCPXX object.

When the service has been initialized, the computer remains in a listening state until it receives a TCP connection request. If a connection request is generated, the object **ServerSocketClientConnect** is executed. When it is running, this object analyzes the port, looking for open connections. If there is any open connection, it is closed, to attend the next connection request (See Fig. 5).

Program Code for TRS232TCPXX object

```
void __fastcall TRS232TCP10::ServerSocketClientRead(TObject
*Sender, TCustomWinSocket *Socket){
Initialize Variables;
Generate file Log.txt;
Generate file Users.txt;
USERNAME:
        Collect New character of username;
        if(character Ascii es “13”){
            Save username in variable;
            Goto PASSWORD;
         } else{
            Goto USERNAME;
         }
PASSWORD:
        Collect New character of password;
        if(character Ascii es “13”){
            Save password in variable;
            Open file Log.txt;
            Open file Users.txt;
            Search string of characters in file Access.txt;
                if(cadena caracteres=true){
                    Message: access granted;
                    Save in file Log.txt;
                }else{
                    Message: access deneed;
                    Save in Log.txt;
                    Close connection;
                }
         } else{
            Goto PASSWORD;
         }
```

Figure 5. Program Code for TRS232TCPXX object.

If no further connections, the system will request a user name and password. At this time, the object **ServerSocketClientRead** is initialized. This object is responsible for analyzing all characters entered for the username and password, until it finds an "Enter" (See Fig. 6).

When the username and password is entered, the program will compare these names with the names defined in the file users.txt. If the username and password are correct, the application displays "access granted". In other case it will show "access denied". A register called log.txt, stores all system activity, gathering the data about users that are accessed, date and time, as well as the IP address.

If the application allows the access to the server, the port will open, allowing traffic between TCP port and COM port selected. At this time, the user is connected to the console of the device, allowing to work without problems. However, if access has not been enabled, the port will remain closed.

Program Code for ServerSocketClientRead object

```
void __fastcall TRS232TCP10:: ServerSocketClientConnect(TObject
*Sender, TCustomWinSocket *Socket) {

     if (ServerSocket->Socket->ActiveConnections>1) {
        Socket->Close();
        return;
    }
    Socket->SendText("\r\nUser:");
    LName=true;
    LPass=false;
    Login=false;
    Password="";
    UserName="";
    ComPort1->Open();
}
```

Figure 6. Program Code for ServerSocketClientRead object.

For each port, there is available an application or service. The services are left enabled to start automatically when the Console server is started. Fig. 7, shows the diagram of our program.

Fig. 8 shows the operation of the protocol and message exchange that are registered when a remote user wants to access to any device connected to the console server.

In it, we can see that the user sends a service request to the server. The service request is a TCP connection to the server. The server sends the response to this request, to establish this connection.

When the connection is established, the server sends the request for username to remote user. In this case, the remote user will respond with a new message, leading the user name information. The server sends the request for password that the user must respond. Finally, the server sends the message of action to take. If the comparison of the characters gathered coincides with one of the entries stored in the file of access, the connection is allowed. Otherwise, access is denied and the application closes the connection.

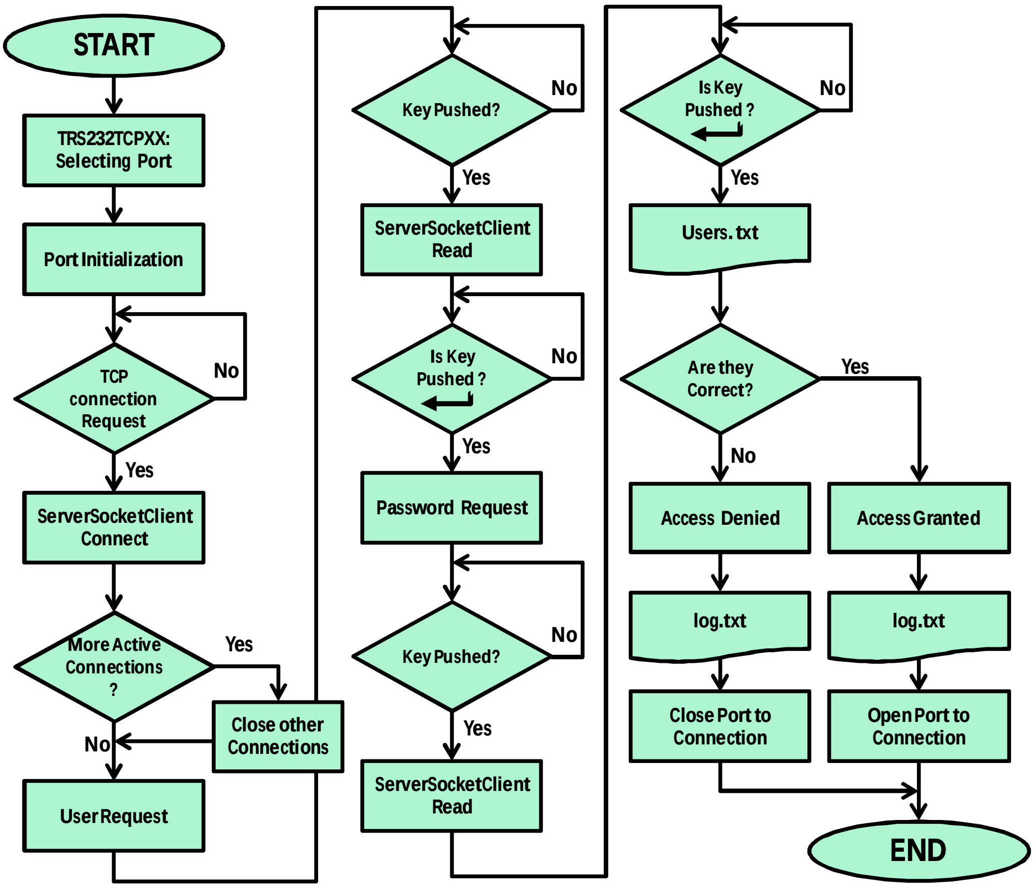


Figure 7. Flow diagram of program

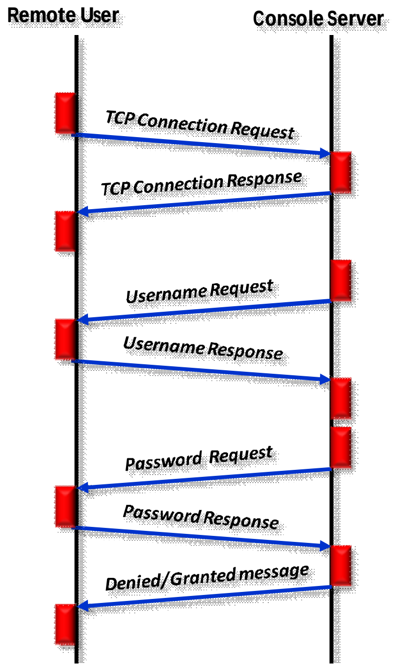
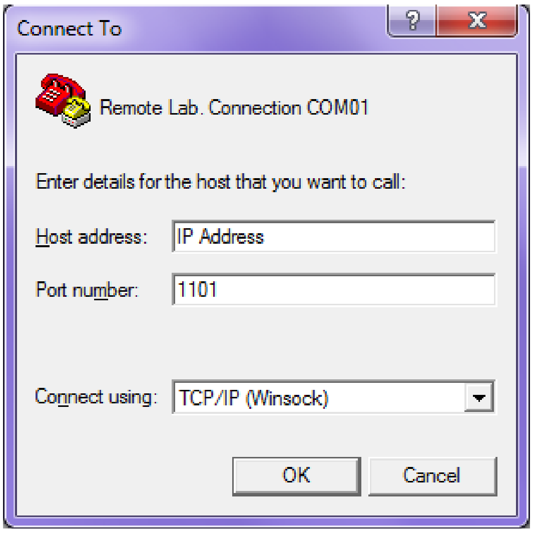


Figure 8. Protocol Operation.



Remote Lab. Connection COM01

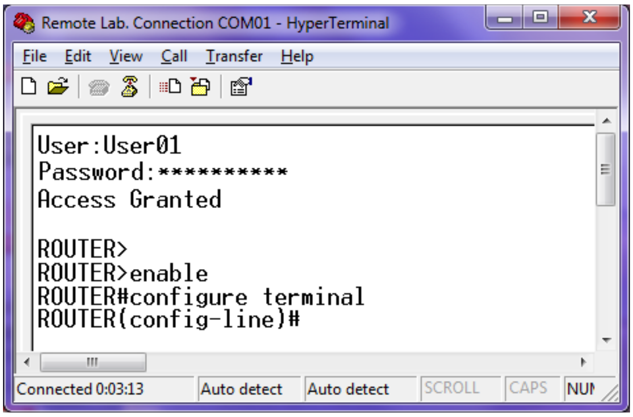
Enter details for the host that you want to call:

Host address: IP Address

Port number: 1101

Connect using: TCP/IP (Winsock)

Figure 9. Connection configuration.


Remote Lab. Connection COM01 - HyperTerminal
File Edit View Call Transfer Help
Connected 0:03:13
Auto detect
Auto detect
SCROLL
CAPS
NUM

```
User:User01
Password:**********
Access Granted

ROUTER>
ROUTER>enable
ROUTER#configure terminal
ROUTER(config-line)#
```

Figure 10. Hyperterminal connection with a router.

Finally, we activate all the services to run when the OS starts. We tried to connect to port COM01. To do this, it is used the Hyperterminal and the application is configured to connect to the COM01 port. To do this, we will define the IP address, the TCP port, which we have defined for COM01 and specify that the connection is done using the TCP protocol. Fig. 9 shows an image of the program

Fig. 10 shows a picture of access to a router. In it, we can see how the application requests the user name and password. As the access is allowed, we can directly access the router configuration.

VI. Conclusion

In this paper, we presented the implementation of a console server for remote management of network devices that can be found in real teaching laboratories. This low cost solution has been designed with the objective of enabling students to do practices on real devices. In addition, the remote access allows students work more time with real devices.

This system could be used also to control applications and devices over the Internet and favor the development and evolution of the field of IoT.

References

[1] J. Lloret, J. R. Diaz, J. M. Jiménez, and M. Esteve. “Aprendizaje Colaborativo en Profesionales de Nuevas Tecnologías”. International Institute of Informatics and Systemics (IIIS). Revista Iberoamericana de Sistemas, Cibernética e Informática.Vol. 2, Iss. 2, July 2005.

[2] J. Lloret, J. R. Diaz, and J. M. Jiménez, “A Collaborative Learning and Evaluation Method In Telematics”, In procceedings of Sefi Annual Conference 2004, Valencia (Spain), September 8-10, 2004.

[3] S. Sendra, A. Canovas,, M. Garcia, and J. Lloret, "Cooperative assessment in the hands on skills of computer networks subjects", in proceedings of IEEE Education Engineering Conference (EDUCON 2010), Madrid (spain), April 14-16, 2010.

[4] M. Garcia, H. Coll, M. Edo, and J. Lloret, "Mixing collaborative learning techniques for practice evaluation in networking,", in proceedings of EAEEIE Annual Conference, Valencia (Spain), June 22-24, 2009.

[5] J. Lloret, J. M. Jiménez, F. Boronat, J. Tomás, and J. R. Díaz. “Utilización de diversas metodologías didácticas para desarrollar las habilidades de los estudiantes de Ingeniería Técnica de Telecomunicaciones”, in proproceedings of Congreso Internacional de Docencia Universitaria e Innovación (CIDUI 2006). Barcelona (Spain), July 5-7, 2006.

[6] J. Lloret, J. M Jimenez, J R Diaz, and G. Lloret, “A remote Network laboratory to improve University classes”, in proceedings of the 5th WSEAS/IASME International Conference on ENGINEERING EDUCATION (EE'08), Heraklion (Greece), July 22-24, 2008

[7] Netlab Home page. Available at: http://cisco.cit.ie/Netlab.html. [Last access: May 15, 2012]

[8] J. Lloret, J. R. Diaz, and J. M. Jiménez, “Creation and Development of an E-Learning Formative Plan”, In procceedings of Sefi Annual Conference 2004, Valencia (Spain), September 8-10, 2004.

[9] Quatech ESC-100D board features. Available at: http://www.dpieshop.com/quatech-esc100d-pci-eight-port-rs232-serial-board-db25-p-204.html [Last access: May 15, 2012]

[10] Sunix's 4079T board features. Available at: http://www.sunix.com.tw/product/4079t.html [Last access: May 15, 2012]

A new Strategy to Improve the Pathfinding in Wireless ad-hoc Networks

Andreas Redmer
Department of Computer Science
University of Rostock
Rostock, Germany
Email: andreas.redmer@uni-rostock.de

Andreas Heuer
Department of Computer Science
University of Rostock
Rostock, Germany
Email: heuer@informatik.uni-rostock.de

***Abstract*—In link-state computer networks it is usual that every node knows the topology of the entire network and can make the routing decisions based on that. One of the protocols in use is OLSR. The OLSR routing protocol implements the algorithm of Dijkstra to find the shortest paths from the nodes to the gateways of the network. For that purpose, Dijkstra's algorithm has to be executed k times, while k is the number of gateways. In this paper, we present a strategy that generalizes all gateways to one gateway. We call this the General Gateway Strategy. Using this, the Algorithm of Dijkstra has to be executed only one time, which significantly increases the performance of the overall algorithm to find the shortest paths to the gateways.**

***Keywords*-Wireless mesh networks; Ad hoc networks; Wi-Fi; Shortest path problem.**

I. Introduction

In link-state computer networks it is usual that every node knows the topology of the entire network and can make the routing decisions based on that. One of the common protocols is OLSR (Optimized Link State Routing). It is a protocol for link-state routing in ad-hoc networks and is described in RFC 3626 [1].

The discovery of topology is specified in OLSR by two kinds of messages: HELLO- and Topology-Control (TC) Messages. The discovery of neighborhood is done by HELLO-Messages, which contain the current direct neighborhood of the sending node. TC-Messages contain information of the entire network, encompassing all nodes and routes to these nodes.

By the propagation of TC-Messages in the network, every node can derive the network topology and can make routing decisions using that knowledge. The quality of the link between two nodes is described by two parameters: link quality (LQ), which is the quality from the current node to the neighbor and neighbor link quality (NLQ), which is the reverse direction. These values are not necessarily equal for the same link.

Some of the nodes in the network share their internet connection and make it available to the rest of the network. These nodes are called gateways. Every node must be able to determine the shortest path to the nearest gateway.

OLSR uses the Algorithm of Dijkstra [2] to find the shortest paths between the nodes. The general problem is, that the Dijkstra Algorithm must always be executed k times, where k is the number of gateways.

Originally Dijkstra's Algorithm was described in [2] to find the shortest path between two nodes in a graph (single-pair shortest path). That means the algorithm will terminate as soon as the shortest path is found. For that purpose it is not necessary to explore the entire graph. In OLSR the Algorithm of Dijkstra is used to determine the shortest path from one gateway to all the other nodes (single-source shortest path). So, the algorithm explores the entire graph and can not be terminated sooner. This implies that the algorithm will always have the runtime that is described by the worst-case complexity. It is important to know that many other improvements of the Dijkstra Algorithm (e.g., the A*-Algorithm [3]) are only aiming on terminating sooner, by finding the destination node faster. So, these improvements are not applicable for the needs in ad-hoc wireless routing protocols. The network-graph can be described as a weighted directed graph (V, E) with

- $V := \{$a set of nodes$\}$
- $E := \{$a set of edges$\}$
- $G(\subseteq V) := \{$a set of gateways$\}$
- $(a, b) \in E$ (with $a \in V$ and $b \in V$) describing an edge from a to b
- $f : E \rightarrow \mathbb{R}$ a function describing the asymmetric distance between two nodes. Thereby the value LQ is defined as $f((a, b))$ and NLQ as $f((b, a))$.

Modern implementations Dijkstra's Algorithm use a Fibonacci-Heap [4] to store the nodes, which reduces the amortized complexity of the worst case to

$$O(|V| \cdot log|V| + |E|) \tag{1}$$

Based on that, the complexity of the procedure to find all the shortest path from all gateways to all nodes is

$$O(|G| \cdot |V| \cdot log|V| + |E|) \tag{2}$$

The strategy described in the following section removes the factor $|G|$ and so reduces the complexity significantly. After that, we present the experimental results, that show the actual enhancements of the new strategy. In the last section, a conclusion and future work is given.

II. A NEW STRATEGY

Figure 1 shows a network graph with four gateways (G_1, G_2, G_3 and G_4) printed in blue. The red nodes are non-gateway nodes. The labels of the edges represent examples of the metric, which is used to describe the distance between two nodes. In this case, a simple additive metric is applied. So, higher numbers represent a higher distance and a worse connection. Low values stand for a good connection. Dijkstra's algorithm must be executed four times there. Each run uses one of the gateways as initial node. In the end all four paths that lead from a node to a gateway will be compared and only the shortest one returned. Each run considers the whole graph. In the example it would be determined four times, that the shortest path from C to E is always across D and has always the costs 2. So, calculations are repetitive.

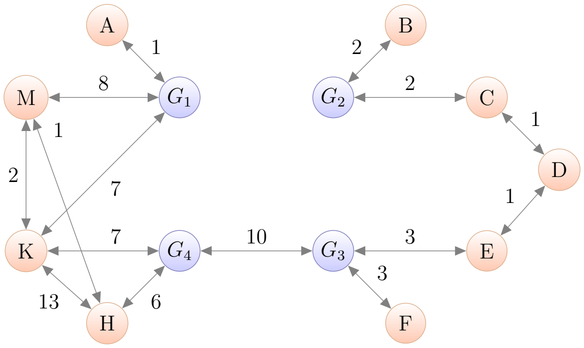


Figure 1. An example for an unmodified network graph

We developed an idea to solve this problem of executing the same calculations multiple times on the same graph. The basic principle for this, is the fact, that a shortest path to a gateway never crosses another gateway. The theoretical perfection would be a graph, where all the gateways are centralized in the network, there are no nodes between the gateways and all the gateways can reach each other with the costs 0. In this case it would be possible to combine all gateways to one gateway. So it would only be necessary to find the shortest paths to this generalized gateway.

In practice that is not the case, because the gateways are spread all over the network and have always costs (> 0) to reach each other. But the direct edges between the gateway are not relevant for our problem, since we have already noticed that a shortest path to a gateway, never crosses another gateway. It is also not necessary to find a path from a gateway to another gateway. So, all the irrelevant edges can be substituted by 0-Edges[1] and so all gateways will be connected by 0-Edges. It does not matter if there has been an edge existing between these gateways before, or not. That also means, that the entire graph does not have to be connected. It is also possible to apply this technique to separated sub-graphs, whereat each sub-graph contains at least one gateway. It also does not matter how many 0-Edges are inserted. It only must be ensured, that every gateway can reach each other gateway with the cost 0. In Figure 2 these 0-Edges are added and the generalized gateway is marked with a dashed line.

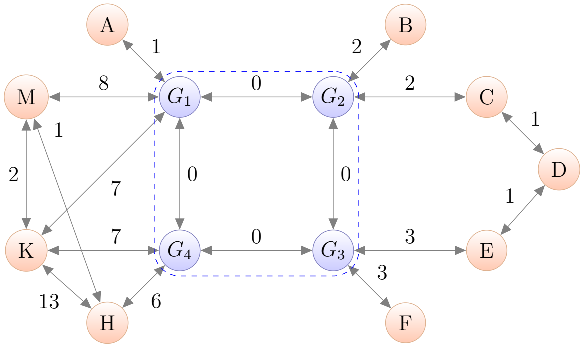


Figure 2. A network graph with generalized gateway

For the newly created graph, the following procedure will be applied:

1) Select a random gateway as initial node.
2) Execute Dijkstra's Algorithm one time.
3) Remove the 0-Edges from the result[2].

For instance, one uses G_1 as initial node in Figure 2. Firstly, Dijkstra's algorithm returns shortest paths from every node to G_1. Amongst others the route:

$$G_1 \leftarrow G_2 \leftarrow C \leftarrow D$$

will be returned for node D. In the beginning of each route there can be several gateways now. All the 0-Edges and the leading gateways have to be removed in the final result e.g., it would be the route:

$$G_2 \leftarrow C \leftarrow D$$

for the node D. This is the shortest path to any available gateway. We called this method General-Gateway-Strategy (GGS). This strategy returns the same result as the multiple execution of Dijkstra's algorithm and the additional selection of the closest gateway (the shortest of all the shortest paths).

III. RESULTS

To evaluate the performance of the GGS we generated one million weighted directed graphs randomly. Each graph was connected and had the characteristics

- $100 \leq |V| \leq 1000000$ and
- $|G| \in \{1, 2, 4, 8\}$.

We implemented the Dijkstra Algorithm using a Fibonacci Heap in Java. For this purpose, a desktop PC was equipped

[1] Exact would be "identity-element-edges", which means that always the identity element of the metric should be used. In additive metrics 0, in multiplicative metrics 1, etc.

[2] All 0-Edges in the beginning of a route between two gateways.

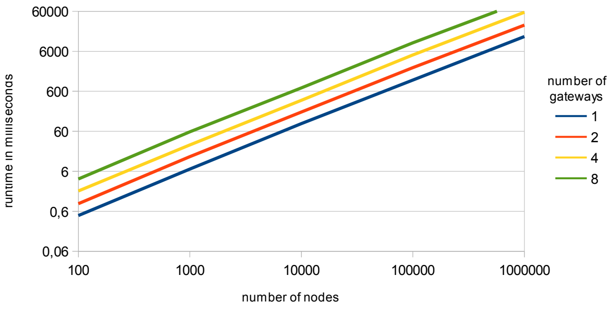


Figure 3. Runtimes without General Gateway Strategy

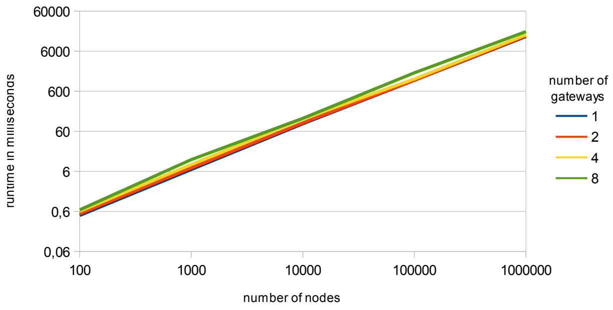


Figure 4. Runtimes with General Gateway Strategy

with a 64 bit dual core 2.2 GHz CPU, 128 kB L1 CPU cache, 512 kB L2 CPU cache and 4 GB DDR II RAM (800 MHz). We ran the Java Virtual Machine OpenJDK 1.9.9 [5] on the operating system Ubuntu Linux 10.04.3 LTS [6] with the 2.6.32-32 Linux Kernel [7].

To measure the runtime we used the Java API Method `System.nanoTime()` [8] directly before and after our algorithm execution. The algorithm was executed several times to avoid deviation due to caching or just-in-time-compilation. The measured times in milliseconds of the endpoints of the test-interval are presented in Table I.

The results of all test runs are plotted in Figure 3. To plot the runtimes (y-axis), a logarithmic axis was chosen in order to provide a better scalability in the plot. The x-axis shows the number of nodes used in the graph. There are four lines representing the number of gateways in the graph ($|G| \in \{1, 2, 4, 8\}$). Based on the complexity of Dijkstra's Algorithm (Formula 1) the results grow as expected exponentially by increasing the number of nodes. Increasing the number of gateways adds a factor growth to the runtimes.

Figure 4 shows the same kind of diagram as Figure 3, but here are the results of the GGS plotted. As expected, all the lines are now very close to the blue line in Figure 3, which represents only one execution of the Dijkstra Algorithm. There is only a very small increment resulting from the raising amount of gateways, which is due to the last step of the GGS. This last step is the removal of the 0-Edges in the final result. In efficient implementations this step might be included in the post-processing of the result and so it might not be noticeable anymore.

As it can be seen in Table I, there are no considerable differences between the runtimes, in networks with only one gateway. But in the case of one million nodes and eight gateways, the GGS takes practically only around 17% (theoretically 12.5%) of the runtime compared with the conventional method that executes the Dijkstra Algorithm eight times. That makes it 5.9 times faster than before. Generally, there is always a small overhead, which results from the postprocessing of the final result. The theoretically possible improvements by the factor k $(= |G|)$ can not be reached.

Table I
Runtimes of Endpoints of the Interval in Milliseconds

	without GGS		with GGS	
nodes	1 gateway	8 gateways	1 gateway	8 gateways
100	0,47	3,86	0,48	0,66
1000000	14022,50	108703,70	13934,73	18494,70

IV. Conclusion and Future Work

We have presented a strategy that generalizes all gateways in an ad-hoc wireless network to one gateway. We call this the General Gateway Strategy. Using this, the Algorithm of Dijkstra has to be executed only one time, which significantly increases the performance of the overall algorithm to find the shortest paths to the gateways.

Our new strategy improves the overall complexity by the factor k (number of gateways). This value can be very high for larger scaled mesh networks. Our measurements showed, that even for small scaled networks (not more than eight gateways), there are improvements in the performance of more around 600%. For larger mesh networks a higher enhancements can be expected. This highly depends on the degree of connectivity of the network graph and on the amount of gateways.

In the future, the GGS can be practically used in data analysis algorithms. In one of our research projects we have captured the topology data over more than one year in the wireless mesh-network. For that purpose one of our nodes in the networks saved the topology data one time per minute into a relational database. To analyze this data it is always the first step to use the Dijkstra Algorithm to calculate the current routes at a particular instant of time. This enables the analyst to perform a risk analysis, a bottleneck analysis, evaluate the existence and quality of alternative routes and more. This information can be used for further network planning, to find measures for the importance of nodes and edges and to enhance the network quality. In the past, we passed other tries to increase the performance of the data analysis, e.g., by using cloud computing [9]. The approach presented in this work, brings an additional performance boost to the data analysis.

Furthermore, the GGS can directly be implemented on devices, which implement the OLSR routing protocol. Since it is fully compatible to the conventional implementations of OLSR, updated devices can be used in existing networks without any restrictions. This will lead to reduced CPU load on the routing node and so a higher throughput of user data.

REFERENCES

[1] T. Clausen, P. Jacquet, C. Adjih, A. Laouiti, P. Minet, P. Muhlethaler, A. Qayyum, and L. Viennot, "Optimized Link State Routing Protocol (OLSR)," *Network Working Group Request for Comments : 3626 Category : Experimental*, 2003.

[2] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.

[3] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems Science and Cybernetics SSC*, vol. 4, no. 2, pp. 100–107, 1968.

[4] M. L. Fredman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms," *J. ACM*, vol. 34, pp. 596–615, July 1987. [Online]. Available: http://doi.acm.org/10.1145/28869.28874

[5] OpenJDK, "Openjdk website," 2012. [Online]. Available: http://openjdk.java.net/ [retrieved: July, 2012]

[6] Canonical, "Official ubuntu linux 10.04.3 download page," Canonical Group Limited, 2012. [Online]. Available: http://mirror.eftel.com/ubuntu-dvd/10.04.3/ [retrieved: July, 2012]

[7] L. Torvalds, "The linux kernel archives," 2012. [Online]. Available: http://www.kernel.org/ [retrieved: July, 2012]

[8] Oracle, "Java api online documentation of the class system," 2011. [Online]. Available: http://docs.oracle.com/javase/1.5.0/docs/api/java/lang/System.html [retrieved: July, 2012]

[9] T. Mundt and J. Vetterick, "Network topology analysis in the cloud," in *ICOMP'11 - The 2011 International Conference on Internet Computing*, July 2011.

Codeset Overlay for Complementary Code Keying Direct Sequence Spread Spectrum

Farzad Talebi
Electrical Engineering Department
University of Notre Dame, IN 46556
e-mail: ftalebi@nd.edu

Thomas G. Pratt
Electrical Engineering Department
University of Notre Dame, IN 46556
e-mail: tpratt@nd.edu

***Abstract*—A method to improve communications reliability of complementary code keying in IEEE 802.11b WiFi systems is proposed that involves the use of an *overlay* signaling dimension that preserves the underlying data rate and power spectrum characteristics of the WiFi signals. The specific overlay technique expands the number of code sets and selects the codeset in a data-dependent manner to provide redundancies that improve the error rate performance of the underlying communications scheme. Sequential detection is employed at the receiver, where symbols are recovered using the redundancies embedded in the overlay. The maximum likelihood sequential detection criterion vectors adopted for the scheme correspond to different memory depths, and in the best case achieve gains that are in excess of 3 dB without any rate loss or bandwidth expansion relative to to the original complementary code keying scheme, and the method can be implemented with reasonable increases in computational complexity.**

Keywords–WiFi; CCK; Spread Spectrum; IEEE 802.11b.

I. Introduction

IEEE 802.11b [1] is based on a direct sequence spread spectrum (DSSS) scheme which uses complementary code keying (CCK) in the higher data rate modes. In the highest data-rate scheme, which achieves 11 Mb/s, each symbol conveys 8 bits– 2 bits are determined by the QPSK symbol while the other 6 bits are conveyed from a set of 64 spreading sequences which we refer to as a codeset.Besides its use in indoor wireless LAN applications, CCK technology has been used for other applications [2]-[5] because it is more robust to harsh propagation conditions than the 802.11a or 802.11g [3].

In this paper, we consider the use of additional CCK codesets in a data-dependent manner to impart redundancies for the protection of the data associated with the underlying communications. The overlay encoding technique is complemented by a receiver processing that employs sequence estimation techniques to leverage the encoded redundancies for enhanced communications reliability. The technology provides a range enhancement that would improve the reach of mobile ubiquitous systems.

The idea of using different code sets has been proposed before, for example to support interference mitigation [6]. In the presented scheme, an overlay based on multiple code sets is used as a mechanism to convey data redundancies for enhanced reliability of the underlying communications. The codesets used in the scheme are designed using a search technique to find codesets exhibiting large mutual distance profiles between spreading sequences from different codesets. In the encoding scheme, the codesets are selected in a data-dependent manner and the resulting redundancies are exploited at the receiver using Maximum Likelihood (ML) sequence detection. We show that a signal to noise ratio (SNR) gain in excess of 3 dB can be obtained using seven extra codesets, each with 64 codes, when a memory length of 3 symbols is employed, in comparison to the original CCK scheme at the cost of increased computational complexity.

The remainder of the paper is organized as follows. Section II describes the system model and includes a description of the stochastic search method employed to synthesize the extra code sets. Section III discusses the proposed data transmission scheme that uses a recursion for selecting codeset indices in a data dependent manner. Section IV describes the sliding window ML detection that is implemented for different memory depths and sliding window sizes, including memoryless (F), 2-symbol detection (Y), 3-symbol detection (Z), and 4-symbol detection (X). Section V discusses simulation results and analytic performance characterizations for small window detection vectors. Finally, the conclusions of the research are presented in Section VI.

II. System Model

A. CCK Modulation

In its highest data rate mode, IEEE 802.11b uses direct sequence spread spectrum (DSSS) in an 8-chip per symbol spreading scheme where the chip rate is

11Mchips/s and where the code is drawn from a codeset comprising 64 codes. The spreading sequences in this pre-defined set are given by (1).

$$\begin{aligned} C_1 &= [C_1(1), C_1(2), \ldots, C_1(8)] \qquad (1) \\ &= [e^{j(\phi_2+\phi_3+\phi_4)}, e^{j(\phi_3+\phi_4)}, e^{j(\phi_2+\phi_4)}, \ldots \\ &\quad -e^{j(\phi_2+\phi_4)}, e^{j(\phi_2+\phi_3)}, e^{j(\phi_2+\phi_3)}, -e^{j(\phi_2)}, 1] \end{aligned}$$

The phases ϕ_2, ϕ_3, and ϕ_4 are selected from the set $\{0,\pi/2,\pi,3\pi/2\}$, and the resulting codes within the set are designated by $C_1(i),\ 1 \leq i \leq 64$.

B. Code Set Expansion

Few design methods for creating additional CCK code sets have been proposed in literature. In one reference, Cotae [8] has proposed a method for designing orthogonal sets of spreading sequences for use in overloaded multi-cell code division multiple access (CDMA) systems. The approach allowed for complex valued spreading sequence designs according to a total weighted square correlation criterion. Furthermore, Xu et al. [9] provides a method to mitigate co-channel interference in cellular communication systems using nearly orthogonal CCK codesets at different cells, but it does not provided any intuition or criteria to help understanding the process of designing the nearly orthogonal codesets. Such methods do not appear to have direct application to our problem and so we proceed with a stochastic search method.

The distance between code sequences will be an important design metric in the synthesis of additional code sets since in additive white Gaussian noise (AWGN) ML detection can be reduced to minimum distance detection [10]. Our goal is to achieve distance profiles comparable to those associated with the original code set. The distance profile between the first and other sequences in the original codeset is plotted in Figure 1.

The method we use to design extra codesets $C_2, \ldots, C_8$ involves a stochastic search to minimize a cost function $f(C_2, \ldots, C_8)$ which is proportional to the probability of error [10]. This cost function is associated with a system using codesets C_2 to C_8, where the cost function is the uniform average of error probabilities associated with all spreading sequences within all codesets:

$$f(C_2, \ldots, C_8) = \sum_{\substack{2 \leq m,n \leq 8 \\ 1 < i,j < 64}} Q\left(\sqrt{\|c_m(i) - c_n(j)\|^2}\right). \qquad (2)$$

In (2), the function Q represents the normal CDF, and $c_m(i)$ represents the i-th code sequence from codeset C_m.

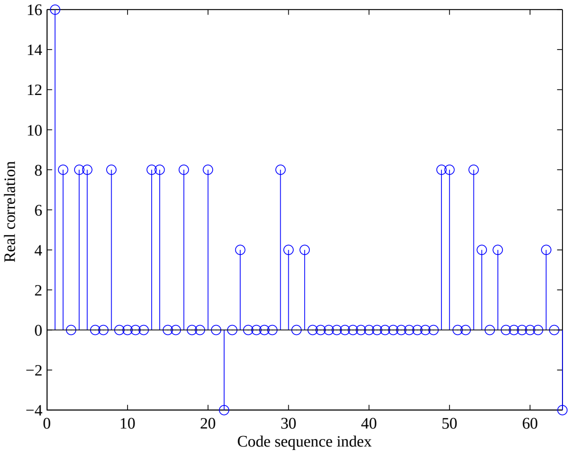


Fig. 1. Cross correlation between the first code set and all 64 original CCK code sequences

Minimization of $f(C_2, \ldots, C_8)$ is achieved via a random search which assumes that permutated and rotated versions of the original CCK code set C_1 can be used as $C_2, \ldots, C_8$. Seven code sets resulting from such a search are given by :

$$\begin{aligned} C_2 &= [C_1(6), -C_1(4), -jC_1(5), -C_1(3), \ldots \\ &\quad -jC_1(8), -jC_1(7), -jC_1(2), -jC_1(1)] \\ C_3 &= [jC_1(5), C_1(8), C_1(7), -jC_1(6), \ldots \\ &\quad -jC_1(2), -jC_1(4), jC_1(1), -jC_1(3)] \\ C_4 &= [-jC_1(2), C_1(8), -C_1(4), -C_1(3), \ldots \\ &\quad -C_1(5), -jC_1(6), C_1(1), -jC_1(7)] \\ C_5 &= [-C_1(6), C_1(4), jC_1(2), -jC_1(3), \ldots \\ &\quad -C_1(1), jC_1(5), -jC_1(7), -jC_1(8)] \\ C_6 &= [-C_1(6), -jC_1(8), -C_1(2), -jC_1(1), \ldots \\ &\quad -C_1(5), -C_1(3), -jC_1(4), jC_1(7)] \\ C_7 &= [-C_1(7), -C_1(4), -jC_1(3), C_1(2), \ldots \\ &\quad C_1(1), -C_1(5), C_1(6), -C_1(8)] \\ C_8 &= [-C_1(7), jC_1(8), -jC_1(6), jC_1(5), \ldots \\ &\quad -C_1(3), -jC_1(4), -C_1(2), -C_1(1)] \qquad (3) \end{aligned}$$

III. Data Dependent Code set Selection

Now, we describe the method proposed for adopting extra code sets into the spreading scheme. This approach involves selection of a data-dependent code set index for each transmitted symbol. Let us assume a packet of length $N_p = 1000$ symbols, which corresponds to a medium-length packet length in WLAN systems. Let $S(k)$ designate the code set index for symbol k in the packet, and let $N(k)$ represents the code sequence index for the k-th symbol, where the code sequence is chosen from the members in $S(k)$. Let us assume that the first and second code set and code number indices are pre-determined for each packet (for the sake of initialization). Subsequent code set indices are selected through a recursion, which in the following example has a memory depth of three symbols:

$$\begin{aligned} S(k) \quad &= \quad \lceil mod(\lceil \frac{N(k-1)}{N_R} \rceil + S(k-1) + \ldots \\ & \qquad N(k-2) + S(k-2), N_R) \rceil \end{aligned} \tag{4}$$

Here, $mod(a, b)$ is the remainder of a divided by b, and N_R is the number of code sets (assumed to be 8 in our analysis). We emphasize that $N(k)$ will only depend on the k-th symbol in the data packet and is not affected by priori code numbers or code set indices. The recursion for choosing the next code set index helps to discriminate against false peaks in the detection process by partitioning indices into groups of N_R, where for a given code sequence (i.e., the data) each group results in the use of a different code set index for the next symbol, thereby achieving data-dependent codeset selection.

IV. Multi-Symbol Processing

To leverage the redundancies achieved through the overlay, multi-symbol detection processing is used at the receiver. The method amounts to a greedy implementation of an exhaustive search to find the best code word index according to a minimum distance (or maximum correlation) criterion which is optimal in a ML detection sense for the case of AWGN, which we assume in our analysis. When the transition between states is unrestricted (as in our case) minimum distance detection of a finite number of possible symbols should be achieved using exhaustive search because a structure (channel code) ruling the sequence of selected indices does not exist. The greedy approach can be implemented in a sliding window manner using different window lengths. Because the inherent memory in the encoding process in (4) is three, we anticipate good performance gain when using detection windows of length 3 or more, which is verified in the simulation results.

Symbol-by-symbol detection corresponds to the conventional detection approach that would be employed in an 802.11b system, and serves as a baseline against which to compare the performance of the overlay approach. Symbol-by-symbol detection employs a metric $F(N_k = j_k)$, which represents the correlation of the received symbol at time k with pre-known symbol N_k-th in code set $S(k)$. At time k the receiver knows $S(k)$ because it depends on previous symbols N_{k-1} , N_{k-2} , $\cdots$. The metric F is computed as follows:

$$F(N_k = j_k) = \sum_{i=1}^{8} 2Re(c_{j_k}(i) c_{rx,k}(i)^*) \tag{5}$$

where the code word c_{j_k} is the j_k-th symbol from the code set defined by the previous symbols, and $c_{rx,k}$ represents the received symbol at time k.

The metric for the two-symbol detection scheme can be written as:

$$\begin{aligned} Y(j_{k-3}) \quad &= \quad F(N_{k-3} = j_{k-3}) + \\ & \qquad \max_{j_{k-2}} \{F(N_{k-2} = j_{k-2} | N_{k-3} = j_{k-3})\} \end{aligned} \tag{6}$$

which reflects the correlation between symbols received at times $k-3$ and $k-2$ based on the decoded estimated of the previous symbols. In (6), $F(N_k = j_k | N_{k-1} = j_{k-1})$ refers to the correlation of the received symbol at time k with the symbol N_k set equal to the j_k-th symbol in $S(k)$ conditioned on the premise that the symbol at time $k-1$, N_{k-1}, is equal to the j_{k-1}-th symbol in code set $S(k-1)$.

By following the same approach, a three-symbol detection metric vector Z can be computed that reflects the correlation between received symbols at times $k-3$,$k-2$ and $k-1$ and is given by (7):

$$\begin{aligned} Z(j_{k-3}) \quad &= \quad F(N_{k-3} = j_{k-3}) + \\ & \qquad \max_{j_{k-2}} \{F(N_{k-2} = j_{k-2} | N_{k-3} = j_{k-3}) + \\ & \qquad \max_{j_{k-1}} \{F(N_{k-1} = j_{k-1} | N_{k-2} = j_{k-2}, \ldots \\ & \qquad N_{k-3} = j_{k-3})\}\} \end{aligned} \tag{7}$$

In a like manner, a four symbol detection metric is achieved using (8):

$$\begin{aligned} X(j_{k-3}) \quad &= \quad F(N_{k-3} = j_{k-3}) + \\ & \qquad \max_{j_{k-2}} \{F(N_{k-2} = j_{k-2} | N_{k-3} = j_{k-3}) + \\ & \qquad \max_{j_{k-1}} \{F(N_{k-1} = j_{k-1} | N_{k-2} = j_{k-2}, \ldots \\ & \qquad N_{k-3} = j_{k-3}) + \\ & \qquad \max_{j_k} \{F(N_k = j_k | N_{k-1} = j_{k-1}, \ldots \\ & \qquad N_{k-2} = j_{k-2}, N_{k-3} = j_{k-3})\}\}\} \end{aligned} \tag{8}$$

Note that in the above derivations, it is assumed that symbols before time $k-3$ are known, which means that symbol errors will propagates through the sequential demodulation of the packet. However, since we are interested in packet error rate (PER) any error is sufficient to result in loss of the packet.

V. Performance Analysis and Simulation Results

A. Analytic Performance Analysis

The symbol error rate associated with the symbol-by-symbol detection over additive white Gaussian noise

TABLE I
SQUARED DISTANCE AND MULTIPLICITY

i	**Squared distance** (d_i^2)	**Multiplicity** m_i
1	8	12
2	12	6
3	16	43
4	20	2

TABLE II
SECOND ORDER SQUARED DISTANCE AND MULTIPLICITY

i	**Squared distance** (d_i^2)	**Multiplicity** m_i
1	8	4
2	10	0.05
3	12	0.5
4	14	1.75

(AWGN) channel can be written as [10]:

$$P_e = \sum_i m_i \; Q(\sqrt{d_i^2}) \tag{9}$$

where m_i represents the multiplicity of neighbors at distance d_i, which is given in Table I for the original CCK code set. The corresponding PER for symbol by symbol detection and for two-symbol detection matches results corresponding to an assumption of independent symbols and can be written as (10):

$$\text{PER} = 1 - (1 - P_e)^{N_p} \approx N_p P_e \tag{10}$$

Using the distance-multiplicity data in Table I analytically-determined results for the PER of memoryless detection using F is presented in Figure 2. However, to evaluate the performance of scheme when using two-symbol detection with vector Y as metric, second order distance distributions of codewords are required that can be obtained numerically thanks to the short length and small number of codewords. This distance distribution is given in Table II; by using the distance distribution, the analytic PER performance of schemes Y can be obtained.

B. Simulation Results

Figure 2 shows the PER performance of different simulated detection schemes. The simulation assumed packets consisting of 1000 symbols, where each symbol (i.e., code) carries 6 bits of information where the codes have 8 chips. The SNR is defined as E_c/N_0, where E_c represents chip energy, and is assumed to be unity, and N_0 is the variance of complex Gaussian noise with variance $N_0/2$ on the in-phase and quadrature (I and Q) channels.

TABLE III
RUNTIME AND MEMORY USAGE FOR PACKET LENGTH=1000

Vector	**Runtime**	**Memory Usage (Bytes)**
F	54 msec	512
Y	195 msec	2560
Z	2.14 sec	100352
X	83 sec	6359040

Performance curves are plotted for simulation results associated with detections based on the F, Y, Z and X vectors. We observe that Y, Z, and X yield approximately 0.5 dB, 2.5 dB, and 3 dB gains. As Table II shows, using vector Y (i.e., using two consecutive symbols for detection) reduces the multiplicity of nearest neighbors at distance $d^2 = 8$, so the gain is not expected to be large because the minimum distance has not been reduced. Though we have not reported the distance distribution for the case of using vector Z, simulation results indicate that the minimum distance is increased to $d^2 = 14$, reflected by the SNR gain (about 2.5 dB) relative to F, which represents the original CCK scheme used in IEEE 802.11b. Detection using vector X results in a gain exceeding 3 dB in comparison to original scheme, but is considerably more complex.

The runtime and memory requirements of the described detection schemes are shown in Table III, which indicates a memory usage growth with factors 5, 39 and 64 for each additional memory depth, while the runtimes grow with factors 4, 10 and 40 respectively. Note that the memory usage is expressed in terms of bytes in Table III. Since the runtime and memory usage are dependent on the implementation, the numbers presented are based on our specific implementation using MATLAB on a computer with a 2GHz Dual Core processor. Implementations using FPGA-based hardware that better exploit parallel calculations are anticipated to yield much better computational efficiencies.

Figure 3 compares the required SNR to achieve a PER = 0.01 for different packet lengths. SNR differences between different schemes are seen to be independent of the packet length. However, the required SNR is seen to increase by approximately 1 dB when the packet size changes from 100 symbols to 2000 symbols. This can be of interest for mobile ubiquitous systems where packet sizes as small as 64 bytes might be used [11].

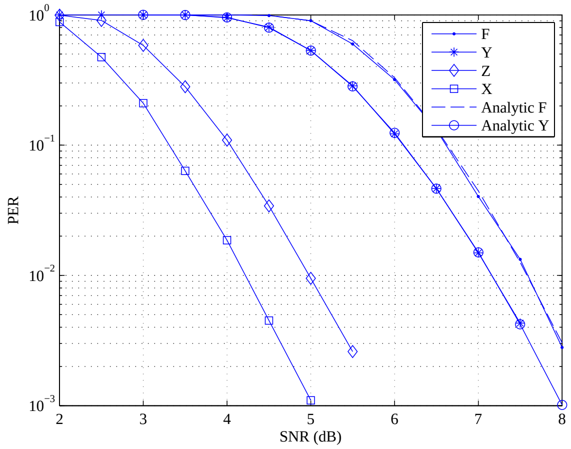


Fig. 2. PER performance of different detection schemes

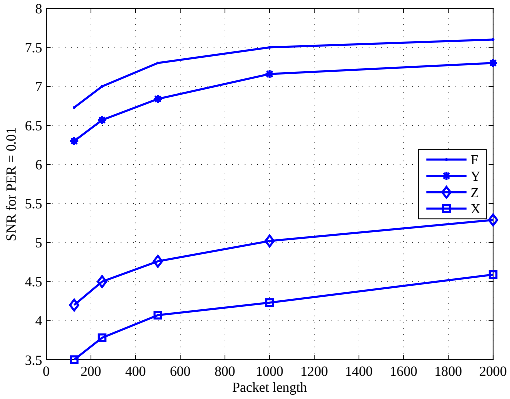


Fig. 3. Required SNR to reach PER = 0.01

VI. Conclusion And Future Work

The use of additional codesets configured as an overlay was shown to improve the error rate performance of an underlying CKK-DSSS signal. The scheme employed the extra code sets that are selected in a data-dependent manner to improve noise immunity without bandwidth expansion or data rate reduction. For encoding involving a recursion that uses three consecutive symbols, we have shown that different gains are achieved depending upon the memory depth assumed at the decoder, where up to more than 3 dB SNR gain is possible when a memory depth of three symbols is employed. The proposed method is seen to exhibit reasonable implementation complexities for most of the processing cases that were considered. Also, it has better performance at shorter block lengths which is achieved without rate loss and forward error correction coding.

A future step in this work is to reduce the computational complexity and memory usage of the detection scheme through algorithm implementation efficiencies and through more efficient numerical representations. For example, our analysis was conducted using double-precision floating point representations. However, other detection/decoding algorithms reported in literature have represented detection vector entries using quantization levels of 4 or 5 bits [12] without significant reduction in performance. Carefully designed quantization schemes can potentially achieve more efficient implementations while simultaneously maintaining a detection performance close to that of the unquantized scheme.

References

[1] Institute of Electrical and Electronics Engineers, “Standard 802.11b-1999, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Higher-Speed Physical Layer Extension in the 2.4 GHz Band,” Jan. 2000.

[2] P. Bhagwat, B. Raman, and D. Sanghi “Turning 802.11 inside-out,” ACM SIGCOMM Computer Commun. Rev., vol. 34, no. 1, pp. 33–38, Jan. 2004.

[3] K. K. Leung, M. V. Clark, B. McNair, Z. Kostic, L. J. Cimini Jr., and J. H. Winters, “Outdoor IEEE 802.11 cellular networks: radio and MAC design and their performance, IEEE Trans. Veh. Technol., vol. 56, no. 5, pp. 2673–2684, Sept. 2007.

[4] A. Silvennoinen, T. Karhima, M. Hall, and S. G. Haggman, “IEEE 802.11b WLAN capacity and performance measurements in channel with large delay spreads,” in Proc. 2004 IEEE Military Communications Conference, vol. 2, pp. 693–696, Oct. 2004.

[5] V. Arneson, K. Ovsthus, O. I. Bentstuen, and J. Sander, “Field trials with IEEE 802.11b-based UHF tactical wideband radio,” in Proc. IEEE Military Communications Conference vol. 1, pp. 493–498, Oct. 2005.

[6] X. Xiaohui, Z. Chao and L. Xiaokang, “Using different orthogonal code sets for CCK modulation to mitigate co-channel interference among WLANs,”, In Proc. IEEE International Symposium on Communications and Information Technology (ISCIT), vol. 2, pp. 885–888, 2005.

[7] T.-D. Chiueh and S.-M. Li, “Trellis-coded complementary code keying for high-rate wireless LAN system,” IEEE Communications Letters, vol. 5, pp. 191–193, May 2001.

[8] P. Cotae, “Multicell Spreading Sequence Design Algorithm for Overloaded S-CDMA,” IEEE COMMUNICATIONS LETTERS, VOL. 9, NO. 12, DECEMBER 2005, pp. 1028–1030

[9] Xiaohui Xu, Chao Zhang and Xiaokang Lin, “Using different orthogonal code sets for CCK modulation to mitigate co-channel interference among WLANs,” IEEE International Symposium on Information Technology and Communications, 2005. ISCIT vol. 2, pp. 885–888.

[10] Todd K. Moon, “Error Correction Coding Mathematical Methods and Algorithms,” John Wiley & Sons, Inc. 2005.

[11] C.-K. Toh, M. Delwar, and D. Allen, “Evaluating the Communication Performance of an Ad Hoc Wireless Network,” IEEE transactions on wireless communications, vol. 1, no. 3, July 2002.

[12] I.M. Onyszchuk, K.-M. Cheung, O. Collins, “Quantization loss in convolutional decoding,” IEEE Transactions on Communications, vol. 4, no. 2, pp. 261–265, 1993.

Analysis and Compensation for HPA Nonlinearity with Neural Network in MIMO-STBC Systems

Oussama B. Belkacem†, Mohamed L. Ammari†, Rafik Zayani† and Ridha Bouallegue†
belkacemoussema@supcom.rnu.tn mlammari@ele.etsmtl.ca rafik.zayani@supcom.rnu.tn ridha.bouallegue@supcom.rnu.tn
†SUPCOM, Innov'Com Laboratory, Carthage University, Tunis, Tunisia

***Abstract*—In order to provide high data rate over wireless channels and improve the system capacity, Multiple-Input Multiple-Output (MIMO) wireless communication systems exploit spatial diversity by using multiple transmit and receive antennas. Moreover, MIMO systems are equipped with High Power Amplifiers (HPA). However, HPA causes nonlinear distortions and affect the receiver's performance. Since a few decades, Neural Networks (NN) have shown excellent performance in solving complex problems like classification, recognition and approximation. In this paper, we present a receiver technique based on NN schemes for the compensation of HPA non linearization in MIMO Space-Time Block Coding (STBC) systems. Specifically, we assess the impact of HPA nonlinearity and NN on the average symbol error rate (SER) and the error vector magnitude (EVM) of MIMO-STBC in uncorrelated Rayleigh fading channels. Computer simulation results confirm the accuracy and validity of our proposed analytical approach.**

***Index Terms*—MIMO(Multiple Input Multiple Output), HPA(High Power Amplifier), NN(Neural Network), SER(Symbol Error Rate), EVM(Error Vector Magnitude).**

I. Introduction

Wireless services are driven by the rising demand to provide high-speed data transmissions (several 100 Mbit/s). A common way to improve the system capacity is to increase the transmission bandwidth. Multiple-Input Multiple-Output (MIMO) has been proposed to develop wireless systems that offer both high capacity and better performance. It has been recognized as a key technology for 4G wireless communications [1], [2].

High-power amplifier (HPA) is a primary block of a wireless communication system, such us Travelling Wave Tube Amplifier (TWTA), Solid State Power Amplifier (SSPA) and Soft-Envelope Limiter (SEL). It operates between the modulator and radio frequency (RF) modules. However, HPA introduces nonlinear distortions to the transmitted signal when operating in nonlinear region [3]. Nonlinear distortions, including amplitude and phase distortions, are introduced into the transmitted symbols, which in turn can cause adjacent channel interference and power loss. These distortions degrade considerably the system performance.

Nonlinear HPA can be described by two kinds of models: memoryless models with flat frequency responses, and memory models with frequency-selective responses [4]. Memoryless HPA models, such as the TWTA, SSPA and SEL, are characterized by their amplitude modulation/amplitude modulation (AM/AM) and amplitude modulation/phase modulation (AM/PM) conversions [5]. On the other hand, HPA may be characterized by more realistic memory models, such as the Volterra, Wiener, Hammerstein and memory polynomial models [3], [4].

To improve the system throughput, the effect of HPA was analyzed in [3] for MIMO systems employing Orthogonal Space Time Block Coding (OSTBC). In [1], authors proposed an adaptive predistortion technique based on a feed-forward Neural Network (NN) to linearize power amplifiers such as those used in satellite communications. Authors in [6] extended the efficient NN Predistorter (NNPD) to MIMO-OFDM systems. In [7], NN technique has gained a great interest in nonlinear MIMO channel identification and authors proposed an efficient nonlinear receiver to compensate the joint effects of HPA nonlinearity and the impact of time-varying MIMO channels. In this paper, we focus on HPA nonlinearity on MIMO-STBC systems. we propose a NN compensator technique to enhance nonlinearity distortion at the receiver. For the outlined transmission chain, we derive the expressions for the average SER and EVM, which are valued for memoryless nonlinear HPA models, considering that the system operates under Rayleigh flat fading channel. The remainder of the paper is organized as follows: Section II introduces the MIMO system model with HPA nonlinearity, the NN scheme is revisited and explained. In Section III, we derive the exact SNR expression in presence of NN compensation scheme and evaluate the system performance in terms of SER and EVM in the case when knowledge

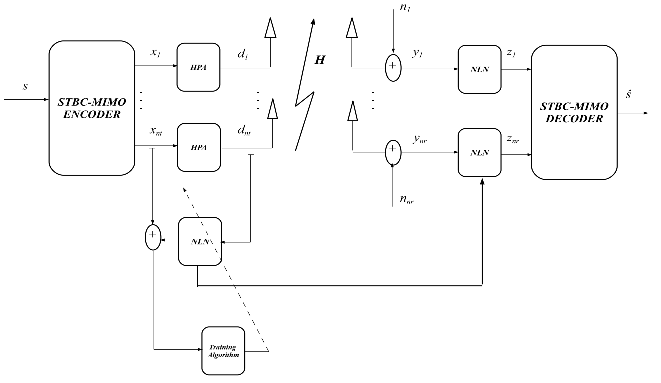


Fig. 1. Block diagram of the considered MIMO-STBC system in the presence of nonlinear HPA and NLN compensation.

of the HPA parameters is available. Numerical results and comparisons are then presented in Section IV. We complete this study by conclusions in Section V.

II. System Model

A. MIMO-STBC and HPA nonlinearity

The block diagram of the considered MIMO-STBC system is shown in Figure 1. The MIMO-STBC system is equipped with n_t transmit and n_r receive antennas. In frequency non-selective block-fading channels, assuming that the x_l are M-QAM modulated symbols of period $\boldsymbol{T}$ and average energy $\boldsymbol{P}_s$ for $l = 1, \cdots, n_t$. the received signal is given by

$$\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{D} + \boldsymbol{N} \tag{1}$$

where $\boldsymbol{Y} \in \mathcal{C}^{n_r \times \boldsymbol{T}}$ is the received matrix, $\mathbf{H} = [h_{k,l}]^{n_r,n_t} \in \mathcal{C}^{n_r \times n_t}$ indicates the $n_r \times n_t$ channel gain matrix with $h_{k,l}$ representing the channel coefficient between the lth transmit and the kth receive antennas and $\mathbf{N}$ is the Additive White Gaussian Noise (AWGN) matrix with *i.i.d.* entries $\sim \mathcal{CN}(0, N_0)$.

The transmitted signal $\boldsymbol{D} \in \mathcal{C}^{n_t \times \boldsymbol{T}}$ has to be amplified at RF through the HPA, which may operate in its nonlinear region, causing amplitude distortion and phase distortion on the input signal [8]. We consider memoryless HPA that can be characterized by their AM/AM and AM/PM conversions. We denote the input signal at the HPA as

$$x_l = r_l e^{j\theta_l} \tag{2}$$

where $r_l(.)$ is the input modulus and $\theta_l(.)$ is the input phase. The signal at the output of the HPA can be expressed as

$$d_l = A_l\,(r_l)\exp\{P_l(r_l) + \theta_l\} \tag{3}$$

where $A_l(.)$ and $P_l(.)$ denote the HPA amplitude conversion (AM/AM) and phase conversion (AM/PM), respectively.

Many models are tailored for a particular type of HPA. TWTA and SSPA as in [9], and SEL as in [10]. The TWTA can be characterized by the Saleh's model [5], which has the advantage of exhibiting greater simplicity and accuracy than other models. The AM/AM and AM/PM conversions can be represented as follow

$$A(r_l) = \tfrac{\alpha_a r_l}{1+\beta_a r_l^2} \quad \text{and} \quad P(r_l) = \tfrac{\alpha_p r_l^2}{1+\beta_p r_l^2} \tag{4}$$

where α_a and β_a are the parameters of the non-linear level, and α_p and β_p are phase displacements. The AM/AM and AM/PM conversions of the SSPA model first and SEL model are the following

$$A(r_l) = \frac{r_l}{\left[1+\left(\frac{r_l}{A_{os}}\right)^{2\beta}\right]^{1/2\beta}} \quad \text{and} \quad P(r_l) = 0 \tag{5}$$

$$A(r_l) = \begin{cases} r_l & r_l \leq A_{is} \\ A_{is} & r_l > A_{is} \end{cases} \quad \text{and} \quad P(r_l) = 0 \tag{6}$$

where β indicates the flexibility of the transition from linear operation to saturation. A_{is} is the input saturation voltage and A_{os} is the output voltage at the saturation point. For simplicity, the HPAs at all the transmitting branches are assumed to exhibit the same nonlinear behavior [3], [11].

According to the central limit theorem, a signal can be approximated as a complex Gaussian distributed random process. From the Bussgang theorem and by extending that to complex Gaussian processes, the output signal at the HPA can be expressed as [12]

$$d_l = K_l x_l + w_l \tag{7}$$

where K_l denotes an arbitrary deterministic complex factor and w_l is a suitably additive zero-mean Gaussian noise uncorrelated with the input signal x_l. The value of K_l is given by [6, eq. (19)]

$$K_l = \frac{1}{2}E\left[d_l^{'}(r) + \frac{d_l(r)}{r}\right] \tag{8}$$

where $d_l^{'}(r)$ denotes the differential of $d_l(r)$. Furthermore, the variance of w_l is given by [12, eq. (37)]

$$\begin{aligned}\sigma_{w_l}^2 &= E[|w_l|^2] = E[|d_l|^2] - |K|^2E[|x_l|^2] \\ &= E[A^2(r_l)] - |K|^2E[r_l^2]\end{aligned} \tag{9}$$

Specifically, for the SEL model, the analytical evaluation of K_l and $\sigma_{w_l}^2$ values, can be obtained using [8] [6, eq. (42)] as follows

$$K_l = \left(1 - e^{-(A_{is}^2/\boldsymbol{P}_s)}\right) + \frac{1}{2}\sqrt{\pi\frac{A_{is}^2}{\boldsymbol{P}_s}}\text{erfc}\sqrt{\frac{A_{is}^2}{\boldsymbol{P}_s}} \tag{10}$$

$$\sigma_{w_l}^2 = \boldsymbol{P}_s\left(1 - e^{-(A_{is}^2/\boldsymbol{P}_s)} - K_l^2\right) \tag{11}$$

In addition, the parameters K_l and $\sigma_{w_l}^2$ for the TWTA and SSPA models has been evaluated in [12, Tab. I].

When the channel gain matrix is perfectly estimated at the receiver, the MIMO-STBC model can be converted into an equivalent single-input single-output (SISO) scalar model, yielding [3]

$$y = c\|\boldsymbol{H}\|_F^2 d + \tilde{n} \tag{12}$$

where $\|.\|_F$ denotes the Frobenius norm, d represents the distorted version of the transmitted symbol with average power $\boldsymbol{P}_s^{HPA} = K_l^2\boldsymbol{P}_s$, c is a code-dependent constant based on the STBC mapping and $\tilde{n}$ is the noise term after STBC decoding with distribution $\sim \mathcal{CN}(0, c\|H\|_F^2 N_0)$. Consequently, the effective SNR at the output of the MRC decoder can be expressed as

$$\begin{aligned}\gamma^{STBC} &= \frac{c^2K_l^2\boldsymbol{P}_s\|\boldsymbol{H}\|_F^4}{\left(cN_0 + c\sigma_{w_l}^2\right)\|\,\boldsymbol{H}\|_F^2} \\ &= \frac{cK_l^2\boldsymbol{P}_s}{N_0 + \sigma_{w_l}^2}\|\boldsymbol{H}\|_F^2 = \frac{c\boldsymbol{P}_s^{HPA}}{N_0 + \sigma_{w_l}^2}\|\boldsymbol{H}\|_F^2\end{aligned} \tag{13}$$

B. Architecture of the applied neural network

In our investigation, a multilayer perceptron is used to compensate the effect of HPA nonlinearity. A Nonlinear Network (NLN) (see Figure 2) is a very interesting model for adaptive equalisation due to its properties such as the parallel distributed architecture, the adaptive processing, the nonlinear approximation, the easy integration in large information processing chains and the efficient hardware implementation [13].

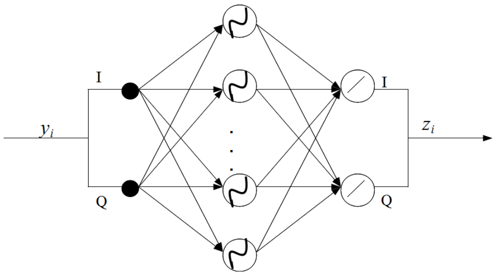


Fig. 2. A NLN multilayer perceptron neural network: This network has two layers, two input signals, one hidden layer, 2 neurons in the output layer, and 2 output signals. (Indexes I and Q refer to the real and imaginary parts, resp.)

Using the structure illustrated in Figure 1, we aim to identify the HPA inverse transfer functions. The complex envelope signals are differentiated and the error sent to the "learning algorithm" bloc reacts on coefficients of NLN. The weights of the NLN Receiver are determined by copying the weights of a trained network NLN.

The multilayer perceptron has two inputs, namely the I and Q components of the input signal complex envelope. The NLN has two outputs that are the compensated signals I and Q signals. Applying an input signal $\boldsymbol{y}_k = [\boldsymbol{y}_{k,I}, \boldsymbol{y}_{k,Q}]^t$, the output of the hidden neuron m [14]

$$v_{k,m} = f(\sum_j w_{j,m} y_{k,j} + b_m) \tag{14}$$

where $w_{j,m}$ is the weight connection between $y_{k,j}$ and the neuron m for $(j = I, Q)$. The function f is a nonlinear activation function (hyperbolic tangent function). In this case, the output of the NLN can be expressed as

$$z_{k,j} = \sum_m u_{m,j} v_{k,m} \tag{15}$$

where $u_{m,j}$ is the weight connecting the neuron m of the hidden layer to the neuron j of the output layer.

The received signal at the output of the neural network block may be modeled as the sum of the transmitted signal x_l and a noise factor caused by the effects of the NLN errors and the channel transmission. Consequently, the output signal $z_{k,j}$ can be approximated by

$$z_{k,j} \simeq x_{l,j} + \sum_m \hat{n}_{m,j} \tag{16}$$

where $\hat{n}_{m,j}$ is the noise of the m-th neuron of the hidden layer. Since f is not a linear activation function, $\hat{n}_{m,j}$ is a non gaussian random variable. However, using the central limit theorem, we can approximate $\sum_m \hat{n}_{m,j}$(the sum of 9 random variable in our case) as a gaussian noise.

During the training sequence, the NLN model uses supervised learning to update the weight parameters in order to minimize a cost function. This function is the sum of squared errors between the unknown system outputs and the HPA inputs (see Figure 2). For the training algorithm, we have chosen the Levenberg-Marquardt algorithm [1]. The contribution of this algorithm is similar to determinate the second-order training speed without having to compute the Hessian matrix. Under the assumption that the error function is some kind of squared sum, the Hessian matrix $\boldsymbol{He}$ can be approximated as

$$\boldsymbol{He} = \boldsymbol{J}^T \boldsymbol{J} \tag{17}$$

and the gradient can be computed as:

$$g = \boldsymbol{J}^T e \tag{18}$$

where e is a vector of network errors and $\boldsymbol{J}$ is the Jacobian matrix that contains the first derivatives of the network errors. This matrix determination is computationally less expensive than the Hessian matrix. The new weight vector w_{n+1} can be adjusted as:

$$w_{n+1} = w_n - \left[\boldsymbol{J}^T \boldsymbol{J} + \mu \boldsymbol{I}\right]^{-1} \boldsymbol{J}^T e \tag{19}$$

The parameter μ is a scalar controlling the behaviour of the algorithm.

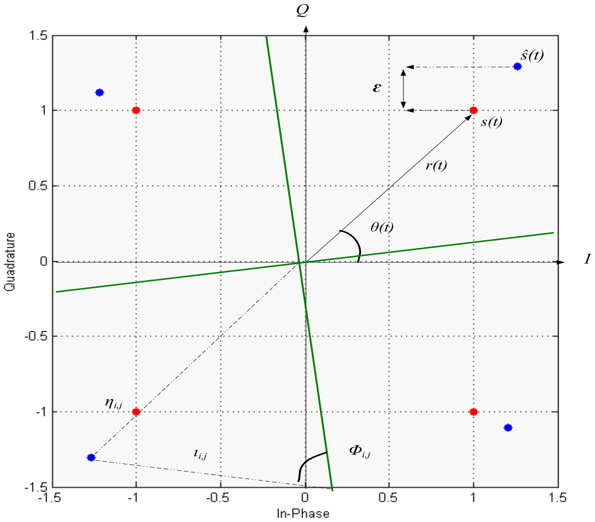


Fig. 3. Rectangular 4-QAM constellation and decision regions with NLN Receiver compensation without noise effect.

III. Performances in Terms of SER and EVM

In this section, we investigate the performance of MIMO-STBC systems over uncorrelated Rayleigh fading channels in the presence of HPA nonlinearity and NLN in terms of BER and EVM.

A. Derivation of the effective SNR Expression

In conventional MIMO-STBC systems, received signals from the NLN elements are combined at baseband. As the number of antenna elements increases, this receiver architecture becomes costly, especially for mobile devices [3]. However, if the signal combining takes place at the RF level, only one receive chain is required, which produces essentially the same output as with the conventional MRC receiver [11]. Hereafter, we consider the approach that combines signals from antenna elements at the RF level. The signal using the

NLN implementation scheme which enhances the cancellation of the distortion signal ŝ can be written as

$$\hat{s} = \hat{r}e^{j\hat{\theta}} \tag{20}$$

where $\hat{r}$ and $\hat{\theta}$ are the amplitude and the phase of $\hat{s}$, respectively. Let ε denotes the errors between the modulus of the input signal and the NLN output patterns. According to the equation (16) and using the central limit theorem, ε can be characterized by a gaussian distribution. For noislyless MIMO-STBC system, the variance of ε will be represented by [14]

$$\sigma_\varepsilon^2 = E\left\{|r_l - \hat{r}_l|^2\right\} = \left(r_l - \rho \sum_i u_i^m f(w_i^m \rho)\right)^2 \tag{21}$$

where ρ is the modulus of the NLN output and m denotes the coefficient of NLN identification. In this case ,the signal $\hat{s}$ can be rewritten as

$$\hat{s} = s + \boldsymbol{\varepsilon} \tag{22}$$

The error ε can be considered as a HPA nonlinearity results in modulus of the input s and output patterns ŝ. An example illustrating such distortion is shown in Figure 3 where we present the constellation and the decision regions of a rectangular 4-QAM without and with HPA-NLN distortion.

We note that the error $\boldsymbol{\varepsilon}$ is uncorrelated with $\boldsymbol{N}$. In this case, the effective SNR at the output of the NLN can be expressed as:

$$\gamma^{STBC-NLN} = \frac{cE[A^2(r)]}{N_0 + \sigma_\varepsilon^2} \|\mathbf{H}\|_F^2 = \alpha\gamma^{STBC} \tag{23}$$

where $\|.\|_F$ denotes the Frobenius norm and γ^{STBC} is the effective SNR at the output of the distortion signal $\mathbf{Y}$:

$$\gamma^{STBC} = \frac{cE[A^2(r)]}{N_0} \|\mathbf{H}\|_F^2 \tag{24}$$

In this case, α can be expressed as:

$$\alpha = \frac{N_0}{N_0 + \sigma_\varepsilon^2} \tag{25}$$

Then, according to (25) into (23) , the effective SNR $\gamma^{STBC-NN}$ can be written as:

$$\gamma^{STBC-NLN} = \frac{cE[B^2(r)]}{N_0} \|\mathbf{H}\|_F^2 \tag{26}$$

where $E[B^2(r)] = \alpha E[A^2(r)] = P_s^{NLN}$ is the average power per symbol at the output of NLN.

B. SER evaluation

In this section, we evaluate the SER performance for the MIMO-STBC systems in the presence of both nonlinear HPA and NLN. Based on decision regions of the distorted version of the transmitted signal and constellation, the SER, can be expressed as a function of the instantaneous output SNR for arbitrary 2-D modulations, using Craig's method [15]

$$P_s(\gamma) = \sum_{i=1}^{M} \sum_{j=1}^{N_{s,i}} \frac{P_{s_i}}{2\pi} \int_0^{\eta_{i,j}} \exp\left[\frac{c_{i,j}\gamma \sin^{2\phi_{i,j}}}{\sin^2(\upsilon + \phi_{i,j})}\right] d\upsilon \tag{27}$$

where M is the number of symbols in the constellation, $N_{s,i}$ is the number of subregions for symbol s_i, P_{s_i} represents the a priory probability that symbol s_i is transmitted, γ is the output SNR of the MIMO-STBC system, and $c_{i,j} = l_{i,j}/\boldsymbol{P}_s^{NLN}$ is the scaling factor. Parameters $l_{i,j}$, $\eta_{i,j}$ and $\phi_{i,j}$ are related to the symbol s_i and the subregion j and is determined by the decision region geometry [3], [15] (see Figure 3).

The average SER using such decision region boundaries can be written as

$$P_s = \int_0^\infty P_s(\gamma) P_{\gamma^{STBC-NLN}}(\gamma) d\gamma \tag{28}$$

where $P_{\gamma^{STBC-NLN}}$ is the pdf of the output SNR for the MIMO-STBC system with NLN compensator under Rayleigh flat fading

$$P_{\gamma^{STBC-NLN}}(\gamma) = \tag{29}$$

$$\frac{2}{\Omega_{k,l}} \sqrt{\frac{\sigma_\Delta^2 \gamma}{c\boldsymbol{P}_s^{NLN}}} \exp\left(-\frac{\sigma_\Delta^2 \gamma}{c\boldsymbol{P}_s^{NLN} \Omega_{k,l}}\right) \tag{30}$$

where $\Omega_{k,l} = E[\lambda_{k,l}^2]$ represents the average fading power and $\lambda_{k,l}$ denotes the path gain. Substituting (27) and(29) into (28), the average SER can be rewritten as

$$P_s = \sum_{i=1}^{M} \sum_{j=1}^{N_{s,i}} \frac{P(s_i)}{2\pi} \int_0^{\eta_{i,j}} \boldsymbol{\Psi}_{\gamma^{STBC-NLN}} \left[\frac{c_{i,j} \sin^{2\phi_{i,j}}}{\sin^2(\upsilon + \phi_{i,j})}\right] d\upsilon \tag{31}$$

where $\boldsymbol{\Psi}_{\gamma^{STBC-NLN}}(j, \omega)$ is the characteristic function of $\gamma^{STBC-NLN}$ and is given by

$$\boldsymbol{\Psi}_{\gamma^{STBC-NLN}}(j, \omega) = \left\{1 - j\omega \frac{cG^2 \boldsymbol{P}_s^{HPA}}{\sigma_\Delta^2}\right\}^m \tag{32}$$

Substituting (32) into (31) and making use of [3], [16], the average SER of MIMO-STBC over uncorrelated Rayleigh flat fading channels with NLN technique can be obtained by

$$\begin{aligned} P_s = \sum_{i=1}^{M} \sum_{j=1}^{N_{s,i}} \frac{P(s_i)}{2\pi} \times \{\, & \eta_{i,j} + \phi_{i,j} - \nu_{i,j} \\ & \times \left[\left(\frac{\pi}{2} + \arctan \lambda_{i,j}\right) \sum_{k=0}^{m\kappa-1} \binom{2k}{k} \right. \\ & \times \frac{1}{4^k \left(1 + \varsigma_{i,j}\right)^k} + \sin\left(\arctan \lambda_{i,j}\right) \\ & \times \sum_{k=1}^{m\kappa-1} \sum_{l=1}^{k} \frac{T_{l,k}}{\left(1 + \varsigma_{i,j}\right)^k} \\ & \left. \times \left(\cos\left(\arctan \lambda_{i,j}\right)\right)^{2(k-l)+1} \right] \Big\} \end{aligned} \tag{33}$$

where

$$\varsigma_{i,j} = c_{i,j} \left(c\boldsymbol{P}_s^{NLN} / \sigma_\Delta^2\right) \sin^2 \phi_{i,j} \tag{34}$$

with

$$T_{l,k} = \binom{2k}{k} \Big/ \left[\binom{2(k-l)}{k-l} 4^l (2(k-l) + 1)\right] \tag{35}$$

and $\nu_{i,j} = \sqrt{(\varsigma_{i,j}/(1+\varsigma_{i,j}))}sgn(\eta_{i,j}+\phi_{i,j})$ and $\lambda_{i,j} = -\varsigma_{i,j}\cot(\eta_{i,j}+\phi_{i,j})$

C. EVM degradation

The EVM is usually used as a parameter for evaluating the effects of imperfections in digital communication systems on the constellation diagram and is an effective method for calculating the system performance [17]. The EVM evaluation is based on the difference between an ideal transmitted constellation point $s(t)$ and the received symbol location $\hat{s}(t)$ at each symbol instant t. By definition [**?**], EVM is the root mean square (rms) value of the magnitudes of the error vectors Γ is expressed as

$$EVM_{rms} = \sqrt{\frac{1}{N_s}\sum_{t=1}^{N_s}|\Gamma|^2} \qquad (36)$$

The residual error vector on sample s is obtained at each symbol instant and is defined as [17]

$$\Gamma = \frac{\hat{s}W^{-t} - C_0}{C_1} - s \qquad (37)$$

where the complex constants C_0, C_1, and W compensate the constellation offset, constellation complex attenuation and the amplitude and offset phase rotation. The normalized EVM can be defined as the ratio of the rms EVM to the averaged symbol power [**?**] [18]

$$EVM_m = \frac{\sqrt{\frac{1}{N_s}\sum_{t=1}^{N_s}|\Gamma|^2}}{\sqrt{\frac{1}{N_s}\sum_{t=1}^{N_s}|s|^2}} = \frac{\sqrt{E\{|\Gamma|^2\}}}{E\{\sqrt{|\boldsymbol{s}|^2}\}} \qquad (38)$$

The residual error vector Γ is obtained by

$$\Gamma = \hat{\boldsymbol{s}} - \boldsymbol{s} = Gc\|\boldsymbol{H}\|_F^2\boldsymbol{w} + \boldsymbol{v} + G\tilde{\boldsymbol{n}} = \Delta \qquad (39)$$

Using the fact that the total noise sample and the interference term are uncorrelated, we can obtain the following expression

$$EVM_m = \frac{\sigma_\Delta^2}{\sqrt{\boldsymbol{P}_s}} \qquad (40)$$

IV. SIMULATION RESULTS

To investigate the performance of the proposed NLN compensation in MIMO-STBC with HPA, a series of Monte Carlo simulations were carried out. The full-rate Alamouti code ($R_c = 1$, $c = 1$) is investigated. We have considered a MIMO system with 2 inputs and 2 outputs with 4-QAM modulation using 10^7 randomly generated symbol blocks. The memoryless selected nonlinear model for HPA is the TWTA one. The NLN (see Figure 2) neural network, is composed of two inputs, nine neurons in the hidden layer (with sigmoid activation function) and two linear neurons in the output layer. In the simulations, we define the input back-off (IBO) as:

$$IBO = 10\log_{10}(\frac{A_0^2}{P_{in}}) \qquad (41)$$

where A_0^2 is the maximum output modulus and P_{in} is the average input power.

In order to identify the best SNR to create the learning data base, we have realised many data bases using various SNR, then we simulated the NLN to quarry out the SER obtained using MIMO-STBC system.

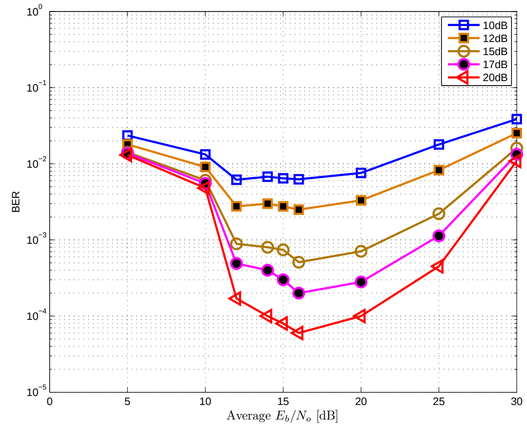


Fig. 4. SER over MIMO-STBC system with NLN correction (Fixed SNR simulation for each curve).

Figure 4 shows the SER performs versus the SNR used in the learning phase with different SNR used in generalization phase. We note from these results that in all cases, the learning data base with SNR around 16 dB offers the best performs. For our simulations, we have used a learning data base with SNR equal 16 dB.

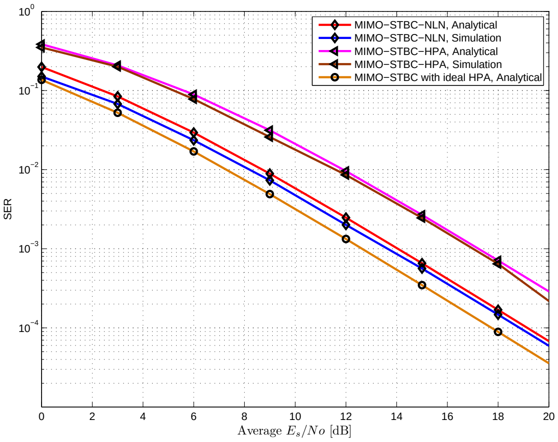


Fig. 5. The SER of signal transmission over MIMO-STBC system.

Figure 5 shows the average SER performances of the considered system for three cases i) MIMO-STBC with ideal HPA which serves as a benchmark. ii) MIMO-STBC-NLN iii) MIMO-STB-HPA without compensation. We show that analytical and simulation results are in perfect match. We note that the NLN is able to improve the system performance.

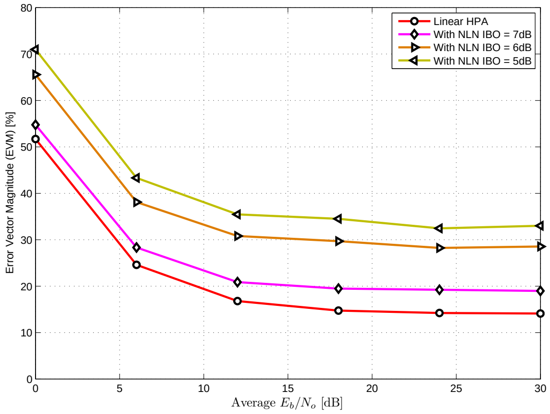


Fig. 6. EVM_m as a function of SNR with the IBO as parameter.

Figure 6 illustrates the EVM of Equation (40) versus SNR and selected IBO values. It indicates that SNR imposes a great impact on the EVM performance when the IBO varies.

V. Conclusion

In this paper, the effects of nonlinear HPA on the performance of the MIMO-STBC system were evaluated when it is operated under Rayleigh fading channel. It was shown that the NLN technique implemented at the receiver is able to compensate the nonlinear behaviour caused by the power amplifier. The simulation results showed that, in the presence of the proposed NLN decreased the used SNR to 14 dB at SEP to 10^{-4} which is an improvement of more than 3 dB compared to HPA without compensation. The system performance was analyzed in terms of effective SNR expression, average SER and EVM. Theoretical results show a close matching with those obtained by simulations for the 4-QAM MIMO-STBC systems.

References

[1] R. Z. R. Bouallegue and D. Roviras, “Adaptive Pre-distortions based on Neural Networks associated with Levenberg-Marquardt algorithm for Satellite Down Links,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2008, pp. 1–8, 2008.

[2] A. J. Goldsmith, S. A. Jafar, N. Jindal, and S. Vishwanath, “Capacity Limits of MIMO Channels,” *IEEE Journal on Selected Areas in Communication*, vol. 21, no. 5, pp. 684–702, June 2003.

[3] Q. Jian and S.Aissa, “Analysis and Compensation of Power Amplifier Nonlinearity in MIMO transmit Diversity Systems,” *IEEE Transactions on Vehicular Technology*, vol. 59, pp. 2921–2931, July 2010.

[4] F. H. Gregorio, “Analysis and Compensation of Nonlinear Power Amplifier effects in Multi-antenna OFDM Systems,” Ph.D. dissertation, Helsinki Univ. Technol, Nov 2007.

[5] A. A. M. Saleh, “Frequency-independent and Frequency-dependent Nonlinear Models of TWT Amplifiers,” *IEEE Transaction on Communication*, vol. 29, pp. 1715–1720, 1981.

[6] R. Z. R. Bouallegue and D. Roviras, “Crossover Neural Network Predistorter for the Compensation of Crosstalk and Nonlinearity in MIMO OFDM Systems,” in *IEEE 21st International Symposium on Personal Indoor and Mobile Radio Communications Information Sciences System*, 2010.

[7] A. Al-Hinai and M. Ibnkahla, “Neural Network Nonlinear MIMO Channel Identification and Receiver Design,” in *IEEE International Conference on Communications, ICC ’08*, 2008, pp. 835 –839.

[8] Q. Jian and S. Aissa, “Impact of hpa Nonlinearity on MIMO Systems With Quantized Equal Gain Transmission,” in *20th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC’09*, 2009, pp. 2891 – 2895.

[9] G. Santella and F. Mazzenga, “A Hybrid Analytical-Simulation Procedure for Performance Evaluation in M-QAM-OFDM Schemes in Presence of Nonlinear Distortions,” *IEEE Trans. Veh. Technol.*, vol. 47, pp. 142–151, Feb 1998.

[10] H. E. Rowe, “Memoryless Nonlinearities with Gaussian Inputs: Elementary Results,,” *Bell Syst. Tech. J.*, vol. 61, pp. 1519–1525, Sept 1982.

[11] J. Qi and S. Aissa, “On The Effect of Power Amplifier Nonlinearity on MIMO Transmit Diversity Systems,” in *Proc. IEEE ICC*, Dresden, Germany, 2009.

[12] D. Dardari and V. T. A. Vaccari, “A Theoretical Characterization of Nonlinear Distortion Effects in OFDM Systems,” *IEEE Trans. Commun.*, vol. 48, pp. 1755–1764, 2000.

[13] M. Ibnkahla, “Applications of Neural Networks to Digital Communications Survey,” *Signal Processing*, vol. 80, no. 7, pp. 1185–1215, 2000.

[14] D. R. H. Abdulkader, F. Langlet and F. Castanie, “Natural gradient algorithm for neural networks applied to non-linear high power amplifiers,” *Int. J. Adapt. Control Signal Process*, vol. 2, no. 16, pp. 557–576, 2002.

[15] J. W. Craig, “A New, Simple and Exact Result for Calculating the Probability of Error for Two Dimensional Signal Constellations,” in *Proc. IEEE MILCOM*, 1991.

[16] M. S. Alouini and M. K. Simon, “An MGF Based Performance Analysis of Generalized Selection Combining Over Rayleigh Fading Channels,” *IEEE Trans. Commun*, vol. 48, pp. 401–415, 2000.

[17] H. Zareian and V. T. Vakili, “Analytical EVM, BER, and TD Performances of the OFDM Systems in the Presence of Jointly Nonlinear Distortion and IQ Imbalance,” *Springer-Verlag*, vol. 64, pp. 753–762, 2009.

[18] A. Georgiadis, “Gain, Phase Imbalance, and Phase Noise Effects on Error Vector Magnitude,” *IEEE Trans Veh Tech*, vol. 2, pp. 443–449, 2004.

Towards Self-Organizing Network Orchestration Management for LTE Mobile Communication Systems

Javier Rubio-Loyola and Hiram Galeana-Zapién
Information Technology Laboratory
CINVESTAV Tamaulipas
Ciudad Victoria, Tamaulipas, México
{jrubio, hgaleana}@tamps.cinvestav.mx

Ramón Agüero
Telematics Engineering Group
Universidad de Cantabria
Cantabria, Spain
ramon@tlmat.unican.es

Abstract—**Self-organizing network (SON) functions of Long Term Evolution (LTE) systems have been traditionally studied in isolation even when it is widely accepted that they need to work together to provide the next generation mobile services and applications. This paper describes a novel QoS- and resource-oriented SON orchestration SON management framework in favor of convergence to trade-offs between service level requirements and network performance targets in LTE systems. In order to orchestrate SON functions of LTE networks it is needed consider standalone optimization processes as well as QoS- and resource-aware tradeoffs between service level requirements and overall network resource performance.**

Keywords – Long Term Evolution; Orchestration; Self-organizing networks; Quality of Service

I. Introduction

Mobile communication systems are continuously evolving to provide higher data rates and to pave the way for ubiquitous, high speed broadband wireless coverage. Nowadays, the Long Term Evolution (LTE) technology [3], [4] is recognized as the most outstanding technology to meet these goals, and it is considered as the next generation mobile technology that will dominate the worldwide mobile ecosystem in the next decade and beyond.

LTE is a complex system where a large number of control mechanisms are executed at different nodes in the architecture to perform resource allocation tasks at different levels of granularity (in terms of the time-scale they operate). Algorithmic solutions for resource management operate in a highly dynamic environment where network resources (e.g., transmission power), services and applications (e.g., offered service quality), and user behavior (e.g., user mobility) experience changes over time. Those changes can degrade network performance and perceived QoS.

The research community has highlighted the need and value of LTE self-organizing capabilities, commonly referred to as Self Organizing Network (SON) capabilities. SON capabilities in the network will lead to higher end-to-end QoS and reduced churn [2] (i.e., a measurable metric of the migration of users from one service provider to another, mostly due to dissatisfaction with perceived QoS), thus allowing for overall improved network performance in terms of network quality and reliability. Finally, SON is hyped to provide higher performance from adapting the network to variations in loading and other dynamic operational conditions.

Although there is a clear definition of the most prominent SON functionalities, their implication with service quality provision and management has not been studied in the literature. The modeling and performance evaluation of SON functions are normally addressed following a standalone approach where a given function is assumed to work independently of other SON functions. The collateral effects among SON functions are neglected to make tractable the design of algorithmic solutions for LTE. As there is mutual dependency among network parameters, conflicting situations among SON functions may take place.

This paper presents a work in progress that aims at investigating novel self-optimization and management solutions to provide enhanced support for QoS of next-generation services in LTE systems. In order to address the interdependency of SON functionalities, this research will develop a coordination management framework aimed to cope with system's instabilities due to potential conflicting decisions among SON management functions. After this Introduction, Section II presents the application domain of the research. Section III presents the foreseen technical approach and Section IV presents the related work. Finally, Section V concludes the paper.

II. Application Domain of the Research

A. LTE SON Functionalities

Self-Organizing Networks (SON) is seen as one of the main promising areas for operators to save on operational expenditures. SON is currently discussed in 3GPP standardization [3]. Furthermore, the NGMN group has made recommendations [4] and 3GPP has written some use cases into the SON standards for LTE release 8, LTE release 9, as well as in release 10 (LTE-A). However, the SON self-optimization algorithms are not standardized or defined step-by-step. Figure 1 illustrates the SON use case functionalities envisioned by the 3GPP. The interested reader can find extended descriptions of all functionalities in reference [3]. This work focuses on Mobility Load Balancing Optimization and Mobility Robust Optimization, whose main goals are briefly described hereafter [3].

Mobility Load Balancing Optimization. The goal of this use case functionality is to optimize cell reselection/handover parameters to cope with the unequal traffic load and at the same time the minimization of the number of handovers (HO) and redirections needed to achieve the load balancing.

Mobility Robust Optimization. Manual setting of HO parameters in current 2G/3G systems is a time consuming task. For some cases, RRM (Radio Resource Management) in one eNodeB can detect problems and adjust the mobility parameters, but there are also examples where RRM in one eNodeB cannot resolve problems. The objective of this use case functionality is to automatically adjust the mobility parameters in those cases that cannot be done by RRM.

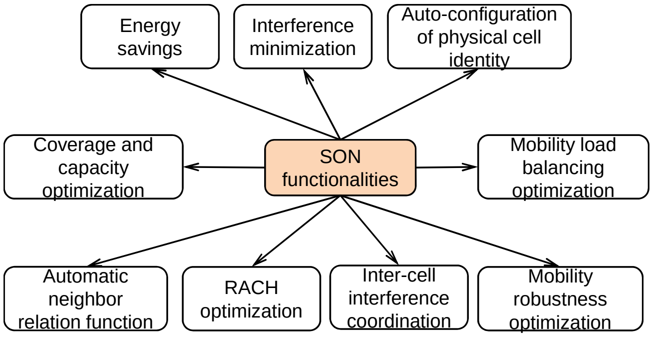

Energy savings
Interference minimization
Auto-configuration of physical cell identity
Coverage and capacity optimization
SON functionalities
Mobility load balancing optimization
Automatic neighbor relation function
RACH optimization
Inter-cell interference coordination
Mobility robustness optimization

Figure 1. SON use case functionalities

Each of the SON functionalities operates at different time scales of network operation, from short-term to long-term, and each will cause changes in network configuration at different levels of granularity. SON functionalities exhibit a certain degree of interdependency and it is proven that some of them may have an impact on the configuration of the same network parameters which can result in conflicting situations.

B. *Orchestration mechanisms between SON functions*

The analysis of the different functionalities of the SON architecture and their specific characteristics has opened a big opportunity area for the implementation of optimization techniques. However, this is not a simple task, every SON-functionality has an interrelation with at least another one, therefore it may be found that an optimal solution for one SON function could be controversial for other functionalities sharing the same control parameters.

For example, SON functionalities Mobility Load Balancing and Mobility Robust Optimization shown in Figure 2 both modify inter-related parameters (Handover parameters). The Load Balancing (LB) algorithms in SON Function 1 may want to decrease the handover offset to optimize the load distribution between cells, while Handover Optimization algorithms in SON Function 2 may want to increase the handover hysteresis to reduce Ping-Pong effects, thus both would eventually change the condition on which handovers are taken [5]. The modified values of the Handover parameters that give the best performance from the perspective of the Mobility Load Balancing may affect negatively the final performance of the Mobility Robust Optimization function.

There is a need to find a way to orchestrate the different functionalities by means of a coordination management entity that can adjust the conflictive SON decisions and that can decide when to intervene in the optimization process of the SON functionalities because every process in LTE technology has a different time scale and it is important to act at the right moment. Orchestration functionalities will take advantage of the best set of values for improving the performance of the network. Moreover, there is a need for this orchestration functionality to consider the operator quality of service concerns during the SON self-* optimization process.

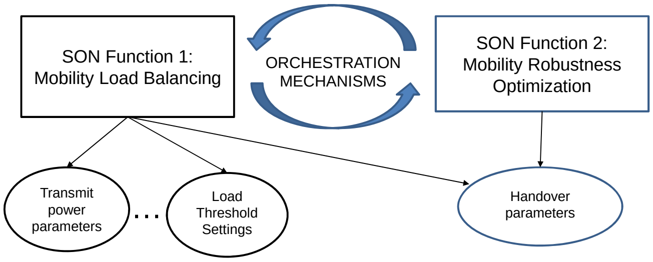

SON Function 1: Mobility Load Balancing
ORCHESTRATION MECHANISMS
SON Function 2: Mobility Robustness Optimization
Transmit power parameters
...
Load Threshold Settings
Handover parameters

Figure 2. SON-Functionalities in the need for orchestration mechanisms

III. FORESEEN TECHNICAL APPROACH

This research aims to investigate novel self-optimization and management solutions to provide enhanced support for QoS of next-generation services in LTE systems. In order to address the interdependency of SON functionalities this research will develop a coordination management framework aimed to cope with system's instabilities due to potential conflict decisions among SON functions. This section presents the foreseen technical approach towards this complex task. Namely, we present the LTE system model and the corresponding optimization problem that we are targeting.

A. *LTE System Modeling*

This section presents the modeling of SON functions in LTE, considering the mapping of service characteristics and requirements with resource management mechanisms responsible of allocating available network resources to users. This research concentrates on the Mobility Robust Optimization and Mobility Load Balancing Optimization SON functions, since they have a high degree of dependency and rely on a common management policy (namely, handover or base station assignment procedure) to achieve their corresponding goals.

The LTE system is modeled attending to the downlink performance of an OFDMA-based cellular network. The considered system model [6], illustrated in Figure 3, consists of N eNodeBs that cover a geographical area in which there are M active users. It is assumed that each user *i* has a minimum data rate requirement, denoted as R^{i}_{min}, which must be satisfied irrespectively of the assigned eNodeB. The arrows connecting users and eNodeB's in Figure 3 indicate possible eNodeB assignment choices.

The overall network uses a single frequency channel with a total bandwidth *BW* that is divided into *K* OFDM subcarriers so that each eNodeB *j* can operate a subset of K_j subcarriers. Radio and transport resources are assumed to be allocated to each user in a single eNodeB, due that LTE systems do not consider macro-diversity support (i.e., only hard-handovers are allowed). The model considers that each eNodeB is constrained by a limited amount of radio resources and a limited amount of transport resources. As to radio resource constraints, each eNodeB in the LTE system is assumed to be able to allocate simultaneously a maximum of K_j subcarriers and to having a maximum downlink transmission power limitation $P_j^{\max}$. The radio channel gain between eNodeB *j* and user *i* is modeled by a vector $\overrightarrow{G_{ij}}=\{G_{i,j,1},\ldots,G_{i,j,K}\}$, where $G_{i,j,k}$ denotes the radio channel gain over subcarrier $k \in\{1,\ldots,K_j\}$.

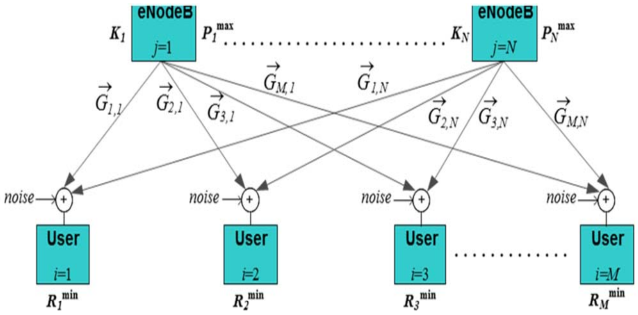


Figure 3. LTE System Model

The amount of radio resources needed to fulfill users' rate service requirements could be quite different depending on the selected eNodeB's. So, for each possible assignment, it is necessary to determine which "resource cost" it has in terms of resource consumption. To that end, we define a radio cost, denoted as α_{ij}, and a transport cost indicated as β_{ij} to quantify resource consumption when assigning user *i* to eNodeB *j*. The eNodeB assignment problem should try to find a *feasible assignment* (i.e., radio costs do not exceed their respective constraints) so that users' rate requirements are satisfied. Additionally, when several feasible eNodeB assignments solutions exist (i.e., there are several ways to allocate all the users without exceeding network resources), we are also interested in finding the "best" of these possible solutions. This can be modeled using the concept of utility function, which allows us to quantify the appropriateness of assigning user *i* to eNodeB *j* by means of a magnitude denoted as u_{ij}, so that $u_{ij} > u_{il}$ would indicate that eNodeB *j* is more appropriate than eNodeB *l* to serve user *i*. As well, $u_{ij} > u_{lj}$ would indicate that is better to assign user *i* to eNodeB *j* than user *l*. Details of utility and resource cost functions envisaged so far are provided in the following.

B. *Resource Cost and Utility function.*

Before presenting the resource cost function definition, we provide a brief review of basic concepts concerning the LTE's air interface evaluation metrics.

In a cellular OFDMA system like LTE, the computation of the signal to interference and noise ratio (SINR) achieved at subcarrier *k* in the receiver of user *i* served by eNodeB *j*, is obtained as follows:

$$SINR_{i,j,k}=\frac{G_{i,j,k}P_{i,j,k}}{I_{i,j,k}+\eta} \qquad (1)$$

where $G_{i,j,k}$ is the radio channel gain between eNodeB *j* and user *i* over subcarrier *k*, $P_{i,j,k}$ is the transmit power of eNodeB *j* on subcarrier *k* allocated to user *i*, η is the thermal noise per subcarrier, and $I_{i,j,k}$ is the co-channel interference power received by user *i* in that subcarrier. The value of the co-channel interference $I_{i,j,k}$ can be computed as:

$$I_{i,j,k}=\sum_{n=1,n\neq j}^{n=N} G_{i,n,k}P_{m\neq i,n,k} \qquad (2)$$

where $P_{i,n,k}$ is the transmit power of interfering eNodeB *n*, on subcarrier *k* assigned to other user *m≠i*. Equation (1) that models the SINR denotes the channel frequency response of user *i* on subcarrier *k*, and the achievable transmission rate $r_{i,j,k}$ on this subcarrier of user *i* assigned to BS *j* is given by:

$$r_{i,j,k}=\frac{BW}{K}\cdot\log_2\left(1+SINR_{i,j,k}\right) \qquad (3)$$

For illustration purposes, if all resources of eNodeB *j* were allocated to user *i*, the maximum achievable rate would be:

$$R_{i,j}^{\max}=\sum_{k=1}^{K_j} r_{i,j,k} \qquad (4)$$

In this context, considering that an eNodeB dynamically shares transmission resources between assigned users by allocating a given amount of subcarriers to user *i*, denoted as K_{ij}, being $K_{ij} < K_j$, during a given amount of transmission time, denoted as ΔT_{ij}, being $\Delta T_{ij} < T_s$, where T_s is a scheduling reference time, we could relate the achievable rate to the amount of used subcarriers and the amount of allocated transmission time to meet users' minimum rate requirements:

$$\frac{K_{ij}}{K_j}\frac{\Delta T_{ij}}{T_s}R_{ij}^{\max}\geq R_i^{\min} \qquad (5)$$

From the previous expression, the radio resource cost is defined directly as:

$$\alpha_{ij} = \frac{R_i^{\min}}{R_{ij}^{\max}} = \frac{K_{ij}}{K_j}\frac{\Delta T_{ij}}{T_s} \leq 1 \tag{6}$$

Note that α_{ij}=1 would mean that serving user *i* in eNodeB *j* makes use of all available radio resources at the eNodeB. Attending to practical considerations, it is considered that there is a limited set of modulation and coding schemes (MCS) that must be used in each subcarrier, thus reducing the output of expressions (3), (4) and (6) to a set of discrete values.

On the other hand, to quantify the appropriateness of each eNodeB assignment, a utility-based framework is used. Different types of utility functions have been used in resource allocation problems. Commonly, a utility function is a non-decreasing function of the amount of allocated resources and its shape (e.g., step, convex, concave or sigmoid are often used) depends on the expected benefit that resource allocation can bring into a given system (e.g., a step function can be used to model a system where allocating resources below a given threshold has no utility at all but the maximum utility is just achieved when reaching this threshold). In our case, we formulate the utility function to reflect the bit rate efficiency of the allocated resources to supporting the data transfer of a user assigned to a given eNodeB. Hence, as to the air interface, the efficiency is directly obtained according to Shannon's law from expression (3) as $\log_2(1+SINR_{ij})$ (the bigger the SINR, the less amount of resources are needed to fulfill user's requirements). Hence, the utility function can be defined as:

$$u_{ij}(SINR_{ij}) = \frac{\log_2(1+SINR_{ij})}{\log_2(1+SNR)} \tag{7}$$

where SNR is the signal to noise ratio achieved in case of no co-channel interference.

C. Optimization Problem Formulation

Using both the resource cost function and utility-based function, it is possible to formulate the base station assignment problem as an optimization problem aiming to maximize the utility of the assignments while not exceeding radio capacity limits at each BS. Defining the BS assignment matrix $B = \{b_{ij}\}_{MxN}$, with i∊{1,…,*M*} and *j*∊{1,…,*N*}, where the assignment indicator variable b_{ij} equals to 1 if user *i* is assigned to BS *j*, or zero otherwise, the BS assignment problem can be formally written as:

$$\max_{ij}\left(\sum_{i=1}^{M}\sum_{j=1}^{N} u_{ij} b_{ij}\right) \tag{8}$$

$$s.t. \quad \sum_{i=1}^{M} \alpha_{ij} b_{ij} \leq 1 \qquad j = 1,...,N \tag{9}$$

$$\sum_{j=1}^{N} b_{ij} = 1 \qquad i = 1,...,M \tag{10}$$

$$R_i \geq R_i^{\min} \tag{11}$$

$$b_{ij} \in \{0,1\} \tag{12}$$

The above optimization problem aims to maximize the total welfare utility, as defined in (8), of the assignments in the system. Under the considered objective function, the assignments that lead to have a most efficient connection, in terms of the bit rate efficiency of the allocated radio resources, are preferred. The set of constraints considered in (9) assures that no more resources than available are assigned to each BS. The second set of constraints (10) is used to indicate that all users need to be assigned to a single BS, while constraint (11) indicates the individual rate required by each user. Moreover, to avoid splitting or partial assignment of users, constraint (12) is used, which however leads to the combinatorial nature of the problem with exponentially growing complexity in the degrees of freedom.

Problem (8)-(12) is a non-linear combinatorial optimization problem since entries in the assignment matrix *B* can only take integer values. Notice that utility and resource cost functions are non-linear functions that depend on the SINR values, which in turn depend on the eNodeB assignment solution because of the co-channel interference, resulting in a mutual dependency. So, both utility and radio resource cost function values are coupled with the assignment of the users in the system, making the eNodeB assignment problem very hard to tackle.

The above optimization problem formulation resembles the behavior of the Mobility Robust Optimization functionality, where the underlying idea is to find the base station assignment for each mobile user so that the overall system's utility is maximized and network constraints and users' requirements are satisfied. This problem formulation is valid whenever a standalone implementation of a single SON function is performed.

As a result, in order to analyze both the Mobility Robust Optimization and the Mobility Load Balancing Optimization in a coordinated framework, a multi-objective optimization should be defined. This latter type of approaches has not been addressed in the literature to simultaneously perform different self-organizing management tasks. With this regard, Figure 4 illustrates the case where two SON functionalities co-exist and work together in an orchestrated manner. More specifically, the Mobility Load Balancing Optimization aims to distribute traffic among the cells in the LTE system, so that unbalance conditions are, at some extent, prevented or mitigated. For instance, this can be achieved by controlling the amount of resources allocated to mobile users (i.e., transmit power that is related to constraint (9) in our system model), and/or manipulating the load threshold values used to trigger appropriate load balancing actions in LTE. In any case, notice that actions performed by SON function 1 indirectly impact on control parameters been directly reconfigured by the SON function 2. With this regard, direct influences can be seen as a controllable behavior, whereas indirect impacts are actually uncontrollable behavior. Therefore, management solutions encompassing orchestration mechanisms are required to allow a seamless integration and coordination of two different SON functions, and particularly to prevent any potential conflict among considered objectives. The orchestration mechanisms can be realized by having a feedback loop between both functionalities, so that the actions taken by a given function are properly send to the input of the other function.

IV. State of the Art

Zhan et al. [7] proposed an algorithm for the Mobility Load Balancing where load balancing actions are triggered by cells experiencing a relatively low load. In practice, the objective of this approach is that an underutilized cell anticipates to an eventual congestion situation in a neighboring cell. Although it is clear that a handover mechanism is used to steer users to the lightly loaded cell, this work does not provide details about the selection choice of uses that are likely to be handed over by the proposed approach.

Sas et al. [8] analyzed the problem of dynamic admission control threshold settings and handover parameters. In this sense, the admission procedure is assumed to have a reserved amount of resources that can be allocated to users for which a handover has been performed. Depending on the perceived network conditions, the proposed solution is able to modify admission control thresholds. The main drawback of this approach is that the policy used to set threshold values does not consider minimum QoS requirement as our research proposal.

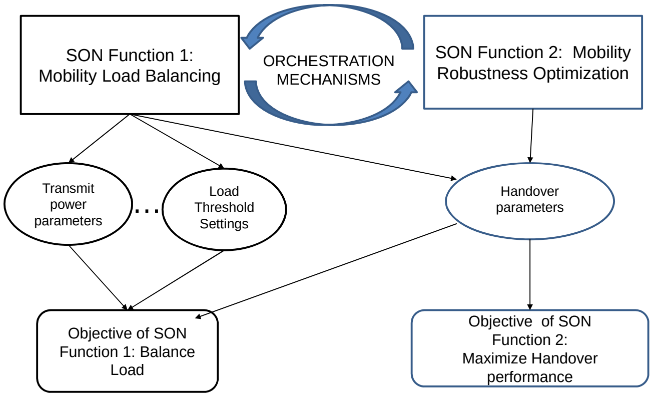


Figure 4. Interdependencies between two SON functions

Hu et al. [9] aim at improving the functionality of Mobility Load Balancing by introducing a handover penalty function in the handover decision making process. Ewe and Bakker [10] proposed a handover optimization procedure performed distributed to base stations of the radio network. However, none of these previous works evaluate or address the combination or analyze the performance of their solutions together with other LTE SON function, e.g. Mobility Robust proposed in this work.

Zhang et al. [11] detected the load conditions of eNodeBs by making use of the sizes and shapes of cellular coverage. Coverage can be adjusted automatically according to load conditions, so as to balance load. However, this work does not take into account quality of service and does not consider the impact that their methods will have on other functionalities that our proposal does.

Handover parameter optimization algorithms are commonly used in the literature in order to tune handover related parameters, namely, hysteresis and time-to-trigger. The work presented by Jansen et al. in [12] falls into this category. However, the work lacks from a formal analysis on how the proposed solution could be integrated with other SON functionalities, which are the advantages of our proposal.

V. Expected Contributions and Concluding Remarks

In order to orchestrate SON functions of LTE networks it is needed consider standalone optimization processes and overall network performance to drive their decisions. It is also needed to find QoS- and resource-aware tradeoffs between service level requirements and overall network resource performance. We have presented a work in progress towards a novel QoS- and resource-oriented SON orchestration management framework to drive the optimization procedures of the Mobility Robust and Load Balancing SON functions, that exploits computational intelligence algorithms in favor of convergence to trade-offs between service level requirements and network performance targets in LTE mobile communication systems environments.

Acknowledgment

This work is partially supported by the LACCIR Project No. R1211LAC005, the project TEC2009 - 14598 - C02 - 02 granted by the MEC Spanish Ministry and partially funded with FEDER funding, and the CONACYT FOMIX Project No. 190552.

References

[1] J. M. Graybeal and K. Sridhar, “The Evolution of SON to Extended SON”, Bell Labs Technical Journal, 2010, pp. 5-18

[2] S. K. Das, H. Lin and M. Chatterjee, "An econometric model for resource management in competitive wireless data networks", IEEE Network, vol.18, no.6, Nov.-Dec. 2004, pp. 20- 26

[3] 3GPP Technical Report 36.902 V9.3.1, “Self-configuring and self-optimizing network (SON) use cases and solutions (release 9)”, March 2011, pp 1 – 21

[4] Next Generation Mobile Networks (NGMN Deliverable), “Next Generation Mobile Networks Use Cases related to Self Organising Network, Overall Description”, May 2007, pp1- 17

[5] L.C. Schmelz, M. Amirijoo, A. Eisenblaetter, R. Litjens, M. Neuland and J. Turk, "A Coordination Framework for Self-Organisation in LTE Networks”, IFIP/IEEE International Symposium on Integrated Network Management (IM), 23-27 May 2011, pp193-200

[6] H. Galeana-Zapién and R. Ferrús, “Design and Evaluation of a Backhaul-Aware Base Station Assignment Algorithm for OFDMA-Based Cellular Networks”, IEEE Transactions on Wireless Communications, vol. 9, no. 10, Oct. 2010, pp. 3226-3237

[7] M. Zhang, W. Li, S. Jia, L. Zhang and Y. Liu, “A Lightly-loaded Cell Initiated Load Balancing in LTE Self-Optimizing Networks”, 6th International ICST Conference on Communications and Networking in China, 2006, pp. 421-425

[8] B. Sas, K. Spaey, I. Balan, K. Zetterberg and R. Litjens, “Self-optimisation of admission control and handover parameters in LTE”, IEEE Vehicular Technology Conference, 2011 pp 3026-3031

[9] H. Hu, J. Zhang, X. Zheng, Y. Yang and P. Wu, “Self-Configuration and Self- Optimization for LTE Networks”, IEEE Communications Magazine, February 2010, pp. 94-100

[10] L. Ewe and H. Bakker, “Base Station Distributed Handover Optimization in LTE Self-Organizing Networks”, in Personal Indoor and Mobile Radio Communications (PIMRC), 2011, pp 243-247

[11] H. Zhang, X. Qiu, L. Meng and X. Zhang, “Achieving Distributed Load Balancing in Self-Organizing LTE Radio Access Networks with Autonomic Network Management”, IEEE GLOBECOM, 2010, pp 454 - 459

[12] T. Jansen, I. Balan, S. Stefanski, I. Moerman and T. Kurner, T.; “Weighted Performance based Handover Parameter Optimization in LTE”, IEEE 73rd Vehicular Technology Conference, May 2011, pp 3267 - 3271

A Musical Feast

How Musical Performance Using Playful Utensils Can Enrich the Cooking and Dining Experience

Cao Yan Yan
Graduate School of Media Design
Keio University
Yokohama, Japan
yanyan.cao@gmail.com

Jeffrey Tzu Kwan Valino Koh
NGS, Keio-NUS CUTECenter
National University of Singapore
Singapore
jtkv.koh@gmail.com

***Abstract*-Playful Utensils is a system of music-enabled, eating and food preparation utensils that facilitate light-hearted interaction and communication in the kitchen and dining room both locally and remotely. In order to understand the use of utensils to support communal behavior in dining situations, three design studies were conducted. Addressing the need for a new direction for food research in HCI, Playful Utensils aims to draw attention away from contemporary kitchen and dining media research, which focuses too acutely on ubiquitous sensory overload, to make apparent the need for socially communicative, emotional assets investment regarding communities in the kitchen and dining space.**

Keywords-dining; playfulness; performance

I. Introduction

Appearing to be and functioning as everyday household cooking and dining utensils such as forks, knives, spoons and chopsticks, Playful Utensils is a system of music-enabled, eating and food preparation utensils that facilitate light-hearted interaction and communication in the kitchen and dining room. When Playful Utensils are activated, they create a ubiquitous wireless mesh network in which each utensil talks to one another via a host server. This enables the musically augmented utensils to be orchestrated by family members at the dinner table or remote locations. The light-hearted interactions will enrich cooking and dining experience and enhance emotional connections among family members.

Traditionally, in Asian households, the kitchen and dining room has been a place for families to connect and engage with one another, yet today's accelerated lifestyle endangers such nurturing activities. The generation gap between children, adults and the elderly is ever increasing, partly due to the fact that the experience of communal cooking and dining by adults is often at odds with the technological pursuits of children. The desire to revive nostalgic sharing among family in modern domestic space poses new challenges and requirements [1, 2, 3].

You might remember playing drums with your chopsticks and bowl and getting a slap from your mother, but many cultures, in fact, practice performative eating and dining and Asian cultures are no exception. We only need to look to Korean Nanta and Japanese Teppanyaki to see that playful cooking and dining has been nurtured by the diversity of Asian cultures and is alive and thriving, sometimes even facilitating big business. Not only do such practices improve the communication between all participants, but it also makes food actually taste better; at least, in a psychosomatic way, as the people involved enjoy the company of one another on a dimension that transcends the basic and sometimes mundane activity of cooking and eating [4].

What if playful interaction was introduced as a new behavioral model to improve communication between children, adults and the elderly within the kitchen and dining room at home? Playful Utensils aims to be a system to promote interaction between members of the family who experience such a generation gap. Everyday cooking and dining utensils become intuitive musical instruments so that the very act of eating and food preparation becomes fun and communally interactive for the whole family.

Working in tandem, a collection of Playful Utensils becomes an orchestra of harmonious, generative musical instruments in which the whole family can play with. Using the traditional and intuitive actions that each utensil was purposely built to function as, ambient dining music can be created adding a new dimension to family-centered interaction in a communal and playful way. This method of playful interaction aims to improve the communication and enjoyment of preparing and eating food in the family household.

The paper is presented in eight sections. In Section II, we presented our motivation along with related works in cooking and domestic space. We present key feature and novelty of Playful Utensils in Section III and show details of three qualitative user studies in Section IV. Section V describes current design and implementation of Playful Utensils, followed by future work and possible scenarios in Section VI, with major contribution of research described in final Section VII.

II. Motivation and Related Works

Rich and robust telecommunication tools such as mobile phone technology, audio/video online chat and email have facilitated our mastery of audio and visual communication. Yet even with these systems in place, we still have some difficulty choosing what to talk about. "Small talk" and discussion about the weather can only take a conversation so far. By injecting light-hearted playful and educational activities into an online conversation, it is our hopes to further engage people, especially within remote communications situations that include cross-generational relationships.

Many works in domestic space involve exploration with robotic cooking assistants, but they do not address the need for interpersonal relationships in regards to the preparation of food, nor the sharing of recipes, meals and traditions across generations [17]. Numerous projects address the need for contextually rich information while cooking without addressing the need for building social capital nor facilitating family intergenerational bonding [5]. These could overload the user with superfluous data. Other systems provide the technology needed to achieve what Playful Utensils aims to do, but does not provide an adequate application in order facilitate family and intergenerational bonding.

Digital media is seemingly rebounding in regards to its shift from the digital and ephemeral, back to the physical and tangible [10, 14, 15]. Synesthetic and multimodal implementations to improve memory have been proposed for augmented object interaction and the television [12], yet sound still continues to be illusive in regards to navigation and accessibility [18]. Yet there is no doubt that sound can enhance family remembering [9].

Other studies include the need to address the heritage of cooking and recipe transmission practice, but most of the time cooking is seen as laborious and often outside the domain of entertainment in many households, leaving young people without the necessary skills to equip themselves for independence [13, 16].

The Playful Toothbrush system developed at the National Taiwan University is most similar to our offering by presenting the user with the opportunity to be persuaded into better habits using ubiquitous computing [7]. Our system also has persuasive qualities, but is geared more towards collaborative interaction between many people, where as Playful Toothbrush concentrates more on the individual user.

Our system aims to build upon all these works in order to address issues of interpersonal relationships in families, activity and alternative means of learning, and address the new direction for food research in HCI [11].

III. Key Feature of playful utensils

A. Playfulness

Mostly referring to animals, the psychologist Gordon M. Burghardt (1984) outlined a working set of characteristics for play [6]. He mentioned among other things that play:

- Has a pleasing effect
- Is sequentially variable
- Is stimulus seeking
- Is quick and energetically expensive behavior
- Involves exaggerated incompetent or awkward movements
- Is most prevalent in juveniles
- Has special "play" signals
- Has a background role in relationships
- Is marked by a relative absence of threat or sub- mission
- Is marked by a relative absence of final consummatory behavior.

Most of his work was derived from the observation of animals, but at least some of these

ISBN: 978-1-61208-236-3

characteristics for play can be applied to the way human beings play. Although children and adults alike can sometimes be seen as “messy beasts” if they do not observe the proper table manners, most of the time children at play can be seen as quite graceful, if you are not the one doing the cleaning up, that is [8].

For the purposes of Playful Utensils, Burghardt’s characteristics of play outline a possibly revolutionary way to socialize at the dinner table. Children love to play and although they may not admit it, adults like to play just as much. Studies have shown that mothers often signal children as young as 3 months that it is time to play [19]. Playful interactions strengthen the bond between children and adults, so it would only make sense that this type of playful understanding could be expressed when using Playful Utensils.

B. Learn by Object

Think back to when you first purchased the mobile phone that is sitting in your pocket. The chances that you actually read the instruction manual for that phone is probably pretty low, unless you work for IEEE or derive enjoyment from comparing low-level technical specifications. A more likely scenario is that you probably learned how to use said phone by pressing some buttons, exploring the menus, generally ignoring the user’s manual, and poked, prodded and played your way to an understanding with the device until you learned enough to satisfy your need of knowledge regarding its functionality relevant to your context.

For children, playing is one of the most effective ways to learn all sorts of useful knowledge. Any object has the potential to teach a myriad of things and children can transform anything into a toy. By using a tangible set of objects, Playful Utensils not only offers an analogue way to learn about cuisine and music, but also teaches children how to interact in tandem, with one another and even with their sometimes less imaginative, adult counterparts. Co-operation is one of the main learning features in the Playful Utensil system and with this tool, children have the chance to teach adults as well as learn from them.

Speaking of adults, for most of us, objects that were once filled with playfulness fade away into the depths of functionality. These everyday objects become invisible until we need them to perform a specific task. By redesigning not only the objects but also the way adults use these objects, we can make something that both children and adults can play with. Even the elderly could benefit from the stimulation that Playful Utensils could provide, as the mere act of eating would activate the system and the interactions designed around it using minimal energy. If grandma had a chance to rock-out with her knife and fork while negotiating her pork-chops, mashed potatoes and green peas, everyone at the dinner table would have a good time.

C. Tools for Knowledge Sharing & Healthy Eating

As a tool, Playful Utensils can be used to associate healthier meals with favourite songs. This could be used to promote nutrition for say, children who do not want to eat their broccoli. By pairing favourite and fun songs with particular dishes in a meal, parents can add another dimension to dining in order to promote healthier eating.

As a learning objective, parents could also pass on the knowledge of certain recipes that would only normally be conveyed through word-of-mouth. Bringing children into the kitchen in the first place is hard enough, but by making the activity of food preparation and cooking more engaging and entertaining, teaching a child or teen how to compose a particular dish could be assisted by an associated song. People could eventually learn how to prepare all sorts of dishes based on the songs that are produced from a recipe.

D. Simple Functions for Complex Lifestyles

Video games such as RedOctane’s Guitar Hero [20] and Harmonix Music System’s Rockband [21] have successfully merged the activities of musical performance and gaming to entice new and meaningful ways for people to interact and engage with technology and one another. Much like a collection of instruments can play compositions collectively as an orchestra, in the Playful Utensils system each utensil has a simple function but when used in tandem can produce recipes and dining experiences of exponential variety.

The arrangement of musical notation can be treated as a recipe to cook a song. A food recipe can be seen as the sheet music to orchestrate a meal. For Playful Utensils, these two linear formulas are interchangeable in a way so that everyday dishes can be prepared and eaten in regards to what songs the family wishes to re-enact. In a similar fashion, picking a favourite song to play together becomes associated with a favourite dish or meal.

Much like the movements in an opera, courses of a meal could also be seen as operatic chapters. As the courses of a meal progress, new movements in the harmony could act as markers for a meals' development.

E. *How to Make & Orchestrate? Other Instrument in the Family?*

A fork is used differently, then a spoon, compared to a knife and so on. By analysing the basic uses and functions of each utensil we can begin to extrapolate a set of triggers based on natural uses specific to the utensil and how they are used for particular eating situations. For instance, one would use chopsticks to eat rice from a bowl in a specific manner that is unique to that action. Eating noodles with the same set of chopsticks in turn expresses a different way to use them, thus offering another dimension of harmonic music triggering.

F. *Music Generation*

Of course, any musician will tell you that actually playing the piano, violin or any other instrument well takes a lifetime of practice and hard work. The initial model of Playful Utensils is not meant to be a pure musical instrument in this respect, as the learning curve for such a device would discourage some people from using them in the first place. In its initial mode, Playful Utensils will simply activate tones and sequences harmoniously as opposed melodically to create interactive ambient music for cooking and dining, using the most natural and intuitive motions that these two activities are already synonymous with.

Concentrating on performing music that doesn't leave a bad taste in everyone's mouth can be a stressful task. Objectives such as playing in key, maintaining synchronicity, timing and so on can actually act adversely to having a pleasurable experience. Considering these problematic outcomes, a central server in which the timing and key are always assisted in an advantageous way monitors each item in the Playful Utensils system. This is achieved by using harmony as the main vehicle to drive musical performance. Even if a member of the family does happen to play "out of step" or accidentally, dissonant notes can actually introduce interesting highlights within the composition. In this manner, “free-styling” and improvisation is actually promoted and rewarded.

Each pair of chopsticks acts as an instrument in an orchestra and is wirelessly connected to a server, much like an individual node in an ad-hoc network. Working in tandem, they create music. MAX-Stream Xbee modules complete the wireless network by broadcasting data to the server. Multiple pairs of chopsticks will stream data synchronously to the server. This enables collaboration in the creation of sound.

IV. Qualitative User Studies

To understand how to support communal behavior in dining situations, and find new opportunities for designing the system, three qualitative user studies were conducted before development our prototype.

The first two studies took into account cross-generational co-cooking scenarios using a computer-mediated environment through the Internet using the teleconferencing tool, i.e: Skype. Shared activity, i.e.: co-cooking game and instructional co-cooking respectively was taken into account in order to ascertain the level of engagement when compared to cross-generational interaction using online tools without a pre-defined activity.

The third gestural study looked at how people eat together. Chopstick usage was analyzed and interpreted into a general family of gestures, which were then used to define the functions and actuations of the current prototype.

The first study implied hypothesis that a communal and synchronized activity supports communication shared between generations over long distances, engagement could be more lasting compared to telecommunication without shared activities.

The second study implied the hypothesis that learning would be more effective using enactive cognitive theories when designing a computer-mediated interaction.

The third study confirmed the assumption that gestural data in using chopsticks are rich for extrapolation.

These studies informed our design for the current prototype of Playful Utensils. They were designed to enable new communication activities, encourage people to join in sound and music creation, and at the same time, encourage communal playing. This will enrich engagement between parties who reside at different locations through gestural behavior, but may not have the means to physically be present to share these experiences.

A. *Qualitative Study One: Cross-Generational Instructional Co-Cooking Game Between Grandmother and Grandchildren*

In the first study, grandmother will teach grandchildren how to assemble their sandwich using real food. The first study consisting of two scenarios was conducted between grandmother and grandchildren in order to test our first hypothesis; whether communal, synchronized and shared activity between generations over long distances could increase engagement duration and quality. As a comparison, both parties played face-to-face for fifteen minutes, and were then separated in the second scenario and continued to play over the Internet using a computer-mediated environment facilitated by Skype.

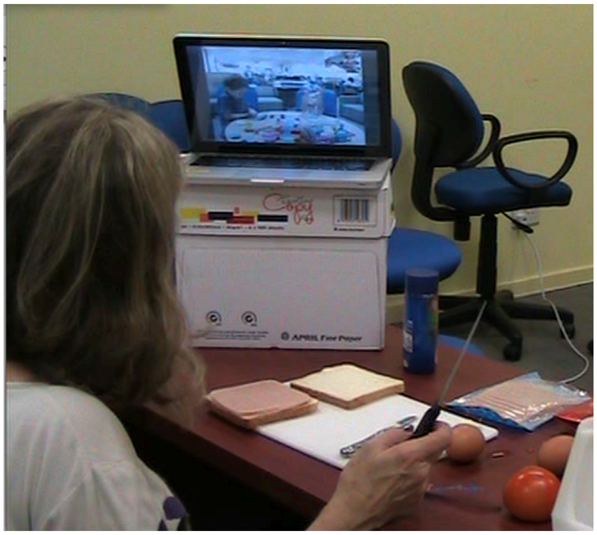


Figure 1. Grandmother instructing grandchildren over the Internet using Skype and real food

1) *Description of Participants*

We chose participants who has experience living apart from their close family members and are used to telecommunication. In this study, the chosen family crosses three generations, the grandmother in her 60's visiting from the USA; two kids – 5-year-old boy and 3-year-old girl. The family has multi-culture background, as father is from Spain and mother is American Hungarian.

It is assumed that each subject has some measure of experience playing with one other and that a socio-political framework involving each family member individually and in tandem has been established prior to this study.

2) *Observations and Learning*

According to observations from parents and grandmother, the children became more engaged with tele-presence communication through task-oriented play.

Communication between children and grandparents were more engaging, they both are strongly engaged with the game and shared the interactive narrative. This differs from the reflections of all parties when recounting conventional interaction previously shared using Skype without a shared activity.

There was some disparity between looking at a video stream and using objects. The girl tended to engage with the grandparent visually through video conferencing, while the boy tended to engage with the make-belief toys and followed audio cues from the grandmother as apposed to using video and visual cues.

People became more engaged with tele-presence communication through task-oriented play. "Small-talk" and "chit-chat" was eliminated, replaced by an interactive narrative shared between parties.

Using a playful interactive learning toy, instructional embodying tasks can assist play, which are normally communicated verbatim, using tangible objects instead.

Toys should assist instruction using multimodal feedback; therefore objects should embody multimodal communicative qualities such as haptic, sound and lights.

This could provide a seamless and engaging experience that moves focus away from mono-directional, instructional learning towards experiential and enactive methods of cognition.

B. *Qualitative Study Two:Cross-Generational Instructional Co-Cooking Using Real Food Between Mother & Daughter via Skype*

The second study was designed to test co-cooking activities via telecommunication channel. The aim of the daughter, who has basic cooking knowledge, wants to learn more sophisticated recipes from mother using a computer-mediated environment.

1) *Description of Participants*

In order to understand how computer-mediated environments affected established relationships, subjects with close relationships were chosen. The chosen subjects were mother aged 58 and daughter aged 26.

Mom lives in Shanghai while daughter lives in Singapore. Although there was no temporal distance, geographically, the distance was real, which added to the experience of the experiment subjects.

The mother had sophisticated skills in cooking, whereas the daughter was familiar with cooking simple dishes. Both mom and daughter have had some experience cooking together before but never through the Internet via a computer mediated environment. Both mom and daughter have had no formal cooking instruction training.

2) *Observations and Learning*

During the exercise, both experimenters noticed and participants reflected that communication was mostly to convey control timing and temperature, order of actions and amount of ingredients.

Both parties we mostly occupied by audio cueing and did not paying attention to the video stream during the cooking phase. This could be because of the physical coordination needed in the cooking process, as well as the attention to things such as colour of food, taste while cooking, etc.

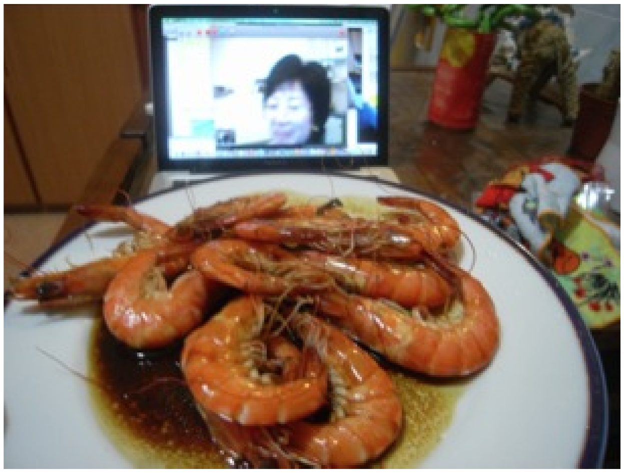


Figure 2. Mother and daughter co-cook using Skype on Internet

Participants naturally developed easy-to-follow expressions for measuring. For example “throw in three spoons of sugar; mix with one spoon of water; warm up the oil for three minutes; now is about right”, etc.

Both mom and daughter mentioned that they wish that the video camera was better placed, as opposed to physically attached to their display on the notebook computer, suggesting that current technology does not adequately support such types of interaction.

Follow up results reveal that daughter has built up confidence for learning more sophisticated recipes using this type of communication. She also expressed confidence in recreating the dish because she has the opportunity to cook together with her mom, albeit through use of the Internet and not face-to-face. Both parties also expressed that the experience was enjoyable and it was good to be engaged with an activity as an alternate means to spend time together using a computer-mediated environment over long distances.

These outcomes support the second hypothesis that learning could be enhanced using enactive and multi-modal activity, even when using a computer-mediated environment.

C. *Qualitative User Study Three: Gestural Study Using Chopsticks*

Finally, in order to understand how users would negotiate food with chopsticks specifically,

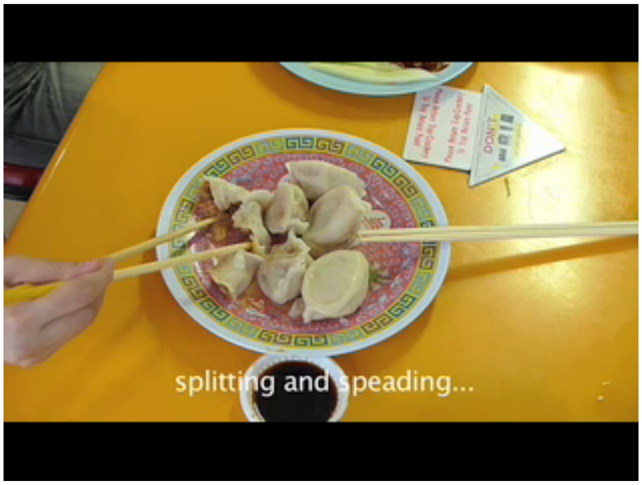


Figure 3. Video still of chopstick use gesture analysis

we conducted an observational study of groups of people eating food with the utensil. The observations we recorded and analysed in order to extrapolate specific gestures that were used with the utensil. These gestures were then mapped to specific actuations in our prototype that triggered sounds produced by our software server-side. Corresponding food attuned to the utensil was used, in this case Chinese food.

V. System Description

Chopsticks embodied the first prototype of the playful utensils system. Chopsticks were chosen for its potentially rich and diverse set of gestures. It also addresses a cultural significance in regards to Asia in which the Playful Utensils system was developed for.

To make each utensil easy to use, only the power supply, gestural data acquisition module and wireless unit are located on the chopsticks.

The server side software sub-system running Cycling '74's Max/MSP is used to analyse gestural data and generate music. More details are introduced in the next section.

A. Hardware Design

The system electronics are divided into two main subsystems. Built from proven, off-the-shelf hardware, the system is very robust. The two main subsystems, which are the utensils and the server, are described below.

1) Utensils

As seen in Figure 4, this is the third iteration prototype, which is lightweight to achieve robust gestural capturing during use, reliable network performance and low power consumption.

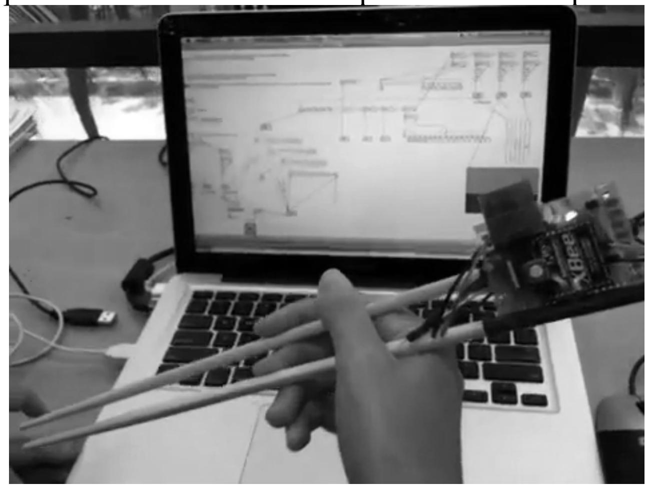


Figure 4. The Playful Utensils prototype system

2) Sensing

After experimentation with different sensing methods such as photometers and linear potentiometers, accelerometer and compass sensing was finalized for the actuation translation of gestures. This decision provides a fluid and easy way for users to interact with the system without having to press a complex series of touch-points as found in a linear potentiometer, and is independent of ambient lighting as photometers often are.

The accelerometer and compass units are connected to a custom-made circuit. We use the Analogue Device ADXL345 3-way accelerometer for our current sensing needs. This provides us with a small, thin, low-powered unit with 13-bit resolution and tilt and dynamic acceleration sensing capabilities.

3) Networking

Also attached to the breadboard is an Xbee wireless module that uses the 802.15.4 protocol stack on the 2.4GHz wideband spectrum, providing the utensils with simple and reliable, low-power, wireless connectivity to the server.

4) Power Supply

A quick-charging, flat, compact and regulated, 3.7v, 110mAh lithium ion battery powers the entire utensil unit.

5) Server

To achieve real-time processing of signals, as human hearing could detect tiny time delays in sound feedback, all data processing is streamed onto a server. An off-the-shelf, late 2010 Macbook Pro computer provides enough

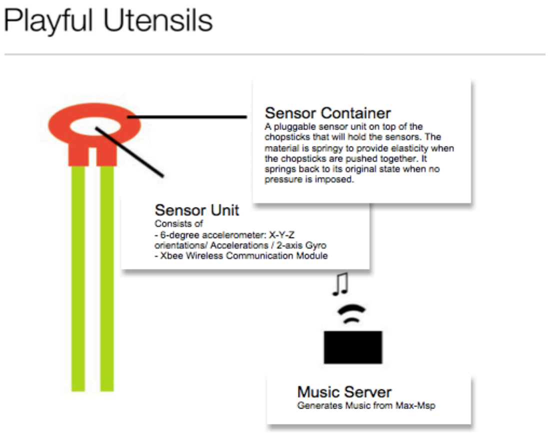
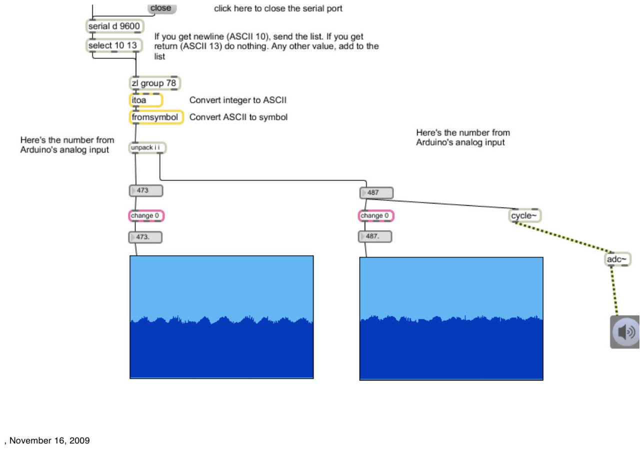

Playful Utensils
Sensor Container
A pluggable sensor unit on top of the chopsticks that will hold the sensors. The material is springy to provide elasticity when the chopsticks are pushed together. It springs back to its original state when no pressure is imposed.
Sensor Unit
Consists of
- 6-degree accelerometer: X-Y-Z orientations/ Accelerations / 2-axis Gyro
- Xbee Wireless Communication Module
Music Server
Generates Music from Max-Msp

Figure 5. Diagram of the Initial Prototype

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-236-3



If you get newline (ASCII 10), send the list. If you get return (ASCII 13) do nothing. Any other value, add to the list

Here's the number from Arduino's analog input

Here's the number from Arduino's analog input

, November 16, 2009

Figure 6. Gestural data acquired from utensil-embedded accelerometer in Max/MSP

processing power for real-time performance, and is a reliable platform to run music generation programs.

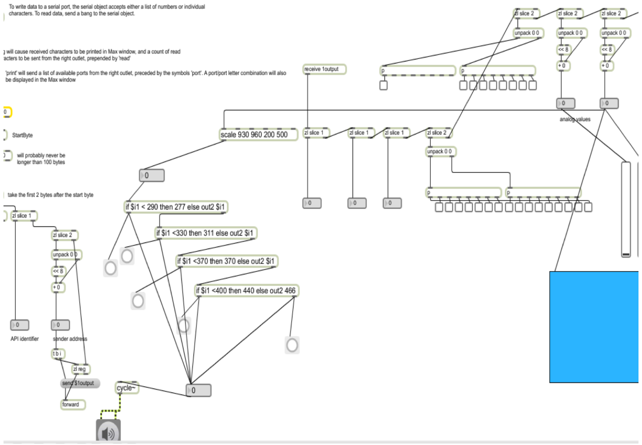
The Playful Utensils system server runs Cycling '74's Max/MSP. The program co-ordinates the utensils, as well as plays audio according to the gestural information. All data is transferred wirelessly to an Xbee receiver attached to the server via USB.

B. *Software Design*

The software sub-system that controls the analysis of gestures and playback of appropriate audio was written in Cycling '74's Max/MSP. Max/MSP. It was chosen because it is a high-level, object-oriented programming environment designed for creative practitioners.

The visually oriented interface is relatively easy to use, is well supported by the software developers who publish it. It also has a strong community of end-users. For these reason it is an ideal development platform for other developers to appropriate and modify our system in order to promote widespread adoption for research purposes.

When food is prepared or eaten with the utensils, the gesture data is recorded by the accelerometer and is transported over serial connection using wireless network connectivity. This data is placed in a list and is decoded then filtered by the software, triggering audio in the pentatonic scale within a ± 2-octave range. A digital-to-analogue converter built into the software platform interprets the data, which then plays the appropriate frequency through output speakers.



To write data to a serial port, the serial object accepts either a list of numbers or individual characters. To read data, send a bang to the serial object.

; will cause received characters to be printed in Max window, and a count of read acters to be sent from the right outlet, prepended by 'read'

'print' will send a list of available ports from the right outlet, preceded by the symbols 'port'. A port/port letter combination will also be displayed in the Max window

will probably never be longer than 100 bytes

take the first 2 bytes after the start byte

Figure 7. The Playful Utensils Cycling '74's Max/MSP patch

VI. Future Work

We are currently looking into the technologies needed to achieve the most nurturing and intuitive user interaction possible for the Playful Utensils system. A prototype has been recently developed and once more affordances in relations to the above-mentioned studies are considered, researchers can further work with families to accurately design the user experience that is needed to express the objectives of the Playful Utensils system.

The initial prototype will also be shared with chefs in order to extrapolate an accurate way to generate sounds within the boundaries of use for each utensil that is explored. Game-like functionality will also be explored. Finally, composers and musicians will be consulted so that researchers can begin to create a new type of cooking and dining musical notation based on the Playful Utensils systems which will hopefully lead to the publication of a Playful Utensils song and recipe book.

VII. Conclusion

We presented a qualitative study regarding intergenerational co-cooking, instruction and leaning using a computer-mediated environment, which lead to the design and implementation of an activity based utensil.

We have learned that playful activity can improve the quality of engagement and communication.

Using these findings, we have developed an initial prototype, which we will use to further study shared, communal and collaborative activity in a co-cooking and co-dining environment using computer-mediated environments.

ACKNOWLEDGMENTS

This research is carried out under CUTE Project No. WBSR-7050000-100-279 partially funded by a grant from the National Research Foundation (NRF) administered by the Media Development Authority (MDA) of Singapore.

REFERENCES

[1] K. Tee, A.J. Bernheim Brush, and K. M. Inkpen, “Exploring communication and sharing between extended families,” Int. J. Hum.-Comput. Stud., vol.67, Feb. 2009, pp. 128–138, doi: 10.1016/j.ijhcs.2008.09.007

[2] S. Lindley, R. Harper and A. Sellen, “Desiring to be in touch in a changing communications landscape: attitudes of older adults,” Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI 09), ACM Press, Apr. 2009, pp. 1693-1702, pp. 10.1145/1518701.1518962

[3] T. K. Judge, C. Neustaedter, and A. F. Kurtz, “The family window: the design and evaluation of a domestic media space,” Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI 10), ACM Press, Apr. 2010, pp. 2361-2370, doi:10.1109/SCIS.2007.357670.

[4] P. Barden, R. Comber, D. Green, D. Jackson, C. Ladha, T. Bartindale et al., “Telematic dinner party: designing for togetherness through play and performance,” Proc. ACM Conf. Designing Interactive Systems (DIS 12), ACM Press, Jun. 2012, pp. 38-47, doi: 10.1145/2317956.2317964

[5] C. Jackie Lee, L. Bonanni, J. H. Espinosa, H. Lieberman, T. Selker, “Augmenting kitchen apliances with shared context using knowledge about daily events,” Proc. ACM Conf. Intelligent User Interface (IUI 06), ACM Press, Jan. 2006, pp. 348-350, doi: 10.1145/1111449.1111533

[6] G. M. Burghardt, The Genesis of Animal Play: Testing the Limits. Cambridge, MIT Media Press, 2006.

[7] Y. Chang, J. Lo, C. Huang et al., “Playful toothbrush: ubicomp technology for teaching tooth brushing to kindergarden children,” Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI 08), ACM Press, Apr. 2008, pp.363-372, doi: 10.1145/1357054.1357115

[8] D. Cohen, The Development of Play. New York, NY, Routledge, 2006.

[9] L.Dib, D. Petrelli, S. Whittaker, “Sonic souvenirs: exploring the paradoxes of recorded sound for family remembering,” Proc. ACM Computer Supported Cooperative Work (CSCW 10), ACM Press, Feb. 2010, pp.391-400, doi: 10.1145/1718918.1718985

[10] D. M. Frohlich, Audiophotography: bringing photos to life with sounds. Dordrecht, The Netherlands, Kluver Academic Publisher, 2004.

[11] A. Grimes and R. Harper, “Celebratory technology: new directions for food research in HCI,” Proc. SIGCHI Conf. Human Factors in Computing Systems (CHI 08), ACM Press, Sep. 2008, pp. 467-476, doi: 10.1145/1357054.1357130

[12] E. Hoven, and B. Eggen, “Informing augmented memory system design through autobiographical memory theory,” Personal Ubiquitous Comput., vol 12, Aug. 2008, pp 433-443, doi: 10.1007/s00779-007-0177-9

[13] L. Yeung, W. Ling, “A study of perceptions of food preperation skills in Hong Kong adolescents,” Journal of the HEIA, vol. 14, no. 2, 2007, pp. 16-24

[14] M. Nunes, S. Greenberg, C. Neustaedter, “Sharing digital photographs in the home through physical mementos, souvenirs, and keepsakes,” Proc. ACM Conf. Designing Interactive Systmes (DIS 08), ACM Press, Apr. 2008, pp. 250-260, doi:10.1145/1394445.1394472

[15] D. Petrelli, S. Whittaker and J. Brockmeier, “Autotopography: what can physical mementos tell us about digital memories?” Proc. SIGCHI Conf. Human Factors in Computer Systems (CHI 08), ACM Press, Apr. 2008, pp. 53-62, doi: 10.1145/1357054.1357065

[16] S. Rumsey, “Cooking, recipes, and work ethic: passage of a heritage literacy practice,” Journal of Literacy and Technology, vol. 10, no. 1, pp. 69-9, Apr. 2009

[17] Y. Sugiura, D. Sakamoto, et al. “Cooking with robots: designing a household system working in open environments,” Proc. SIGCHI Conf. Human Factors in Computer Systems (CHI 10), ACM Press, Apr. 2010 pp. 2427-2430, doi: 10.1145/1753326.1753693

[18] Whittaker, S., Hirschberg, B. Amento, L. Stark, M. Bichianni, P. Isenhour et al., “SCANMail: a voicemail interface that makes speech browsable, readable and searchable,” Proc. SIGCHI Conf. Human Factors in Computer Systems (CHI 02), ACM Press, Apr. 2002, pp. 275-282, doi: 10.1145/503376.503426

[19] L. M. Youngblade, J. Dunn, “Individual differences in young children’s pretend play with mother and sibling: links to relationships and understanding of other people’s feelings and beliefs,” Child Development, vol. 66, Oct. 1995, pp.1472-1492, doi:10.1111/j.1467-8624.1995.tb00946.x

[20] Guitar Hero http://hub.guitarhero.com/

[21] Rock Band http://www.rockband.com/

[22] Skype http://www.skype.com

HDR Video Compression Using High Efficiency Video Coding (HEVC)

Yuanyuan Dong, Panos Nasiopoulos
Electrical & Computer Engineering Department
University of British Columbia
Vancouver, BC
{yuand, panos}@ece.ubc.ca

Mahsa T. Pourazad
TELUS Communications Inc. &
University of British Columbia
Vancouver, BC
pourazad@icics.ubc.ca

Abstract—**It is only a matter of time before High Dynamic Range (HDR) video content becomes commercially available. It is necessary, therefore, to develop proper video compression standards that address the peculiarities of this content and enable the introduction of this technology to the consumer market. So far there is no dedicated standard for HDR content. This paper investigates the performance of the emerging High Efficient Video Coding (HEVC) standard on HDR content and compares it with that of the H.264/AVC standard. Performance evaluations show that HEVC outperforms the H.264/AVC standard by 22.47% to 58.61% in terms of bitrate or 1.02 dB to 4.88 dB in terms of PSNR in the case of HDR content.**

Keywords-HEVC; H.264/AVC; HDR compression

I. Introduction

The human visual system is able to adapt to light conditions at approximately 10,000,000,000:1 contrast or dynamic range, and at a single time instant, human eyes can perceive a dynamic range at the order of 100,000:1 [1]. Contrary to the wide range of light intensity allowed by the human vision system, only a range between the order of 100:1 to 1000:1 – known for this reason as "Low Dynamic Range (LDR)" - is supported by the majority of existing capturing and display devices. A new-generation of imaging systems promises to overcome this restriction by capturing and displaying high dynamic range (HDR) images and videos which contain information that covers the full visible luminance range and the entire color gamut [2].

In order to fully capture and represent the color space and dynamic range visible to human eyes, many solutions have been proposed in recent years. One solution is to combine multiple LDR videos captured at different exposure levels. Recently capturing HDR videos has become even more feasible, due to the availability of novel sensors that allow capturing multiple exposures.

The display industry has also started to take note of the potential of HDR technology. In recent years important developments in HDR display and projection technology have been made. Prototypes of HDR display are built with dynamic ranges of well beyond 50,000:1 according to [3]. Moreover to ensure smooth transition from LDR to HDR service, the backward compatibility with current low dynamic range display systems has been investigated. At the introductory phase of HDR systems, HDR displays (that accept 10-bit or 12-bit signals) and LDR systems (that accept only 8-bit data) will coexist. Thus, the broadcasters should provide both LDR and HDR signals for consumers. To efficiently allow for this overlap, a number of tone-mapping operators have been developed which converts 10-bit or 12-bit high dynamic range content to the 8-bit low dynamic range signal. Simple tone-mapping methods utilize the tone-mapping curve for all the pixels in an image [4] [5]. More sophisticated tone-mapping algorithms consider the local features of each pixel and use local operators to perform tone reproduction [6] [7].

As with all HDR technologies for capture and display, HDR compression is a topic worth more research attention as it is going to enable efficient transmission of HDR video. The transmission of HDR content requires provisions beyond those used in transmission of conventional LDR content. So far there is no dedicated video coding standard for HDR content. Some of the existing video coding standards allow coding of LDR video content with more than 8-bits per pixel. These encoders are optimized to compress video content with the statistical distributions of LDR video with more detail than the traditional LDR video (the same dynamic range though). However, HDR content differs from LDR content as it uniquely has higher color bit-depth with more details in high intensity (brightness) as well as low intensity regions. Overall, this introduces more texture and information, which result in large amounts of data.

To this day, the majority of compression efforts related to HDR have focused on separating HDR to a LDR stream and an enhancement layer both coded with existing 8-bit based standards [8]. This approach comes at a cost of low compression efficiency, but it ensures backward compatibility with the existing LDR displays, and allows reconstruction of the HDR content for HDR displays [6]. Only some very preliminary studies have been done on direct compression of HDR content. The method proposed in [9] adapts HDR signals to the JPEG-2000 coding requirements while the one described in [10] is developed around MPEG.

Among the existing video coding standards the H.264/AVC is the most advanced and efficient video compression standard (developed by the Joint Video Team (JVT) of the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG)). Recently, the international community for standardization has considered a new generation of video compression technology, known as High Efficient Video Coding (HEVC).

This standard is offering substantially higher compression capability than the existing H.264/AVC standard. Current comparison results show that HEVC offer superior compression performance compared with H.264/AVC.

The performance of the HEVC standard on HDR content has not been taken into account at the time of developing this standard and all the tests were conducted using LDR content. Given the difference in properties and characteristics between LDR and HDR content, it is important to consider how HEVC will perform on HDR video and from these tests try to identify challenges and additions or changes to the new standard.

In this paper we investigate the compression performance of HEVC on HDR video content, to examine if HEVC has the potential to be used as a platform for a devoted HDR compression scheme. We conduct experimental tests on HDR content and compare the performance of HEVC with that of H.264/AVC standard. Comparable experiment settings of two codecs are introduced which could also be used in other similar tests.

The remainder of the paper is structured as follows. Section II provides a brief background on the formats of HDR content and the specifications of high efficiency video coding technology, Section III presents the details of our experiment, Section IV discusses the results and future work and Section V includes conclusions.

II. BACKGROUND

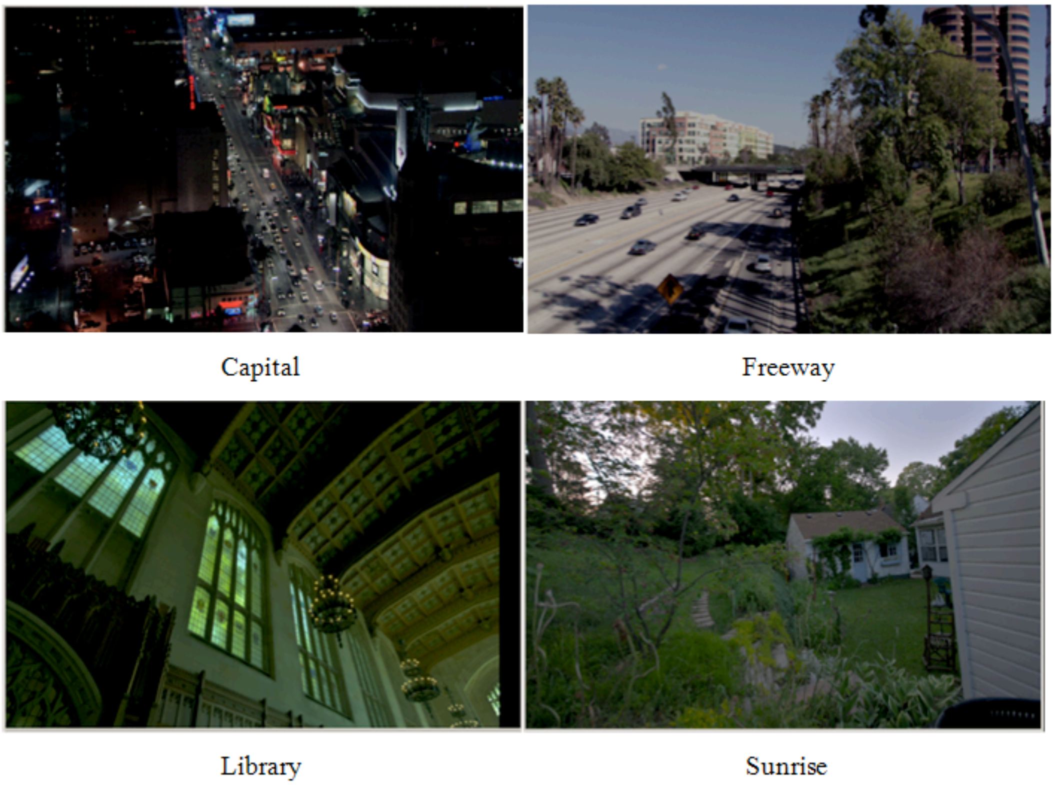
This section provides a brief background on the HDR content and the emerging high efficiency video coding technology.

A. *High Dynamic Range (HDR) Format*

HDR imaging offers the opportunity of capturing, storing, manipulating, and displaying dynamic real-world lighting. HDR signals preserve colorimetric or photometric pixel values (such as CIE XYZ) within the visible color gamut and allow for intra-frame contrast to reach the magnitude of 10^6:1, without introducing contouring, banding or posterization artifacts caused by excessive quantization. The photometric or colorimetric values, such as luminance ($\text{cd}\cdot\text{m}^{-2}$) or spectral radiance ($\text{W}\cdot\text{sr}^{-1}\cdot\text{m}^{-3}$), span to a much larger range of values than the luma and chroma values (gamma corrected) used in typical video encoding (JPEG, MPEG, etc.). In order to represent the dynamic range of intensities found in a real life scene, we need to use more than the typical 8-bits for each color. An intuitive solution is to represent the pixel value with floating point numbers to cover the larger dynamic range. One shortcoming of using floating-point numbers is that compression of HDR content becomes challenging since floating-point numbers are not optimal for compression compared to integer values. The other issue is that the precision error of floating point numbers varies across the full range of possible values. For these reasons, several file formats have originally been proposed for storing HDR data, including the Radiance RGBE (.hdr) [13], OpenEXR (.exr) [14], and LogLuv TIFF (.tiff) [15]. The RGBE format assigns four bytes to represent each pixel: one byte used for the mantissa of each of the RGB channels and the remaining one byte is used as a shared exponent. The exponent byte together with the mantissa part is able to represent a value of a very large range. On the other hand, OpenEXR spends 16 bits for each of the RGB channels: a sign bit, five bits for exponent and ten bits for mantissa. The LogLuv TIFF format represents the data in the logarithmic domain and supports 32 bits per pixel using one sign bit, 15 bits to encode the log scale of the luminance, and 8 bits for each of the two chrominance channels. These three formats are considered nearly lossless and require high data rate.

B. *High Efficiancy Video Coding (HEVC)*

The recent advances in technology have made it possible to capture and display video material with ultra-high definition (UHD) resolution. To enable transmission of large amounts of data the ISO/IEC Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG) established a Joint Collaborative Team on Video Coding (JCT-VC) with the objective to develop a new high-performance video coding standard. A formal Call for Proposals (CfP) on video compression technology was issued in January 2010, and 27 proposals were received in response to that call [16]. The evaluations that followed showed that some proposals could reach the same visual quality as H.264/MPEG-4 AVC High profile at only half of the bitrate and at the cost of two to ten times increase in computational complexity. Since then, JCT-VC has put a considerable effort towards standardization of a new compression technology known as the High Efficiency Video Coding (HEVC), with the aim to significantly improve the compression efficiency compared to the existing H.264/AVC high profile. Generally speaking, HEVC is a block-based compression scheme, similar to H.264/AVC, with some new features. Some of the key elements of HEVC compared to H.264/AVC are: flexible block structure (recursive quad-tree partitioning and block sizes up to 64x64 pixels), more intra prediction modes (35 in total), improved motion vector estimation, and different integer transforms allowing non-square transforms. HEVC also includes two new filters (Sample Adaptive Offset (SAO) and Adaptive Loop Filter (ALF)) to undo the distortion introduced in the main steps of the encoding process (prediction, transform and quantization) [11]. The effort for standardization of HEVC is still ongoing, and it is expected to be finalized by July 2012. So far, the objective comparison results reported in [11] show that the current HEVC design outperforms H.264/AVC by 29.14% to 45.54% in terms of bitrate or 1.4dB to 1.87dB in terms of PSNR. Subjective comparison of the quality of compressed videos – for the same (linearly interpolated) Mean Opinion Score (MOS) points - shows that HEVC outperforms H.264/AVC, yielding average bitrate savings of 58% [12]. Note that all the reported performance evaluations are based on LDR content.



Capital Freeway

Library Sunrise

Figure 1. Snap shot of the test sequences.

Conventionally, video compression techniques have considered only 8 bits-per-pixel (bpp) input videos, yet HDR videos require 10-14 bpp. The current design of HEVC provides the necessary capabilities to handle LDR videos of up to 14 bpp without clipping the bit depth during the encoding process. This allows us to encode HDR content using HEVC. However, the compression performance might not be optimal, since HEVC is optimized to compress video content with the statistical distributions of LDR video but not HDR video.

III. EXPERIMENT

In this paper, our objective is to test the performance of HEVC for compressing HDR content, and compare it with that of H.264/AVC. The following subsections elaborate on the details of our experiment.

A. *Test sequences*

For our experiment, four test sequences are selected from the database provided by JVT of ISO/IEC MPEG & ITU-T VCEG [17] [18]. These test videos are in YUV 4:2:0 format, with a resolution of 1080p and a frame rate of 50 fps. The dynamic range of two of the videos is 10 bits and that of the other two is 12 bits. Fig. 1 shows a snap shot of the four test sequences. The specifications of the test sequences are summarized in Table I.

TABLE I HDR TEST SEQUENCES

Name	Bit Depth	Resolution	Frame Rate
Capital	10	1920x1080	50 fps
Freeway	10	1920x1080	50 fps
Library	12	1920x1080	50 fps
Sunrise	12	1920x1080	50 fps

These test sequences have been generated form high dynamic range video content that was originally stored in floating point format and in a linear RGB space. The representation of the sequence was created by first normalizing the RGB values to the set [0, 1]. Then these normalized values were converted to the YCbCr format using the ITU-R BT.709 reference primaries. Chroma planes were subsampled by a factor of two in each dimension using the given separable filter (refer to [17] for more details). Finally, the resulting 4:2:0 YUV file was quantized linearly with a rounding operation to create the test sequences [18].

B. *HEVC configuration*

To evaluate the performance of HEVC on HDR content we used the High Efficiency Video Coding Test Model 5 (HM 5.0) [19]. Note that HM 5.0 was the latest available

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-236-3

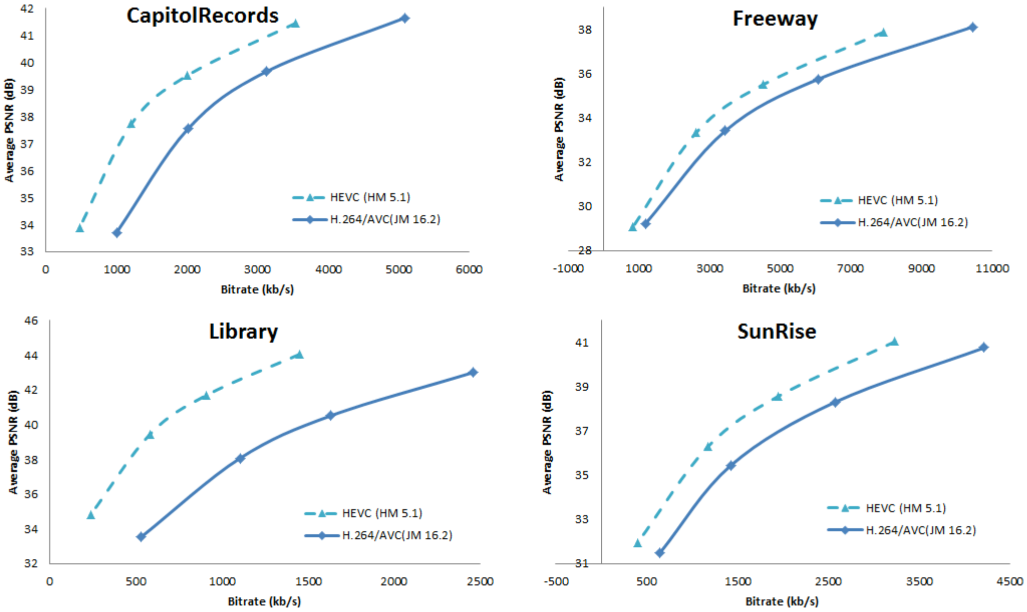


Figure 2. RD curves for HDR content.

HEVC Test model at the time of conducting this experiment. To enable the highest possible compression performance, the Random Access High Efficiency (RA-HE) configuration is used in our experiment: Hierarchical B pictures, Group of Picture (GOP) length of 8, ALF (Adaptive Loop Filter), SAO (Sample Adaptive Offset) and Rate Distortion Optimized Quantization (RDOQ) were enabled. In order to obtain a reasonable span of Rate-Distortion (RD) curves, the following Quantization Parameters (QPs) were used: 28, 32, 36, and 44. QP is the parameter, which controls the quantization step size, and in turn decides the level of quantization error involved during compression. A higher QP value leads to a larger quantization step size and worse video quality.

C. H.264/AVC configuration

In our experiment, the performance of HEVC is compared with the state-of-the-art video compression standard H.264/AVC (JM 16.2). To accommodate HDR content, the configuration of H.264/AVC was set to High 4:4:4 Profile, which accepts up to 14 bits. In our experiment we used hierarchical B pictures, GOP length of 8, CABAC entropy coding and RDOQ enabled. These settings were recommended for comparing H.264/AVC to HEVC by MPEG/VCEG in the Joint Call for Proposals (for more details check the Alpha anchor in [20]). The same QP settings as those in the HEVC case are used for H.264/AVC.

All the above-mentioned configuration settings were chosen to ensure a fair comparison between HEVC and H.264/AVC. However, these codecs are so different and have different tools and configuration options. As a result, aside from the necessary changes and above-mentioned settings, the default settings are used for the rest of available options.

IV. RESULTS AND DISCUSSION

To evaluate the performance of HEVC versus H.264/AVC for coding HDR content, we conducted our experiment using the infrastructure provided in the previous section. Fig. 2 shows the RD curves for all the test sequences and Table II lists the average PSNR improvement and average PSNR savings achieved by HEVC over the H.264/AVC standard.

As it can be observed, HEVC outperforms H.264/AVC by 22.47% to 58.61% in terms of bitrate (with same PSNR) or 1.02 dB to 4.88 dB in terms of PSNR (with same bitrate). Our results show that the compression efficiency of HEVC when applied to HDR content is dramatically higher than H.264/AVC and seems to follow the performance already witnessed for LDR content. In future work, we will include a new rate-distortion optimization process with HEVC that features an updated signal model and coding parameters derived specifically for HDR content. Moreover, subjective tests using an HDR display will be conducted to evaluate the performance of the two standards and the proposed schemes.

TABLE II AVERAGE COMPRESSION IMPROVEMENT

Name	Average PSNR Improvement	Average Bitrate Saving
Capital	1.26 dB	42.12 %
Freeway	1.02 dB	22.74 %
Library	4.88 dB	58.61 %
Sunrise	1.79 dB	32.60 %

V. CONCULUSION

This paper compared the performance of the current HEVC test model with the state of the art compression standard, H.264/AVC, for compressing HDR content. Configuration settings used for this study were chosen carefully to represent similar scenarios and ensure a fair comparison. Our experiment results show that HEVC outperforms H.264/AVC by 22.47% to 58.61% in terms of bitrate (with same PSNR) or 1.02 dB to 4.88 dB in terms of PSNR (for the same bitrate).

The current progress of HEVC is proved to be promising and HEVC has the potential to replace H.264/AVC as the next state-of-the-art compression standard. Our study confirms that HEVC does not only offer superior compression performance for LDR content but also HDR videos. The compression improvement in the HDR case is in line with that of LDR. However the overall saving differs among different sequences (content dependent). Service providers could greatly benefit from HEVC due to more efficient use of bandwidth. It is worth mentioning that HEVC's high compression performance comes at the price of increased coding complexity compared to H.264/AVC.

With the rapid growth in the multimedia industry, it is only a matter of time before HDR videos become widespread. HEVC and future compression standards should be optimized accordingly to fully capitalize on this upcoming trend.

REFERENCES

[1] J. A. Ferwerda, “Elements of Early Vision for Computer Graphics,” IEEE Computer Graphics and Applications, vol. 21, no. 5, pp. 22–33, 2001.

[2] E. Reinhard, G. Ward, S. Pattanaik, and P. Debevec, “High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting,” Morgan Kaufmann, 2005.

[3] H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and A. Vorozcovs, “High dynamic range display systems”, ACM Trans. Graphics, vol. 23, no. 3, pp. 760-768, 2004.

[4] G. W. Larson, H. Rushmeier, and C. Piatko, “A Visibility Matching Tone Reproduction Operator for High Dynamic Range Scenes,” IEEE Transactions on Visualization and Computer Graphics, vol. 3, no. 4, pp. 291-306, 1997.

[5] F. Drago, K. Myszkowski, T. Annen, and N. Chiba, “Adaptive logarithmic mapping for displaying high contrast scenes,” Computer Graphics Forum, vol. 22, no. 3, 2003.

[6] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, “Photographic tone reproduction for digital images,” ACM Trans. Graphics, vol. 21, no. 3, pp. 267-276, 2002.

[7] F. Durand and J. Dorsey, “Fast bilateral filtering for the display of high-dynamic-range images,” ACM Trans. Graphics, vol. 21, no. 3, pp. 257-266, 2002.

[8] R. Mantiuk, A. Efremov, K. Myszkowski, and H.-P. Seidel, “Backward compatible high dynamic range mpeg video compression,” ACM Trans. Graphics (Proc. SIGGRAPH), vol. 25, no. 3, pp. 713–723, 2006.

[9] R. Xu, S.N. Pattanaik, and C.E. Hughes, “High-Dynamic-Range Still-Image Encoding in JPEG 2000,” IEEE Computer Graphics and Applications, vol. 25, no. 6, pp. 57–64, 2005.

[10] R. Mantiuk, G. Krawczyk, K. Myszkowski, and H.-P. Seidel, “Perception-Motivated High Dynamic Range Video Encoding,” ACM Trans. Graphics (Proc. SIGGRAPH04), vol. 23, no. 3, pp. 730–738, 2004.

[11] M. T. Pourazad, C. Doutre, M. Azimi, and P. Nasiopoulos “HEVC: The New Gold Standard for Video Compression,” IEEE Consumer Electronic Magazine, vol.1 , issue 3, pp. 36-46, July 2012.

[12] G. J. Sullivan, J.-R. Ohm, F. Bossen, and T. Wiegand, “JCT-VC AHG report: HM subjective quality intestigation,” JCTVC-H0022, February 2012.

[13] G. Ward, “Real pixels,” Graphics Gems II, pp. 80–83, 1991.

[14] R. Bogart, F. Kainz, and D. Hess, “Openexr image file format,” ACM SIGGRAPH, Sketches & Applications 2003.

[15] G. W. Larson, “Logluv encoding for full-gamut, high-dynamic range images,” Journal of Graphics Tools, vol. 3, no. 1, pp. 15–31, 1998.

[16] G. J. Sullivan and J.-R. Ohm, "Recent developments in standardization of high efficiency video coding (HEVC)," Proc. SPIE, vol. 7798, 2010.

[17] Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, “New Test Sequences in the VIPER 10-bit HD Data,” JVT-Q090, October, 2005.

[18] Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, “Donation of Tone Mapped Image Sequences,” JVT-Y072, October, 2007.

[19] K. McCann, B. Bross, W.-J. Han, S. Sekiguchi, and G. J. Sullivan, “High Efficiency Video Coding (HEVC) text specification draft 5,” JCTVC-G1102, November 2011.

[20] ISO/IEC JTC1/SC29/WG11, “Joint Call for Proposals on Video Compression Technology,” N1113, January 2010.

Development of a 3D Stereo Comic Creation Tool and a Display System for the 3D Android Smartphone

Shigeru Sasaki, Masafmi Furuta
Department of Human Information Systems
Teikyo University
Utsunomiya, Japan
e-mail: sasaki@ics.teikyo-u.ac.jp
11M107@uccl.teikyo-u.ac.jp

Seiichi Tanaka, Marika Kishi, Yui Takifuji
Department of Art
Bunsei University of Art
Utsunomiya, Japan
e-mail: seiichi-tnk@jcom.home.ne.jp
kishi.m.ggc@gmail.com
takifuji.yui.1217@gmail.com

***Abstract*—In recent years, three-dimensional (3D) stereo content has become familiar, and portable information devices equipped with a naked-eye 3D liquid-crystal display, such as smartphones, have appeared on the market. In addition, images created by conventional hand drawing, such as comics, are being created more frequently using a computer. In this study, we developed a 3D stereo comic creation tool in which a stereo image is created from pictures drawn on layers. We also developed a 3D stereo comic viewer application program that displays 3D pictures on a smartphone. This 3D stereo comic content creation technique is not intended to replace techniques based on 3D computer graphics, but is a useful, additional technique for the creation of 3D stereo comic content.**

Keywords-3D stereo comics; smartphone; naked-eye 3D liquid-crystal display; Android application; Java3D.

I. INTRODUCTION

In recent years, three-dimensional (3D) stereo content has become familiar, and portable information devices equipped with a naked-eye 3D liquid-crystal display, such as smartphones, have appeared on the market. Many of the 3D stereoscopic images for these devices are created using 3D computer graphics (CG). However, special technology and knowledge are needed for the creation of 3DCG, and it takes considerable time and effort.

In addition, images conventionally created by hand drawing, such as comics, also are being created more frequently using a computer. When creating a hand-drawn picture on a computer, a part of the picture is drawn on a transparent sheet called a layer, and the technique of stacking them to create one picture is used. Usually, even if a lot of comic creators have the skills to hand draw pictures on layers, they rarely have the skills to create 3DCG. The authors have succeeded in creating 3D stereo comics by adding depth to individual layers. Although this technique differs from the conventional manner of creating 3D stereo images using 3DCG, it is expected that 3D stereo pictures that incorporate the benefits of hand drawing will be created using our method.

The remainder of the paper is organized as follows. In Section II, we describe the state-of-the-art in 3D stereo comic creation. Section III describes the aim of our research. Section IV describes outline of the system. Section V describes creation procedure for 3D stereo images. In Section VI we describe questionnaire survey we conducted. Section VII provides discussions. Finally we summarize our result in Section VIII.

II. STEATA-OF-THE-ART IN 3D STEREO COMIC CREATION

Stereoscopic 3D images from a work in a comic book are created by the method using 3DCG. Creation of these works is mainly carried out as animation [1]. There are also many works that contain 3D images that were created from hand-drawn pictures, and hand-drawn animations have been converted into stereoscopic 3D images by adding depth to individual pixels [2][3].

While the creation of an animated movie is a large-scale project carried out by a lot of people, the work of creating comics is on a small scale and is usually performed by one person or a small group of people.

In Japan, many comic magazines are published every day. These techniques are unsuitable for converting a lot of hand-drawn comics into stereoscopic images immediately. Moreover, it is desirable that you can read numbers of newly released stereo comics on a ubiquitous portable device.

In this research, although it was not a precise but a simple stereo effect, we propose an easy technique of creating stereoscopic images. And we also create a viewer program for displaying those stereo comic contents. Such a tool has not been developed until now and it is useful for the comic creators who are going to make stereoscopic contents personally or in a small group of people.

III. AIM OF RESEARCH

The aims of this study are to provide a technique by which comic creators can easily develop 3D comics and an application through which people can easily enjoy 3D comics—like a digital book—on a smartphone equipped with a 3D liquid-crystal display. To this end, we developed a 3D stereo comic creation tool, in which stereo images are created from pictures drawn on layers. In addition, we developed a 3D stereo comic viewer application program

ISBN: 978-1-61208-236-3

that displays 3D pictures on a smartphone with various effects. Using these tools, we converted comic content into 3D stereo images and displayed these images on a smartphone.

IV. Outline of System

A. 3D stereo image creation tool

In our 3D stereo image creation tool, a graphics file (in Photoshop format) that contains layers is read, and depth is added to each layer. With this tool, depth can be set for a layer in three ways: by setting a fixed depth to a layer, by setting different depths for the four corners of a layer, and by making the color of each pixel of a grayscale image correspond to depth. When depth is added to four corners of a layer, the depth of each pixel of an image is given by bilinear interpolation.

After setting the depth, the spatial relation of a layer can be previewed.

Finally, the tool creates a 3D stereo image using the depth information and saves the image file in jpeg format. The stereo image files were output in side-by-side-half format, because the smartphone used in this research could handle only stereo images in this format. However, the tool can also save stereo images in a standard side-by-side format for other applications. The 3D stereo-image creation tool was made as a Java application. The spatial arrangement of layers can be previewed by 3DCG. In this program, a PSD parser [4] is used for reading files saved in Photoshop format (PSD), and Java3D is used for the 3D preview.

B. Viewer application program

The viewer application program is intended for displaying the 3D stereo images created with the 3D stereo image creation tool on a smartphone equipped with a naked-eye 3D liquid-crystal display. This application can apply a number of effects to comics. The first effect is to place a 2D image on a 3D stereo image, in which process parallax can be added to the 2D image. The second effect is to animate the 2D image displayed on the 3D image. The third effect is to play a sound file when images are displayed. Operations are performed by tapping and flicking. A page is turned over by flicking, and another image—such as a word balloon—is displayed or a sound file is played by tapping. The comic image files are saved in a folder for each work on a microSD card. The viewer program was created as an Android application. In this program, the SHARP SDK AddOn [5] distributed by SHARP was used.

V. Creation of 3D Stereo Images

The procedure for the creation of 3D stereo images is shown below.

A. Save image data in Photoshop format

Pictures with different depths are drawn on different layers. However, even if the order of depth and layer is not in agreement, 3D stereo-image creation can still be performed. When adding depth by using a grayscale image, the grayscale image layer should be placed immediately before (above) the target layer. An example of a grayscale layer for adding depth is shown in Fig. 1. A 3D preview of depth given by a grayscale image is shown in Fig. 2.

The image data are then saved in PSD format.

B. Read a PSD file into the 3D stereo comic creation tool

A PSD file is read into the 3D stereo comic creation tool.

C. Set depth for each layer

There are input columns for setting the depth for each layer. A fixed depth or different depths for each of the four corners can be input. If a check box is unchecked, the layer will not appear on the 3D stereo image. When using a grayscale image for setting depth, the value of the grayscale color to the amplitude of depth is set. The main window of a 3D stereo comic creation tool is shown in Fig. 3.

D. Check spatial arrangement of a layer in 3D view

After setting the depth, the spatial arrangement of a layer can be checked by a 3DCG preview, as shown in Fig. 4.

E. Save 3D stereo-image data in side-by-side-half format

Finally, image data is saved in a side-by-side-half format. A sample of a stereo image is shown in Fig. 5. The image files should be saved in a separate folder for each work. A numbering system (such as 1.jpg, 2.jpg …) is used for naming the image files.

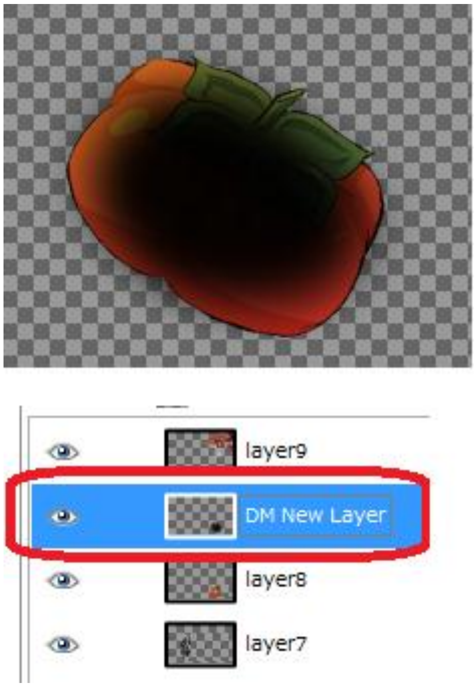


Figure 1. Example of a grayscale layer that adds depth

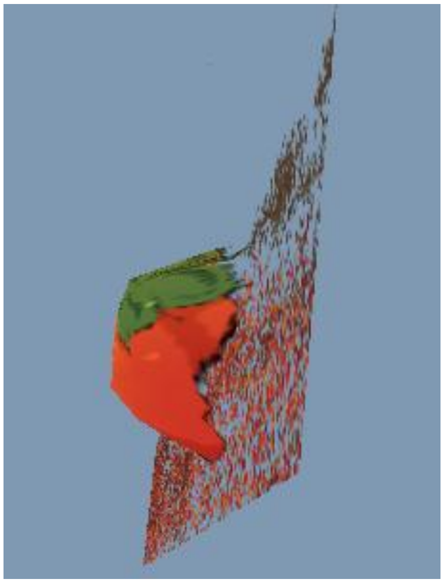
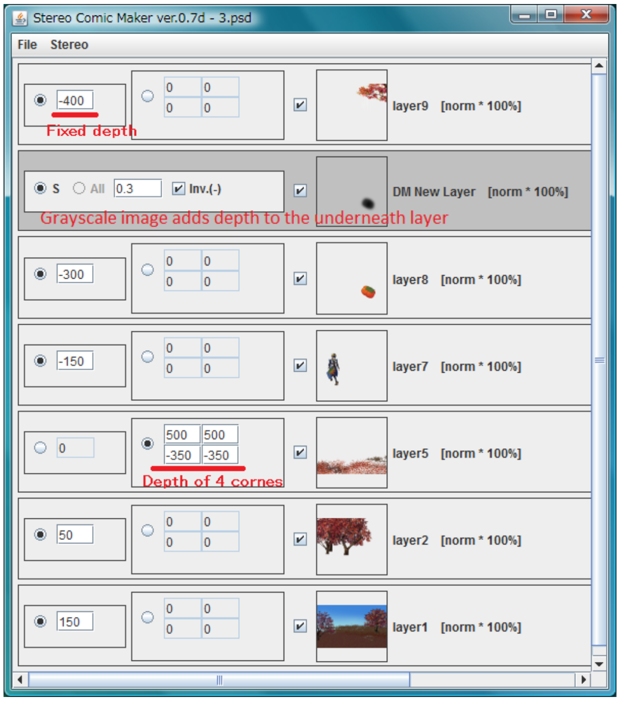


Figure 2. Depth added by grayscale image



● -400	○ 0 0 / 0 0	☑	layer9 [norm * 100%]
● S ○ All 0.3 ☑ Inv.(-)		☑	DM New Layer [norm * 100%]
● -300	○ 0 0 / 0 0	☑	layer8 [norm * 100%]
● -150	○ 0 0 / 0 0	☑	layer7 [norm * 100%]
○ 0	● 500 500 / -350 -350	☑	layer5 [norm * 100%]
● 50	○ 0 0 / 0 0	☑	layer2 [norm * 100%]
● 150	○ 0 0 / 0 0	☑	layer1 [norm * 100%]

Fixed depth

Grayscale image adds depth to the underneath layer

Depth of 4 cornes

Figure 3. Main window of the 3D stereo image creation tool

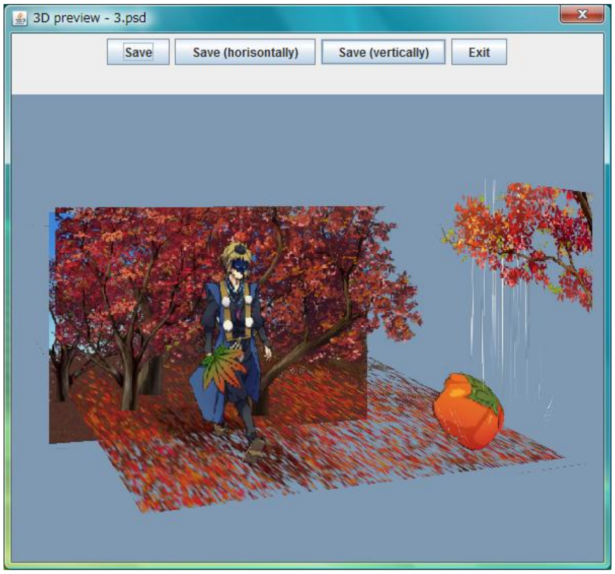


Figure 4. Preview of an entire image in the preview window

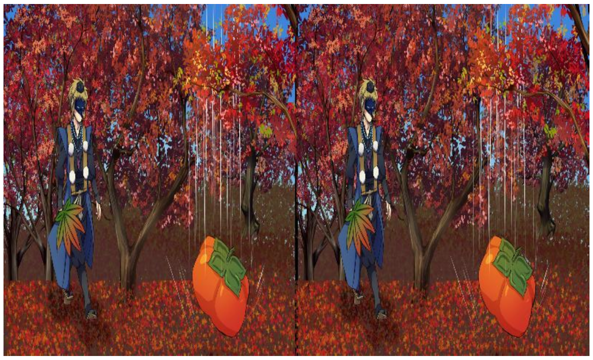


Figure 5. Main window of the 3D stereo image creation tool

VI. QUESTINAIRE SURVEY

After we created 3D comic content using the tool developed in this research, we asked people to view the contents on a smartphone and answer a questionnaire survey. Responses were received from 57 persons, most of whom were under 20 years. In answer to the question whether they found the 3D stereo comic interesting, 91% replied “interesting” or “interesting to some extent.” In answer to the question on whether they would like to continue to read 3D stereo comics, 78% said “yes.” From these responses, we can say 3D stereo comics made by adding depth for individual layers of a Photoshop file are accepted in a friendly, positive manner.

VII. Discussion

The 3D stereo comics created by the tool developed in this research offer the advantage of allowing 3D comic content to be created using conventional comic-drawing techniques.

For comic creators who are already using computers in their work, it would be very easy for them to create 3D stereo comics from their existing digital image data. Two undergraduate students at the University of Art, who are co-authors of this paper, created two 3D comic works by converting 2D images into 3D images using our tool. Although they set the depth numerically, they did not set the depth using grayscale images. We need to add a function that helps the generation of grayscale-depth image easily.

Using this tool, 3D stereo comics can be created without losing the merit of a hand-drawn picture. This 3D stereo comic content creation technique is not intended to replace techniques based on 3DCG computer graphics, but be a useful, additional technique for the creation of 3D stereo comic content.

VIII. Conclusion

In this study, we developed a tool to create a 3D stereo-image from pictures hand drawn on separate layers. We also developed an Android application that displays images created using this tool on a smartphone with a 3D stereo liquid-crystal display. From responses to a questionnaire survey, it seems that our 3D stereo comic content creation method will prove useful as a technique for creating 3D stereo images.

Acknowledgment

The authors would like to thank Enago [6] for the English language review.

References

[1] J. Gaudiosi, “3D wow!,” Computer Graphics World, Volume 32, Issue 9, 2009, pp.31-34

[2] K. Tucker-Fico, E. Goldberg, K. Koneval, D. Mayeda, R. Neuman, Riley, and M. Schnittker, “Design and realization of stereoscopic 3D for disney classics,” ACM SIGGRAPH 2011 Talks, art. no. 12, 2011, pp.12

[3] M. Salvati, M. Kinoshita, Y. Katsura, K. Anjyo, T. Yotsukura, and H. Uchibori, “Developing tools for 2D/3D conversion of Japanese animations,” ACM SIGGRAPH 2011 Talks, art. no. 14, 2011, pp.14

[4] PSD-parser download page, http://blog.alternativaplatform.com/en/2007/07/09/parser-psd-formata/ (last access July 10 2012)

[5] SHDevelopersSquare3D, https://sh-dev.sharp.co.jp/ android/modules/download/?/api_stereo3dlcd (last access July 10 2012)

[6] Enago, www.enago.jp (last access July 10 2012)

Internet of Things to Provide Scalability in Product-Service Systems

Danúbia Espíndola, Nelson Duarte Filho,
Silvia Botelho, Jônata Carvalho
Center for Computational Science
Federal University of Rio Grande
Rio Grande, RS, Brazil
{danubiaes, dmtnldf, silviacb,
jonatatyska}@furg.br

Carlos Eduardo Pereira
School of Electrical Engineering
Federal University of Rio Grande do Sul
Porto Alegre, RS, Brazil
cpereira@ece.ufrgs.br

***Abstract*—In a convergence context between Internet of Things - IoT and Product-Service Systems - PSS, this paper presents a way to integrate concepts and build computing systems to assist PSS. The conceptual basis and a architecture for a middleware to do this convergence was investigated. A prototype of the middleware was built and was ran in a simulated case study in the marine industry. It was concluded that the adopted ideas are feasible and an implementation of the system has viability under performance point of view, in the industry of construction and assembly.**

Keywords; Product-Service Systems; Internet of Things; Human-Computer interaction; Ubiquitous and Pervasive Computing; Marine Industry.

I. INTRODUCTION

Product Service Systems, put simply, are when a firm offers a mix of both products and services, in comparison to the traditional focus on products. PSS, as defined in [1], is "a marketable set of products and services capable of jointly fulfilling a user's needs". PSS can be realized by smart products. In [2], there is also a definition as follows: A PSS is pre-designed system of products, services, supporting infrastructures and necessary networks that is a so-called dematerialized solution to consumer preferences and needs. It has also been defined as a "self-learning" system, one of whose goals is continual improvement.

Researches on Product-Service are not recent. In the computing communities, the first papers date from the 60's [3]. However, terms like servicizing, dematerialization, green industry, functional economy and other concepts aiming at sustainability and competitiveness in production systems are still recent topics of discussion and challenges for the academy [4][5]. Challenges about the implementation and modeling of PSS systems give rise the research area called Service Engineering (SE) [6].

Service Engineering, term introduced by Bullinger in 1995, adopts a technique and systematic approach to the design and development of services using models, methods and tools [7]. One of the major challenges in SE is to deploy service architectures that allow users to incorporate new added-value services to products during the whole product's lifecycle.

Consumers desire new tools to interact with companies offering new services and they want to be able to customize and co-create value. In traditional methods of adding value, value creation is usually a flow from producers to consumers. In co-creation and customization, this distinction disappears once the consumers are engaged in the processes of both defining and creating value [8].

Furthermore, the customization and co-creation are part of a journey of customer's experiences using the products. The initial value is set by producers based on their assumptions on customers' expectations before they purchase the product. As users start interacting with the products and have their personal evaluations resource-based approach to a knowledge-based one, which takes into account users models and perspectives, has to occur.

However, the adoption of this new paradigm for (re)adding value, (re)use and customization, based on the different expectations from distinct agents, require new technologies for its implementation. New Information and Communication Technologies (ICT) infrastructures for acquiring and processing of information, such as smart devices, human-computer interfaces and computational models are required. These infrastructures must describe the relationship between product and service, and manufacturers and consumers, as well as allow the exchanging of knowledge among these agents throughout the product lifecycle. The lack of tools for this purpose is evidenced by the low number of studies in the literature in the Service IT area [7] and, at least so far, it is also due to the difficulty of embedding computing and communication in all elements of PSS.

Besides being useful in PSS context, tools and ICT infrastructures should consider the social issues and the value adding during the whole lifecycle (time domain), in order to lead a general taxonomy that encompasses technology, people and business. Thus, ICT play a fundamental role in supporting the "how to deploy" this transformation from resource-oriented paradigm to knowledge-oriented paradigm. In this sense, new ICT must be considered. For example, pervasive computing (or infiltrating) coupled with mixed reality techniques for human-computer interaction [9][10], might be valuable for discovering new horizons for PSS solutions. The importance of pervasive computing in this context arises from Weiser assertion in 1991: "the most profound technologies are those that disappear…" Also known as ubiquitous computing, the pervasive computing provides an environment highly integrated to user where the perception of being using computers is minimal [11]. It shows clearly that this characteristic is crucial regarding to this approach, since the usage and knowledge about ICT should not be a required condition from its users.

To enable a combination of pervasive computing with the necessity of embedding ICT on all devices of PSS, the Internet of Things (IoT) has emerged as a possible and potential solution. The IoT is a novel paradigm that makes possible the pervasive presence around us of a variety of things or objects. Some devices, such as Radio-Frequency IDentification (RFID) tags, sensors, actuators, mobile phones, and so on, permit, through unique addressing schemes, the interaction and cooperation between things (objects) to reach common goals [12].

Such technological elements, in which information can be collected, processed and accessed from "things", allow the computational modeling of previously disconnected systems. For example, situations, activities and processes that were present only in the social domain can be virtualized in digital frameworks by shaping and aggregating the circulating information among their agents.

In a convergence context between IoT and PSS, this paper presents a way to integrate the concepts and build computing systems to assist PSSs. The conceptual basis and a architecture for a middleware to do this convergence was investigated. A prototype of the middleware was build and was ran in a simulated case study.

The prototype addresses a simulated case study in the marine industry. This choice was made because in this kind of industry the use of information and communication technologies are still scarce and authors have access to a large shipyard that is being installed near of its University laboratories.

In what follows, it is presented the conceptual issues of a model and an architecture of a middleware which aims to be feasible and scalable. A prototype of the middleware is described and finally a simulated case study for validation of the approaches is presented.

II. Conceptual design

The conceptual model that follows should help readers to identify the potential use of IoT in PSS as a solution for development of services in SE.

One of the initial ideas with emergence of PSS systems was to extend the product lifecycle through the addition of services such as maintenance and upgrading [3]. Currently, the focus has changed from selling products to selling functions and services, which could even be downloaded and upgraded remotely – at customers location – allowing a customization of the purchased goods. Within this new context, new research topics related to this product “servicification” have emerged. These methodologies should provide means to evaluate the economic and environmental impacts using the product “servicification”. Service-oriented approaches have been considered as a potential candidate for handling this problem in PSS studies and SOA (Service-oriented architectures) have been often adapted for use in PSS, in which the service provider is the manufacturer and the customer is the receiver.

Considering the elements involved in PSS and following the definition that a service is an activity [13], for the service development and deployment a provider module and a receiver module are required. The provider module must, as the name already implies, provide the service based on product information. This module is implemented in software and can use a set of programming techniques. Among the possible techniques that can be used, intelligent agents with support to expert systems for composition and delivery of services appear as an interesting option. Fig. 1 depicts an high level conceptual design and aims to provide an overview of modules involved in IoT and PSS in order to allow a better understanding of the issues that must be addressed.

Another PSS element that needs to be considered for the composition and delivery of services is the product lifecycle data. Information about products’ features and operational status is taken into account during the creation of services. Product data can be static and dynamic. Static data are generally about product characteristics and are usually related to the BoL (Begin of Life) lifecycle. These data do not change as the time goes by and can be acquired from CAD design models of the product. Dynamic data are data that change in time and correspond to use and end stage product lifecycle (MoL – Middle of Life and EoL – End of Life).

Dynamic data can be either acquired at periodic time intervals or when an event happens. To generate knowledge from these data an intermediate processing stage is necessary. For example, data related to the product’s energy consumption can be calculated periodically based on the temporal acquisition. This information can be used to generate a monitoring service of energy consumption costs for a product or a system. An example of dynamic data which is event-based is the occurrence of a failure or error, where the user has to be warned about the need of maintenance or replacement.

Finally, for the services provision is necessary to understand and to visualize the different service types. According to [14] the PSS can be classified into five types:

- Support services – are services to help the product design. Here, the term Services can be a combination of product/service, only product or only service;
- Sale services – provide explanation about usage, training and maintenance and are offered during the product sale as an additional value through the product functions;
- Use-based services – are services such as leasing, renting, pooling or sharing;
- Maintenance services – are services based on lifecycle data of product to extend her life or to avoid breakdowns in the operation;
- EoL services – are services to support recycling, reuse and refurbishing services.

So, an IoT interface must be defined to support these service types. This interface must be far beyond the identification of objects on the network. Its use intends that the system becomes as smart as possible without user

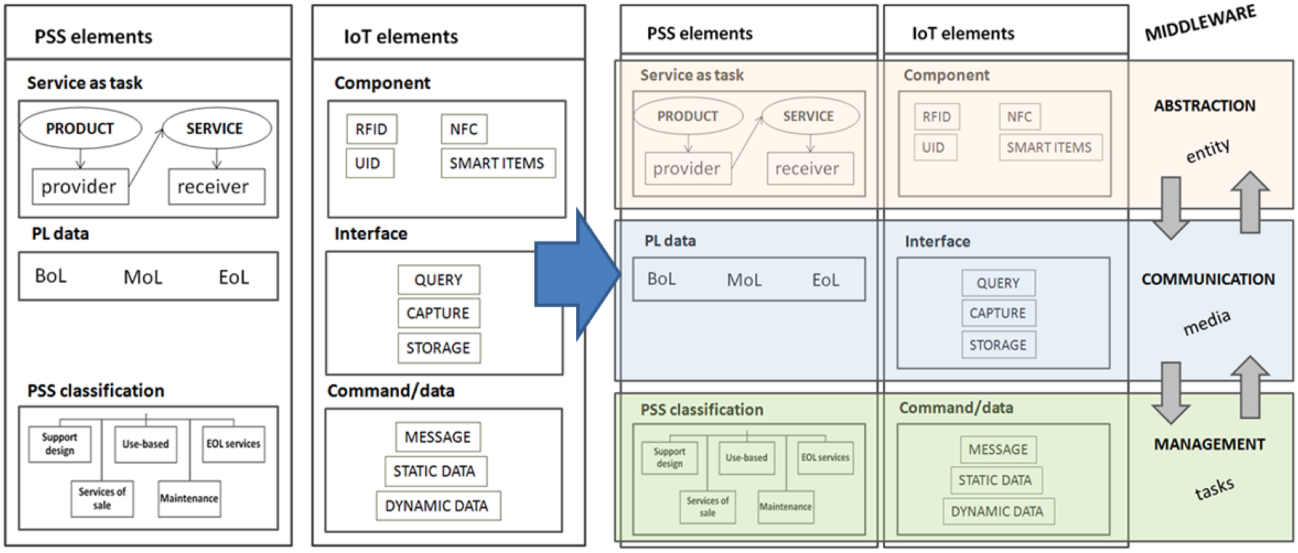


Figure 1. Elements of areas (left). The conceptual design (right).

intervention. That is, by having a pervasive computing system that makes users interact minimally with the devices.

It should allow devices to communicate with each other in order to generate knowledge for system operators and end users. The use of embedded components in products allows the automatic acquisition of information without any intervention from operators/customers. The idea is that devices like RFID, smart items and so on, do automatic data acquisition and store these data in repositories to support the creating of intelligent environments. The repositories should organize the informations in an intelligible way for that pervasive systems can supply services based on these data.

The basic elements for implementing the IoT paradigm are: the embedded components, the communication interface and the command/data module. The embedded components are devices such as RFID, UID, NFC or Smart items used for identification and data transmission to the processing module. Data transmission requires a communication channel. Issues about protocols, processing, energy, distance and storage should be considered when the embedded component is chosen. These components embedded in products constitute the pervasive computing that becomes the environment intelligent.

The other element of IoT paradigm is the information that is transmitted/exchanged by IoT components. This information can be data or commands. Aspects about information management in devices without processing capability should be considered. In addition, acquisition commands, such as those of command/data module of IoT must guarantee the processing of static and dynamic data from product lifecycle.

It is worth mentioning that the IoT paradigm is recent and the basic structure of representation of this technology still differs considerably [15] [16] [17] [18]. Researches that explore solutions for standardization are required in order to achieve a consensus. Although the paradigm is under construction and it is still new, the IoT is a result from multiple visions about the Internet of the future, what began to be discussed around 2005 with things like RFID and ubiquitous computing [19] [20] [21].

It is still important to consider the basics aspects about components like RFID, NFC, UID to be used for applying the IoT paradigm. For example, the spread of RFID usage was mainly due to their low cost and small size. However issues of energy (batteries and harvested), processing and communications (asymmetric and peer-to-peer) have motivated researches on networks like WSN (wireless sensor networks) and RSN (RFID sensor networks) and their applications on IoT. A detailed study comparing these technologies is presented by Atzori [16]. The main characteristics of RFIDs are: they usually do not have processing power, the communication is asymmetric, the communication range is around 10m, the energy is harvested, the useful life time is indefinite and the ISO 18000 standard is used.

III. A MIDDLEWARE FOR THE USE OF IOT IN PSS SYSTEMS

Based on the following premises: (i) re-use, re-customization and adding value are key elements for implementation of PSS, (ii) issues of social domain (environmental, social and economical aspects) hardly integrated PSS models; (i) and (ii) are associated with the acquisition, processing and use of large amounts of information distributed in time and space between different players through the lifecycle. A middleware to use the Internet of Things as a technological support to provide the effective service deployment is proposed. The IoT can provide a self-configuring network of wireless sensors, which can interconnect all space-time information. Re-use, customization, social information from players and things, such as location, status and preferences can be tracked and updated in real time.

So, an IoT-based middleware is presented to enable the implementation of the PSS paradigm throughout the product/service lifecycle. It connects services and objects/things tracked by RFIDs, sensor and actuator networks, mobile phones, etc, virtual elements and human-computer advanced interfaces. The middleware considers the following aspects in order to provide general solutions:

- Architecture – event-driven and bottom-up organization with elements of the semantic web, aiming to handle exceptions and issues non-deterministics associated with the design, manufacturing and information obtained from the customers;
- Intelligence - the network should be non-deterministic and open. The real or virtual entities should be self-organized or intelligent, capable of acting in context dependent manner;
- Time issues - the system includes several parallel and simultaneous real time events. Time will no longer be used like a common dimension and linear, but will depend on each product-service. The middleware should be based on parallel and real time processing;
- Complex system - the middleware supports the complexity associated with the large number of different connections and interactions between players and its ability to integrate new actors during all lifecycle;
- Space considerations - the middleware provides the location of acquired information. An new definition for space-time patterns is essential.

In summary, the proposed middleware provides conditions for acquisition, processing, visualization and usage of different information that comes from the various players involved in the lifecycle of each product-service and enables tracking, sensing, data acquisition and data query, even in real time of "things" (people and all kinds of devices). This way, the middleware will support the construction of applications targeted to PSSs, whose requirements of the social domain are not neglected during the technical modeling of product-service relationship.

The proposed middleware aims to provide solutions for three key aspects to deployment and operation of PSS: abstraction (acquisition and processing data), communication (among different devices) and management (providing service as task), each one represented by a module layer of the middleware.

The first step of an IoT implementation in PSS systems is to embed devices like RFID, NFC, etc, on the products and distribute them in the environment. The integrations between devices and product will represent entities. These abstractions are necessary to represent elements in an object-oriented model.

After having represented the entities, the next step is to define the communication between them. Each entity must be able to identify new entities entering the environment and be able to use a communication media to transfer data of product and environment for the management module (better explained bellow). For this, standardization of messages is necessary in order to provide communication between different entities. The communication module will allow the data acquisition and through the communication interface will transmit data about the product lifecycle. The communication interface implements functions like query, capture and storage of information that comes from BoL (Begin of Life), MoL (Middle of Life) and EoL (End of Life) cycles. Figure 2 represents the layers of middleware.

Another component of the middleware is the management layer. The management layer is responsible for activites like information treatment, service composition and application interface. The management issues must provide means to handle static and dynamic data and trigger commands to other layers. The data acquired about product lifecycle should be processed to allow the developing of different types of service and present them in application interfaces. The application interfaces will present the services in web pages.

The three layers of the proposed middleware encompass the five aspects presented above and ensure an scalable open solution. The architecture aspects are provided by the conceptual structure of the middleware. The intelligence and the complexity of system are considered by the management layer, where the intelligence is implemented by the information treatment module. The space considerations are addressed by the abstraction layer. i.e., by embedding tags in products and using IoT components one can identify the presence (location) of the entities in the environment. The time aspects are solved by the communication layer.

To implement the proposed model of the middleware an abstraction was adopted. In the abstraction every entity has a communication interface and when an entity communicates with other, in context of a service without user interaction, a smart environment emerges. The set of smart environments is called Hyper-environment (H-ENV). The next section presents the H-ENV architecture based on entities, tasks and media to describe the middleware implementation done.

IV. THE H-ENV ARCHITECTURE AND MIDDLEWARE IMPLEMENTATION

Based on the technological environment depicted in the last section, the following components co-exist:

- Elements, referenced in this work as real elements, such as animate beings or inanimate objects present throughout the lifecycle of the product/service (staff, customers, tools, equipment, consumers, etc.);
- Ubiquitous and pervasive technologies that allow even non-existent or inanimate elements in the environment, to become devices - things capable of presenting interactive behaviors, perceive, make decisions and act in pervasive environments (e.g., RFID technologies, smart devices, sensor networks, PDAs, among others associated with equipments and devices);
- Communication and interface technologies that enable the integration of real elements, virtual elements and

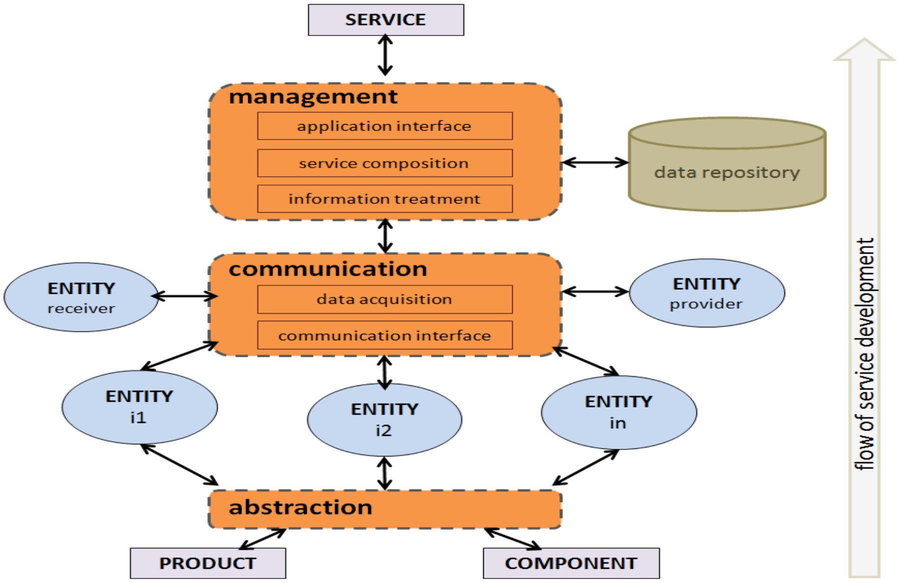


Figure 2. Middleware layers.

devices (such as internet protocols and technologies: wireless technologies - Bluetooth, etc.).

A possible conjecture about the architecture definition allows the construction of mixed environments, called hyper-environments, composed of real elements, avatars and things.

The hyper-environment must enable context-sensitiveness resulting from technological integration of acquisition, processing and use of large amounts of information, distributed in time and space between different players through the PSS lifecycle. Thus, in order to better characterize the different levels of organization that can arise in a hyper-environment, it is necessary to present some concepts that will be used throughout this section.

Hyper-environments elements. Real elements, avatars and things arise throughout the PSS lifecycle. These elements and their interrelationships compose the basic components of hyper-environments, as follows.

Media. Any medium, together with their properties, able to carry information, this way enabling communication.

Entity. Real elements, things or avatars associated with products/services that have technology able to assure their identifications to be perceived by others.

Tasks. Data processing activities performed by entities motivated by consumer's intention or for some specific purpose.

Organizing the Environment. Throughout the product lifecycle it can be established to exchange information from different sources and different spatial and temporal levels of abstraction.

Basically, the following levels of organization emerge:

Smart Environment. Is the basic unit. The entities aim to achieve common goals (intelligence). The tasks associated with the different entities are available on the hyper-environment. For example, one can cite a virtual-pervasive smart environment associated with the maintenance avatars, equipment repair - things, remote technicians, etc.)

Interface. The interface is responsible for the availability of the tasks inherent to a smart environment to extern communication. Smart peers also centralize the information of each smart environment.

Hyper-environment. Set of interfaces (or smart peers), merging different smart environments (real and/or pervasive and/or virtual) that intentionally and sporadically are related to execute specific goals.

An hyper-environment is responsible to combine informations in space-time throughout the PSS lifecycle. These information can be used to develop new services. Different actors involved with the services and products can exchange information during lifecycle.

To present the performance of the prototype of middleware that will be showed in the next section, some components that constitute the hyper-environment must be presented here. So, the main interfaces and classes implemented in Java are described below.

Entity class, with the following methods:

- **Start**: has the address parameter or a list of addresses of other entities in order to notify the entry of a new entity in the hyper-environment;
- **GetEntityList**: requests the list of all the entities present in the hyper-environment;
- **GetServiceList**: requests the list of tasks available in the hyper-environment;

- **CreateTask**: allows that the entity provides services to the hyper-environment, receiving as parameters: name, description and input signals.

Task class. The Task class has basically two methods, namely (i) StartTask which receives name, description and task list of input data, and (ii) TaskHandler that receives a function which transforms information received by task**.**

Smart Environment Class. This class allows the creation of an intelligent environment. When associated with a hyper-environment it is treated as a different entity represented by a smart peer.

In the adopted peer-to-peer network model the methods were implemented using the basic primitives of the JXTA [22] [23] [24] and the architecture and protocols of the PUCC [25] [26]. The entities and smart environments were implemented by Groups of JXTA peers-to-peers. Another special type called Rendezvous peer was responsible for communication of entities in a network with entities located in external networks. This type of peer is used to implement the definition of smart peer in the H-ENV architecture. The hyper environment is represented by a special group called the World Peer Group, which includes all JXTA peers. They are also used for the discovery mechanisms of JXTA and for the discovery treatment of new entities, tasks, and smart environments.

Due to the use of task concept (services) in the description of the H-ENV architecture, the prototype specification was performed based on Service-Oriented Architecture (SOA) [27]. However, an implementation that enables the use of Event-Driven Architecture (EDA) [28] was also build to provide means for an entity receive warnings through events.

V. Validation

Conceptual issues of a model and an architecture of a middleware to converge computational aspects of PSS and IoT were addressed. To validate the searched ideas there was necessary verify if the middleware is feasible and scalable.

A prototype of the middleware was then build and a simulated case study, in the marine industry was ran. This choice was made because in this kind of industry the use of information and communication technologies are still scarce, and authors have access to a large shipyard that is being installed near the University premises. The major constructs of the shipyard will be eight platforms hulls for use in exploring recent oil reserves discovered in Brazil, the so called pre-salt.

The first phase realised was the design of a hyper environment capable of modeling the production process of the industry in focus, as well as their byproducts. Different environments present on all lifecycle of a ship were then modeled. Some predominantly virtuals (associated with design and monitoring tools), other ubiquitous (installed in the workshops or aboard of ships already in operation), as shown in Table I. In each one of these environments, called smart environments, the data acquisition was done by RFID readers installed in tools, equipments, supplies and workers, by sensors embedded in measuring devices, by camcorders, by cell phones, by PDAs, and by tablets ported by workers, etc. Also, there was computers to access network servers that provide various kinds of services, called smart peers. The servers have databases containing design libraries, equipment descriptions, staff of construction, operation and maintenance of the ships, navigation maps, maintenance charts, among other.

The described elements to acquire, process and store data constitute the hardware of the focused IoT. The software that runs over this hardware was a prototype of the middleware described in the previous sections. So, the prototype was a distributed system.

In order to examine the adequacy of the H-ENV in a scenario like this, tests of computational performance were conduced. The boot-time of the computational processes that represent peers of the H-ENV and the RTT (Round Trip Time) of the messages exchanged by them were mesured. The system’s capacity to react on situations of joining and leaving elements in the environments were also evaluated. For example, the reaction time of the system to the inlet or outlet of a worker or a equipment in its premises was measured. The round trip time (RTT) was used to evaluate the capacity of the system to transport informations between their components.

To assess the boot-time, the runtimes of Start method were measured. To obtain the RTT, the runtimes of GetEntityList method were measured. Both methods belong to the Entity Class.

The simulations involved forty smart environments (twenty ubiquitous and twenty virtual). In ten of the ubiquitous environments, it was simulated the existence of a machine with a sensor and an actuator. In each one of ten smart ubiquitous environments a computer monitor was used to show all entities present in the hyper-environment. Finally, in each one of the twenty virtual smart environments it was installed a computer monitor which shows the avatars of ten employees chosen to be tracked. In one of these virtual smart environments a database capable of providing information about the entities present in hyper environment was simulated.

The simulation was performed in two laboratories, each one with 20 desktop computers connected through a 10Mbps wireless LAN (Local Area Network). The labs are connected by a university internet provider via a 1Gbps network. Each desktop computer runs Linux and have an Intel i3 processor, 4GB of main memory and a 500GB HD.

TABLE I. DESCRIPTION OF ENVIRONMENTS THROUGHOUT THE SHIP LIFECYCLE

Description	Predominant environment	Lifecycle stage	Peers
CAD/CAM design, supervision and scheduling project systems	Virtual	BoL	Tools and libraries for projects, equipment, personnel building
Yard plates, panels lines, block assembly lines, paint shops, edification area	Ubiquitous	Bol	RFID, video cameras, tablets and PDAs fitted sheets, trollers and workers
Comissioning, supervision and Control of the areas of deck, engine room, cargo holds	Ubiquitous	MoL	RFID readers, cameras, PDAs and tablet computers installed in equipment, devices, and workers
Dismiss and Conversions (oil tanker into an FPSO for example)	Ubiquitous	EoL	RFID readers, cameras, PDAs and tablet computers installed in equipment, devices, and workers

In each laboratory, there were three cell phones, a temperature sensor and an actuator connected through the serial port of a computer to control a air conditioner equipment.

The 20 ubiquitous environments were simulated in one laboratory and the 20 virtual environment in the other.

In order to evaluate the scalability of the middleware, it were simulated 1000 entities in the hyper-environment, twenty-five in each one of the smart environments. Table II shows the boot-time to some of these entities, according to their entering order in the system. The average of the boot-time was 52s with a standard deviation of 28s, performed through ten samples.

Regarding the ability to disseminate information, the middleware was evaluated by monitoring the method GetEntityList. The response times of a request sent to all entities present in the hyper-environment was measured. An average of 1050ms RTT in the implementation of the middleware using version 2.6 of the JXTA, running on the JVM version 1.6, using safe pipes of HTML was obtained.

TABLE II. BOOT-TIME OF THE PEERS.

Entity	Startup (seconds)
1	39
2	45
3	41
4	61
...	...
1000	75

Considering that the simulations were consistent with the operational conditions under which the possibility of use of the middleware was glimpsed and considering that the average startup time and RTT of 1,000 peers consumed solely by the JXTA protocols are around 10,000ms and 200ms [29] respectively, the results of about 60,000ms and 1,050ms obtained by the H-ENV implementation were adequate. It must be noted that they encompass the processing time consumed by other operations beyond communication.

With these results it can be concluded that the adopted ideas are feasible and a implementation of the system has viability under performance point of view in the application depicted.

By the way, at this point, we emphasized that the application envisioned in this work was thought to be used in a restricted environment, to be eventually implemented in the premises of the Brazilian petroleum company - PETROBRAS and under its full control. In short, the system prototyped was addressed to be used strictly in the offshore platforms that will be build to the PETROBRAS in the shipyard cited above. In this conditions, the apparent big problem domain with a relatively simple middleware solution is mitigated.

VI. CONCLUSION AND FUTURE WORK

The paper presented the use of IoT to provide a scalable architecture to PSS. The IoT use enables to obtain information throughout the lifecycle of the system, including data about the developers and users of the PSS, through networks of pervasive objects. The use of IoT as an ICT solution can assist the construction of PSS by using information feedback loops in any lifecycle stage. It means that aspects of any phases: Begin of Life, Middle of Life and End of Life can be automatically considered to influence aspects of any stage, in cycles of any size. These possibilities are mediated by applications that use databases and models to describe the system dynamic and to offer the necessary subsidies for manual or automatic decision-making.

In this way, the contributions of this work intends to start discussions about a conceptual design for the IoT use in PSS systems. So, an approach to identify the main problems to be addressed in the context of IoT-PSS, in order to design a midlleware to be used in various areas of PSS applications was proposed.

Finally, stands out the emerging need for new methodologies, models and tools to improve PSS building for general use. However, the use of IoT is a possibility to obtain informations about all stakeholders of the system in real time, during all its lifecycle, through RFIDs, mobile phones, sensors networks and so on, thus representing a differential to building a PSS.

REFERENCES

[1] C. Van Halen, C. Vezzoli, and R. Wimmer (2005). Methodology for Product Service System Innovation. ISBN 90-232-4143-6.

[2] B. Cope and D. Kalantzis (2001). Print and ElectronicText Convergence. Common Ground. ISBN:1-86335-071-3.

[3] O. Mont (2004). Product-service systems: Panacea or Myth? Doctoral dissertation, The International Institute for Industrial Environmental Economics, University of Lund, Sweden.

[4] E. Heiskanen, M. Jalas, and A. Kärnä (2000), The Dematerialisation Potential of Services and IT: Futures Studies Methods Perspectives, Workshop on Futures Studies in Environmental Management, Turku, Finland.

[5] H. Chen, R. Kazman, and O. Perry, (2010), From Software Architecture Analysis to Service Engineering: An Empirical Study of Methodology Development for Enterprise SOA Implementation, In IEEE Transactions on Services Computing, vol. 3, no. 2, pp. 145-160

[6] C. Rolland, M. Pinheiro, and C. Souveyet, (2010), An Intentional Approach to Service Engineering, In IEEE Transactions on Services Computing, vol. 3, no. 4, pp.292-305.

[7] G. Pezzotta, S. Cavalieri, and P. Gaiardelli (2009), Product-Service Engineering: State of the Art and Future Directions, In Proceedings of the 13th IFAC Symposium on Information Control Problems in Manufacturing Moscow, Russia, June 3-5, 2009, pp. 1329-1334.

[8] O. Isaksson, T. Larsson and A. Rönnbäck (2009), Development of Product-Service Systems: Challenges and Opportunities for the Manufacturing Firm, In Journal of Engineering Design Special Issue on Product-Service Systems, 2009, pp.329-348.

[9] I. Barakonyi and D. Schmalstieg (2006), Ubiquitous Animated Agents for Augmented Reality, In IEEE/ACM International Symposium on Mixed and Augmented Reality, 2006. ISMAR 2006.

[10] J. York and P. Pendharkar (2004), Human–computer interaction issues for mobile computing in a variable work context, Int. J. Human-Computer Studies 60, pp. 771-797.

[11] U. Hansmann (2003), Pervasive Computing: The Mobile World. Springer. ISBN 3540002189.

[12] L. Atzori, A. Iera, and G. Morabito, (2009), The Internet of Things: A survey, Int. J. Computers Networks, pp.2787-2805.

[13] T. Tomiyama (2001), Service Engineering to Intensify Service Contents in Product Life Cycles, pp. 613, 2nd International Symposium on Environmentally Conscious Design and Inverse Manufacturing (EcoDesign'01), 2001.

[14] O. Mont (2002), Clarifying the concept of product–service system, Journal of Cleaner Production, 2002, pp. 237-245.

[15] D. Uckelmann, H. Mark, and F. Michahelles (2011), an Architectural Approach Towards the Future Internet of Things, D. Uckelmann (Eds.), 2011.

[16] L. Atzori, A. Iera, and G. Morabito (2010), The Internet of Things: A survey, Computer Networks, v. 54, n. 15, 28 Oct. 2010, pp. 2787-2805.

[17] M. Presser and A. Gluhak (2009), The Internet of Things: Connecting the Real World with the Digital World, EURESCOM mess@ge – The Magazine for Telecom Insiders, vol. 2, 2009, http://www.eurescom.eu/message, retrieved: aug, 2012.

[18] A. Katasonov, O. Kaykova, O. Khriyenko, S. Nikitin, and V. Terziyan (2008), Smart semantic middleware for the internet of things, in: Proceedings of the Fifth International Conference on Informatics in Control, Automation and Robotics, Funchal, Madeira, Portugal, May 2008, pp. 169-178.

[19] E. Fleisch and F. Mattern (2005), Das Internet der Dinge: Ubiquitous Computing und RFID in der Praxis: Visionen, Technologien, Anwendungen, Handlungsanleitungen. Springer, Berlin, 2005.

[20] N. Gershenfeld, R. Krikorian, and D. Cohen (2004), The internet of things, Scientific American 291 (4), 2004, pp. 76–81.

[21] L. Srivastava (2006), Pervasive, ambient, ubiquitous: the magic of radio, in: European Commission Conference ‘‘From RFID to the Internet of Things”, Bruxelles, Belgium, March 2006.

[22] L. Gong (2001), Jxta: a network programming environment. Internet Computing, IEEE, 5(3), pp. 88-95.

[23] H. Park, E. Paik, and N. Kim (2009). Architecture of collaboration platform for ubiquitous home devices. Mobile Ubiquitous Computing, Systems, Services and Technologies, International Conference on, pp. 301-304.

[24] L. Barolli and F. Xhafa (2010). Jxta-overlay: A p2p platform for distributed, collaborative and ubiquitous computing. Industrial Electronics, IEEE Transactions on, pp. 2163-2172.

[25] H. Sumino, N. Ishikawa, S. Murakami, T. Kato, and J. Hjelm (2007). Pucc architecture, protocols and applications. In Consumer Communications and Networking Conference, 2007. CCNC 2007. 4th IEEE, pp. 788-792.

[26] K. Fukushima, Y. Tanaka, H. Kato, and N. Ishikawa (2010). Home network system for gas appliances using pucc technologies. In Consumer Communications and Networking Conference (CCNC), 2010 7th IEEE, pp. 1-5.

[27] D. Krafzig, K. Banke, and D. Slama (2004). Enterprise SOA: Service-Oriented Architecture Best Practices (The Coad Series). Prentice Hall PTR, Upper Saddle River, NJ, USA.

[28] H. Taylor, A. Yochem, L. Phillips, and F. Martinez (2009). Event-Driven Architecture: How SOA Enables the Real-Time Enterprise. Addison-Wesley Professional, 1st edition.

[29] E. Halepovic and R. Deters, JXTA Performance Study, in Proc. PACRIM 2003, pp. 149-154.

Improving TelEduc Environment with GPL Software

The visualization content case for computers and mobile devices

André Constantino da Silva
Institute of Computing (PG)
UNICAMP IFSP
Campinas, Brazil Hortolândia, Brazil
acsilva@ic.unicamp.br

Heloísa Vieira da Rocha
Institute of Computing, NIED
UNICAMP
Campinas, Brazil
heloisa@ic.unicamp.br

***Abstract*—e-Learning environments are proposed to support teaching and learning activities using the Web infrastructure. One example is TelEduc, a GPL e-Learning environment which its developers are mostly students involved in academic projects, making it difficult to have a long-period team with expertise to maintain the code and evolve the software with new requirements and to new and desirable contexts, like mobile access. One solution is use third-party software to improve the software and minimize the code maintain efforts, solution applied in the TelEduc environment to visualize attached documents on computers and smart phones without installed appropriated software, an open problem in the environment before this work. The main result is a TelEduc version with visualization file for desktops and smart phones. Another result is identification of issues dealt when integrates with third party software like impact on usability, accessibility and performance, license compatibility and easily to integrate, update and change.**

Keywords-e-Learning environment; Mobile user interfaces; Mobile learning; software integration.

I. INTRODUCTION

Devices like smart phones are been increasingly popular; most of them have touch screen displays and access to the Internet. Therefore, Web sites and applications, initially developed to be used with keyboard, mouse and a medium size display, are been accessed by small touch screen devices. Tablets are been popularized too, and some models allow interaction with the Web applications using fingers and others models using special pens.

One kind of Web application is e-Learning environments, as Moodle [1], SAKAI [2] and TelEduc [3]. They are computational applications with tools to support teaching and learning activities though the Web. These tools allow users to create content, communicate with other users and manage the virtual space.

The actual version of these environments take advantages of the available Web resources and offers textual content with images, audios and videos in a hypertext document to transmit the content. Tools like chat, forums, portfolios, repositories are widely used. And tools that explore the audio and video resource to user communication, like instant messenger and video-conferences, are become common too.

Due to the quickly Internet growth (number and types of devices accessing it) and evolution (number of technologies that can be used to develop applications to use it), it is hard to a small development team updates and evolves complex systems with their own solutions. A lot of new requirements are requested by users, bugs and errors are found every day, and a lot of technologies are available, requiring expert developers.

One strategy is found a development community with different knowledgeable and skill members that together maintain and evolve the system, like communities organized to evolve Moodle and SAKAI environments. TelEduc project is organizing its community [4], but some solutions need to be used to evolve the code until the community growth enough. Another strategy is get third-party software available that meets the requirements and integrate them.

We applied this strategy to attend a user-reported problem: visualization of attached items in the environment. This is an old problem in the TelEduc and other e-Learning environments (e.g., Moodle and SAKAI): the user needs to download the file, write permission to save the file and have the appropriated software to view it. To solve this, we integrated TelEduc with FlexPaper to easily the document view.

This integration generated some questions about the use of third-party software in the environment and a list of issues to be analyzed, like the distributed license. Other relevant issue is about data responsibility, like where the data will be persist, who will ensure the data availability, backup and restore mechanisms.

Section II presents the TelEduc Project with a brief historical view and the needs for use third-party software. Section III describes the TelEduc e-Learning environment its tools, features and problems. Section IV describes how to avoid the elicited problems in Section II and III through integration with existing free software, and Section V summarizes the dealt problems in this research. Section VI presents a conclusion and future works.

II. TELEDUC PROJECT

The TelEduc environment was conceited in the end of 90, born with the Cerceau´s Master dissertation (1998), with professor Heloísa Vieira da Rocha as advisor, using constructivism as base [5,6] for teacher´s continuance formation with situated learning bases [7] or contextualized learning [8] bases. In 2001 February, the first free version was released over GNU General Public License (GPL), an unprecedented fact in the Brazilian Educational Software scenario. Many public and private institutions adopted the TelEduc as platform, maximizing the TelEduc user's community, and consequently, the development demand. This fact culminated in the release of TelEduc version 3.0 in March 2002. The version 3.0 was completely redesigned and optimized, reason for TelEduc project was awarded by ABED (Brazilian Association for Distance Education) in the "Research and Development about Distance Learning" category. In August 2011, TelEduc version 4.3 was released, with its user interface redesigned to improve user tasks and be more similar than the most visited Web sites.

The TelEduc development is mainly performed by a team situated at NIED/UNICAMP, composed of undergraduates, plus research graduate students of IC/UNICAMP. This is a small team, considering the complexity of the environment, responsible for developing new features, code maintenance and management of existing servers. The next Section shows some TelEduc features and problems.

III. TELEDUC E-LEARNING ENVIRONMENT

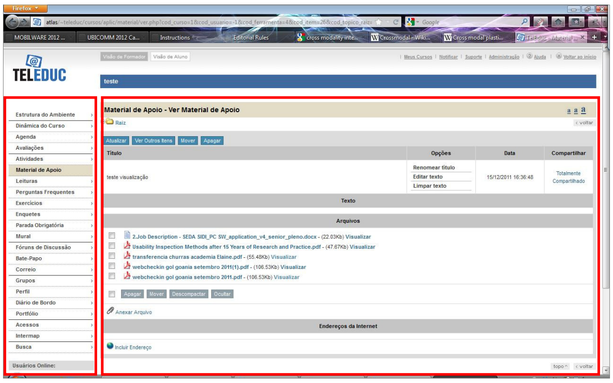
The TelEduc environment [3] is a teaching and learning environment developed by Nucleus of Applied Informatics in Education (NIED) of State University of Campinas (UNICAMP), adopted by many public and private institutions, like UNICAMP through Ensino Aberto project.

TelEduc is a system that aggregate administration, management and communication tools designed to support teaching and learning activities. Some tools allow content creation, other ones allow synchronous or asynchronous communication among users, participant and course management and administrative features. The TelEduc course page is structured in two parts: the left one (Figure 1a) has a list of all tools available and in the right one (Figure 1b) the content of the selected tool.

Figure 1a shows TelEduc instance for a course where the teacher dispose the Couse Dynamic, Agenda, Readings, Support Material, Activities, Chat, Mail, Discussion Forum, Frequently Asked Questions, Portfolio, Groups and other tools available. Figure 1b shows the user interface to visualize a Support Material item, where the teacher can change the title, content, attach or remove files or links and see and write comments, and the student can see the item, download the attached files and visit the posted links.

To provide content, TelEduc uses the Web infrastructure, more specifically, hypertext with images, links, audios and videos. All these media can be published as content in tools like Agenda, Support Material and Readings.

Agenda is the ongoing course home page and shows the course´s program for a given period (daily, weekly, etc.). Agenda is an important tool because organize the activities that must be done in a specific period, similar teachers do in the beginning of a presence class. The Support Material is a tool that provides an area for the file storage and sharing among course participants, named Support Material Area. To store an item in this area, the user needs to be a coordinator or an instructor. When the user stores an item in support material area, she can specify the type of access (for example, i) not shared, ii) shared only with users who have instructor role or iii) shared with all participants). Users with student role can access the stored items published with all participant sharing type, read their content, visit their links and download their attached files.



Material de Apoio - Ver Material de Apoio

Título	Opções	Data	Compartilhar
teste visualização	Renomear título Editar texto Limpar texto	15/12/2011 16:36:48	Totalmente Compartilhado

Texto

Arquivos

□ 2.Job Description - SEDA SIDI_PC SW_application_v4_senior_pleno.docx - (22.03Kb) Visualizar
□ Usability Inspection Methods after 15 Years of Research and Practice.pdf - (47.67Kb) Visualizar
□ transferencia churras academia Elaine.pdf - (55.48Kb) Visualizar
□ webcheckin gol goania setembro 2011(1).pdf - (106.53Kb) Visualizar
□ webcheckin gol goania setembro 2011.pdf - (106.53Kb) Visualizar

Endereços da Internet

(a) (b)

Figure 1 – Course page at TelEduc environment (a) tools available; (b) content of selected tool.

The Readings, Activities and Frequently Asked Questions tools have features and user interface similar with Support Material tool, but different purposes. Readings tool is used to publish relevant documents, like books, magazines, news and articles. The Activities tool is an area to publish activities to the accomplished during the course, like home work descriptions. And the Frequently Asked Questions tool contains a list of the most frequently questions done by the participants during the course and their respective answers.

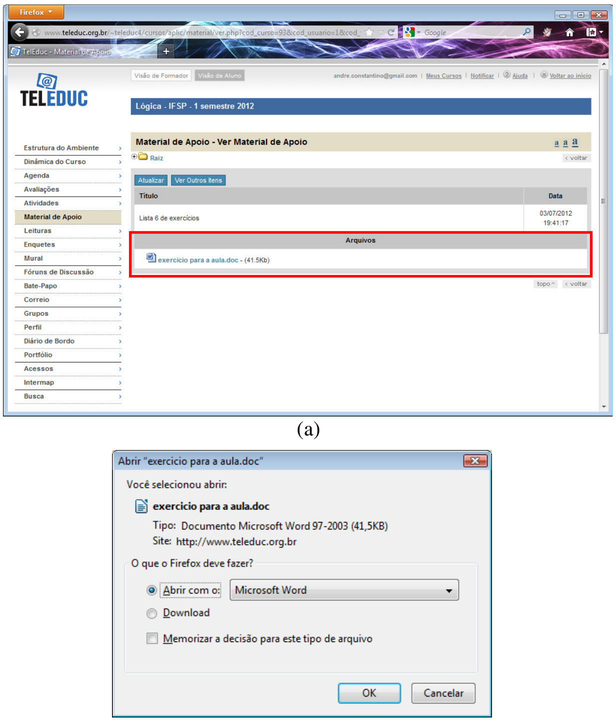
Tools like Discussion Forums and Mail are used to participant communication, supporting text message change in asynchronous mode. To synchronous communication, there is Chat tool, with some features similar with Web chat sites.

The Portfolio is a communication tool that aims to promote the collaboration among participants through the sharing of "items" (documents, presentations, programs, links, etc.). So the Portfolio tool provides an area to item storage and sharing for each participant (user or group of users) within a course.

The Bulletin Board tool is a dedicated space where all the participants can post information considered relevant to the course content.

The Agenda, Activities, Support Material, Readings, Bulletin Board, Discussion Forums, Mail and Portfolio allow users to create text content using a text editor. Details about it are in Section IV.

The Agenda, Activities, Support Materials, Readings, Mail and Portfolio tools allow attaching files into them items in similar way, following these steps: the user goes to Files area, click in the link "Attach File", click in the "Select file..." button, browse into the computer directories, choose one file, click on "Open" button and waiting for the uploading process. The environment response for the correct uploading process is writing the feedback message "File attached successfully".



Lógica - IFSP - 1 semestre 2012

Material de Apoio - Ver Material de Apoio

Título	Data
Lista 6 de exercícios	03/07/2012 19:41:17

Arquivos

exercicio para a aula.doc - (41.5Kb)

(a)

Abrir "exercicio para a aula.doc"

Você selecionou abrir:

exercicio para a aula.doc

Tipo: Documento Microsoft Word 97-2003 (41,5KB)

Site: http://www.teleduc.org.br

O que o Firefox deve fazer?

● Abrir com o: Microsoft Word

○ Download

□ Memorizar a decisão para este tipo de arquivo

(b)

Figure 2. Step to view an attached file in TelEduc: (a) item visualization; (b) open or save dialog..

To download the file, the users need to go to the Files area, click in the file name (Figure 2a), wait for the browser´s “Open” dialog, choose “Open” or “Save” (Figure 2b). If the user chooses “Open”, the used computer needs to have the correct software to visualize and edit the file. If the user chooses “Save” option, the browser will download the file and save it into the download folder. The user will need open the download folder to open the file or change the file to another folder. In TelEduc version 4.3, there is no feature to preview the content file, as Moodle version 2.3 and Sakai 2.8. Moodle allows the instructor creates some spaces to visualize a content (e.g., a .PDF file), but it is necessary have the appropriate program installed to view. For attached files there is no preview feature.

In mobile devices, to view attached file it is necessary to have installed the appropriate software, an application to visualize the file. But, for some files formats there are no free applications or the device have a limited number of installed application.

Motivated by the lack of a solution for the described problem and focused to minimize the user efforts to view the content, we developed this work. Maybe the steps to visualize and edit a downloaded file are easy for Web experts, but it is not so simple to new users.

Since the e-Learning environments need to be easily to use for users with different levels of Web experience, the usability is an important nonfunctional requirement. TelEduc was designed to have good usability in an iterative design-evaluation process, and the main focus is the user interface does not impair teaching and learning activities. The accessibility is another nonfunctional requirement desired for TelEduc, to allow impaired people to use the environment without meet barriers or obstacles.

Since the e-Learning environments are available on the Internet, this software can be accessed by mobile devices nowadays and the developers need to study how to allow all features into these devices. Access the environment in anywhere and anytime is one of the biggest attractions, but research is necessary to design good user interfaces that take into account the restrictions imposed by the devices.

The e-Learning environments are proposed to be generic so that they can be support any kind of educational content in any learning context. The solution adopted by the developers to reach out the generic state is dispose features that allow the users upload any kind of file; so, sometimes, the environment is used like a repository server and a lot of interactions problems appear. Users spend your time uploading and downloading files and need specific software installed into their machines to visualize the file. Sometimes, it is not possible to have this software installed, or the users just want to visualize quickly the content without save the file. So a solution is allow users preview the file in the browser, but it is not trivial to implement.

To guide what file extension to focus, we analyzed the files attached in the TelEduc at NIED servers. At NIED servers there are a total of 50 finished courses and 50 incoming courses running over TelEduc 4.x, and a total of 443 finished courses and 77 ingoing courses running over the TelEduc version 3.x. The users usually attach files with DOC or PDF extension for text content, PDF or PPT extension for presentations, MP3 or WAV for audio, and AVI or WMV for video. Figure 3 shows the quantity of each type file in the courses at NIED/UNICAMP servers since 2009 until 2012. Analyzing the extension of 9,019 attached files, 39.7% of all files are text documents at DOC, DOCX, ODT, RTF or TXT format. PDF documents are 23.5%, and

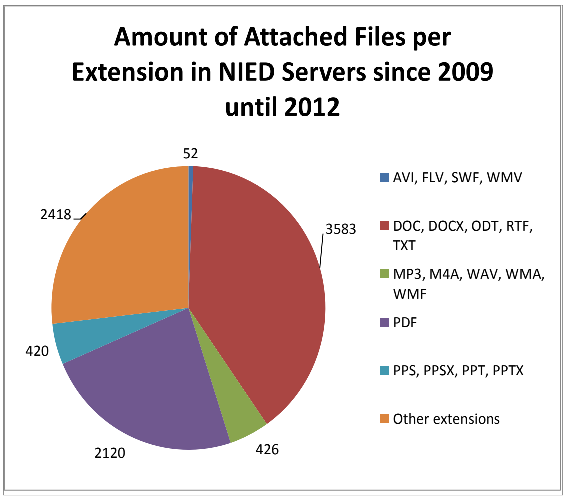


Figure 3. Amount of attached file per extension at NIED/UNICAMP TelEduc 4.x server since 2009 until 2012.

4.6% are presentations archive in PPT, PPTX, PPS and PPSX format. Audio files are 4.7% and the formats are MP3, M4A, WAV, WMA and WMF. Video files are 0.5% in AVI, FLV, SWF and WMV formats. It is important to consider that in this analysis, we did not consider audio and videos of other sites, like YouTube, that usually are posted like links instead of attached files.

Considering the small development team and the need to allow visualization of attached documents, as well as allow access from mobile devices, we used third-party platforms to integrate with TelEduc environment. These platforms are described in the next section.

IV. WEB PLATFORMS FOR SHARING CONTENT

Using third-party software integrated to TelEduc is not a new task. To allow users write rich text instead of simple plain text, the CKEditor was integrated with TelEduc (Figure 4) since version 3.0. So, the TelEduc team did not develop the text editor, minimizing coding efforts.

CKEditor [9] is a WYSIWYG text editor to be used inside web pages, bringing to the web application common editing features found on desktop editing text applications. Some of them are basic and advanced styling, block-quoting, colors, copy and paste features, build advanced links, insert images and Flash content. Its main purpose is, instead of having plain text fields to write content, having a richer content and a richer user experience. In the version 3.0 of CKEditor are possible insert videos from YouTube, so the user only needs get the video´s URL. In this case, there are some problems like data responsibility that will be discussed on Session V.

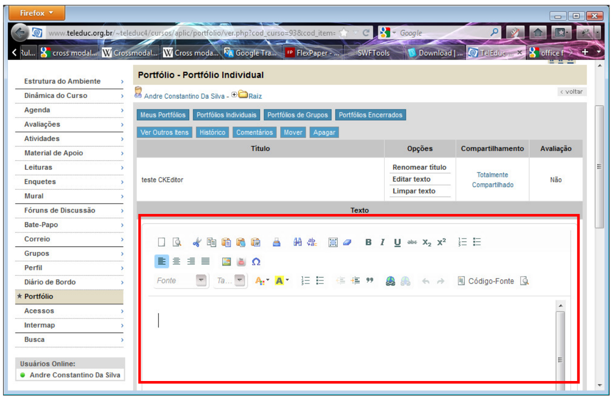
In version 3.0, the developers improved the accessibility, turned it compatible with screen readers, a better navigation with the keyboard and high contrast support. It is also compatible with Internet Explorer 6+, Firefox, Safari, Google Chrome and Opera.

The flexibility and performance are two important features of CKEditor. It is possible to customize the editor interface and behavior, choosing tool bar elements, interface colors or adding features. The editor loads fast and its features run.

CKEditor is distributed under the GPL, LGPL and MPL Open Source licenses, so the use of these tree licenses turn this software more flexible to be adopt in another Open Source software, as the developers intend:

> *"Because CKEditor is licensed under flexible Open Source and commercial licenses, you'll be able to integrate and use it inside any kind of application. This is the ideal editor for developers, created to provide easy and powerful solutions to their users."*

There are some solutions available into Web with different licenses and different mode of use, too. We analyzed free solutions, like Flexpaper [10]. Flexpaper is an open source web based document viewer allowing publishing documents to multiple platforms, making it possible to display and interact with PDF files in web pages.



Título	Opções	Compartilhamento	Avaliação
teste CKEditor	Renomear título Editar texto Limpar texto	Totalmente Compartilhado	Não

Figure 4. CKEditor in TelEduc environment.

So, it is possible to distribute the CKEditor with the application code in one package. These easier the software download and configuration.

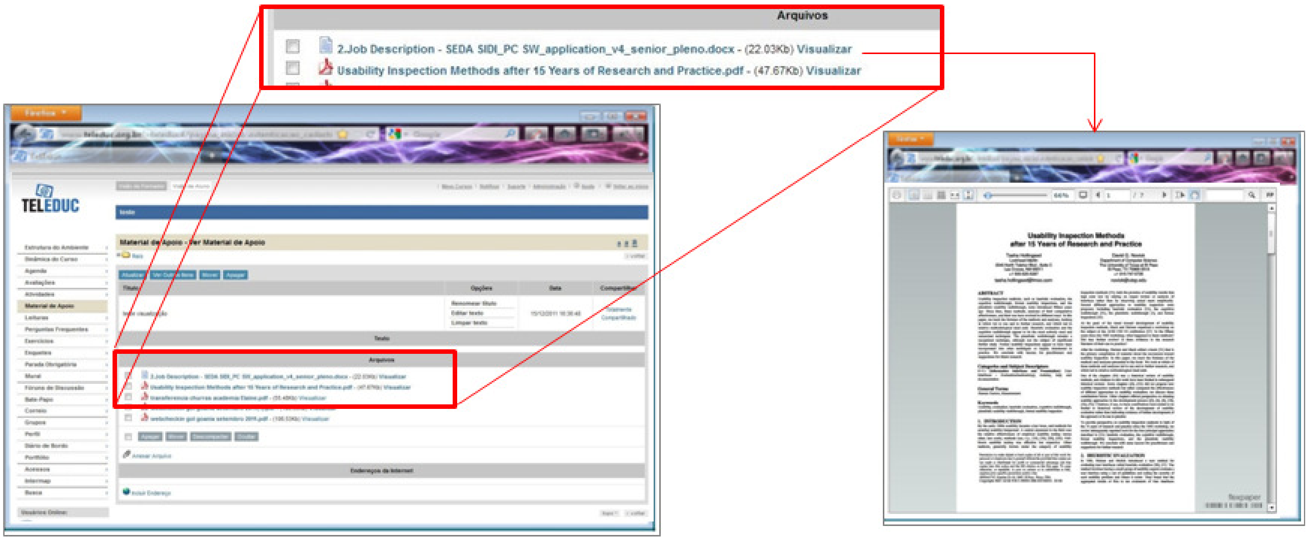
To show visualize documents into browsers without the need of downloading. Some of these solutions is proprietary, but are solutions developed with free of use license and open

PDF files into Web browsers and without Adobe Reader installed on the device, Flexpaper needs to convert the PDF file to a SWF file. So, to see the document into Web browser is necessary to have Flash Player installed. To covert PDF to SWF, the Flexpaper uses SwfTools [11], an open source tool that can be automated if needed, and it is built in PHP language. The SWFTools have features to convert JPEG, PDF, PNG, GIF and WAV files to SWF files.

But some files upload is in another format, like DOC file format, as Figure 3 shows. In this case, it is necessary to use another software to convert DOC files to PDF files, like the CUPS-PDF [12] or WVPDF [13]. So, integrating CUPS-PDF and PDF2SWF it is possible to have a more powerful document viewer.

Flexpaper allows distributing the code with the Web application, but some sites expose the HTML code of a given resource to be copied and pasted in the Web application. For example, YouTube, a video sharing platform, the Issuu [14], a publishing platform for reading documents, the Soundcloud [15], a sharing sound platform, and the Slideshare [16], a platform to sharing presentations, videos, documents and webinars. All these platforms are Flash or HTML5 based technologies, so mobile phones with support for these technologies can access the content provided by these platforms.

The Issuu and Soundcloud platform have applications for mobile phone available at iPhone´s and Android´s market, allowing the user access the content in the device. Flexpaper does not provide an app for devices, instead of it uses Adobe Flash and needs to be installed on the server. A php script was created to, when user request file visualization (Figure 5a), the respective file is copied to a temporary folder and a convert command is executed to create the respective SWF file. After the conversion, the server response the user request sending a HTML code to visualize the created SWF file (Figure 5b). So the user can see the document in the Web


Arquivos

2.Job Description - SEDA SIDI_PC SW_application_v4_senior_pleno.docx - (22.03Kb) Visualizar

Usability Inspection Methods after 15 Years of Research and Practice.pdf - (47.67Kb) Visualizar

TELEDUC
Usability Inspection Methods after 15 Years of Research and Practice

(a) (b)

Figure 5. Integration between Flexpaper and TelEduc accessed on desktop: (a) an item with attached files; (b) file preview using Flexpaper.

browser. The access can be done into an Android-based phone (Figure 6a and 6b) with similar steps.

Flexpaper integration is simpler, since the Flexpaper code can be distributed with Web application code, so both can run in the same server. Issuu, Youtube, Soundcloud and Slideshare have different approaches to integrate, but similar steps for the user, turn more easier to see the document content.

V. Issues About Web Software Integration

The first one issue analyzed was meeting requirements. So, we selected some Web platforms focused on content visualization and studied them, how to integrate them and the licenses compatibility.

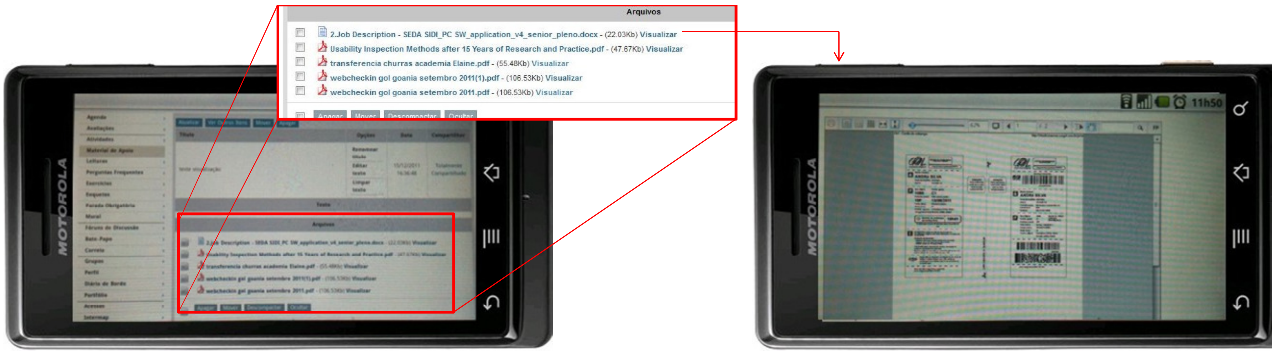
The TelEduc environment was designed to have good usability, performance, and have accessibility studies to improve its access by impaired people. So, when the code is modified, e.g., during the integration between TelEduc and Flexpaper, it is necessary to verify the impact on usability, accessibility and performance from TelEduc. Is not desirable increase the usability, accessibility or performance problems, so it is necessary investigate these nonfunctional requirements on platform and after integration.

The TelEduc was implemented to be compatible with the most used browsers, so it is desired to, after the integration, does not minimize the compatible browsers number.

The flexibility to adapt the platform was considered another important issue, since this impact on usability and gives for users a unity sense.

More than be easily to integrate, it needs to analyze the platform updating and changing. Since Web applications are evolving quickly, it is important to update the version used in the integration. Since, sometimes, the platform can be discontinued, so it is necessary replace with another one.

TelEduc has three languages, Portuguese (Brazilian and Portugal), Spanish and English. So, one issue is considering if the platform offers user interfaces on these languages.


Arquivos

2.Job Description - SEDA SIDI_PC SW_application_v4_senior_pleno.docx - (22.03Kb) Visualizar

Usability Inspection Methods after 15 Years of Research and Practice.pdf - (47.67Kb) Visualizar

transferencia churras academia Elaine.pdf - (55.48Kb) Visualizar

webcheckin gol goania setembro 2011(1).pdf - (106.53Kb) Visualizar

webcheckin gol goania setembro 2011.pdf - (106.53Kb) Visualizar

(a) (b)

Figure 6. Integration between Flexpaper and TelEduc accessed on Android-based phone: (a) an item with attached files; (b) file preview using Flexpaper.

Changing the language it is important too, e.g., if the TelEduc interface is in English, the platform must be in English too.

Before integration, it is necessary to analyze the platform license to verify law violation and to choose the better integration form. If the platform cannot be delivery with the environment code, it is possible create a plugin and give instructions about how to configure the integration.

Platforms like YouTube, Issuu and Slideshare, need to have the file into their servers. So, these platforms are data servers, too. It is important to analyze data ownership issues and if this data violate some law. It is not a trivial task, since the content generated sometimes cannot be available to all Internet users due to intellectual property or special owner restrictions.

VI. Conclusion And Future Work

Developing e-Learning environment is not a trivial task: the software has a variety of users and roles, many features and must be technologically updated and get advantages from Web to facilitate the teaching and learning activities. The Web evolves quickly, a lot of technologies are proposed, requiring for projects have high expertise members. One example is the Web applications, either e-Learning environments, are been accessed by mobile phones. But how teams without expertise in these technologies can improve them software? One approach is integrating the application with third-party platforms available in Web.

To explore this solution, we chose the content visualization problem in the e-Learning environment for desktop and mobile devices. Files in the environment need to be downloaded to be viewed, and the device needs to have installed appropriated software to show the content. TelEduc version 4.3, Moodle version 2.3 and Sakai version 2.8 have this problem.

There is a lot of third-party software to content visualization, like Flexpaper, YouTube, Issuu, Soundcloud and Slideshare. An environment integrated with these platforms can allow documents, audio, video and presentations visualization more easily to the users and better usability. This motivated this work.

Analyzing the number of documents uploaded in TelEduc at NIED servers from 2009 until 2012, the DOC, DOCX, ODT, RTF, TXT and PDF files correspond to 63.24% of all uploaded documents. So in this work we decided to integrate Flexpaper to allow visualizing these type of files. Integration with YouTube, Issuu, Soundcloud and Slideshare will allow users visualize almost all uploaded files using desktop and mobile phones.

In this work, we challenged some problems and questions, and we generated a list of issues to think about when integrate two Web applications: features; usability; accessibility; performance; license compatibility; browsers compatibilities; flexibility to adapt; easily to integrate, update and change; user interface languages; data location and data ownership.

Using platforms as Flexpaper, Issuu, Soundcloud and Slideshare is possible to increase the e-Learning environment with features hard to implement and maintain. Doing an appropriate relationship, it can be beneficial for both software, one use the platform and increase the number of users, turning the platform more popular and receive code contribution, and the another, who increase its features and turn easier the users tasks.

The integration among TelEduc and content platforms is not the only one necessity to turn this environment accessible on mobile phones. For example, it is necessary study the influence of changing the mouse by touch screens (and some coding solutions like use of JavaScript's onMouseOver events) and medium size display to a small one.

Acknowledgment

The authors thank the CAPES and CNPq for financial support and for FAPESP through TIDIA-Ae Project (n. 05/60572-1), which provided the equipment.

References

[1] Moodle Trust. Moodle.org: open-source community-based tools for learning. Available at <http://moodle.org>. Accessed in July 2012.

[2] SAKAI Environment. Sakai Project | collaboration and learning - for educators by educators. Available at <http://sakaiproject.org>. Accessed in July 2012.

[3] TelEduc Environment. TelEduc Ensino à Distância. Available at <http://www.teleduc.org.br>. Accessed in July 2012.

[4] F. L. Arantes, A. C. da Silva, L. L. de Oliveira, and F. M. P. Freire, "(Re)Projetar para Crescer – Reestruturação do Site do TelEduc Centrado na Comunidade", Proceeding of International Free Software Workshop, July 2012, in press.

[5] S. Papert, "Constructionism: a new opportunity for elementary science education". Proposal to The National Science Foundation. Massachusetts: Cambridge, 1986.

[6] J. A. Valente, "Computadores e conhecimento: repensando a educação". Campinas, Brazil: Gráfica da UNICAMP., 1993.

[7] J. Lave and E. Wenger, "Situated Learning – Legitimate peripheral participation". Cambrigde: Cambrigde University Press, 1991.

[8] J. A. Valente, "Formação de Professores: Diferentes Abordagens Pedagógicas". In O Computador na Sociedade do Conhecimento, J. A. Valente, Ed. Campinas, Brazil: UNICAMP/NIED, 1999, pp. 131-156.

[9] CKSource - Frederico Knabben. CKEditor - WYSIWYG Text and HTML Editor for the Web. Available at <http://www.ckeditor.com>. Accessed in July 2012.

[10] Devaldi Ltd. FlexPaper - The web based pdf viewer solution. Available at <http://flexpaper.devaldi.com/>. Accessed in July 2012.

[11] SWFTools. Available at <http://wiki.swftools.org>. Accessed in July 2012.

[12] CUPS-PDF. Available at < http://www.cups-pdf.de/ >. Accessed in July 2012.

[13] D. Lachowicz, C. McNamara. wvWare, library for converting Word documents. Available at <http://wvware.sourceforge.net>. Accessed in July 2012.

[14] Issuu. Issuu - You Publish. Available at <http://www.issuu.com/>. Accessed in July 2012.

[15] Soundcloud Ltd. Dashboard on SoundCloud - Create, record and share your sounds for free. Available at <http://www.soundcloud.com>. Accessed in July 2012.

[16] SlideShare Inc. Upload & Share PowerPoint presentations and documents. Available at <http://www.slideshare.com>. Accessed in July 2012.