UBICOMM 2013

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

ISBN: 978-1-61208-289-9

September 29 - October 3, 2013

Porto, Portugal

UBICOMM 2013 Editors

Wolfgang Narzt, Johannes Kepler University - Linz, Austria

Ann Gordon-Ross, University of Florida, USA

UBICOMM 2013

Foreword

The Seventh International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2013), held between September 29 and October 3, 2013 in Porto, Portugal, continued a series of international events meant 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 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.

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 place 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.

We take here the opportunity to warmly thank all the members of the UBICOMM 2013 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 who dedicated much of their time and efforts to contribute to UBICOMM 2013. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the UBICOMM 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that UBICOMM 2013 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of mobile ubiquitous computing.

We are convinced that the participants found the event useful and communications very open. We hope that Porto, Portugal, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

UBICOMM 2013 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 2013 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 2013

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 2013 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 2013 Technical Progam Committee

Afrand Agah, West Chester University of Pennsylvania, USA
Rui Aguiar, Universidade de Aveiro, Porutgal
Tara Ali-Yahiya, Paris Sud 11 University, France
Mercedes Amor, Universidad de Málaga, Spain
Timothy Arndt, Cleveland State University, USA
Mehran Asadi, Lincoln University, U.S.A.
Sergey Balandin, FRUCT, Finland
Matthias Baldauf, FTW Telecommunications Research Center Vienna, Austria
Michel Banâtre, IRISA - Rennes, France
Aurelio Bermúdez Marin, Universidad de Castilla-La Mancha, Spain
Daniel Bimschas, University of Lübeck, Germany
Carlo Alberto Boano, University of Lübeck, Germany
Bruno Bogaz Zarpelão, State University of Londrina (UEL), Brazil
Jihen Bokri, ENSI (National School of Computer Science), Tunisia
Sergey Boldyrev, Nokia, Finland
Diletta Romana Cacciagrano, University of Camerino, Italy
Jose Manuel Cantera Fonseca, Telefonica Investigacion y Desarrollo, Spain
Juan-Vicente Capella-Hernández, Universidad Politécnica de Valencia, Spain
Davide Carboni, CRS4 Research Center - Sardinia, Italy
Rafael Casado, Universidad de Castilla-La Mancha, Spain
Everton Cavalcante, Federal University of Rio Grande do Norte, Brazil
José Cecílio, University of Coimbra, Portugal

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., USA
Jorge Pereira, European Comission, Belgium
Yulia Ponomarchuk, Kyungpook National University, Republic of Korea
Daniel Porta, German Research Center for Artificial Intelligence (DFKI) - Saarbrücken, Germany
Ivan Pretel, DeustoTech - Deusto Institute of Technology, Spain
Elmano Ramalho Cavalcanti, Federal Institute of Education, Science and Technology of Pernambuco, Brazil
Juwel Rana, Luleå University of Technology, Sweden
Maurizio Rebaudengo, Politecnico di Torino, Italy
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
Kouichi Sakurai, Kyushu University, Japan
Johannes Sametinger, Institut für Wirtschaftsinformatik, Austria
Luis Sanchez, Universidad de Cantabria, Spain
José Santa Lozano, University of Murcia, Spain
Andrea Saracino, University of Pisa, Italy
Zary Segall , Royal Institute of Technology, Sweden
Sandra Sendra Compte, Universidad Politecnica de Valencia, Spain
Anton Sergeev, St. Petersburg State University of Aerospace Instrumentation, Russia
MªAngeles Serna, University of Castilla-La Mancha, Spain
Shih-Lung Shaw, University of Tennessee, U.S.A.
Kazuhiko Shibuya, The Institute of Statistical Mathematics, Japan
Catarina Silva, Polytechnic Institute of Leiria, Portugal
Luca Stabellini, The Royal Institute of Technology - Stockholm, Sweden
Radosveta Sokullu, Ege University, Turkey
Álvaro Suárez Sarmiento, University of Las Palmas de Gran Canaria, Spain
Kåre Synnes, Luleå University of Technology, Sweden
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Tetsuji Takada, The University of Electro-Communications - Tokyo, Japan
Kazunori Takashio, Keio University, Japan
Yoshiaki Taniguchi, Osaka University, Japan
Adrian Dan Tarniceriu, Ecole Polytechnique Federale de Lausanne, Switzerland
Markus Taumberger, VTT Technical Research Centre of Finland, Finland
Saïd Tazi, LAAS-CNRS, Université de Toulouse / Université Toulouse1, France
Tsutomu Terada, Kobe University, Japan
Stephanie Teufel, University of Fribourg, Switzerland
Lei Tian, University of Nebraska-Lincoln, USA
Marco Tiloca, University of Pisa, Italy
Chih-Cheng Tseng, National Ilan University, Taiwan
Ardian Ulvan, University of Lampung, Indonesia
Jean Vareille, Université de Bretagne Occidentale - Brest, France
Dominique Vaufreydaz, INRIA Rhône-Alpes, France
Miroslav Velev , Aries Design Automation, USA

Copyright Information

Table of Contents

Indoor Navigation by WLAN Location Fingerprinting

Reducing Trainings-Efforts with Interpolated Radio Maps

Dutzler Roland & Ebner Martin
Institute for Information Systems and Computer Media
Graz University of Technology
Graz, Austria
r.dutzler@student.tugraz.at, martin.ebner@tugraz.at

Brandner Robert
Department of System Development
evolaris next level GmbH
Graz, Austria
robert.brandner@evolaris.net

***Abstract*—Due to the fact that smartphones are today already used by about one out of seven persons worldwide and their capabilities concerning hardware and sensors are growing, many different indoor navigation solutions for smartphones exist. The solution presented in this paper is based on Wireless Local Area Network Location Fingerprinting. Fingerprinting is a method where signals at a number of specific points are measured once and stored in a database that is needed to determine the position later on. Measuring each and every fingerprint makes the off-line phase a complex and very time-intensive process, especially for big buildings. The bigger the building, the higher is the effort to create the map needed for the on-line phase to determine the position of a device. In order to cope with this complexity, an approach for optimizing the off-line phase is realized. The system substantially lowers the number of positions at which fingerprint measurements have to be taken by identifying ideal positions. All other non-measured fingerprints are determined by using a form of the Log-Distance Path Loss Model.**

Keywords-Indoor Navigation; Location Fingerprinting; WLAN; Smart-phone; Log-Distance Path Loss Model

I. INTRODUCTION

Wireless Local Area Network (WLAN) is the most widely used and studied technology for realizing an Indoor Positioning System (IPS). There are also other ones, such as Bluetooth ([1], [2]), Infrared ([1], [3]), the earth's magnetic field [4] or dead reckoning approaches [5], which can also be combined to make positioning more accurate. WLAN based IPSs have two basic advantages compared to other systems: usually a number of WLAN Access Points (APs) are already deployed in buildings and no additional hardware has to be installed to realize an IPS based on WLAN. Also, most smartphones as well as other mobile devices like tablets support WLAN by itself.

WLAN IPS are usually based on Location Fingerprinting (LF). LF approaches are generally divided into two phases: an off-line or learning phase and an on-line or positioning phase.

In the off-line or learning phase, the target environment is split up into a grid of a certain width. At each and every grid point (GP) reference measurements of the surrounding APs are taken. These reference measurements are called Fingerprints (FPs) and consist of a location information as well as the Received Signal Strengths (RSS) of the different APs at this position. The collected FPs are stored in a so called Radio Map (RM).

In the on-line or location phase, a mobile device, which is aware of the stored RM of the target environment, can be used to determine the position. The device measures the signal strengths of the APs and compares it to the FPs stored in the RM in order to find its current position. There are many ways to calculate the position. A very simple is to compare the measured FP to the FPs stored in the RM and to find the three most similar ones by its mean square error. The position of the device then can be determined by calculating the balance point of the triangle that is made up by the location information of these three FPs.

Many research projects focus on improving the accuracy of the LF approach. The studies in [6] focus on improvements during the off-line or learning phase. The authors pointed out that additional direction information for every RSS measurement at one FP location might be helpful. The evaluation showed that a map of averaging RSS measurements from four directions to one FP showed the best values concerning accuracy and memory load, comparing to a map having a separate FP measurement for four directions at one location and another one considering a lower amount of all samples.

Kaemarungsi and Krishnamurthy [7] did some research on factors which influence the RSS values. They investigated that the user shadowing the signal from a device has an impact on the measured RSS values and so does the user's orientation. Measured RSS values at the same location tend to be different when examined over time, for example a day.

Zhao et al. [8] implemented a Differential Evolution (DE) algorithm to optimize the number and locations of APs for WLAN IPSs. It maximizes the variety of the FPs to improve location determination. Their tests show that symmetrical placement is worse than the calculated one by the DE algorithm. According to the authors it is better to place the APs in a 'zigzag' pattern. An increasing number of APs stops improvements when the system proposes to place them closely together.

It can be summarized that a general advantage of the WLAN LF approach is that it deals very well with factors that have influence on the RSS at a certain point, e.g. attenuation caused by walls. However, there is one major drawback for this method. The off-line phase is a complex and very time-intensive process, especially for big buildings. The bigger the building, the higher is the effort to create the map needed for the on-line phase to determine the position

 ISBN: 978-1-61208-289-9

of a device. The number highly depends on the buildings size and on the accuracy the final system shall achieve. Common grid widths can range from one to five meters. For a large shopping center and a grid width of five meters the number of reference FPs that have to be measured for the RM can easily grow up to 2500 or more. In order to cope with these complexities, a new, very promising approach for optimizing the off-line phase has been realized.

Section II describes the goals of the research work. Section III explains how the implemented algorithm works. The evaluation and evaluation results are described in section IV. The last section concludes the research and points out problems that have to be investigated in future studies.

II. Objectives

The research work of this publication aims to develop a prototype application for an Android smartphone, which can be used for indoor navigation in large public buildings like shopping centers, hospitals or museums. The application shall determine the position of a device as accurate as possible and work robustly. In order to overcome the complexity of measuring each and every FP for the RM in the off-line phase, an algorithm has been implemented that substantially lowers the number of FP measurements. The algorithm suggests ideal GPs at which reference measurements should be taken on the basis of the building's plan, the building's dimensions, the exact course of the walls and the exact places of APs.

The idea behind the algorithm is that the fingerprint for each and every GP which distance to a certain AP is crossed by the exact same walls can be interpolated easily and exactly by measuring only one of them. Hence, the signals for these GPs travel through the same walls, their attenuation and other signal influence factors are nearly the same. Their signal strengths are of course dependent on the distance the GP is away from the AP. The next chapter explain the algorithm in more detail.

III. Method

The algorithm basically consists out of two parts:

- Part 1 – Determine Ideal GPs
- Part 2 – Interpolate FPs

In order to determine the GPs at which measurements have to be taken and subsequently interpolate the non-measured fingerprints, the algorithm needs the following information:

- True scale building's plan
- Dimensions of the building
- Exact course of walls
- Exact positions of APs
- GPs

A. Part 1 – Determine Ideal GPs

Part 1 of the algorithm determines ideal GPs at which measurements have to be taken. The requirement of good GPs depends on the number of APs, the grid density and hence the number of GPs, and the number of unique walls an AP crosses on the path to the GPs.

The algorithm takes the total of all GPs and determines for every GP and every AP the walls, which are crossed by the connecting line between GP and AP. All GPs whose connecting lines to a certain AP are crossed by the same walls, are grouped up and one of them gets suggested as ideal GP, which has to be measured, and is used to interpolate all other GPs of the group in part 2. The ideal GP is the one that is closest to the central point out of the GPs of one group, to calculate all surrounding ones as good as possible. Fig. 2 shows an example for one suggested GP and one AP. The solid line points to the suggested GP. All other GPs, which are marked by the dotted arrows, will be interpolated by the measurement of the suggested one, because the distances between these and the suggested GPs are crossed by the same wall.

B. Part 2 – Interpolate Fingerprints

After the suggested ideal GPs have been measured, in part 2 of the algorithm the non-measured FPs are determined by interpolation, with the help of the Log-Distance Path Loss Model for Line-of-Sight (LOS) environments, as described in [9]. The signals of all GPs which are covered by the exact same walls from an AP have nearly the same attenuation factors. So if the signal strength of one of those GPs is measured, the signal strength of all other surrounding GPs which are covered by the same walls can easily and very precisely be calculated by using a signal attenuation model for LOS environments, the Log-Distance Path Loss Model (1).

$$\overline{PL}(d) = \overline{PL}(d_0) + 10n\log\left(\frac{d}{d_0}\right) \qquad (1)$$

The wanted value PL(d) is the path loss or signal strength at a distance d. PL(d_0) is a measured signal strength or path loss at a reference distance d_0, whereas n is the path loss exponent. The exponent varies for every environment and is usually evaluated using empirical data. For LOS environments the path loss exponent is 2. The distances shall be given in meters. Fig. 1 shows how the ideal grid points are selected using pseudo code.

```
function FINDGRIDPOINTS()
Input: accessPoints, gridPoints, walls ∈ List
Output: foundGPs ∈ List
    for all accessPoints do
        wallHashMap ← newHashMap < String, List < Point >> ()
        for all gridPoints do
            for all walls do
                if wall.intersects(gridPoint, accessPoint) then
                    wallHash ← wallHash + wall.hashCode()
                end if
            end for
            wallHashMap.add(wallHash, gridPoint)
        end for
        for all wallHashMapEntries do
            goodGridPoint ← selectGoodGridPoint(wallHashMapEntry.value)
            if !foundGPs.contains(wallHashMapEntry.key) then
                foundGPs.add(wallHashMapEntry.key, goodGridPoint)
            end if
        end for
    end for
    return foundGPs
end function
```

Figure 1. Pseudo Code – Find Ideal Grid Points

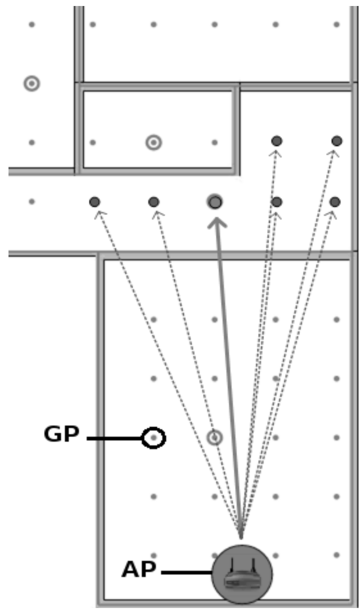


Figure 2. Part 1 – Determine Ideal GPs Example

The described method is another similar approach compared to the dominant path model described in [12]. The dominant path to any GP is the shortest path which crosses the lowest number of walls. The idea is that the signal is attenuated less when the dominant path crosses only one wall compared to the direct path that crosses three walls. The result achieved by the method described in this paper is at least the same or even more precisely. Difference is that the dominant path model suggests only one GP to measure for every AP in one room, but it also assumes that the attenuation along a long path is less compared to a direct short path.

Fig. 3 shows the differences between the calculated RSS attenuation in a LOS environment and real recorded attenuation.

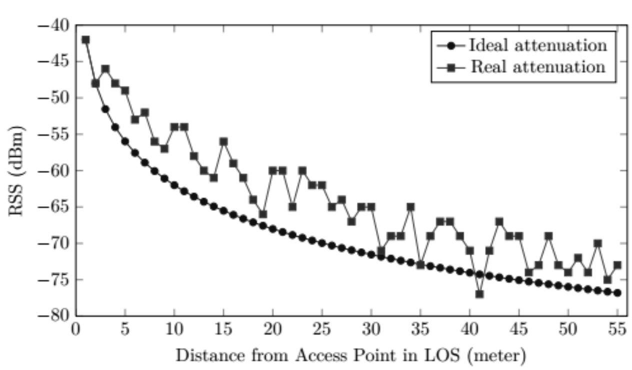


Figure 3. Differences between calculated RSS attenuation and real recorded attenuation

It is evident that the real recorded attenuation is smaller than the ideal, calculated signal attenuation.

IV. EXPERIMENTS AND RESULTS

The evaluation of the prototype aimed to investigate the performance of the algorithm in complexity-savings as well as accuracy. The tests were done in three different surroundings: two Non-Line-of-Sight (NLOS) environments - one office building and a flat - and one LOS environment, a tennis court. For each test scenario two RMs were created to compare the accuracy. One RM where all GPs of the test environment were measured and one RM that used our optimization algorithm to create it. In each environment some Test Points (TPs) were randomly picked and afterwards the average error rate in meters is determined for all implemented positioning algorithms [10]:

- Simple Triangulation Algorithm (STA)
- Weighted Triangulation Algorithm (WTA)
- K-Nearest-Neighbor Algorithm (KNN)
- Weighted K-Nearest-Neighbor Algorithm (WKNN)

To make the evaluation as accurate as possible and hence ensure the integrity of all distances including distances to APs, GPs and TPs were exactly displayed by the application and measured in the real environment using a digital range measure. Fig. 4 shows a screenshot of the activity of the application that is used for creating the RM. The distances from the selected GP to the surrounding walls are displayed.

The creation of the RMs as well as the evaluation of the algorithm in the following test scenarios were done within one day. All FPs in the learning phase were measured in four directions and averaged with the user heading in the given direction and holding the smartphone in front of the chest. The TPs in the evaluation phase were measured in one random direction, also with the user holding the smartphone.

The following hardware was used for doing the evaluation:

- 4 APs – Cisco Linksys Wireless-G Broadband Router (WRT54GL)
- Smartphone – HTC Evo 3D (X515m)
- Silverline Ultrasonic Digital Range Measure

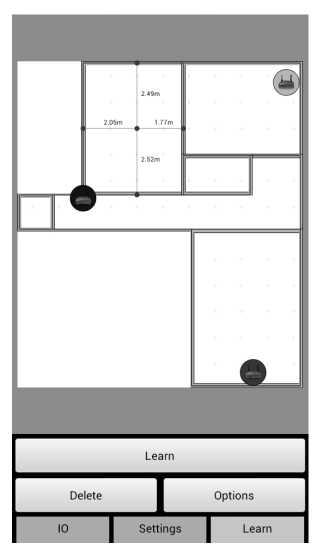


Figure 4. Screenshot of the Application

 ISBN: 978-1-61208-289-9

C. *Test Scenario I (Flat)*

Test scenario I is a typical small flat with two rooms, a kitchen, a bathroom, a toilet and a corridor of the size of 10.5m x 12.5m and a total area of 77.0m². In that scenario a grid width of 1.0m and 3 APs were used, which results in a total of 66 GPs. The number of ideal GPs that were suggested by the optimization algorithm was 22, which results in savings of 66%.

The evaluations were done at ten equally distributed TPs. The overall accuracy achieved by the Measured-RM (M-RM) was slightly better with used all algorithms. At two TPs the positioning accuracy reached with the Interpolated-RM (I-RM) was even better. This could be due to measuring inaccuracies caused by the test person covering signals from surrounding APs. The mean accuracy averaged over all TPs and algorithms by the M-RM was 1.66m and 1.89m with the I-RM. If we keep in mind that only one-third of all GPs were measured for the I-RM the error increase of 23cm is a very encouraging result. Fig. 5 shows the mean positioning error averaged over all TPs for each algorithm and every RM. The performances reached by the different algorithms are nearly the same for the M-RM. For the interpolated one, the WTA achieved the best overall accuracy. Fig. 6 shows the absolute RSS differences between the M-RM and the I-RM. The highest difference at that scenario is at the marked circled position in Fig. 6. The absolute RSS difference at this point is -13 dBm. Hence, the overall differences between both RMs are low.

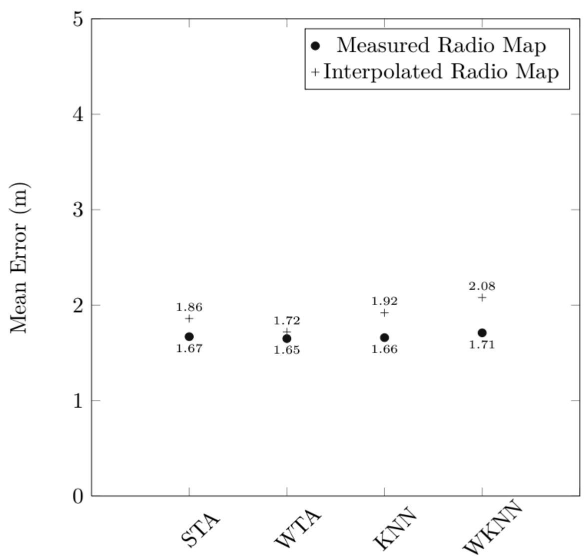


Figure 5. Mean Positioning Error of all Algorithms of Test Scenario I

D. *Test Scenario II (Office Building)*

Test scenario II is a typical office environment with some small offices and meeting rooms. The environment is 21.0m x 15.0m in size and covers an area of about 200m². A grid width of 1.3m and 3 APs were used in this scenario, which results in a total of 85 GPs. The optimization algorithm suggested 47 GPs, which are savings of 45%.

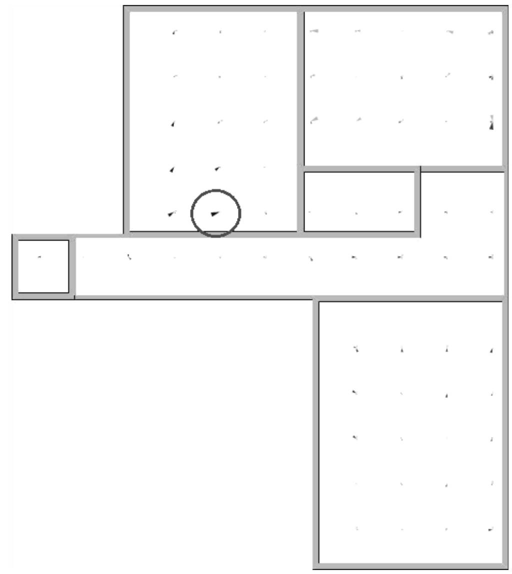


Figure 6. Absolute Fingerprint Differences between M-RM and I-RM of Test Scenario II

To evaluate the accuracy between the fully M-RM and the I-RM created with the optimization algorithm, measurements at 12 equally distributed TPs were taken. The results are similar to the results of test scenario I. The overall mean error for the M-RM was 2.11m and 2.54m for the I-RM. This means that the error that occurred due to only measuring about the half of all GPs and interpolating the other ones is 43cm or about 20% higher. Fig. 7 shows the mean errors of all implemented algorithms for both Radio Maps. The absolute FP differences are shown in Fig. 8.

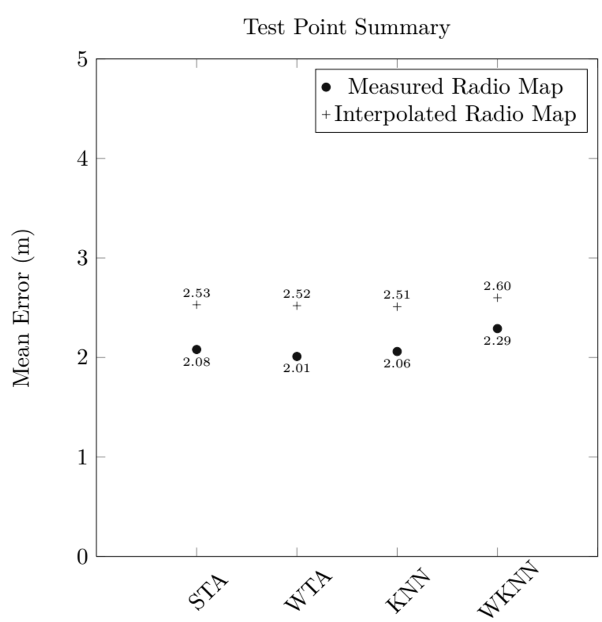


Figure 7. Mean Positioning Error of all Algorithms of Test Scenario II

Figure 8. Absolute Fingerprint Differences between M-RM and I-RM of Test Scenario II

E. Test Scenario III (Tennis Court)

The last test aimed to test the performance of the algorithm in a fully LOS environment and was done outdoors on two tennis courts. The area was 36.0m x 36.0m in size which result in a total area of 1296m². For the evaluation 4 APs and a grid width of 5.0m was used, which leads to a total number of 49 GPs. Because they are no walls in that scenario the optimization algorithm suggested only one GP in the mid of the environment for the I-RM, which are savings of 98%.

Measurements from 13 equally distributed TPs were taken. The overall accuracy for both RMs were substantially lower compared to the results from Test Scenario I and II. The mean error for the M-RM was 4.44m and 1.14m higher for the I-RM, which is 5.58m. Fig. 9 shows the mean errors for all algorithms and both RMs. Fig. 10 again shows the FP differences between the M-RM and the I-RM.

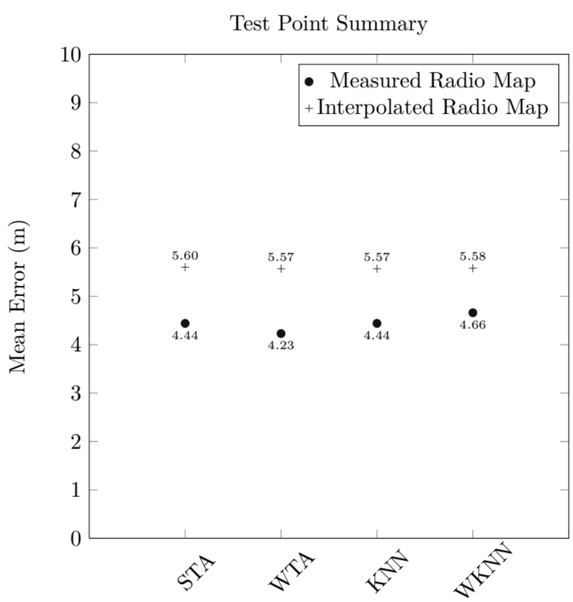


Figure 9. Mean Positioning Error of all Algorithms of Test Scenario III

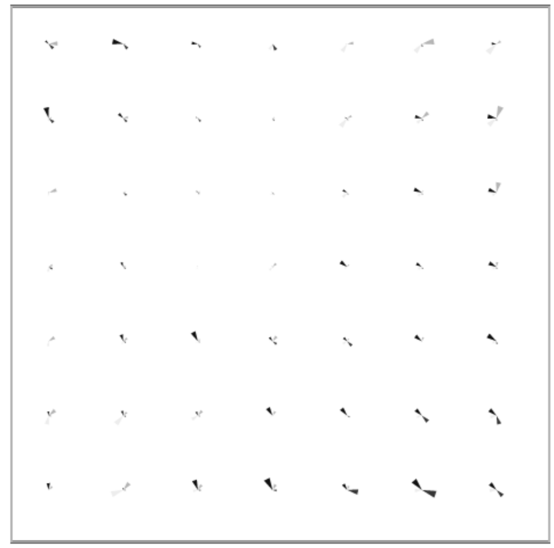


Figure 10. Absolute Fingerprint Differences between M-RM and I-RM of Test Scenario III

F. Summary and Overview

Table I gives an overview over the test scenarios, the environment they represent, their dimensions, the total area they cover, the number of APs that were used for the evaluation and the selected grid width. Table II then summarizes the performance of the implemented optimization algorithm. It states the complexity savings as well as the accuracy differences between the M-RM and the I-RM.

TABLE I. OVERVIEW TEST SCENARIOS DESCRIPTION

Scenario	Environ ment	Dimensions	Area	# APs	Grid Width
I	NLOS	10.5m x 12.5m	77.0m²	3	1.0m
II	NLOS	21.0m x 15.0m	200m²	3	1.3m
III	LOS	36.0m x 36.0m	1296m²	4	5.0m

TABLE II. OVERVIEW TEST SCENARIOS EVALUATION RESULTS

Scenario	# GP	# Ideal GP	Savi ngs	∅ Acc. M-RM	∅ Acc. I- RM	∅ Acc. Deviatio n
I	66	22	67%	1.66m	1.89m	+23cm
II	85	47	45%	2.11m	2.54m	+43cm
III	49	1	98%	4.44m	5.58m	+1.14m

The evaluation in NLOS environments showed a very satisfying overall performance. The mean error in test scenario I and a grid width of 1.0m for the M-RM of all positioning algorithms is 1.66m. The mean error for the I-RM is 1.89m and hence only 23cm or 13.98% higher. At test scenario II, another NLOS environment, and a grid width of

1.3m the mean error for the M-RM is 2.11m and 2.54m for the I-RM, which is a difference of 43cm or 20.37%, although there were high deviations at some TPs of nearly up to 5m (in some cases only for the I-RM or even both RMs). At test scenario III, a LOS environment, and a grid width of 5m the mean error for the M-RM is 4.44m and 1.14m or 25.60% higher, which is 5.58m, for the I-RM. In that scenario there were also some TPs at which the mean error grew up to over 8m. A further test aimed to determine the influence of the grid width on the positioning accuracy pointed out that a grid width of 2.5m is best. Lower as well as higher grid width has a negative impact on the resulting positioning accuracy.

V. CONCLUSION AND FUTURE WORK

The overall results are promising although there are some points that have to be further investigated. The WLAN LF approach has some basic restraints compared to other solutions. The RM of one environment is usually created once but the actual signal strengths may change over time because they are influenced by many factors. Furthermore every change in the environment can have an impact on the signal distribution of the WLAN routers. Walls, doors or even new or differently placed furnitures will change the received signal strengths at some points in the environment.

The tests were carried out using the same routers as well as the same devices for every test. One test case shows that different devices receive slightly different signal strengths from the same AP at the same place, depending on the built-in hardware. The evaluations do not care about what happens when different or other routers are used. Different routers have different antennas and hence do not cover the same area or have the same signal attenuation over the same distance. If different routers are used for creating one RM, a constant value might have to be added for every single AP to make the optimization algorithm accurate.

The approach was only designed and tested in 2-dimensional areas. It can be used for multi-story buildings by identifying the storey by the strongest signals, but the currently developed algorithm does not consider APs located on different storeys for optimization.

The evaluation revealed some incidents that might have an influence on the accuracy and have to be further investigated and improved. One very basic but important thing is the distribution and number of APs, which was topic of the studies [8] and [11].

At test scenario 3 there is only one GP out of a total number of 49 GPs used for the I-RM. The area is 36m x 36m in size. At some TPs the mean error for the I-RM is dramatically higher compared to the measured one. A final test pointed out that measuring more GPs (for example at least every 10 meters) and lowering the grid width improve the results arbitrarily.

Another issue that has to be reconsidered in detail is the Log-Distance Path Loss Model and the path loss exponent. An evaluation showed that the exponent could be slightly lower, probably this also depends on the APs.

The research work shows that the implemented optimization algorithm for the off-line learning phase of the WLAN LF approach is a legitimate way to reduce the complex, time-consuming way of establishing a RM for big buildings. The GPs at which measurements have to be taken can be substantially lowered and in addition an acceptable positioning accuracy reached compared to a fully M-RM. As already mentioned the evaluations revealed some factors that have to be considered and explored in future work to increase the positioning accuracy: The number as well as the distribution of APs, the maximum distance between two measured grid points to make the interpolation more accurate, the influence of different hardware – mobile devices as well as routers - and the calculation using the Log-Distance Path Loss Model. Especially the influences on the Path Loss exponent should be a starting point for further work.

ACKNOWLEDGMENT

I would like to express my gratitude to evolaris next level GmbH for offering and allowing me the work on this exciting topic of research.

REFERENCES

[1] Y. Gu, A. Lo, and I. Niemegeers, "A Survey of Indoor Positioning Systems for Wireless Personal Networks", IEEE Communications Surveys and Tutorials, Volume 11, Number 1, 2009, pp. 13 – 32.

[2] G. Fischer, B. Dietrich, and F. Winkler, "Bluetooth Indoor Localization System", Workshop on Positioning, Naviagtion and Communication, 2004, pp. 147 – 156.

[3] R. Want, A. Hopper, V. Falcão, and J. Gibbons., "The Active Badge Location System", ACM Transactions on Information Systems (TOIS), Volume 10, Issue 1, 1992, pp. 91 – 102.

[4] IndoorAtlas, "Ambient magnetic field-based indoor location technology", Technical Report, 2012.

[5] S. Beauregard and H. Haas, "Pedestrian dead reckoning: a basis for personal Positioning", 3rd Workshop on Positioning, Navigation and Communication (WPNC), 2006, pp. 27- 35.

[6] H. Leppäkoski, S. Tikkinen, and J. Takala, "Optimizing Radio Map for WLAN Fingerprinting", Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS), 2010, pp. 1-8.

[7] K. Kaemarungsi and P. Krishnamurthy, "Properties of Indoor Received SignalStrength for WLAN Location Fingerprinting", Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous), 2004, pp. 14-23.

[8] Y. Zhao, H. Zhou, and M. Li, "Indoor Access Points Location Optimization Using Differential Evolution", International Conference on Computer Science and Software Engineering, 2008, pp. 382-385.

[9] J. Mulligan, "A Performance Analysis of a CSMA Multihop Packet Radio Network", Master Thesis, Polytechnic Institute of New York, 1997.

[10] H. Liu, H. Darab, P. Banerjee, and J. Liu, "Survey of Wireless Indoor Positioning Techniques and Systems", IEEE Transactions on Systems Man and -Part C: Applications and Reviews, Volume 37, Number 6, 2007, pp. 1067 – 1079.

[11] O. Baala, Y. Zheng, and A. Caminada, "Toward Environment Indicators to Evaluate WLAN-Based Indoor Positioning System", International Conference on Computer Systems and Applications, 2009, pp. 243 – 250.

[12] H. Wang, H. Lenz, A. Szabo, J. Bamberger, and U. Hanebeck, "Enhancing the Map Usage for Indoor Location-Aware Systems", International Conference on Human-Computer Interaction (HCI 2007), 2007, pp. 151 – 160.

Facilitating Utilization of Public Transportation for Disabled Persons by an Open Location-Based Travel Information System for Mobile Devices (VIATOR)

Wolfgang Narzt
Department of Business Informatics – Software Engineering
Johannes Kepler University
Linz, Austria
wolfgang.narzt@jku.at

Abstract — **In this paper, we present an open and extensible platform for mobile devices that guides disabled (i.e., blind or physically impaired) persons on their journeys using public transportation. The main objectives of this platform are to (1) provide up-to-date intermodal travel information (i.e., connecting vehicles, schedules, and delays) across different transportation companies, (2) automatically, without manual interaction, reroute to alternative transportation options, if required, (3) visualize physical obstacles and offer hints for each particular target group (e.g., stairs vs. elevators for wheelchair users), and (4) provide an open interface for leaving and consuming self-created location-bound hints (e.g., blind people guide blind people). The platform is called VIATOR (lat.: "the traveler") and has been developed in cooperation with three transportation companies and three research institutes in Austria. This paper presents the technical architecture, shows first results of a prototypical implementation and gives user experiences from blind and mobility impaired people.**

Keywords-VIATOR; Public Transportation; Disabled People; Mobile, Location-Based Servce Platform.

I. Introduction

People with disabilities are disadvantaged when using public means of transportation. Wheelchair users, for instance, are severely limited by stairs and raised vehicle entrances, whereas blind or visually impaired people are reliant on tactile lines or proper path descriptions in order to be able to maneuver on their own at public transportation nodes. Commercially available travel information- and navigation systems for public transportation inform their users about departure times, platforms, and type of vehicles. However, they lack instructions for handicapped persons on the way to and from public vehicles particularly at larger transportation nodes. Ponderous and inflexible content management mechanisms at the backend of such systems exacerbate maintenance for this task. Moreover, many of the systems use proprietary protocols and are unable to exchange data with or connect to competing transportation companies for a closed information chain throughout a journey.

In the course of the research project ways2go [1], within the framework of the strategic initiative IV2Splus funded by the Austrian government (FFG), a prototype for an open and extensible location-based travel information system for disabled persons has been developed guiding each target group through stations regarding their needs. It consolidates diverse travel information systems and provides navigation instructions from arbitrary sources, even from the users themselves (self-organizing content management) using a mobile client prototype with accessible User Interface (UI) elements.

The technical basis for the implementation of this research topic was a mobile location-based and context-sensitive information-, communication-, and collaboration system (Digital Graffiti, see [2], [3] and [4]) developed by the University of Linz, in association with Siemens Corporate Technology in Munich, and the Ars Electronica Futurelab also in Linz, which enables its users to place and consume information at arbitrary locations using state-of-the-art mobile tracking-enabled cell phones. Travel information (including up-to-date actual data of delays, cancellations or detours, etc.) was provided, classified and made generic for shared networking by the project partners Austrian Federal Railways Company (ÖBB), Upper Austrian Transport Association (OÖVG) and Linz AG (a Local Traffic Line Service Provider in Linz). In cooperation with the Department Integriert Studieren at the University of Linz and the Central European Institute of Technology in Vienna (CEIT Alanova) new paradigms for barrier-free interaction have been created, not only guiding users, but also offering them an instrument to provide self-created content for other users. Blind people shall be able to annotate their way for other blind people regarding their special needs.

The paper is structured as follows: Section 2 deals with selected points of state-of-the-art methods and technology. Section 3 gives an insight into the proposed system architecture for VIATOR. Section 4 illustrates the prototype and finally, Section 5 concludes the paper, assesses the preliminary results critically and prospects future work.

II. Related Work

The FH Joanneum in Austria focuses on unaided free movement for people with disabilities using public transportation in a project called NAVCOM [5]. Blind persons are required to find the right vehicle or to signal their wish to enter or leave a vehicle. The authors propose a WLAN-based system communicating between public transportation vehicles and smartphones, an extension to navigation systems for pedestrians, the functionality of which ends at the entrance door of the vehicles. In general, this project group investigates the potentials of technical support for navigating hand-

 ISBN: 978-1-61208-289-9

icapped people in public spaces. Within the project ways4all (funded by the Austrian government), they explore indoor navigation for visually impaired using Radio Frequency Identification (RFID) and navigation instructions with relative coordinates [6][7].

A scientific consortium consisting of four universities in Japan has developed a different approach supporting blind travelers in their navigation. A device with a haptic interface (they call it the Future Body-Finger [8]) enables blind persons to perceive their environment, measuring distances to obstacles using infrared sensors.

MoViH [9] is an approach that goes a step beyond. It tries to identify both mobility and hindering factors for persons with visual or hearing impairments. The outcome of the findings is a catalogue of effective and efficient measures to be depicted in recommendations and standards supporting public transport companies in planning environments considering special needs of blind or acoustically disabled persons. BIS – Barrier Information System [10] especially focuses on the requirements of wheelchair users. The project aims at developing a barrier-free interactive routing system in close coordination with the target group, technology experts and administrative and political stakeholders throughout the research process in order to calculate and visualize the most suitable ways to go for wheelchair users.

Target-oriented automatic delivery of information to the traveler (e.g., for indicating a transfer or delay) is the next field of investigation. First approaches try to inform travelers about entry and exit maneuvers, as an assistive application from the Polytechnic Institute of Leiria, Portugal does [11]. Most systems, though, do not inform their users about changes in the time schedule, once the trip has been calculated. Travel information has to be requested from scratch at every transition point or is difficult to handle due to a complex system of rules across transportation providers. Until today, to the authors' knowledge, there is no automatic mobile travel information system that continuously guides the passenger during his journey and context-sensitively keeps him up-to-date considering transfers or delays.

As a summary, we already recognize a series of isolated research subjects dealing with navigation aspects for disabled persons or closed information chains. The novelty in the VIATOR project is the approach for constructing traffic support systems, i.e., it utilizes a smart location-based information system and a simple processing procedure for triggering specific actions due to regional closeness of its users to selected locations, in contrast to complex rule systems discovered in related projects.

III. SYSTEM ARCHITECTURE

The name of this location-based information system is Digital Graffiti [2][3][4], a framework that stores 3D spatial information elements (e.g., text, images, sound, videos, links, or even executable code) in a central database. Every stored data tuple (i.e., a geo-position and an attached information element – we call it a Digital Graffito) is additionally characterized by a visibility space and a set of recipients. It is transferred when any of the recipients steps into its visibility space. For related developments, see, for example, [12].

As a special feature, the platform offers automatic control of electronic actions (e.g., opening a gate, starting or stopping a machine, triggering a measurement or transaction) without any additional manual action, when a given device is in the vicinity of a Digital Graffito containing executable code (see [13]), presuming adequate access privileges of a person, a device, or a software system.

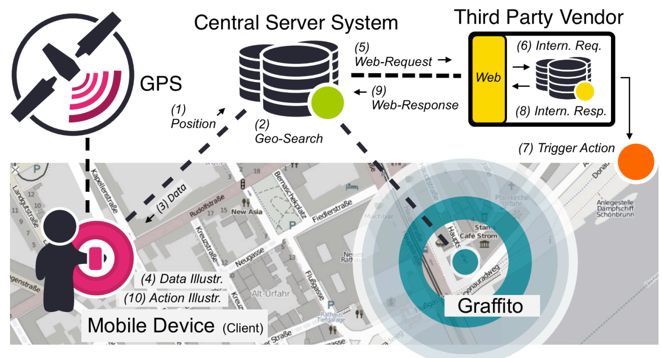


Figure 1. System Architecture for Proximity-Triggered Code Execution

Figure 1 illustrates the common principles of the software architecture enabling proximity-triggered code execution: A central server system stores geographically linked information in appropriate fast traversable geo-data structures (e.g., r-trees). Clients repetitively transmit their own (commonly by GPS-based) position to this 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 text or pictures, it is immediately displayed on the output device of the client (4). The basic idea for executing code is to store program fragments inside these information elements instead of text or binary picture data. Therefore, we propose a web-service-based mechanism, which is both effective and simple to extend: The information elements contain URL or XML-based web-requests to a remote web-service, which is the actual component to execute the code. When a client receives information containing a URL or web-request, it is resolved (5), 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 was executed successfully or not (10).

This mechanism is fairly simple and handles standardized HTTP-requests supported by a majority of currently utilized mobile platforms. 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 of data security for the vendors. The architecture provides the basis for a high degree of extensibility to third-party systems, for the number and variety of electronic connections is unforeseeable and simultaneously enriches the potentials of such a service.

We have conducted experiments on this innovative interaction paradigm: We have put a proximity-triggered code element in front of an electronically controllable gate at the

University Campus. An authorized person approaching the element automatically triggers the execution of the contained code, which causes the gate to open. Admittedly, this use-case is more of a proof of concept than a practical application, for GPS-data are not accurate enough to differentiate between two cars in a row, however, it demonstrates the potentials of the service enabling its users to initiate any electronically controllable actions just by their physical presence.

Applying this service as the technological basis for a mobile public transportation guide means utilizing the location-based action control mechanism for up-to-date calculations regarding transportation schedules. The transportation companies have to provide standardized interfaces for requesting their schedules, enabling their users to site-specifically and automatically perceive appropriate departure information when they arrive at a station or stop (similar to the big screens showing the departing trains). The user is consequently able to pick the desired destination by a single click and, thus, to anonymously specify a route.

Concerning intermodal travel routes, the system does not invent routing algorithms and scheduling procedures from scratch. Instead, it utilizes existing services and triggers them on demand in the same way as the time schedule example given above. So, the system continuously (and, in particular, location-sensitively) re-initiates calculations regarding the selected route by the location-based action control mechanism at neuralgic maneuver points during the journey, i.e., at train stations, bus or tram stops, etc. In general, neuralgic points are particular locations along the route where users may enter or leave means of public transportations. This means that a user automatically triggers the route planning service of the appropriate vendor at spatial proximity to his next stop and is informed about his schedule and connecting means of transportation, giving the passenger a continuous information chain during his journey.

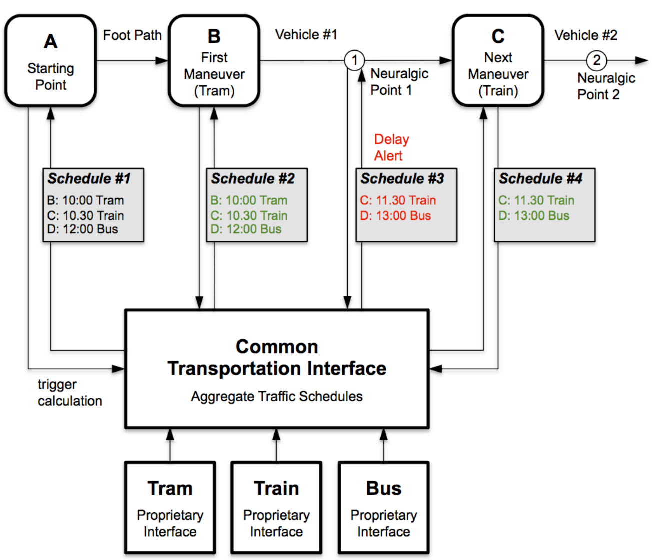


Figure 2. Intermodal Route Recalculations at Neuralgic Points

Figure 2 gives an impression on the procedure with a fictive example: A user starts his journey at an arbitrary location (point A), walks by foot to a tram station (point B), takes the tram to the main railway station (point C), uses the train, and finishes his journey with a bus (point D, not depicted in the figure, anymore). At point A, he initiates the calculation procedures (the only manual interaction by specifying the desired destination), i.e., the system collects all available data concerning the given route from different transportation providers (via a Common Transportation Interface) and offers a sequence and a schedule of the involved vehicles. Along the route, the system places active Digital Graffiti at neuralgic locations, which automatically trigger recalculations when the user arrives there.

So, while staying at point A the system continuously (i.e., with a time-controlled delay) recalculates the same route over and over again, offering the user the chance to specify his journey in advance without considering time schedules. The system always has the most up-to-date travel information for its users.

When leaving point A, the next neuralgic point along the route (i.e., point B in our example) initiates recalculations. As this point is also a maneuver point (i.e., where the user must enter or leave a vehicle due to his schedule) the system not only provides an up-to-date route but also informs the user on the departure time by an alert. In our example, the first recalculations (schedule #2) produce the same data as the initial calculations, indicating that there is no delay or change in the schedule.

While travelling with the tram, the same mechanism is repetitively applied at every neuralgic point (exemplarily shown by point #1). The corresponding schedule reveals a delay for the subsequent train (i.e., the new schedule deviates from the last one) and automatically postpones the trip with the connecting means of transportation after the newly calculated arrival time of the train. When schedules differ, the system automatically notifies its users by an alert. Alerts are also fired when the user is requested to get off at his last stop before changing means of transportations. The signal to leave is triggered by location and is therefore independent from any delays in the time schedules.

Principally, this procedure with repetitive recalculations (done by external schedule calculation services) as shown in Figure 2 is the main idea and a new approach for implementing an automatic mobile location-based travel information service. A “Common Transportation Interface” consolidates proprietary interfaces from different transportation providers and delivers up-to-date schedules every time a user approaches his next neuralgic point along his route.

In terms of barrier-free routing the same mechanism is applied: A handicapped person entering a station or stop automatically triggers navigation calculations due to his/her user profile (i.e., due to the type of disability), selected route and schedule. Users may either receive navigation instructions along tactile lines or get the quickest route to the next elevators. These calculations are externally sourced out to special route computing services with centralized data collections to be updated either via Content Management System (CMS) operators or by the users of the system themselves, who are able to edit these instructions due to their experiences on-site via the inherited mechanism of the basic system for editing Digital Graffiti information elements.

Figure 3 illustrates the principle: Every neuralgic point stores a set of additional active information elements addressed to specific profile groups, e.g., blind people, wheelchair users, etc., which automatically trigger profile-related footpath calculations for the way to the next vehicle. Please note that not every point necessarily stores information for each profile group. However, every user is capable of adding or updating information. Also, note that, for simplicity and scalability reasons, these information elements are directly linked to neuralgic points; they do not have their own position, e.g., in between two bus stops.

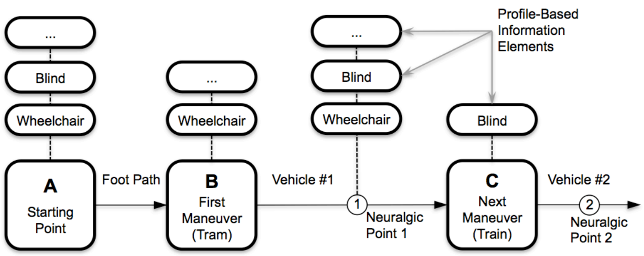


Figure 3. Profile-Based Navigation Instructions at Neuralgic Points

As a summary, the basic mechanism for barrier-free intermodal navigation providing up-to-date travel information and offering an active notification- and feedback instrument for direct interaction and self-organization of the information content is based on location-based action control, a function available from the basic platform Digital Graffiti. So, the VIATOR system does not contain complicated new algorithms for managing complex collaboration of different information providers. Instead, it utilizes a simple mechanism applied for all tasks, which makes the entire system easy to understand, maintain and extend.

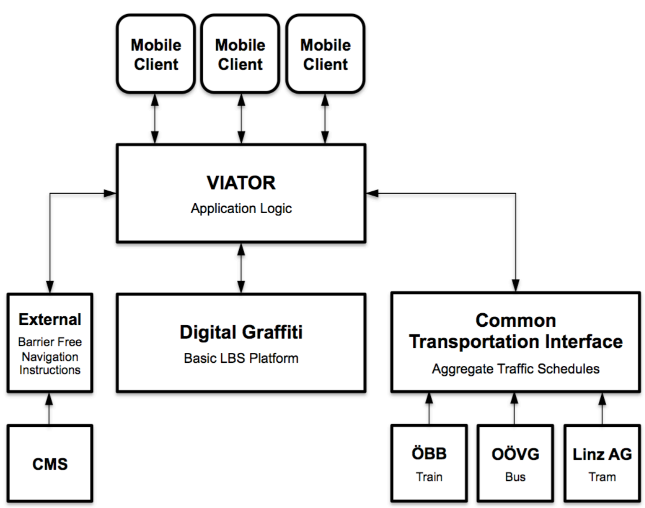


Figure 4. VIATOR system architecture (simplified excerpt)

Figure 4 gives an impression on the collaboration of involved components, omitting details due to space limitations: The application logic is encapsulated within the main VIATOR server component, communicating to its basic platform Digital Graffiti, which remains completely unchanged in its elementary behavior and is executed as a separate process. The mobile clients connect to the main VIATOR server component, which delegates all Digital Graffiti-related functions to its underlying platform, enabling the clients to create and consume location-based information elements. The system is extended by two separate processes consolidating travel information from different transportation providers on the one hand (Common Transportation Interface) and providing barrier-free navigation instructions on the other (left side). The latter can either be managed through CMS or via the VIATOR application logic indirectly by the clients.

IV. PROTOTYPE AND EXPERIMENTS

VIATOR has been prototypically implemented using JBoss AS7 Java Enterprise Application for the server components and Android for the clients. We have conducted several tests in the city of Linz arbitrarily using different means of transportation from the involved transportation companies. Figure 5 gives an impression on a typical test run:

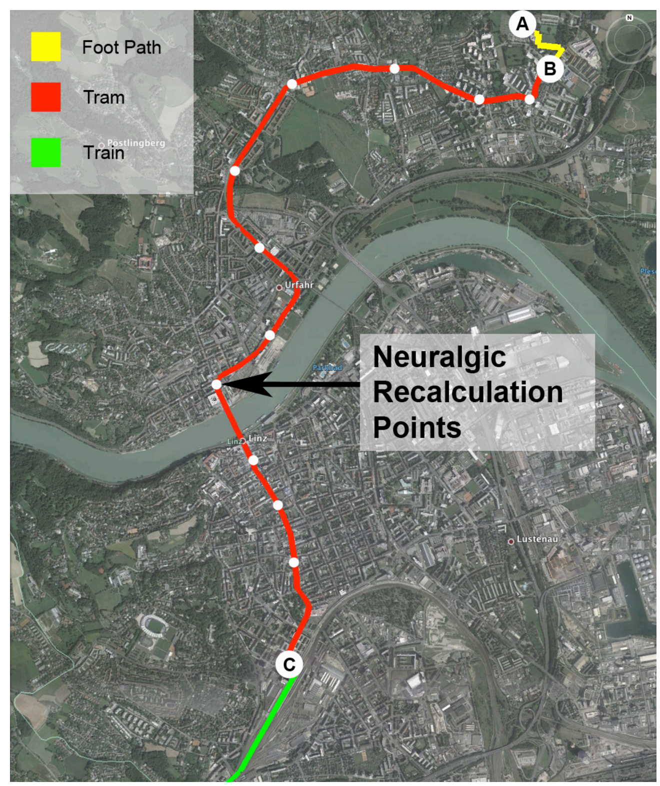


Figure 5. Test Run for Route Recalculations at Neuralgic Points

We started our journey at the University (A), walked by foot to the next tram and bus station (B) and continued either by tram or by bus (both options are possible for this route) to the main rail station (C), where we stopped our journey (for

 ISBN: 978-1-61208-289-9

convenience and time reasons). However, this abbreviated trip is enough for evaluating the mechanism when more than one vehicle is involved. The picture also gives an idea of the position of neuralgic points along our test route. Please note that the points are closer along the tram line than along the train line. This is due to the distance of stops and stations for the different means of transportation.

At point A, the user enters the desired destination (see Figure 6a). The user can either type (and auto-complete) a destination or select from a list (considering the transportation options due to the user's current location). After clicking the START button, a summary of the input, the next station and all subsequent maneuvers are listed (b).

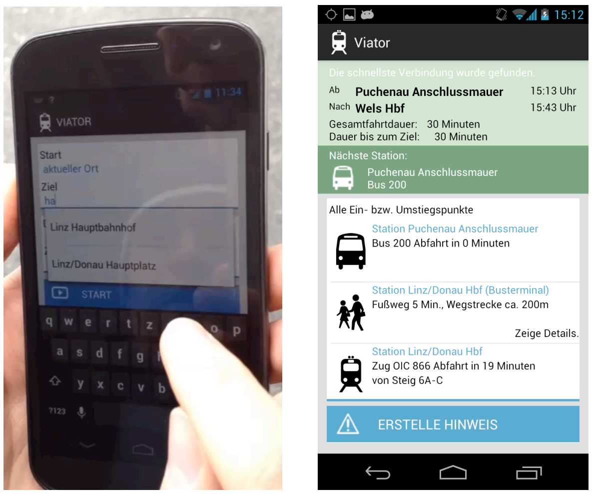


Figure 6. VIATOR UI: (a) Enter Destination, (b) Display Route

As the user interface in all figures is kept in German, here is a brief translation: The upper input field in Figure 6a contains the current location ("aktueller Ort") determined by the location sensors of the smartphone. The user inputs two letters for the destination ("Ziel") and is offered a list of options. The light green section in Figure 6b always displays a summary of the specified route. Departure ("Ab") and destination ("Nach") are listed in combination with departure and arrival times and a summary of the total travel time ("Gesamtfahrtdauer") and the remaining time ("Dauer bis zum Ziel"). The dark green section always shows the next maneuver task for the traveler (in this case, the user has to take a bus with number 200 at the given bus stop). Finally, the remaining white section lists all subsequent maneuver tasks including footways.

So far, VIATOR looks similar to existing travel information systems. The first difference is noticeable, though, when unforeseen events occur and the route changes (i.e., either means of transportation or the schedule). In these cases, VIATOR actively reacts, informs the traveler about the change and calculates new options without manual intervention. Figure 7a gives an example of a delay immediately announced via an acoustic signal and a red marquee text on the top. When the app is in foreground, the user is directly confronted with the change.

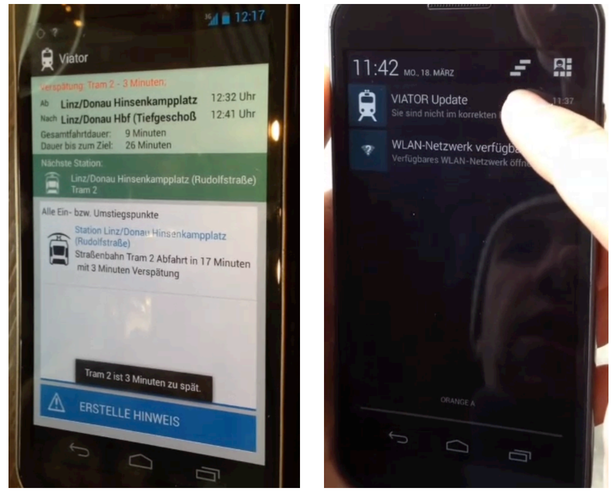


Figure 7. Alerts: (a) Detour Info, (c) Notification Bar Alert

In Figure 7a, the user is informed in the black bottom bar that the alternative vehicle chosen is also 3 minutes late. When the app is in background mode, VIATOR utilizes the Android notification bar for alerts and enables the user to get back to the app via a single tap (Figure 7b).

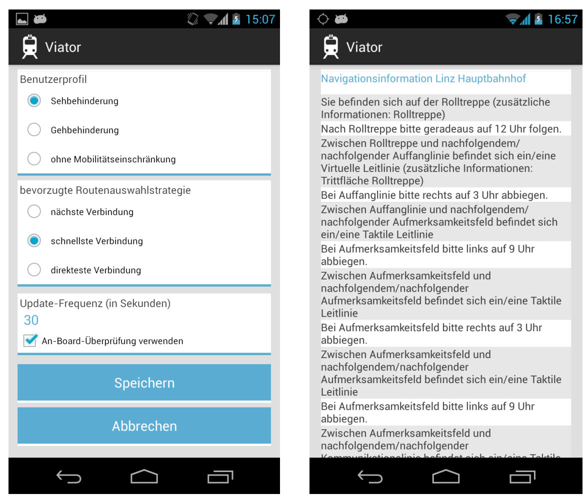


Figure 8. (a) Select Impairment, (b) Maneuver Instructions for Blind

In terms of barrier-free usage, VIATOR offers to select the type of impairment as part of the user profile. The prototype contains three options for blind and disabled persons and for persons without disabilities (see top of Figure 8a). The middle part of Figure 8a offers to select the preferred route calculation method: next, fastest, or direct. At the bottom, the user can select the update frequency – indirectly influencing battery consumption. Figure 8b shows a (pretended unstructured) flow of text representing navigation instructions for blind people. A visual format is unnecessary, though, because this text is meant to be read by screen readers and is only considered for the blind. Exemplary translation: "Nach der Rolltreppe bitte geradeaus auf 12 Uhr folgen" means: "After the escalator go straight ahead, direction 12 o'clock".

Figure 9a gives an impression of an editing tool enabling its users to create or update navigation instructions (self-organization). Users may provide a subject ("Name"), categorize the information ("Kategorie") and place it to a desired location ("Ort"). Subsequent travelers perceive this information automatically when they arrive at this location. The screenshots show a simple implementation of the principle of self-organization. Of course, voice memos or voice recognition would probably be the preferred technology.

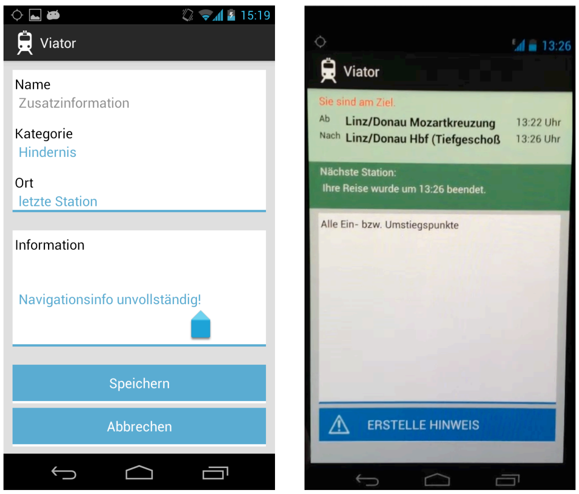


Figure 9. VIATOR UI. (a) Editing Tool, (c) "Get Off" Alert

When the traveler arrives at a maneuver point (either for transfer or at the final destination), VIATOR signalizes to get off ("Sie sind am Ziel", see Figure 9b). This feature distinguishes VIATOR from conventional navigation systems, like Google navigation, location-sensitively (and therefore at the exact maneuver time) telling the traveller what to do next without any manual interaction by its users.

V. CONCLUSION AND FUTURE WORK

The first subjective impressions (no systematic investigations were conducted so far) prove that the general idea of a location-based travel information system for disabled persons is working, however improvements due to experienced weak points are still to be incorporated. VIATOR (1) provides up-to-date intermodal travel information concerning connecting vehicles, schedules, and delays across different transportation companies, (2) automatically, without manual interaction, reroutes to alternative transportation options, (3) guides handicapped people due to their needs (e.g., maneuver instructions along tactile lines for blind people), and (4) provides an open interface for leaving and consuming self-created location-bound hints.

However, the tests have also revealed the first weak points of the system: As an example, the navigation instructions for blind people are in need of improvement (particularly when we consider different types of visual impairment and different walking behavior). We also miss a more accurate method for locating users inside buildings (referring to approaches of indoor navigation done by [5][6][7][8]). In a next step, the consortium focuses on the "bridge" between two connecting public vehicles for handicapped people, their personal preferences regarding navigation instructions (i.e., turn-by-turn vs. "environmental" instructions, etc.), investigating the potentials of categorizing different profiles for the same target group for more adequate support.

REFERENCES

[1] Ways2go, "Innovation and Technology for Evolving Mobility Needs," Funding Programme of the Austrian Federal Ministry for Transport, Innovation and Technology, www.bmvit.gv.at/en/innovation/mobility/ways2go/, retrieved: May, 2013.

[2] 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, 2011, pp. 355-377.

[3] W. Narzt and G. Pomberger, "Digital Graffiti – a Smart Information and Collaboration System," International Conference on Electronic Engineering and Computer Science, EECS 2013, Elsevier IERI Procedia (ISSN: 2212-6678), Beijing, China, 2013, in press.

[4] W. Narzt and 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), Essen, Germany, 2011, pp. 1239-1243.

[5] W. Bischof, E. Krajnc, M. Dornhofer, and M. Ulm, "NAVCOM – WLAN Communication between Public Transport Vehicles and Smart Phones to Support Visually Impaired and Blind People," 13th Intl. Conference, ICCHP 2012, Austria, July 11-13, 2012, Proceedings, Part II, 2012, pp 91-98.

[6] M. Kiers, W. Bischof, E. Krajnc, and M. Dornhofer, "ways4all - Indoor Navigation without absolute Coordinates for the Visually Impaired and Blind People," FHK Conference Vienna, 2011.

[7] M. Kiers, W. Bischof, M. Dornhofer, and E. Krajnc, "A New Approach for an RFID Indoor Positioning System Without Fixed Coordinates for Visually Impaired and Blind People," IPIN2010 Conference, Zürich, 2010.

[8] K. Ito, Y. Fujimoto, J. Akita, R. Otsuki, A. Masatani, T. Komatsu, M. Okamoto, and T. Ono, "Development of the Future Body-Finger: A Novel Travel aid for the Blind," 2nd Intl. Conf. on Ambient Computing, Applications, Services and Technologies, AMBIENT 2012, Spain, 2012, pp. 60-63.

[9] MoViH, "Mobility for the Visually and Hearing Impaired in Public Transport," Hilfsgemeinschaft der Blinden und Sehschwachen Österreichs, Funded ways2go project, 2011, http://www.wu.ac.at/prodman/about/research/en/projects/trade/mobility_for_the_visually, retrieved: July 2013.

[10] BIS, "Barrier Information System," Funded ways2go project, 2012, http://www2.ffg.at/verkehr/projekte.php?id=844&lang=en&browse=programm, retrieved: July 2013.

[11] J. Silva, C. Silva, L. Marcelino, R. Ferreira, and A. Pereira, "Assistive Mobile Software for Public Transportation," Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, UBICOMM 2011, Lisbon Portual, November 2011, pp. 309-313.

[12] W. Ait-Cheik-Bihi, M. Bakhouya, A. Nait-Sidi-Moh, J. Gaber, and M. Wack, "A Platform for Interactive Location-Based Services," Procedia Computer Science, Volume 5, 2011, pp. 697-704.

[13] W. Narzt and H. Schmitzberger, "Location-triggered code execution - dismissing displays and keypads for mobile interaction," UAHCI '09: 5th Intl. Conf. on Universal Access in HCI. Part II. Berlin, Heidelberg: Springer, 2009, pp. 374–383.

Analytical Modeling of Partially Shared Caches in Embedded CMPs

Wei Zang and Ann Gordon-Ross*
Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, 32611, USA
weizang@ufl.edu & ann@ece.ufl.edu
*Also with the NSF Center for High-Performance Reconfigurable Computing (CHREC) at the University of Florida

***Abstract*—In modern ubiquitous devices, optimizing shared last-level caches (LLCs) in embedded chip multi-processor systems (CMPs) is critical due to the increased contention for limited cache space from multiple cores. We propose cache partitioning with partial sharing (CaPPS) to reduce LLC contention and improve utilization. CaPPS can reduce the average LLC miss rate by 25% and 17% as compared to baseline configurations and private partitioning, respectively. To facilitate fast design space exploration, we develop an analytical model to quickly estimate the miss rates of all CaPPS configurations with an average error of only 0.73% and with an average speedup of 3,966X as compared to a cycle-accurate simulator.**

Keywords—cache partitioning; analytical modeling

I. Introduction

Many chip multi-processor systems (CMPs) leverage shared last-level caches (LLCs) (e.g., second-/third-level), such as the ARM Cortex-A, Intel Xeon, and Sun T2 [1][10][11]. To improve cache utilization, LLCs should be large enough to accommodate all sharing cores' data, but long access latencies and high power consumption typically precludes large LLCs from embedded systems with strict area/energy/power constraints. Since battery-operated devices (e.g., cell phones, tablets, laptops, etc.) have limited energy reserves and satisfying the applications' quality-of-services (QoSs) is typically required, optimizing small LLCs' performance is significantly more challenging due to contention for limited cache space.

Shared LLCs afford high cache utilization and no coherence overhead, however, high contention and unfair cache utilization degrades performance. A core's LLC occupancy (utilized space) is flexible and dictated by the core's application's demands. Cores with high LLC requirements occupy a large LLC area and cause high, potentially unfair, contention. For example, streaming multimedia applications occupy the LLC with a large amount of single-accessed data and unfairly evict the other cores' data, thus increasing LLC miss rates. For example, this unfair cache utilization is common in mobile systems when a local music/movie player and other web-service applications are co-executed.

To eliminate shared LLC contention, cache partitioning [5][15][18] partitions the cache, allocates *quotas* (a subset of partitions) to the cores, and optionally configures the partitions/quotas (e.g., size and/or associativity [15][18]) to the allocated core's requirements. Each core's cache occupancy is constrained to the core's quota to ensure fair utilization. *Set partitioning* partitions and allocates quotas at the cache set granularity and is typically implemented using operating system (OS)-based page coloring [12]. However, due to this OS modification requirement, hardware-based *way partitioning* is more widely used. Way partitioning partitions and allocates quotas at the cache way granularity [15][18]. However, way partitioning for shared LLCs typically uses *private partitioning*, which restricts quotas for exclusive use by the allocated core only and can lead to poor cache utilization if a core does not occupy the core's entire allocated quota.

In this paper, we propose to improve way partitioning's cache utilization using cache partitioning with partial sharing (CaPPS). CaPPS improves cache utilization via *sharing configuration*, which enables a core's quota to be configured as private, partially shared with a subset of cores, or fully shared with all other cores. Whereas sharing configuration increases the design space and thus increases optimization potential, this large design space significantly increases design space exploration time. To facilitate design space exploration, we develop an offline analytical model to quickly estimate cache miss rates for all partitioning and sharing configurations, which enables determining LLC configurations for any optimization that evaluates cache miss rates (e.g., performance, energy, energy delay product, power, etc.). The analytical model probabilistically predicts the miss rates when multiple applications are co-executing using the *isolated cache access distribution* for each application (i.e., the application is run in isolation with no co-executing applications). Although several previous works [3][4][6] have developed analytical models to predict shared LLC contention offline, these works' caches where completely shared by all cores and did not consider partial sharing, which vastly increases the design space and thus optimization potential. Due to CaPPS's extensive design space, experiments reveal that CaPPS can reduce the average LLC miss rates by as much as 25% and 17% as compared to baseline configurations and private partitioning, respectively. The analytical model estimates cache miss rates with an average error of only 0.73% and is 3,966X faster on average than a cycle-accurate simulator.

II. Related Work

Since CaPPS uses way partitioning and we developed an analytical model to predict the shared ways' cache contention, we compare our work with prior work in these areas.

For way partitioning, Qureshi and Patt [15] developed utility-based cache partitioning (UCP) that used an online monitor to track the cache misses for all possible numbers of ways assigned to each core. Greedy and refined heuristics determined the cores' quotas. Varadarajan et al. [18] partitioned the cache into small direct-mapped cache units, which were privately assigned to the cores and the cache partitions had configurable size, block size, and associativity. Kim et al. [13] developed static and dynamic cache partitioning for fairness optimization. Static cache partitioning used the cache access's stack distance profile to determine the cores' requirements. Dynamic cache partitioning increased/decreased the cores' quotas in accordance with the miss rate changes between evaluation intervals. Private LLCs also benefit from way partitioning. In CloudCache [14], the private caches were partitioned, but a core could share nearby cores' (limited access latencies) private caches. MorphCache [17] partitioned the level two and level three caches and allowed subsets of cores' private caches to be merged and fully shared by the subset. Although some of these prior works in private LLC partitioning [14][17] enabled a core to share other cores' quotas, CaPPS is more flexible than these works by enabling a *portion/all* of a core's quota to be shared with *any* subset of cores.

Prior works on analytical modeling to determine cache miss rates targeted only fully shared caches. Chandra et al. [3] proposed a model using access traces for isolated threads to predict inter-thread contention for a shared cache. Reuse distance profiles were analyzed to predict the extra cache misses for each thread due to cache sharing, but the model did not consider the interaction between cycles per instruction (CPI) variations and cache contention. Eklov et al. [6] proposed a simpler model that calculated the CPI considering the cache misses caused by contention by predicting the reuse distance distribution of an application when co-executed with other applications based on the isolated reuse distance distribution of each application. Chen and Aamodt [4] proposed a Markov model to estimate the cache miss rates for multi-threaded applications with inter-thread communication.

Analytically predicting the cache miss rate for CaPPS is more challenging than prior works, since in CaPPS, only the interleaved LLC accesses of other cores that pollute the partially shared ways affect the core's miss rate. Determining the effects of these interleaved accesses on the miss rate introduces extensive complexity.

III. CACHE PARTITIONING WITH PARTIAL SHARING

To accommodate the LLC requirements for multiple applications co-executing on different cores, CaPPS partitions the shared LLC at the way granularity and leverages sharing configuration to allocate the partitions to each core's quota. To facilitate fast design space exploration, an analytical model estimates the cache miss rates for the CaPPS configurations using the applications' isolated LLC access traces. We assume that the cores execute different applications in independent address spaces, thus there is no shared instruction/data address or coherence management, which is a common case in mobile systems running disparate applications and is similar to assumptions made in prior works [3][6].

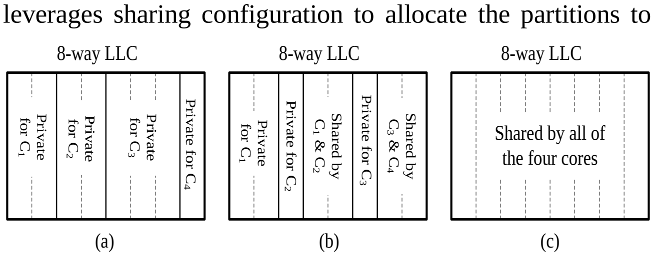


Figure 1. Three sample sharing configurations: (a) the cores' quotas are private; (b) some ways are partially shared with a subset of cores; and (c) the entire LLC is fully shared with all cores.

A. Architecture and Sharing Configurations

CaPPS's sharing configurations enable a core's quota to be configured as private, partially shared with a subset of cores, or fully shared with all other cores. Fig. 1 (a)-(c) illustrates sample configurations, respectively, for a 4-core CMP (C_1 to C_4) and an 8-way LLC: (a) each core's quota has a configurable number of private ways; (b) the cores' quotas are partially shared with subsets of cores; and (c) all of the four cores fully share all of the ways.

CaPPS uses the least recently used (LRU) replacement policy, but we note that the analytical model can be extended to approximate estimations for other replacement policies, such as pseudo-LRU. To reduce the sharing configurability with no effect on cache performance and to minimize contention, cores share an arbitrary number of ways starting with the LRU way, then second LRU way, and so on since these ways are least likely to be accessed. For example, in Fig. 1 (b) two of C_1's ways are shared with $C_{2.}$, therefore, C_1's two most recently used (MRU) blocks are cached in C_1's two private ways, and the two LRU blocks are cached in the two ways shared with C_2 and these two LRU blocks are the only replacement candidates for C_2's accesses. Maintaining this LRU ordering and determining replacement candidate can be easily implemented using a linked list or systolic array implementation [7] for conventional LRU caches with the integration of column caching [5] to achieve low hardware overhead and without increasing the cache access time. Since the hardware implementation is straightforward and is not the focus of this paper, we omit the implementation details for brevity.

B. Analytical Modeling Overview

For applications with fully/partially shared ways, the analytical model probabilistically determines the miss rates using the isolated cache access distributions for the co-executing applications. These distributions are recorded during *isolated access trace processing*. The isolated LLC access traces can be generated with a simulator/profiler by running each application in isolation on a single core with all other cores idle. For applications with only private ways, there is no cache contention and the miss rate can be directly determined from the isolated LLC access trace distribution.

Fig. 2 illustrates the contention in the shared ways using sample time-ordered isolated (C_1, C_2) and interleaved/co-executed (C_1&C_2) access traces to an arbitrary cache set from cores C_1 and C_2. C_1's and C_2's accesses are denoted as X_i and Y_i, respectively, where i differentiates accesses to unique cache blocks. The first access to X_3 and the second access to X_1 occurred at times t_1 and t_2, respectively. C_1's

 ISBN: 978-1-61208-289-9

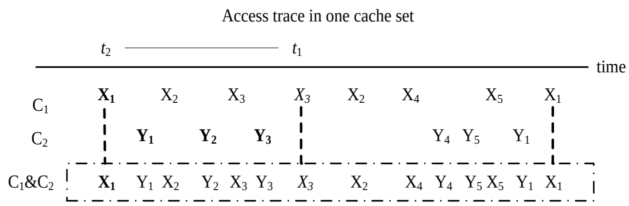


Figure 2. Two cores' isolated (C_1, C_2) and interleaved (C_1&C_2) access traces for an arbitrary cache set.

second access to X_1 will be a cache hit if C_1's number of private ways is greater than or equal to five because four unique blocks are accessed between the two accesses to X_1. Alternatively, if C_1's number of private ways is smaller than five and C_1 shares ways with C_2, X_1's hit/miss is dictated by the interleaved accesses from C_2. For example, if C_1 has six allocated ways and two of the LRU ways are shared with C_2, X_3 evicts X_1 from C_1's private way into a shared way. Therefore, C_2's accesses between t_1 and t_2 dictates whether X_1 has been evicted from the cache or not. If C_2's accesses between t_1 and t_2 evict two or more blocks into the shared ways, X_1's second access will be a cache miss.

In order to determine the contention effects to C_1's miss rate, C_1 and C_2's number of accesses n_1 (Section III.D.a) and n_2 (Section III.D.b), respectively, during the time period (t_1, t_2) must be estimated. Since the number of blocks R from n_2 evicted into the shared ways dictates whether C_1's blocks (e.g., X_1 in Fig. 4) are still in the shared ways, we calculate the probability $p(n_2, R)$ that R number of blocks are evicted into the shared ways (Section III.D.c) to estimate C_1's miss rate (Section III.D.d).

C. *Isolated Access Trace Processing*

To accumulate the isolated cache access distribution, we record the *reuse distance* and *stack distance* for each access in the isolated LLC access trace, which can be obtained using a stack-based trace-driven simulator [9]. For an accessed address T that maps to a cache set, the reuse distance r is the number of accesses to that set between this access to T and the previous access to any address in the same block as T. The stack distance d is the number of unique block addresses, or *conflicts*, in this set of accesses. For example, in Fig. 2, C_1's second access to X_1 has $r = 7$ and $d = 4$.

In each cache set, we accumulate the number of accesses N_d for each stack distance d ($d \in [0, A]$), where A is the LLC associativity. We accumulate the number of accesses with $d > A$ in N_A together with the number of accesses with $d = A$, since all accesses with $d \geq A$ are cache misses in any configuration. Given this information, for any access, the probabilistic information for the access' stack distance is $p(d < d_i) = (\sum_{d=0}^{d=d_i-1} N_d)/(\sum N_d)$ and $p(d \geq d_i) = 1 - p(d < d_i)$, ($\forall d_i \in [1, A]$). For all of the accesses for each d, we accumulate a histogram of different r and calculate the average $\bar{r}$ over all r.

The analytical model uses the base (best case) CPU cycles $Cycles_{base}$ to calculate the CPU cycles required to complete the application when co-executed with other applications. $Cycles_{base}$ assumes that all LLC accesses are hits. An application's total number of CPU cycles $Cycles_{exe}$ are recorded in the isolated execution to calculate $Cycles_{base}$ using $Cycles_{base} = Cycles_{exe} - m_{exe} \cdot LLC_{latency}$, where m_{exe} is the number of LLC misses in the application's isolated execution and $LLC_{latency}$ is the delay cycles incurred by an LLC miss.

Since the access distributions across the cache sets are different, the distributions are individually accumulated and recorded for each set to estimate the number of misses in each set's accesses. Since the analysis is the same for all cache sets, we present the analytical model for one arbitrary cache set.

D. *Analysis of the Shared Ways' Contention*

First, we describe the analytical model to analyze the shared ways' contention for a sample CMP with two cores C_1 and C_2 and then generalize the analytical model to any number of cores. A sharing configuration allocates K_{C_1} number of ways to core C_1, where K_{P,C_1} ways are private and the remaining K_S ($K_S = K_{C_1} - K_{P,C_1}$) ways are shared with core C_2. K_{C_2} and K_{P,C_2} similarly denote these values for C_2. For C_1, all accesses with a stack distance $d \leq K_{P,C_1} - 1$ result in cache hits and all accesses with $d \geq K_{C_1}$ are cache misses. The cache hit/miss determination of the accesses where $K_{P,C_1} \leq d \leq K_{C_1} - 1$ depends on the interleaved accesses from C_2, and the following subsections elaborate on the estimation method for these accesses. If C_1 only has private ways, then $K_{P,C_1} = K_{C_1}$, and these estimations are not required since the number of misses for C_1 can be directly calculated using $\sum_{d=0}^{d=K_{C_1}-1} N_{d,C_1}$.

a. *Calculation of* n_1

For an arbitrary stack distance D in $[K_{P,C_1}, K_{C_1} - 1]$, the associated $\bar{r}$ was determined during isolated access trace processing. This subsection presents the calculation of n_1 for C_1's accesses with stack distance D based on $\bar{r}$.

Fig. 3 depicts C_1's isolated access trace to an arbitrary cache set, where the second access to X_1 has a stack distance D and reuse distance $\bar{r}$. X_3's access evicts X_1 from C_1's private ways, therefore, the numbers of conflicts before and after X_3 are $(K_{P,C_1} - 1)$ and $(D - (K_{P,C_1} - 1))$, respectively. $Conf_i$ denotes the first access of the *i*-th conflict with X_1. We denote the number of accesses before X_3 as n_0, which can be any integer in $[K_{P,C_1} - 1, \bar{r} - (D - K_{P,C_1}) - 2]$. After

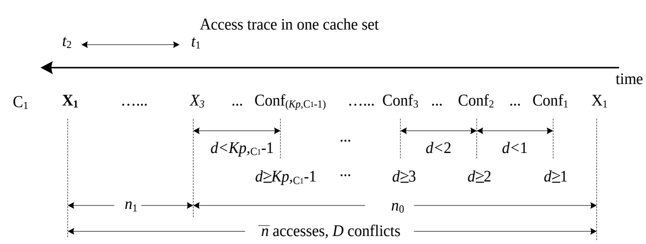


Figure 3. C_1's isolated access trace to an arbitrary cache set for calculating $\boldsymbol{n_1}$.

determining the probability $p(n_0,(K_{P,C_1}-1))$ for each n_0 (where $K_{P,C_1}-1$ indicates the number of conflicts in the n_0 accesses), we can calculate n_0's expected value $\bar{n}_0$ for the evaluated configuration's associated K_{P,C_1} using $\bar{n}_0 = \sum(n_0 \cdot p(n_0,(K_{P,C_1}-1)))$, and n_1's expected value is: $\bar{n}_1 = \bar{r} - \bar{n}_0 - 1$.

For a particular $n_0 \in [K_{P,C_1}-1, \bar{r}-(D-K_{P,C_1})-2]$, the probability is:

$$p\left(n_0,(K_{P,C_1}-1)\right) = p(E_A,E_B|E_C) = \frac{p_{before}(E_A)\cdot p_{after}(E_B)}{p_{total}(E_C)} \quad (1)$$

where E_A is the event that the n_0 accesses have exactly $(K_{P,C_1}-1)$ conflicts and E_B is the event that the n_1 accesses have exactly $(D-(K_{P,C_1}-1))$ conflicts. $p_{before}(E_A)$ and $p_{after}(E_B)$ are the occurrence probabilities of E_A and E_B, respectively. E_C is the event that the $\bar{r}$ accesses have exactly D conflicts and $p_{total}(E_C)$ is the probability of E_C's occurrence, which is the summation of $(p_{after}(E_B)\cdot p_{before}(E_A))$ for all n_0. To calculate $p_{before}(E_A)$ and $p_{after}(E_B)$, we examine the sufficient conditions that E_A and E_B occur. In Fig. 3, the first access following X_1 must be different from X_1 (for $D>0$), which is $Conf_1$ satisfying $d \geq 1$, since $Conf_1$ has at least one conflict: X_1. The second conflict $Conf_2$ satisfies $d \geq 2$, since $Conf_2$ has at least two conflicts: $Conf_1$ and X_1. The accesses between $Conf_1$ and $Conf_2$ satisfy $d<1$ since these accesses can only be $Conf_1$. $Conf_3$ satisfies $d \geq 3$ since $Conf_3$ has at least three conflicts: $Conf_2$, $Conf_1$, and X_1. The accesses between $Conf_2$ and $Conf_3$ satisfy $d<2$, since these conflicts can only be $Conf_2$ or $Conf_1$, etc. Similarly, $Conf_{Kp,C1-1}$ satisfies $d \geq (K_{P,C_1}-1)$ and the accesses between X_3 and $Conf_{Kp,C1-1}$ satisfy $d < (K_{P,C_1}-1)$. Thus, defining $\vec{a}=(a_1,a_2,\ldots,a_{K_{p,C_1}-1})$ where $a_i \in [0, n_0-(K_{P,C_1}-1)]$, $p_{before}(E_A)$ is:

$$p_{before}(E_A) = \left\{\prod_{i=1}^{i=K_{P,C_1}-1} p(d\geq i)\right\}\cdot\left\{\sum_{\forall\vec{a}\in S_a}\left(\prod_{i=1}^{i=K_{P,C_1}-1} p(d<i)^{a_i}\right)\right\} \quad (2)$$

where S_a is a set including all $\vec{a}$ satisfying $\sum a_i = n_0 - (K_{P,C_1}-1)$. Similarly, defining $\vec{b}=(b_0,b_1,\ldots,b_{D-K_{p,C_1}})$ where $b_i \in [0, n_1-(D-K_{P,C_1}+1)]$, $p_{after}(E_B)$ is:

$$p_{after}(E_B) = \left\{\prod_{i=0}^{i=D-K_{P,C_1}} p\left(d\geq i+K_{P,C_1}\right)\right\} \cdot\left\{\sum_{\forall\vec{b}\in S_b}\left(\prod_{i=0}^{i=D-K_{P,C_1}} p\left(d<i+K_{P,C_1}\right)^{b_i}\right)\right\} \quad (3)$$

where S_b is a set including all $\vec{b}$ satisfying $\sum b_i = n_1 - (D - K_{P,C_1}+1)$.

b. Calculation of n_2

To determine the contention effect from C_2, the expected number of accesses $\bar{n}_2$ from C_2 is estimated based on the ratio of the number of cache set accesses from C_1 and C_2 per cycle:

$$\frac{\bar{n}_1}{\bar{n}_2} = \frac{\sum N_{d,C_1}/\widehat{Cycles}_{C_1}}{\sum N_{d,C_2}/\widehat{Cycles}_{C_2}} \quad (4)$$

where $\sum N_{d,C_1}$ and $\sum N_{d,C_2}$ are the total number of LLC accesses from C_1 and C_2, respectively. $\widehat{Cycles}_{C_1}$ is the number of CPU cycles required to execute the application on C_1 when C_2 is co-executing another application, and $\widehat{Cycles}_{C_2}$ is similarly defined. $\widehat{Cycles}_{C_i}$ can be calculated using $Cycles_{base}$ and the number of LLC misses $\hat{m}$ estimated with the contention:

$$\widehat{Cycles}_{C_i} = Cycles_{base} + \hat{m}\cdot LLC_{latency} + delay_{bus\_contention} \quad (5)$$

where $delay_{bus\_contention}$ is the delay imposed by the shared bus contention from the higher level caches (closer to the CPU) of each core to the shared LLC. $delay_{bus\_contention}$ is derived by calculating the *bus contention probability* that another core is sending a read/write request to the LLC and the LLC is returning that core's requested block simultaneously with the evaluated core's bus request. The bus contention probability is dictated by each core's bus request probability, which is equal to the total number of bus requests generated from the core's higher level cache misses divided by $\widehat{Cycles}_{C_i}$.

c. Calculation of $p(n_2,R)$

$p(n_2,R)$ is the probability that R number of blocks are evicted from C_2's private ways in the n_2 accesses. Directly using the expected n_2 to calculate $p(\bar{n}_2, R)$ will introduce a large bias (approximate 10% error) in the estimated LLC miss rate, since different values of n_2 result in different hit/miss determinations and using one expected value $\bar{n}_2$ will estimate all n_2 as hits/misses. Thus, we model n_2 using a Poisson distribution $p(n_2)=Poisson(n_2,\lambda)$, where λ is $\bar{n}_2$ if the LLC is accessed randomly. However, since the LLC's accesses are generally not random and not uniformly distributed in time (which makes (4) valid), we use an empirical variable e to adjust λ to $\lambda=\bar{n}_2/e$. Our experiments indicated that $e=5$ was appropriate for our training benchmark suite, which contains a wide variety of typical CMP applications, and is thus generally applicable. Since the range of n_2 is infinite in the Poisson distribution, and n_2 with very small $p(n_2)$ has minimal effect on the miss rate estimation, we only consider the n_2 with $p(n_2)>0.01$ and calculate the associated $p(n_2,R)$.

To calculate $p(n_2,R)$ for an arbitrary n_2, R is determined by evaluating the n_2 accesses in chronological order with an initial value of $R=0$. If there is one access with $d > K_{P,C_2} + current\ R$, fetching this address into C_2's private ways will evict one block into the shared ways and thus R is incremented by 1. Therefore, we can calculate $p(n_2,R)$ inductively:

$$p(n_2,R) = \begin{cases} p(n_2-1,R-1)\cdot p\left(d\geq K_{P,C_2}+(R-1)\right), & R=n_2 \\ p(n_2-1,R)\cdot p\left(d<K_{P,C_2}+R\right) \\ \quad +p(n_2-1,R-1)\cdot p\left(d\geq K_{P,C_2}+(R-1)\right), & R<n_2 \\ p(n_2-1,R)\cdot p\left(d<K_{P,C_2}+R\right), & R=0 \end{cases}$$

with the initial case $p(n_2=0,R=0)=1$.

d. Calculation of the LLC Miss Rates

Considering the impact of R to the accesses with stack distance $d \in [K_{P,C_1}, K_{C_1}-1]$, the number of cache hits for C_1 is:

$$\widehat{h_{C_1}} = \sum_{d=0}^{d=K_{p,C_1}-1} N_{d,C_1} + \sum_{d=K_{p,C_1}}^{d=K_{C_1}-1} \left(N_{d,C_1} \cdot \sum_{\forall n_2: p(n_2)>0.01} \left(\left(\sum_{R=0}^{R=K_{C_1}-d-1} p(n_2,R) \right) \cdot p(n_2) \right) \right) \quad (7)$$

After accumulating $\widehat{h_{C_1}}$ for all cache sets, the number of LLC misses $\widehat{m_{C_1}}$ and the LLC miss rates can be determined.

Finally, we generalize the analytical model to estimate the LLC miss rate for any core C_i when j additional cores (denoted as C_j) share cache ways with C_i by calculating the expected number of accesses $\bar{n}_{C_j}$ from the additional cores during the time (t_1, t_2) and then estimating $p(n_{C_j}, R_{C_j})$ similarly as estimating $\bar{n}_2$ and $p(n_2, R)$ for C_2. The generalized expression of (6) is:

$$\widehat{h_{C_i}} = \sum_{d=0}^{d=K_{p,C_i}-1} N_{d,C_i} + \sum_{d=K_{p,C_i}}^{d=K_{C_i}-1} \left(N_{d,C_i} \cdot ph \right) \quad (8)$$

where:

$$ph = \sum_{\forall \vec{C} \in S_C} \left(\prod_{C_j \in \vec{C}} \left(p(n_{C_j}) \cdot p(n_{C_j}, R_{C_j}) \right) \right) \quad (9)$$

where $\vec{C} = (n_{C_1}, n_{C_2}, \ldots, n_j)$ with $p(n_{C_j}) > 0.01$ and S_C is a set including all $\vec{C}$ satisfying $\sum R_{C_j} \leq K_{C_i} - d - 1$.

According to (5), a circular dependency exists where $\widehat{Cycles}$ is used to estimate $\hat{m}$ and $\hat{m}$ is used to calculate $\widehat{Cycles}$. The solution cannot be represented using a closed form, thus we iteratively solve for $\hat{m}$. The initial value of $\hat{m}$ is acquired assuming there is no contention (i.e., all K_{C_i} number of ways are privately used by C_i), and $\hat{m}$ is used in (5) to calculate the initial value of $\widehat{Cycles}$. $\widehat{Cycles}$ is provided back into the analytical model to update $\hat{m}$ and the new $\hat{m}$ is used to update $\widehat{Cycles}$. This iterative process continues until a stable $\hat{m}$ (with a precision of 0.001%) is achieved. Experimental results indicated that only four iterations were required for the results to converge.

The analytical model's runtime complexity depends on the evaluated sharing configuration and the isolated cache access distribution for each application. Due to the large number of complex and interdependent variables and unknowns, the complexity of the model is intractable, thus in our experiments, we evaluate the analytical model's measured execution time.

TABLE I. CMP SYSTEM PARAMETERS

CPU	2 GHz clock, single thread
L1 instruction cache	Private, total size of 8 KB, block size of 64 B, 2-way associativity, LRU replacement, access latency of 2 CPU cycles
L1 data cache	Private, total size of 8 KB, block size of 64 B, 2-way associativity, LRU replacement, access latency of 2 CPU cycles
L2 unified cache	Shared, total size of 1 MB, block size of 64 B, 8-way associativity, LRU replacement, access latency of 20 CPU cycles, non-inclusive
Memory	Total size of 3 GB, access latency of 200 CPU cycles
L1 caches to L2 cache bus	Shared, width of 64 B, 1 GHz clock, first come first serve (FCFS) scheduling
Memory bus	Width of 64 B, 1 GHz clock

IV. EXPERIMENT RESULTS

We verified the advantages of CaPPS as compared to two baseline configurations and private partitioning. We also verified the accuracy of our estimated LLC miss rates obtained via the analytical model and evaluated the analytical model's ability to determine the optimal (minimum LLC miss rate) configuration in the CaPPS design space. Additionally, we illustrate the analytical model's efficiency by comparing the time required to calculate the LLC miss rates as compared to using a cycle-accurate simulator that generates the exact cache miss rates for all configurations.

A. *Experiment Setup*

We used twelve benchmarks from the SPEC CPU2006 suite [16], which were compiled to Alpha_OSF binaries and executed using "ref" input data sets. Due to incorrect execution, we could not evaluate the complete suite. Even though our work is targeted towards embedded systems, we did not use embedded system benchmark suites since these suites contain only small kernels, which do not sufficiently access the LLC, and do not represent our targeted embedded CMP domain. Since complete execution of the large SPEC benchmarks prohibits exhaustive examination of the entire CaPPS design space, and since most embedded benchmarks have stable behavior during execution, for each SPEC benchmark, we performed phase classification using SimPoint [8] to select 500 million consecutive instructions with similar behavior as the *simulation interval* to mimic an embedded application with high LLC occupancy.

We generated the exact cache miss rates for comparison purposes using gem5 [2] and modeled four in-order cores with the TimingSimple CPU model, which stalls the CPU when fetching from the caches and memory. Each core had private level-one (L1) instruction and data caches. The unified level-two (L2) cache and all lower level memory hierarchy components were shared among all cores. We modified the L2 cache replacement operation in gem5 to model CaPPS. TABLE I shows the parameters used for each system component. Since four cores shared the eight-way LLC (i.e., L2 cache), CaPPS's design space had 3,347 configurations.

Before CaPPS simulation, we executed each benchmark in isolation during the benchmark's simulation interval and recorded the isolated LLC access traces and the CPU cycles $Cycles_{exe}$. For CaPPS simulation, we arbitrarily selected four benchmarks to be co-executed, which formed a benchmark set, and we evaluated sixteen benchmark sets. Since the four benchmarks' simulation intervals were at different execution points, we forced the four cores to simultaneously begin executing at each benchmark's associated simulation interval's starting instruction using a full-system checkpoint. The full-system checkpoint was created by aggregating the *isolated-benchmark checkpoints*, which were generated by fast-forwarding the benchmark to the starting instruction of the benchmark's associated simulation interval when the benchmark was executed in isolation.

For each simulation, the system execution was terminated when any core reached 500 million instructions. Due to varying CPU stall cycles across the benchmarks, at the termination point, not all cores had completed executing the simulation interval. However, this termination approach guaranteed that the cache miss rates reflected a fully-loaded system (i.e., full LLC contention since all cores were running during the entire system execution). Since we focused on the cache miss rates rather than the absolute number of cache misses, the incomplete benchmarks' execution had no impact on the evaluation. Similarly, due to statistical predictions, the applications are not required to begin execution simultaneously to garner accurate results.

Although our experiments used only four cores and the LLC was a shared 8-way L2 cache, the analytical model itself does not include any limitations on the number of cores, the hierarchical level of the LLC, or the cache parameters (e.g., total size, block size, and associativity for our experiments).

B. *CaPPS Evaluation*

To validate the advantages of CaPPS, we compared CaPPS's ability to reduce the LLC miss rate as compared to two baseline configurations and private partitioning, since shared LLC partitioning in previous works [13][15][18] only provided private partitioning.

Fig. 4 depicts the average LLC miss rate reductions for CaPPS's optimal configurations (the configurations with minimum average LLC miss rate in CaPPS's design space) as compared to two baseline configurations: 1) *even-private-partitioning*: the LLC is evenly partitioned using private partitioning (first bar); and 2) *fully-shared*: the LLC is fully shared by all cores (second bar). Across all benchmark sets, the average and maximum average LLC miss rate reductions were 25.58% and 50.15%, respectively, as compared to *even-private-partitioning*, and 19.39% and 41.10%, respectively, as compared to *fully-shared*.

The third bar in Fig. 4 depicts the average LLC miss rate reductions for CaPPS's optimal configuration as compared to private partitioning's optimal configuration, which is the configuration with minimum LLC miss rate in the private partitioning's design space consisting of 35 configurations—approximately 1% of CaPPS's design space. Across all benchmark sets, the average and maximum reductions in CaPPS's average LLC miss rates as compared to private partitioning were 16.92% and 43.02%, respectively.

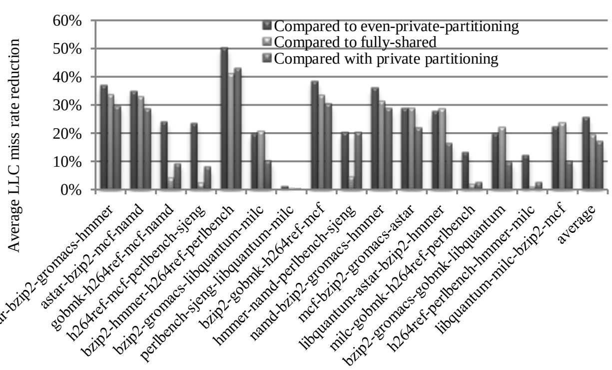


Figure 4. Average LLC miss rate reductions for CaPPS's optimal configurations compared to *even-private-partitioning*, *fully-shared*, and private partitioning.

C. *Analytical Model's Accuracy Evaluation*

For each benchmark set, we compared the average LLC miss rate for the four cores determined by the analytical model with the exact miss rate determined by gem5 for each configuration in CaPPS's design space. We calculated the average and standard deviation of the miss rate errors across the 3,347 configurations. Fig. 5 depicts the results for each benchmark set. The black markers indicate the average miss rate errors and the gray-shaded upper and lower ranges are the corresponding standard deviations. Averaged over all sixteen benchmark sets, the average miss rate error and standard deviation are -0.73% and 1.30%, respectively.

Since the analytical model's cache miss rates are inaccurate, we compared the absolute difference between the LLC miss rates of the analytical model's minimum LLC miss rate configuration and the actual minimum LLC miss rate configuration as determined via exhaustive search. Comparing with an exhaustive search is appropriate for evaluating the analytical model's efficacy, which is only affected by the estimated miss rate errors in determining the optimal configuration. The results indicate that fourteen out of sixteen benchmark sets' differences were less than 1% and the maximum and average differences over all benchmark sets was negligible, 1.3% and 0.36%, respectively.

D. *Analytical Model's Time Evaluation*

To evaluate the execution time efficiency of the analytical model, we compared the time required to estimate the LLC miss rates (including the time for isolated trace access generation) for all configurations in the CaPPS design space as compared to using gem5. We implemented the analytical model in C++ compiled with O3 optimizations. We tabulated the *user time* reported from the Linux *time* command for the simulations running on a Red Hat Linux Server v5.2 with a 2.66 GHz processor and 4 gigabytes of RAM. Fig. 6 depicts the speedup of the analytical model for each benchmark set as compared to gem5. Over all benchmark sets, the average speedup is 3,966X, with maximum and minimum speedups of 13,554X and 1,277X, respectively. For one benchmark set, the time for simulating all 3,347 configurations using gem5 was approximately three months, and comparatively, the analytical model took only

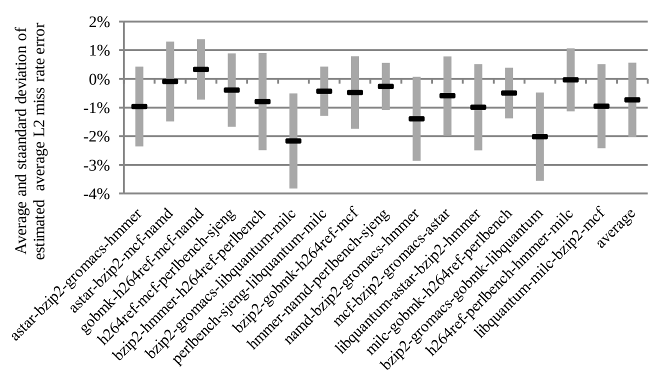


Figure 5. The average and standard deviation of the average LLC miss rate error determined by the analytical model.

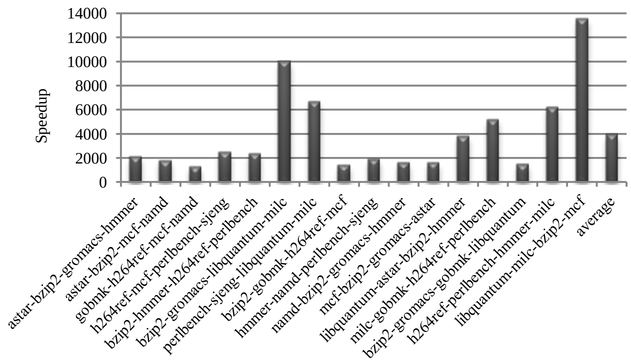


Figure 6. The analytical model’s simulation time speedup compared to gem5.

two to three hours.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented cache partitioning with partial sharing (CaPPS)—a novel cache partitioning and sharing architecture that improves shared last-level cache (LLC) performance with low hardware overhead for chip multi-processor systems (CMPs). Since CaPPS affords an extensive design space for increased optimization potential, CaPPS can reduce the average LLC miss rate by as much as 25% and 17% as compared to baseline configurations and private partitioning, respectively. To quickly estimate the miss rates of CaPPS’s sharing configurations, we developed an offline, analytical model that achieved an average miss rate estimation error of only 0.73%. As compared to exhaustive exploration (since no heuristics exist) of the CaPPS design space to determine the lowest energy cache configuration, the analytical model affords an average speedup of 3,966X. Finally, CaPPS and the analytical model are applicable to CMPs with any number of cores and place no limitations on the cache parameters.

Future work includes extending the analytical model to optimize for any design goal, such as performance or energy delay product, leveraging the offline analytical results to guide online scheduling for performance optimizations in real-time embedded systems, including accesses to shared address space, incorporating cache prefetching in our analytical model, and extending CaPPS to proximity-aware cache partitioning for caches with non-uniform accesses.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation (CNS-0953447). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] ARM Cortex-A Series, http://www.arm.com/products/processors/cortex-a/index.php.

[2] N. Binkert, et. al. The gem5 Simulator, http://gem5.org [retrieved: Feb., 2013].

[3] D. Chandra, F. Guo, S. Kim, and Y. Solihin, “Predicting Inter-Thread Cache Contention on a Chip Multi-Processor Architecture”, In Proceedings of HPCA, Feb. 2005, pp. 340-351.

[4] X. E. Chen and T. M. Aamodt, “A first-order fine-grained multithreaded throughput model”, In Proceedings of HPCA, Feb. 2009, pp. 329-340.

[5] D. Chiou, D. Chiouy, L. Rudolph, S. Devadas, and B. S. Ang, “Dynamic Cache Partitioning via Columnization”, Computation Structures Group Memo 430. M.I.T. 2000.

[6] D. Eklov, D. Black-schaffer, and E. Hagersten, “Fast Modeling of Shared Cache in Multicore Systems”, In Proceedings of HiPEAC, Jan. 2011, pp. 147-157.

[7] J. P. Grossman, “A Systolic Array for Implementing LRU Replacement,” Project Aries Technical Memo ARIES-TM-18, AI Lab, M.I.T., Cambridge, MA, 2002.

[8] G. Hamerly, E. Perelman, J. Lau, and B. Calder, “SimPoint 3.0: Faster and More Flexible Program Analysis”, Journal of Instruction-level Parallelism, 2005, pp. 1-28.

[9] M. D. Hill and A. J. Smith, “Evaluating Associativity in CPU Caches,” IEEE Trans. on Computers, Vol. 38, No. 12, 1989, pp. 1612-1630.

[10] Intel Core Duo Processor, http://ark.intel.com/products/family/22731.

[11] K. Johnson and M. Rathbone, “Sun’s Niagara Processor”, NYU Multicore Programming, 2010.

[12] R. E. Kessler and M. D. Hill, “Page Placement Algorithms for Large Real-indexed Caches”, ACM Trans. on Computer Systems, Vol. 10, No. 4, 1992, pp. 338-359.

[13] S. Kim, D. Chandra, and Y. Solihin, “Fair Cache Sharing and Partitioning in a Chip Multiprocessor Architecture”, In Proceedings of PACT, Sep.-Oct. 2004, pp. 111-122.

[14] H. Lee, S. Cho, and B. Childers, “CloudCache: Expanding and Shrinking Private Caches”, In Proceedings of HPCA, Feb. 2011, pp. 219-230.

[15] M. Qureshi and Y. Patt, “Utility-Based Cache Partitioning: A Low-Overhead, High-Performance, Runtime Mechanism to Partition Shared Caches”, In Proceedings of MICRO, Nov. 2006, pp. 423-432.

[16] SPEC CPU2006. http://www.spec.org/cpu2006 [retrieved: Sep., 2011].

[17] S. Srikantaiah, E., T. Zhang, M. Kandemir, M. Irwin, and Y. Xie, “MorphCache: a Reconfigurable Adaptive Multi-level Cache Hierarchy for CMPs”, In Proceedings of HPCA, Feb. 2011, pp. 231-242.

[18] K. Varadarajan, et al., “Molecular Caches: A Caching Structure for Dynamic Creation of Application-specific Heterogeneous Cache Regions”, In Proceedings of MICRO, Nov. 2006, pp. 433-442.

Challenges for e-Learning Environments in m-Learning Contexts

A survey about the hardware, software, and educational dimensions

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 applications that use the Web infra-structure to support teaching and learning activities. Their user interfaces were designed to have good usability using a desktop computer with keyboard and mouse as input devices and a high resolution medium-size display and speak loud as output devices. Devices, such as tablets and smartphones, have computational power enough to render Web pages, allowing mobile users navigate through the e-Learning environments and closing the e-Learning environments to the m-Learning contexts. But to have an effective use in the mobile context, the e-Learning environments need to be analyzed in three dimensions: hardware, software, and educational. This paper presents a survey about works in merging e-Learning environments and m-Learning and discusses about challenges such hardware and software integration, integration between the browser and the environment and changes on the teaching and learning activities.**

Keywords-Human-Computer Interaction; Electronic Learning Environment; Mobile Devices; Interaction Styles.

I. INTRODUCTION

e-Learning environments, such as Moodle [1], SAKAI [2], TelEduc [3], Ae [4], are applications that use the Web infra-structure to support teaching and learning activities. The e-Learning environments are designed to support a variety of users and learning contexts, but they are designed to conventional computers, usually equipped with keyboard and mouse as input and a medium screen and speakers as output; a limited interaction style for nowadays devices. These modalities and the technology shape the teaching and learning activities done in the e-Learning environments; they focus on reading and writing skills. Despites these technology impact, the e-Learning environments need to have good usability, accessibility, performance, security, availability and other software attributes.

Devices, such as smartphones and tablets, are becoming increasingly popular; most of them have touch screen displays, access to the Internet and enough computing power to process Web pages. So, Web sites and Web applications, initially developed to be used with keyboard, mouse and a medium size display, are being accessed by small touchscreen devices. This can be considered as another aspect of accessibility, so the environments´ development teams are building solutions to provide access on mobile devices, and the use of mobile devices with educational purposes is called m-Learning. Two motivations for m-Learning are to allow participants interact anytime and anywhere with the content and with each other.

The Internet was built to connect many technologies, so allowing mobile users to access e-Learning environments make the boundary between e-Learning and m-Learning not so clear. This brings some challenges for e-Learning environments developers and m-Learning practitioners. This paper presents a survey about the works to merge these fields. Section II presents a literature review about e-Learning, m-Learning and e-Learning environments. Section III presents the three dimensions to analyze the use of e-Learning environments to support mobile users. Section IV presents challenges to have an effective use of the e-Learning environments in the m-Learning context. Section V presents final considerations.

II. E-LEARNING AND M-LEARNING

e-Learning refers to the use of Information and Communication Technologies (ICT) in education. So, e-Learning is any distributed learning experience through the Internet, Intranet, Extranet, CD or DVD-ROM, because the e-Learning main concern is related with the teaching form and not with technology [5]. The e-Learning term refers broadly the terms Web-based learning, Internet-based learning, online learning, distributed learning, and computer-based learning. e-Learning is suited to distance learning, but it can also be used in conjunction with face-to-face teaching, calling blended learning.

Online systems that support e-Learning through the Web are called e-Learning environments or Virtual Learning Environments (VLE) or Learning Management Systems (LMS). Moodle [1], SAKAI [2], TelEduc [3] and Ae [4] are examples of e-Learning environments.

The first e-Learning environments are designed to integrate the content with communication tools arranged on the Internet (such as chat, mail and forum). They evolved increasing the number of tools for content publishing, participant´s communication and course administration, taking advantages of the Web to offer content with text, images, audios and videos in a hypertext document. Tools like chat, forums, portfolios, repositories are widely used, and tools those explore the audio and video resource to user communication, such as instant messenger and video-conferences, are becoming common among the environments.

 ISBN: 978-1-61208-289-9

The environments´ development teams are building solutions to provide access on mobile devices, and the use of mobile devices with educational purposes is called m-Learning.

m-Learning is any kind of learning that happens when the learner is not at a fixed, predetermined location, or learning that happens when the learner takes advantage of the learning opportunities offered by mobile technologies [6]. m-Learning Technology includes handheld computers, MP3 players, notebooks, mobile phones and tablets; devices that have owner autonomy and it is easy to portable. Two motivations for m-Learning is allow the participants interacting anytime and anywhere with the content and with each other.

The recent technology forwards brings a broad of new applications made to support many areas, such Education. One of the mobile devices that are gained repercussion in this scenario is the Tablet PC (Tablet Personal Computer), a device with height as similar to a notebook and had an input device similar a pen. The paper and pen metaphor implies that tasks performed before in paper, like draw or manuscript writing, can be more natural in the Tablet PC than in the another computing devices. Resuming, the Tablet PC has the following hardware characteristics: (i) Pen sensitive screen; (ii) Screen that allows different positions; (iii) Wireless network access by Wireless Local Area Network (WLAN) and bluetooth technology; (iv) Microphones and embedded loudspeakers; (v) Keyboard (some models the keyboard are detachable); (vi) Batteries. Usually, the screen size ranges from 9 inches to 12 inches.

Other devices used in m-Learning are smartphones, which have touchscreen with computation power enough to render Web pages. Usually, the screen size ranges from 3 inches to 5 inches. Resuming, the smartphones have the following hardware characteristics: (i) touch sensitive screen; (ii) Screen that allow two different positions (landscape and portrait); (iii) Wireless network access by WLAN (Wireless Local Area Network), by 3G and bluetooth technology; (iv) Microphones and embedded loudspeakers; (v) Virtual keyboard (some models have a physical keyboard); (vi) Batteries.

Gay *et al*. [7] suggest that the introduction of wireless computing resources in learning environments can potentially affect the development, maintenance, and transformation of learning communities.

It is possible to use e-Learning environments for support m-Learning and three kind of solution are emerging: i) specific device application; ii) web site specific for mobile devices; and iii) improve the web site for mobile and desktop access [8]. Each solution has its vantages and disadvantages due the hardware and software restrictions of the mobile device.

Building specific device application allows designing a suitable user interface for the device and taking advantages of smartphone´s features, such as touchscreen and camera, but needs develop an application for each mobile platform. So, the applications to be developed needs specific knowledge programming skills and increases the code lines number to maintain. Moodle community offers the Moodle App [9] and Moodbile [10], two native mobile applications with versions for the most popular smartphone´s platforms.

Moodle, since version 2.1, offers a Web site specific to mobile devices, an example for the second type of solutions for access e-Learning environments in mobile devices. Building a specific Web site to mobile device allows designing a suitable user interface for mobile devices taking account some common characteristics, such small touchscreen, but depends of the browser to access some platform features, such GPS, and increases the code lines number to maintain too.

The latter solution considers that smartphones and tablets have enough computational power to render Web pages and to do some adaptation if it is necessary, and offer the same user interface for any device. But the interaction styles may vary, so to design this kind of user interface it is necessary to do some usability studies to found barriers or user interaction problems for each interaction style. Disadvantages of this solution are to depend of browsers to use the mobile features and the difficult of consider many interaction styles in the same user interface. Da Silva, Freire and da Rocha [8] point out some problem that happen when a user interface designed to be used with specific interaction hardware is used with other interaction hardware.

Another e-Learning environment characteristic is to be used in many of learning contexts, e.g., teacher training, undergraduate courses, and team training in all areas of knowledge. We call this property as learnability. But, the actual hardware of conventional computers increases the difficulty to use the environment to produce content for any area and support student activities, e.g., to post a mathematic exercise that needs write formulas to resolve, the user need to use a specific software and post the file since the e-Learning environments do not support directly this kind of content.

The e-Learning environments need to be usable and accessible for many users in many social, physical, technological and learning contexts. So, e-Learning environments need to be evaluated in the usability, accessibility, mobility and learnability dimensions.

III. Dimensions for Allowing e-Learning and m-Learning Merging

Khan [11] defines eight dimensions for a good e-Learning environment: Resource Support, Ethical, Institutional, Pedagogical, Technological, Interface Design, Evaluation and Management. Since in our work we want to use e-Learning environments in m-Learning contexts, the Technological, Interface Design, and Pedagogical dimensions need to be discussed due the impact caused by the device changing. We will join the Technological and Interface Design dimensions, since there is an intrinsic relation between them, and our focus is on mobile users. The Pedagogical dimension is related about the teaching and learning activities developed by the teachers and done by the students.

For technological dimension, we propose the e-learning environments needs to be evaluated about their usability, accessibility, mobility, and other software requirements, such performance, scalability and availability. Due the diversity of users whom may use the e-Learning environments, these systems need to have good usability so that the user interface does not prejudice the teaching and learning activities, so the user interface cannot be a barrier between the student and the content or the users and their goals or injure the course activities.

Nielsen [12] defines usability as a combination of five elements: easy to learning, efficient, easy to remember, low probability of users do mistakes and user satisfaction. Nielsen proposes a method for evaluating the user interface usability, and other methods can be found on the Human-Computer Interaction literature [13] [14].

Accessibility, another important requirement, is about to allow disabled people use the environment, and can be understood as to be accessed by anyone, whatever their hardware or software. So to have a high level of accessibility, a Web page needs to be accessed by computers or any other computational device, such the smartphones and Tablet PCs, which can have assistive technology installed. Some methods can be found to evaluate the application accessibility, mainly for Web pages and applications, e.g., evaluating the interface conformance with the W3C guidelines [15].

In the educational dimension, the teacher needs to dispose content, to plain activities, to ask questions of the students about the activities and the content, to evaluate activities and other tasks. Pedagogically, the teacher needs to select the tools to be used to dispose the content and to run the activities. Usually, due the e-Learning environment tools and the hardware used to interact, a desktop computer, the activities are discussions by Forums, writing individual or collaborative texts. The texts may have pictures, but it is similar as works done in presence education, where the work can be delivered in a printed format. This kind of work does not take advantages of multimedia and the hardware available in the mobile devices. We discuss these problems in the next sections as challenges for e-Learning environments in the mobile contexts.

IV. CHALLENGES FOR E-LEARNING ENVIRONMENTS IN THE M-LEARNING

About the technology issues, it is need to analyze the hardware and the software and, since there are a variety of computing devices, aspects of human-machine interaction (such as ergonomics) and the device characteristics need to be considered. The most visible problem in browsing e-Learning environments using mobile device is the user interface adaptation. One of the adaptation issues is the content readability. For tablets it is not a big problem, since it is possible render the web page without injure de readability; the size of the tablet´s screen is not so small compared to the desktop display. But, in the case of smartphones, the screen size is not large enough to ensure readability, so it is necessary to have a page adaptation or techniques to visualize the entire page and use zooming to see the page details.

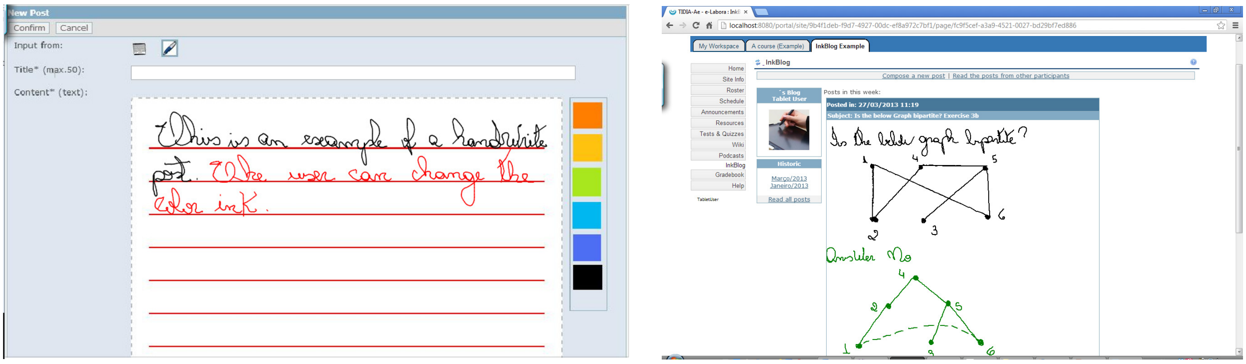
The user interface adaption is not a trivial task, a challenge not only for e-Learning environments, but for the Human-Computer Interaction, is to build system with adaptive user interface. Bickmore and Schilit [16] present a heuristic-based approach for Web page adaptations to be rendered in mobile devices with small screens, but their work does not consider audio and video adaptation. Zhang [17] purposes a framework to do content adaptation for systems accessed by a sort of devices (multidevices), increasing accessibility and doing a distribution optimization over the network. Oliveira and da Rocha [18] purpose priorities for consistence in the adaptation to maintain the same conceptual model for mutidevices, whom done a study case over an e-learning environment. Pyla *et al.* [19] discuss about the task migration between devices: the user starts the task in one device, go to another one to perform some sub-task, and change the device until the task be done.

But, Web applications should consider the Web architecture model, i.e., the client-server, to have a better adaptation. Client is responsible to render the user interface through a browser. It is in the client side that the user interacts with the system using input and output hardware. The server is responsible to process client´s requests and data persistence, but it knows few about the input and output devices in client side. Since the server is responsible to produce the user interface code to be displayed on the client, the server needs to know about the user device, the user location and user preferences to do a better adaptation. The generated interface needs to have good usability.

For example, smartphones are good to read and write small texts, such as post-it notes. The tablet with touchscreen allows users have mobility and a good readability, but it is not so efficient to write texts. The tablets equipped with pen sensitive screen should be used to write formulas or do sketches, instead of only as a pointing device. Da Silva and da Rocha [20] propose the InkBlog tool, a blog tool that receives input data from stylus in a pen sensitive tablet so that the user can handwriting her posts (Fig. 1a).

The InkBlog is a tool for the Ae e-Learning environment, and take advantages of the Pen-based computing to allow users interact with the e-Learning environment with a pen. The authors comment that using the InkBlog is possible a better support to disciplines such Graph Theory and Computer Theory. Without the InkBlog, the user needs to use paper and pencil to resolve an exercise and use specific hardware, such scanner, to digitalize the solution. Or the user needs to use a special application to draw a graph. In both solutions, the user posts the picture as an attached file in weblog post. Using the InkBlog the user can sketch the graph direct on the weblog tool through direct manipulation (Fig. 1b). InkBlog was tested in iPhone (Fig. 2) and Android devices using the stock browser in each device. Both devices display correct the posts, but due the platform does not distinguish between touch and stylus press, it is not possible to handwrite a post in these smartphones; both devices recognize the input as page scrolling. This is another challenge, the web pages depends on the browser, because the browser is responsible to the capture of the user input and the page rendering. So, the hardware and the browser needs be more integrated to enhance the Web application.

 ISBN: 978-1-61208-289-9



(a) (b)

Figure 1. Using InkBlog to (a) handwriting a post and (b) to share a solution for an exercise about Graph Theory. Both pages rendered by Chrome Browser.

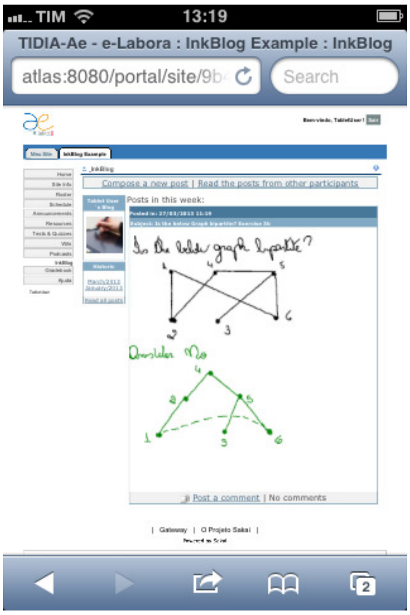


Figure 2. InkBlog rendered by stock browser on iPhone.

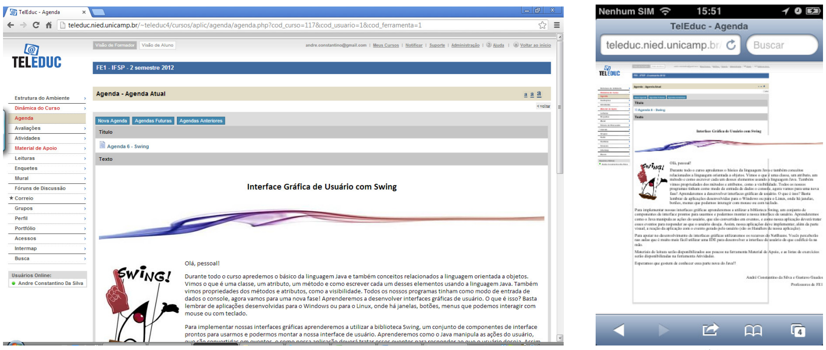
In the case of use e-Learning environments on mobile devices, is important to study the interaction problems that happen, since the actual environments was designed to be used in a desktop computer, and mobile phones have other interaction devices. da Silva, Freire, Arruda and da Rocha [21] present some problems when the users uses an mobile phone to interact with the Teleduc environment using an Android-based mobile device and a iPhone. Da Silva, Freire and da Rocha [8] present some problems when user uses an android-based smartphone and a Tablet PCs. They present a taxonomy about the interaction problems: cross-platform problem, cross-modality problem and platform and modality-independent problem. These problems are barriers or difficulties to the user navigate thought the environment. So one challenge is identify these problems and correct.

The environments need to have good usability and accessibility, but the content created inside the environments needs to. Since the environments are used by a diversity of people, most of them do not have knowledge about Web accessibility, they can create content with low accessibility in mobile devices. Fig. 3 shows an agenda on a course in TelEduc created by the teacher to describe about the topics and activities to be done in a week. Some problems when the agenda is rendered in iPhone (b) can be viewed with compared with the desktop computer (a). So, another challenge is to develop authoring tools that easily create accessible contents and to develop features to allow the mobile users to visualize the published content [22].

Allowing access by mobile devices is not the only challenge to facing it. The mobile devices have specific hardware that can be used to produce content and to use to interact with the environment. The camera can be used to take photos or do videos and the microphone can be used to create audio files to be published by the teacher or by the students as content in the e-Learning environment. To allow this it is necessary to have a better integration between the mobile hardware, the browser and the environment to easily allow users create content using camera, microphone or any other input device. These devices can too be used as communication if integrated with communication software. The actual versions of e-Learning environments only dispose an action button that trigger a dialog box where the user can choose the photo or video he want upload. Depending on the solution adopted to integrate the environment in the mobile device, it is possible to turn the publishing task easier. In the case of a specific device application it is possible to customize the mobile Operation System and include sharing option in the photo and video gallery, similar as Facebook app does (Fig. 4). So, the mobile user has options to publish the media in her virtual space on the e-learning environment, like a Portfolio item on TelEduc environment, or a resource item on SAKAI and Ae.

The TelEduc e-Learning environment has a notification tool that send e-mails for the course participants describing what happen since the last logging. This feature could be integrated with the mobile device advisor management system. So one more challenge is identify e-Learning features that can be integrated with mobile devices features.

The e-Learning environment has data about the course participants, whom can be added on the mobile device



(a) (b)

Figure 3. A content created by a user in TelEduc with accessibility problems displayed on desktop computers (a) and on a mobile device (b).

Contacts tool to turn easily sent messages between the participants. Another one feature is synchronizes the activities deadline or important dates to the mobile device´s Calendar tool.

Allowing student to create multimedia content in activities, the teacher needs to select the best media or a set of them to planning the activities. The student´s works not need be more only texts, allowing them to produce videos, audios, photos or a multimedia work. If the teachers allow user to use the mobile phone, it is important elaborate activities that explore the mobile hardware, such take photos, make videos or record an audio. The student can write some text too, but it is important notice (and this is must be clear on the activity) that usually it is not expected a small texts for student´s activities; since writing on smartphones can be a difficult task, the teacher must specify how long the text must be. This perception impacts on the pedagogical use of the e-learning environment.

The actual user interface design techniques take account just a limit set of input and output hardware, limited to the context, such as techniques to design user interface for desktop or for mobile platforms. But, there is a lot of input or output hardware in these devices and these techniques are asked to consider all of them. Some input and output devices are: touchscreen, microphones, pen sensitive screen, touchpad, TrackPoint, accelerometers, joysticks, loudspeakers, small screen, large screen, printers, etc. One solution to deal with this variety of devices is use multimodal techniques on the e-Learning environment´s user interface. Multimodal interaction is a research proposal to turn the interaction between humans and machines more natural, i.e., more close to the interactions between two humans, and have the benefits to increase the usability, flexibility and convenience [23]. According to Oviatt [24] “multimodal interfaces process two or more combined user input modes (such as speech, pen, touch, manual gesture, gaze, and head and body movements) in a coordinated manner with multimedia system output”. But, developing multimodal interaction systems is a complex task [25]; da Silva and da Rocha propose the IAel environment, an e-Learning environment with multimodal user interface, discuss how multimodal architecture changes the e-Learning environment architecture. Despite these works, the challenge is to have an

Figure 4. The integration between iPhone´s Photo Gallery and Facebook App.

e-Learning environment to deal with the variety of input and output modes in the device, using it in a proper way.

V. Final Considerations

e-Learning is the use of the Web structure to support teaching and learning activities. There is some Web-based software, the e-Learning environments, which support these activities in courses. Due the technology development, it is possible to use mobile devices to teaching and learning activities, calling m-Learning. The mobile devices can access e-Learning platform, but to an effective course accessed by mobile devices it is necessary that the e-Learning does not have barriers or difficulties for the mobile devices, and the content and activities must be shaped to consider the mobile devices and their hardware, such camera and microphone.

We believe that e-Learning and m-Learning will merge in the case of have an environment to support the teaching and learning activities and brings benefits on learning. But to plain an effective course, it is important to consider the hardware and software that the students will used. So, the technology needs to be considered in the instructional design process, together with the educational issues. Future works are detail each part of these dimensions.

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>. [retrieved: Jul. 2013]

[2] SAKAI Environment, “Sakai Project | collaboration and learning - for educators by educators,” available at <http://sakaiproject.org>. [retrieved: Jul. 2013]

[3] TelEduc Environment, “TelEduc Ensino à Distância,” available at <http://www.teleduc.org.br>. [retrieved: Jul. 2013]

[4] Ae Project. “Ae - Aprendizado Eletrônico Environment,” available at <http://tidia-ae.iv.org.br/>. [retrieved: Jul. 2013]

[5] J. R. Lima and Z. Capitão, e-Learning e e-Conteúdos, Portugal: Inova, 2003.

[6] M. Sung, J. Gips, N. Eagle, A. Madan, R. Caneel, et al., “Mobile-IT Education (MIT. EDU): m-learning applications for classroom settings,” in Journal of Computer Assisted Learning, vol. 21, May 2005, pp. 229–237.

[7] G. Gay, M. Stefanone, M. Grace-Martin, and H. Hembrooke, “The effects of wireless computing in collaborative learning environments,” in International Journal of Human–Computer Interaction, vol. 13, n. 2, 2001, pp. 257–276, doi: 10.1207/S15327590IJHC1302_10.

[8] A. C. da Silva, F. M. P. Freire, and H. V. da Rocha, “Identifying Cross-Platform and Cross-Modality Interaction Problems in e-Learning Environments,” Proc. of the 6th International Conference on Advances in Computer-Human Interactions (ACHI 2013), IARIA, Feb. 2013, pp. 243-249.

[9] Moodle Trust, “Moodle App,” available at: <http://docs.moodle .org/dev/Mobile_app>. [retrieved: Jul. 2013].

[10] M. J. Casany, M. Alier, E. Mayol, J. Piguillem, N. Galanis, et al., “Extending Moodle Services to Mobile Devices: The Moodbile Project,” Proc. of the 6th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2012), IARIA, Sep. 2012, pp. 24-28.

[11] B. H. Khan, “A framework for Web-based learning,” in Web-based training, B. H. Khan, Ed. Englewood Cliffs, NJ: Educational Technology Publications, 2001, pp. 75-98.

[12] J. Nielsen, Usability Engineering, EUA: Morgan Kaufmann, 1993, 362 p.

[13] M. G. Helander, T. K. Laundauer, and P. V. Prabhu (Editors), Handbook of Human-Computer Interaction, Amsterdam, Elsevier Science B. V., 1997.

[14] J. Rubin, Handbook of Usability Testing: How to Plan, Design and Conduct Effective Tests, New York: John Wiley & Sons, 1994, 330 p.

[15] W3C, “Web Content Accessibility Guidelines (WCAG) Overview,” available at: <http://www.w3.org/TR/WCAG/>. [retrieved: Jul. 2013]

[16] T. W. Bickmore and B. N. Schilit, “DIGESTOR: Device-independent Access to the World Wide Web,” Proc. of 6th International World Wide Web Conference, Elsevier Science Publishers Ltd., Sep. 1997, pp. 1075-1082, doi: 10.1016/S0169-7552(97)00026-3.

[17] H. Zhang, “Adaptive Content Delivery: A New Research Area in Media Computing,” Keynote Speech on Proc. of the 2000 International Workshop on Multimedia Data Storage, Retrieval, Integration and Applications, Jan. 2000. Available at: <http://research.microsoft.com/apps/pubs/default.aspx?id= 68791 >. [retrieved: Jul. 2013]

[18] R. Oliveira and H. V. da Rocha, “Consistency Priorities for Multi-device Design,” Proc. of 11th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT), Springer-Verlag, Sep. 2007. pp. 426-429.

[19] P. Pyla, M. Tungare, and M. Perez-Quiñones, “Multiple user Interfaces: Why consistency is not everything, and seamless task migration is key,” Proc. of CHI 06 Workshop on the Many Faces of Consistency in Cross-platform Design, vol. 24, CEUR-WS, Apr. 2006. pp. 55-60.

[20] A. C. da Silva and H. V. da Rocha, “InkBlog: A Pen-Based Blog Tool for e-Learning Environments,” in Bridging Disciplinary Boundaries: Issues in Informing Science and Information Technology, vol. 10, May 2013, pp. 121-135. ISBN: 1932886699.

[21] A. C. da Silva, F. M. P. Freire, A. V. P. de Arruda, and H. V. da Rocha, “Interaction Problems Accessing e-Learning Environments in Multi-touch Mobile Devices: a case study in TelEduc,” Proc. of the 7th International IADIS Conference on e-Learning, IADIS Press, Jul. 2013, pp. 199-206.

[22] A. C. da Silva and H. V. da Rocha, “Improving TelEduc Environment With GPL Software: The Visualization Content Case for Computers and Mobile Devices,” Proc. of the 6th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2012), IARIA, Sep. 2012, pp. 222-227.

[23] D. Lalanne, L. Nigay, P. Palanque, P. Robinson, J. Vanderdonckt, et al., “Fusion Engine for Multimodal Input: A Survey,” Proc. of the 11th International Conference on Multimodal Interfaces (ICMI-MLMI´09), ACM Press, Nov. 2009, pp. 153-160, doi: 10.1145/1647314.1647343.

[24] S. L. Oviatt, “Advances in Robust Multimodal Interface Design,” in IEEE Computer Graphics and Applications, vol. 23, n. 5, Sep. 2003, pp. 62-68, doi: 10.1109/MCG.2003.1231179.

[25] B. Dumas, D. Lalanne, and S. Oviatt, “Multimodal Interfaces: A Survey of Principles, Models and Frameworks,” in Human-Machine Interaction, D. Lalanne and J. Kohlas, Eds. Berlin: Springer Berlin/Heidelberg, 2009, pp. 3-26, doi: 10.1007/978-3-642-00437-7_1.

Achieving Cooperative Sensing in Automotive Scenarios Through Complex Event Processing

Christoph Steup, Sebastian Zug, Jörg Kaiser
Department of Distributed Systems
Otto-von-Guericke-University
Magdeburg, Germany
Email: {steup, zug, kaiser}@ivs.cs.ovgu.de

***Abstract*—The actual trend towards more driver assisting systems in automotive scenarios induces a need for additional sensory data from the environment. Those perceptions can be generated by the inherent sensor systems or by external transducers. However, these data may be affected by faults and uncertainties, which may not be obvious to other cars receiving the data. Additionally, the management and dissemination of the data between the plethora of cars on the road is a big challenge. Therefore, this paper proposes an extension to existing complex event detection systems to support fault and uncertainty aware collaborative applications. An automotive scenario was chosen to illustrate the new approach. Consequently, the requirements of automotive scenarios were described and existing complex event systems were evaluated against these requirements. The extended system is called a complex event processing system, since it tries to enable sensor processing in an highly dynamic event system. To reach this goal additional attributes are introduced to the basic event schemes. Finally the processing steps are adopted to cope with the new quality attributes validity and uncertainty.**

Keywords*—Autonomous vehicles; Command and control; Fault-tolerance; Sensor networks.

I. Introduction

On of the important prerequisites for ubiquitous computing is the availability of environment data to allow context-aware computation and behaviour. Today, there is an enormous amount of this environment information potentially available, e.g., from traffic information infrastructure, floating car data, mobile devices and from instrumented smart spaces. Cooperatively perceiving environmental conditions and situations is a crucial component for improving the coordination of mobile entities like , e.g., team robots, autonomous transportation systems and cars. Recently, the use of autonomous air vehicles is considered that are partly controlled through collaborative sensing and coordination [1][2]. The locally perceived events of the environment need to be interpreted, related and combined to recognize more complex events and situations. The relations may be in the temporal domain , e.g., detecting that multiple subsequent events belong to a certain trajectory and in the spatial domain , e.g., predicting a jam or a collision situation of multiple vehicles. In this paper, we will describe our approach of complex event processing along a traffic scenario.

Actual approaches like PeerTIS by Rybicki et al. [3] or SOTIS by Wischoff et al. [4] try to enable collaboration between multiple cars. However, both only deal with the low level data dissemination between cars. To efficiently support collision warning systems, active collision avoidance and adaptive cruise control, a higher abstraction is needed since all cars will have different sensory equipment as well as different features regarding communication and automated control. An early example of a using floating car date to support an overtaking action is provided in [5]. In this example, Jini [6] is exploited to discover and use the front camera of a preceding car to see whether the road is free. Although using remote sensor information, this is a singular event establishing a unique channel between the cars. A promising approach for complex event detection systems emerged from the wireless sensor systems and the active database communities.

To evaluate the different approaches we will introduce an example scenario. It is shown in Figure 1. In this scenario some of the cars are able to communicate (C_{R1}, C_{R2} and C_{R2}) and others are not (C_{L0} and C_{R0}).

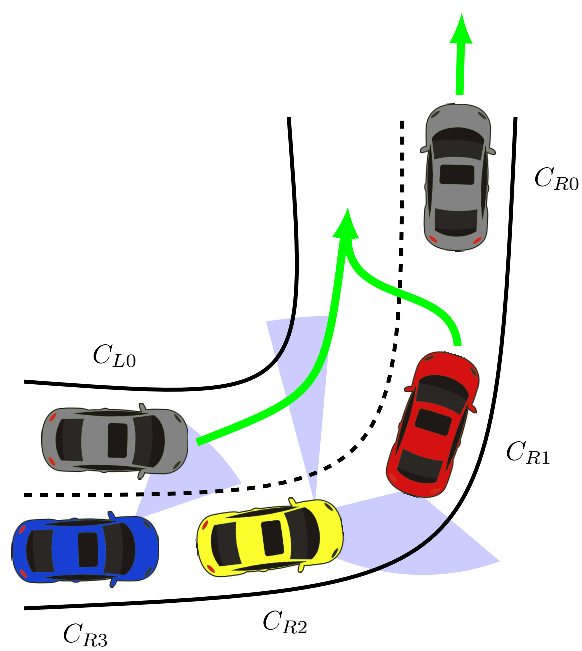


Fig. 1. Example of a dangerous situation in an automotive scenario. Car C_{R1} wants to change lane to overtake C_{R0} while car C_{L0} is driving with high speed on the left lane.

Car C_{R1} wants to overtake car C_{R0} and needs to change lane. However, this lane change is considered a dangerous manoeuvre, therefore C_{R1} wants to know if other cars may

 ISBN: 978-1-61208-289-9

collide during and after the manoeuvre. To detect possible collisions cars need to check their position against the position and speed of other cars. However, some cars cannot communicate and therefore their positions need to be observed and disseminated by the surrounding cars. In the example cars C_{R2} and C_{R3} might be able to detect the speed and the position of the endangered car C_{L0}. To achieve this they need to combine their individual position estimations of car C_{L0} to a speed estimation. The result can then be checked against the manoeuvre trajectory of car C_{R1}. Depending on the result a collision warning event may be issued.

This intuitive approach leads to an directed acyclic graph (DAG) of events and their combination. The event DAG of this example scenario is shown in Figure 2.

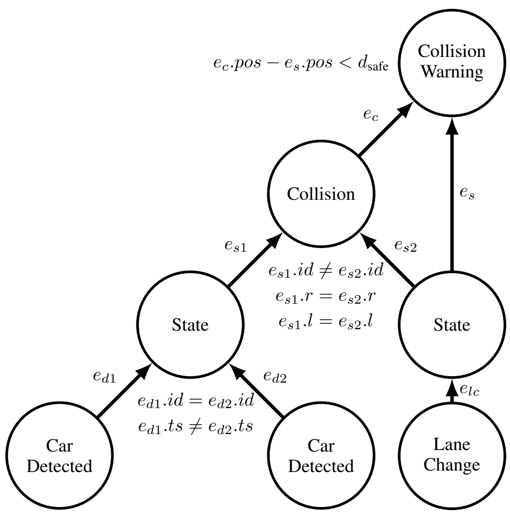


Fig. 2. The complex event DAG induced by the example scenario described in Section I. The nodes refer to types of events used in the systems. The links represent the individual named events travelling through the network.

Each event is a node in this graph. The nodes without any inputs are considered to be primitive events, which are generated by sensors. The top node represents the final event the application is interested in: the collision warning. To support the transition of the primitive events towards the final collision warning multiple combinations of intermediate events are needed. This leads to the definition of complex event detection systems (CEDS). These CEDS are researched mainly in the context of wireless sensor networks (WSN) and active databases.

The following Section II evaluates some typical system against this scenario. Afterwards we extend the CEDS through sensor processing descriptions to Complex Event Processing Systems(CEPS) in Section III . This Section is followed by an approach towards the distribution of the processing within the network in Section IV. The paper ends with a conclusion in Section V.

II. State of the Art

Generally, the existing complex event detection system can be categorized into query based systems and stream based system. The main difference is the considered use case. Query based system efficiently allow single queries to the network. Examples for these systems are Snoop by Chrakravarthy et al. [7] and Abstract Events by Katsiri et al. [8]. Snoop is limited to an active database context, which only supports centralized evaluation of the expressed complex events. However, the language imposed by Snoop is very rich. Abstract events on the other hand only support a very basic language based on temporal first order logic, but allow distribution within the network.

The stream based systems handle periodic events and therefore consider network issues like bandwidth and latency. Examples for these systems are Sensid by Krantz [9] and Solar by Chen et al. [10]. Sensid focus on wireless sensor networks with limited resources. Therefore, the expressiveness is limited and the evaluation of the complex event is centralized in one node. Solar on the other hand considers distributed detection of events and also specifies rules to drop specific events in case of overload. The description of the composition is XML-based, but limited to basic operations.

So far, complex event detection systems focus on the pure existence of events and neglect the contents or the specific attributes of the detected events. They introduce extensive languages to describe sets of events that generate the complex event. However, the generation of the attributes of the new event is almost always limited to basic operations like average, min or max. Considering the needed steps in the example scenario there are multiple steps that cannot be expressed efficiently with the described complex event detection systems. The state estimation of car C_{L0} for example needs to compute the current velocity based on two detection events. To do this, it needs to evaluate the position as well as the time between two events to approximate the velocity.

Another problem emerges from the fact that sensor readings are not perfect. This fact is usually not considered in the related work. Sensor data is affected by transient errors and uncertainties, which may have a varying impact on the sensor data. Some complex event detection systems like the one of O'Keeffe [11] tries to handle uncertainties by the definition of detection policies, which may influence the system decisions in uncertain situations. However, this system only considers timing uncertainties originating from network latencies and no value errors. Brade et al. [12] defined ways to express the trustworthiness of sensor data based on the error probabilities of the sensor. This can be exploited to extend the detection policies of O'Keeffe.

Liebig et al. [13] considered the problem of uncertainty in the timestamps of events. To cope with this uncertainty they expressed the time stamp as an interval and formulated appropriate algebraic operations to replace a simple time stamp in computations. However, the proposed mechanism can not be transferred to other data like positions or sensory values easily.

To overcome the described short comings of the existing systems, we propose the transition from a pure complex event detection system to a complex event processing system, which applies sensory processing steps onto the sensory inputs contained in the events. The following Section will describe our approaches towards a fault and uncertainty aware complex event processing for collaborative sensing.

III. Complex Events for Sensor Data Processing

As complex sensor events are used to represent arbitrary sensor data their structure needs to be appropriate and flexible. Sensor values are a representation of the current state of the environment and therefore are only valid for a limited time as described by Kopetz in [14]. The accuracy of these time-value entities drops over time. Therefore, the time stamp of the sensory information is essential in processing sensory inputs. As described by Liebig et al. [13] an interval representation of the time stamp is appropriate to convey this aging information over multiple nodes in the network. The resulting time stamp $[ts, \pm\alpha_{ts}]$ is defined through the approximate time stamp (ts) and the distance to the interval bounds (α_{ts}).

Most of the sensing equipment available today has a limited range as well as a limited observation area. These information is crucial for processing, because only sensors observing the correct area are useful. Therefore, representation of the observation area as well as the position and orientation of the sensor is needed to support remote processing of sensor data. Like the time representation, these values are subject to uncertainties since localization algorithms are limited in accuracy. The representation of position is dependant on the application. Sensor network applications may benefit from GPS-compatible 2D positions to directly convey their information to the internet. Whereas aeronautic applications need a position representation containing three dimensional coordinates and attitude.

In the case of the example scenario described in Section I, the position may be described as a triple of road segment (r), lane (l) and linear position within the segment (o). In this case, it is sufficient to describe only the offset as an interval with an uncertainty of α_o, because the road and lane can be deduced quite efficiently using maps and camera systems.

The data each sensor produces needs to be described as a general event scheme, since the set of available sensors is dynamic and not known on design-time. If we consider car C_{R3} to use a camera based system to identify overtaking vehicles, the resulting sensor data may be described through the unique identification number id of the vehicle. The identification needs to be deduced from the camera's picture in a variation of light conditions, which may impact the detection's effectiveness. Therefore, a validity value v representing the error probability based on the work of Brade et al. [12] needs to be attached to the value as well. For simplicity an uncertainty abstraction of car identifications is omitted at this point.

Putting together the descriptions of time, space and content we can formulate an event scheme of an identifying vehicle detector as:

$$E_{\text{D}} = ([ts, \pm\alpha_{ts}], r, l, [o, \pm\alpha_o], id, v) \tag{1}$$

In consequence, a general event scheme for sensory data can be derived. The general scheme consists of a time stamp, which is described as an interval together with the position of the event. The position of an event can be derived from the sensor observation area and the current position of the sensor. Since localization mechanism are subject to a defined uncertainty, this value needs to be appended to the event scheme too. Additionally, the positions may be represented differently based on the application's scenario. The sensor's values are represented based on the sensor data sheet as described by IEEE standard 1451 [15] or the MOSAIC framework [16]. Finally, a validity value is attached representing the error probability within the sensor.

$$E_{\text{Sensor}} = ([ts, \pm\alpha_{ts}], [\text{pos}], [\text{sensor data}], v) \tag{2}$$

IV. Distributed Complex Event Processing

The combination of multiple input events towards a complex event needs a domain specific language as they have been introduced in Section II. However, as mentioned these languages lack the support for special sensor fusion operations. As an example, the collision warning system scenario described in Section I may be used. To detect a collision it is necessary to compute the intersection point of two cars extrapolating their current states (speed v_i and position o_i). The time to collision can be expressed as:

$$\Delta t_{\text{collision}} = \frac{o_1 - o_0}{v_1 - v_2} \tag{3}$$

All described systems are not able to compute such a new attribute value for the resulting event. Therefore, these schemes cannot express the concept of virtual sensors as proposed by the MOSAIC sensory middleware [16]. Additionally they consider the data of the events to be perfect, which may result in failures detecting an event.

A collision warning event needs to be issued whenever the time to collision is smaller then a defined safety time: $\Delta t_{\text{collision}} < t_{\text{safe}}$. However, variations of the speed of the two vehicles may have a huge impact on the space and time estimation of a detected collision. Therefore, these uncertainties may change the outcome of the detection.

In the described example, the important car C_{L0} had no mechanism to supply its current state to other cars, therefore the state of this car needed to be derived from other sensor readings. This deduction lowers the quality of the provided sensory data since the resulting speed (v_1 or v_2) may be susceptible to a higher uncertainty or in the case of unreliable detection, to lower validity. Through the provided validity information the impact of this deduction can be estimated and propagated from one processing step to another. Finally, our approach allows the application to describe its quality requirements in terms of validity or uncertainty predicates.

To supply the necessary validity and uncertainty values they need to be automatically determined during the processing of the individual events. To support this, we need to extend the processing description from a simple event set comparison to an applications specific function which handles uncertainties and fault probabilities. We consider the well-known Event-Condition-Action (ECA) rule mechanism with the action being an event generating function.

Such an ECA rule may be defined for the deduction of the speed of non-communicating cars as:

$$\{E_D, E_D\}, (e_0.id = e_1.id) \wedge (e_0.ts \neq e_1.ts) \rightarrow E_S : f_s(e_0, e_1) \tag{4}$$

This rule contains as input events two detection events E_D if both events detected a car with the same id $e_0.id = e_1.id$ and the timestamps of both detections is different $e_0.ts \neq$

 ISBN: 978-1-61208-289-9

$e_1.ts$. The result of this rule will be a state event of type E_S created through the application of function f_s to the input events (e_0, e_1).

The application of these rules within the network creates a directed acyclic graph of complex events processed to finally deliver the collision warning event. The resulting graph including the predicates of the combination operations is visible in Figure 2.

It is important to mention, that the resulting graph heavily depends on the current events available and their attributes. Especially if validity or uncertainty predicates are used, additional invocations of sensor fusion rules may be needed to increase the validity or decrease the uncertainty to pass the event filter.

A general ECA rule will accord the following pattern:

$$\{E_0, ..E_j, ..E_m\}, p_0[\vee|\wedge]p_i[\vee|\wedge]p_n \rightarrow E_k : f(e_0, .., e_j, .., e_m) \quad (5)$$

Therefore, the rule consist of a set of input events of type $\{E_0, ..E_j, ..E_m\}$ as well as some predicates p_i combined through either or ($\vee$) or and ($\wedge$). The resulting event will be of type E_k and is created through the function $f(e_0, .., e_j, .., e_m)$. The function itself is defined to be side-effect free, so that the repositioning of a processing step within the network only needs event dissemination.

There already exists some work on the effective positioning of processing nodes within a network. The CED system of O'Keeffe [11] tackles the problem for overlay networks using spring relaxation algorithms like the one of Pietzuch et al. [17]. However, these algorithms need to be tested in highly dynamic environments like automotive scenarios.

V. CONCLUSION

This paper proposes an extension of the complex event detection mechanism to support sensor driven collaborative applications. The requirements of automotive scenarios are described and existing complex event systems are evaluated against these requirements. To support fault and uncertainty awareness needed in these applications the event schemes and detection steps of classical complex event detection systems are extended, which lead to a new approach called Complex Event Processing.

Applications based on the new approach are able to express individual quality requirements against their input events. Contributing to this classical sensor fusion algorithms may be exploited in a distributed fashion. Related to the automotive scenario applications like CWS systems may enhance their performance through these quality attributes.

The algorithm of the classic complex event detection systems for WSN need to be evaluated against the requirements of the automotive domain. Especially the highly dynamic network topology and the short connection times between cars may impose additional challenges to the processing placement subsystem. The definition of a basic set of event schemes and combination operations for automotive scenarios lead to a domain specific language easing the development of applications of the automotive domain. Finally the described basic event scheme and ECA rule system needs to be evaluated against multiple scenarios to detect short comings in the expressiveness.

ACKNOWLEDGEMENTS

This work is (partly) funded and supported by the EU FP7-ICT program under the contract number 288195 "Kernel-based ARchitecture for safetY-critical cONtrol" (http://www.karyon-project.eu).

REFERENCES

[1] D. Steitz, J. Anderson, and K. Rochon, May 2013, [retrieved: Jul, 2013]. [Online]. Available: http://www.nasa.gov/directorates/spacetech/centennial_challenges/uas/index.html

[2] A. Casimiro, "Karyon to improve safety and performance of smart vehicle coordination," Nov. 2011, [retrieved: Jul, 2013]. [Online]. Available: http://www.karyon-project.eu/wp-content/uploads/2011/11/KARYON_Press_Release_English.pdf

[3] J. Rybicki, B. Scheuermann, M. Koegel, and M. Mauve, "PeerTIS: a peer-to-peer traffic information system," in *Proceedings of the sixth ACM international workshop on VehicularAr InterNETworking*, ser. VANET '09. New York, NY, USA: ACM, Sep. 2009, pp. 23–32.

[4] L. Wischoff, A. Ebner, H. Rohling, M. Lott, and R. Halfmann, "SOTIS - a self-organizing traffic information system," in *Vehicular Technology Conference 2003-Spring: Proceedings of the 57th IEEE Vehicular Technology Conference*, vol. 4, Apr. 2003, pp. 2442–2446.

[5] N. Nils Gura, A. Held, and J. Kaiser, "Proactive services in a distributed traffic telematics application," in *GI-Workshop: Mobile communication over wireless LAN:Research and applications*, Sep. 2001, pp. 585–592.

[6] K. Arnold, R. Scheifler, J. Waldo, B. O'Sullivan, and A. Wollrath, *Jini Specification*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1999.

[7] S. Chakravarthy and D. Mishra, "Snoop: An expressive event specification language for active databases," in *Data & Knowledge Engineering* , vol. 14, no. 1, Nov. 1994, pp. 1–26.

[8] E. Katsiri, J. Bacon, and A. Mycoft, "An extended Publish/Subscribe protocol for transparent subscriptions to distributed abstract state in sensor driven systems using abstract events," in *DEBS Proceedings of the International Workshop on Distributed Event-based Systems*, Edinburgh, Great Britain, May 2004, pp. 68–73.

[9] M. Kranz, "SENSID: a situation detector for sensor networks," Ph.D. dissertation, University of Western Australia, Jun. 2005.

[10] G. Chen, M. Li, and D. Kotz, "Design and implementation of a large-scale context fusion network," in *In First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services (Mobiquitous)*, 2004, pp. 246–255.

[11] D. O'Keeffe, "Distributed complex event detection for pervasive computing," University of Cambridge, Computer Laboratory, Tech. Rep. UCAM-CL-TR-783, Jul. 2010. [Online]. Available: http://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-783.pdf

[12] T. Brade, J. Kaiser, and S. Zug, "Expressing validity estimates in smart sensor applications," in *ARCS International Conference on Architecture of Computing Systems 2013*, Jan. 2013.

[13] C. Liebig, M. Cilia, and A. Buchmann, "Event composition in time-dependent distributed systems," in *COOPIS Conference on Cooperative Information Systems: Proceedings*, Sep. 1999, pp. 70—78.

[14] H. Kopetz and K. Kim, "Temporal uncertainties in interactions among real-time objects," in *Reliable Distributed Systems, 1990. Proceedings., Ninth Symposium on*, 1990, pp. 165–174.

[15] Institute of Electrical and Electronics Engineers, "IEEE Standard for a Smart Transducer Interface for Sensors and Actuators - Common Functions, Communication Protocols, and Transducer Electronic Data Sheet (TEDS) Formats," in *IEEE Std 1451.0-2007*, 2007, pp. 1–335.

[16] S. Zug, A. Dietrich, and J. Kaiser, "An architecture for a dependable distributed sensor system," in *IEEE Transactions on Instrumentation and Measurement*, vol. 60 Issue 2, Feb. 2011, pp. 408 – 419.

[17] P. Pietzuch, J. Ledlie, J. Shneidman, M. Roussopoulos, M. Welsh, and M. Seltzer, "Network-aware operator placement for stream-processing systems," in *ICDE Proceedings of the 22nd International Conference on Data Engineering*, 2006, p. 49.

Enabling Mobile Access to Distributed Recycling Knowledge

Sönke Knoch
German Research Center for Artificial Intelligence
Research Department Intelligent User Interfaces
Saarbrücken, Germany
soenke.knoch@dfki.de

Alexander Kröner
Georg Simon Ohm University of Applied Sciences
Department of Computer Science
Nuremberg, Germany
alexander.kroener@th-nuernberg.de

***Abstract*—From the consumer perspective, classifying a product regarding its environmental impact is a difficult task because relevant knowledge is usually not only diverse, but also distributed over several information sources. In this work, an analysis of mobile "green" applications formed the basis of a mobile application, which aims at providing all recycling-related information in-situ. Its domain model integrates recycling knowledge from several information sources and is capable of decomposing a product into its elementary parts. The mobile application enables the user to initiate interaction with this model over three different ways of describing a product. Beside insights concerning information access and user interaction, a first evaluation of the prototype indicates that the employed fused domain model may outperform results achieved with a traditional approach to web-based information search concerning recycling information.**

Keywords-Sustainability; decision support; mobile computing; case study

I. INTRODUCTION

Limitation of natural resources affects everyday decision making in diverse ways: indirectly through increasing costs for products, e.g., based on oil, or directly due rationale insight and ecological awareness. Unfortunately, such *sustainable decision making* is a non-trivial task for various reasons. For instance, a product has to be chosen that should be "easily" recyclable. From the viewpoint of sustainability, recycling is affected by materials the product is consisting of, the recycling process for decomposing the product, the extent such decomposition is possible, and even the (potentially future) context that determines efforts needed to insert the product into the recycling process.

In order to make an informed decision, the decision maker may have to acquire all of that knowledge – and to fuse it: While there are efforts towards the integration of sustainability-related information along the supply chain [1], community-driven information sources may provide additional hints [2], and of course, the current context (e.g., location) may bias the sustainability of a decision.

This complexity partially explains why expert advice in-situ may increase people's will to do such decisions [3]. Information has to become more available [4], and be explained to the user [5]. Thus, it is little surprising that there exists a considerable amount of "green" mobile applications, which seek to support their user in-situ in solving tasks related to sustainability.

This article reports on user feedback concerning a mobile application and a linked information service, which address decision making concerning consumable products based on recycling-related information. In the following, Section II reviews typical characteristics of such mobile applications. Then, Section III reports on a data mashup, which seeks to fuse different kinds of recycling-related knowledge originating from potentially distributed sources. It produces a domain model, which is accessed by a mobile information service, which combines the services of various previously reviewed applications at a single point. Afterwards, Section IV reports on user feedback concerning the interaction with the new service in comparison with alternative ways of acquiring similar information. Finally, the article closes in Section V with a summary of achieved results and an outlook on future work.

II. RELATED WORK

In 2011, a preparatory internal study addressed the state-of-the-art of mobile applications supporting sustainable decision making. The survey comprised mobile applications offered at the Android Market and the Apple App Store. Search terms were "energy consumption", "energy efficiency", and "green life" and led to a result of 23 relevant mobile applications in the Android Market and 25 mobile applications in the Apple App Store. The result was sorted into four categories:

Promotion (4 mobile applications). Mobile applications in this category, typically, promote energy saving technologies, such as solar energy systems, low-energy devices of certain product classes (e.g., fridges, air conditioning systems, etc.), or energy saving techniques (e.g., monitoring tools and programmable thermostats). For example, the mobile application Lennox [6] calculates the energy savings achievable by a new air conditioning system, provides product information and directs the user to the next local dealer.

Education and Information (20 mobile applications). References, encyclopedia, decision support systems, and games form a category on its own. The majority of such mobile applications provide information in form of references, tips, or links and news collections. For example, the mobile application "this is green" [7] offers information that is thematically organized by a picture of a layout of a common one family house. If the user tabs on the garage he will find information on fuel consumption of the car, if he

 ISBN: 978-1-61208-289-9

tabs on the bathroom information on how to save water is provided. The application “low carbon life” [8] is a collection of little games that tries to teach the user, e.g., how to use the washing machine in an efficient way and how to recycle trash that occurs in a common household.

Calculators (9 mobile applications). Other mobile applications support the user in calculating balances concerning sustainability-related factors. They can be distinguished in mobile applications meant for the private and for the business domain. The former ones focus on an individual’s habits and objects, e.g., flights and TV. The latter ones focus on business branches such as architecture or lamp industry. In general, the user has to enter data manually into the respective mobile application, which is a major difference to mobile applications classified as “monitoring and controlling”. For example, the “green footprint calculator” [9] is filled manually with data such as monthly bills (oil, gas, and electricity), number of flights, and recycling behavior. Once filled with this data, the mobile application calculates the yearly carbon footprint and visualizes it with a maximum of six green trees if the carbon footprint is very good/small. The application “MeterRead” [10] captures energy consumption. The number of kilo watts is synchronized manually with the electrical meter over a graphical meter that looks similar to the one that can be found in households. After data gathering, the mobile application provides a prediction for the consumption over the next 30 days.

Monitoring and Controlling (15 mobile applications). Finally, there are mobile applications which connect to energy consuming devices in the private and the business domain. In the private domain, they focus on devices common for an individual’s environment, e.g., house, car, and mobile phone. In the business domain, such mobile applications focus on branches, e.g., IT, manufacturing industry, and facility management. For example in the private domain, the “power tutor” [11] analyzes system and power usage of the mobile device and provides chart views e.g., for the consumption of the LCD, CPU, and Wi-Fi. The “green gas saver 1.0” [12] shows the greenest way of acceleration in a car and a lot of mobile applications visualize energy consumption (electricity, oil, and gas) and provide remote control features (e.g., switch on/off, timer configuration, etc.) and alarms when consumption exceeds a defined threshold, e.g., “GSH ienergy” [13] from the business domain and “DONG Energy eFlex” [14] to control home environments in the private domain. Community features are included in some mobile applications, where the user’s green performance can be compared to the performance of the user’s friends.

General observations included that mobile applications for sustainable decision making were either highly specialized (focus on product advertisement or industrial applications) or very general (dictionaries, household / lifestyle consulting). Furthermore, the reviewed applications rely on data from a single information source, which does not reflect diverse and distributed character of such information mentioned in the beginning. Finally, despite the mobile platform, there was little use of the mobile sensing capabilities.

This article reports on how these gaps could be addressed for a specific application scenario: an “Eco-Advisor” should support consumers in ranking products according to their environmental impact, and in making informed decisions concerning recycling options regarding a product at hand using information from distribute recycling knowledge. An overview of related work in the area of mobile mashups was already provided in previous work (cf. [15]).

III. FUSION OF RECYCLING KNOWLEDGE

According to the previously introduced classification of related work, the Eco-Advisor could be categorized in the first place as an “information and education” service, which includes aspects of a “calculator”. While the service as such could be employed also for user support in non-mobile scenarios, its particular focus is on decision support concerning a product “at hand”.

Therefore, the service has to support the user in establishing a link between the subject of interest – a physical product instance – and relevant information concerning this individual artifact. This information may originate from distributed sources, and may differ in format and semantics. It may describe aspects of the artifact, this kind of artifacts, resources used for creating the artifact, and related services. Efforts needed in performing this task strongly depend on the way data are organized and structured by the service – its domain model.

A. Requirements

As the mobile application is meant to provide information for products, its domain model has to be capable to represent a product’s most important properties. The model is kept as generic as possible because it is a storage for all kinds of data, structured and unstructured.

A product is defined in an economic sense as the output which is the result of the transformation that was initiated by humans. This transformation consumes scarce resources, such as materials and energy. In this article, we will focus on physical products and exclude virtual products, such as information or services.

Three core requirements for the model were identified:

- **Requirement 1:** The domain model has to carry information in form of various data patterns from distributed sources on an abstract and a concrete level and is open for extensions.
- **Requirement 2:** The domain model has to enable a disassembly of products in terms of kind and amount of materials included in the product’s (current) physical form.
- **Requirement 3:** The domain model has to match the interaction implemented by the mobile application.

Requirement 1 asks for a domain model which supports the mapping of a product at hand to recycling-related information. As recycling information is not provided by all manufacturers, such information can be found on the abstract level in the absence of manufacture specific information. If product specific information is available, it is stored on the

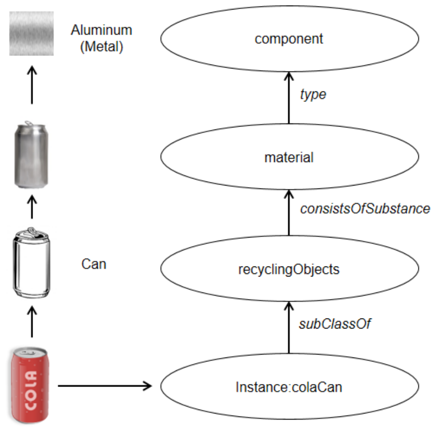


Figure 1. Ontology representation of the product structure.

instance level. Additionally, the model has to ensure a degree of extensibility which allows an adaption for specific needs. The last criterion is related to the open/close design principle from object-oriented programming. To integrate data from distributed sources, the model has to be able to carry data in heterogeneous patterns, and to make information available in a unified format.

Requirement 2 demands a domain model able to reveal a product's components and materials down to an elementary resource level. For example, a beverage can consists of aluminum, which is a chemical element in the boron group with the symbol Al, the third most common element, and most abundant metal in the Earth's crust. Such information can be employed by the service in order to perform calculations involving a product's durability, kind of resources used, and recycling potential. Thus, while a resource used within a product may be scarce, this may be less crucial if the resource can be extracted with limited efforts during recycling for later reuse.

Requirement 3 demands that the domain model supports the particular kind of user-product-service interaction that forms the background of the envisioned kind of support. In order to communicate with the service, the user has to communicate the product to be investigated. Ideally, this product will be at hand, and even be shipped with a label (e.g., Radio-frequency identification (RFID) referenced as ISO 14443, Quick Response (QR) Code referenced as ISO 18004) describing the individual product instance. Other situations may widely differ, and require the user to describe in some way the product. Therefore, the overall system provides the user with three ways of initiating interaction with the service (search by text, search by category, and search by image). The domain model has to reflect this diversity with an organization, which facilitates information retrieval starting from unique identifiers, visual features, keywords and product categories.

B. *Designing the Domain Model*

The assembly information on a product was modeled in the Ontology Web Language (OWL) [16]. In the model shown in Figure 1, a product is an instance of a sub class of

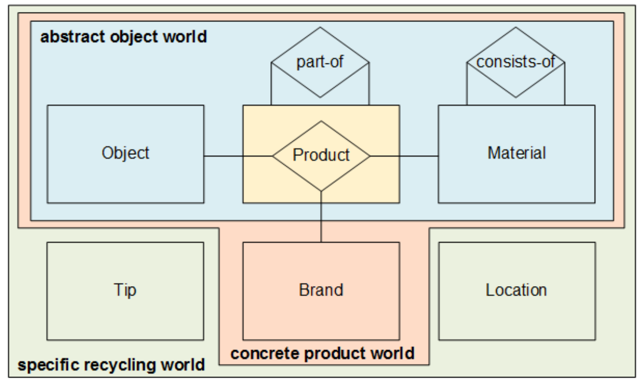


Figure 2. Entity Relationship Model (ERM) of the domain model (most relations and attributes are faded out).

recycling objects, which consist of one or multiple substances of a certain type.

According to Requirement 1, the final domain model is open for extensions; it was developed as an onion layered architecture. In the innermost layer lies the core, most abstract model which is the nucleus of the model that is visualized in Figure 2, the "abstract object world". Objects consist of different Materials, the bill of materials, and have thereby a certain composition (Requirement 2). This kind of product assembly is discussed for electromechanical products by Rachuri et al. [17], an extension of the Core Product Model 2 (cf. Fenves et al. [18]) that covers a product's function, form, and behavior. The entities in the next layer, the "concrete product world", form the world of products and contain all entities from the object world. Objects are manufactured differently by different companies under different Brands. The combination of the entities Brand, Object, and Material forms a Product. These two worlds, the object and the product world, represented by the two innermost layers can be transferred on numerous use cases where product data is involved. Two kinds of products are allowed: products with a structure of certain materials and products that provide a structure under a certain brand. All products can contain sub-modules. This hierarchical modelling approach, indicated by the part-of relation, allows the subordination of sub-products which are produced under a different brand by a certain supplier. A similar classification hierarchy was provided by Pels [19], which distinguishes between product instances, classes, and types to reduce the complexity of product models. In a similar way substances, contained in a material are modeled which allows the decomposition of a product in its most atomic elements. In the outermost layer, the most specific one ("specific recycling world"), the entities for the use case at hand are modeled and set in relation to the entities in the other layers. The entity Tip contains creative recycling tips, the transformation of old objects into something new, for Products, Objects, and Materials. Location contains recycling points where Products, Objects, and Materials can be recycled. The specific (recycling) world is open for more extensions to extend the Object and Product worlds according to specific needs. The decision for an onion layered design of the domain model supports extension of

 ISBN: 978-1-61208-289-9

the model: it is possible to add layers for specializing the model and to remove layers for generalizing the model. A similar way of abstraction was provided by Lee et al. [20], which proposed a generic and independent multilevel product model that is divided into data, model, and metamodel level.

To support the interaction (Requirement 3), textual definitions from WordNet [21] are used to identify the entities Object, Material, and Brand that are denoted as things following the notion "Internet of Things". This kind of identification allows text searches on the IDs and users to find the Object, Brand, or Material of interest. The relation among those three entities allows the presentation of related Materials and Brands when an Object is searched, the presentation of related Brands and Objects when a Material is searched, and the presentation of related Objects and Materials when a Brand is searched. Related products from the overlapping of all three entities can be presented. Additionally to the concept of definitions, word forms – a set of synonyms – are assigned to Objects, Materials, and Brands, respectively. These synonyms support a query expansion mechanism that guarantees search results for a set of valid search terms. For example, "Al" leads to the same result as "aluminum", "aluminium", or "atomic number 13". Recycling Tips are assigned to Objects and Materials. A product taxonomy is used to categorize Products, which allows a search for products by category. Products have additional attributes which are amount and unit. This allows for storing information on the quantity of materials which are obstructed in one object. Locations own the additional fields latitude and longitude to store the GPS position.

C. *Implementation*

The mobile mashup was realized as a mobile application that combines the contents of multiple heterogeneous and distributed information sources. It was decided to integrate all such source into one database, which allows faster query responses and limits the access to one interface. For that reason, the ontology model depicted in Figure 1 was transferred to a relational database according to the ERM in Figure 2. The integration and adaptation of information, for example recycling tips were retrieved from World.org [2], was described from a technical perspective in [15] by the authors of this paper. In the database, per default, each entry consists of the tuple {ID, Name, Description, Image}. The ID is a unique identifier, Name depicts the designation of the data entry, and Description contains a long text that helps to characterize the thing. An Image visualizes the entity and can be stored in form of a file path. Each entity is expandable by additional attributes that might be appended to the 4-tuple.

The application that runs on a mobile device with Internet connection communicates with the backend that runs as a web service and is accessible over a REST interface. The user interacts with the mobile device and things – in our scenario for the experiment an aluminum can, a plastic or a glass bottle. Three ways of interaction were realized, search by text, search by category, and search by image. Search by image was realized by using the IQEngines API, which delivered acceptable results (in most cases the labels and not

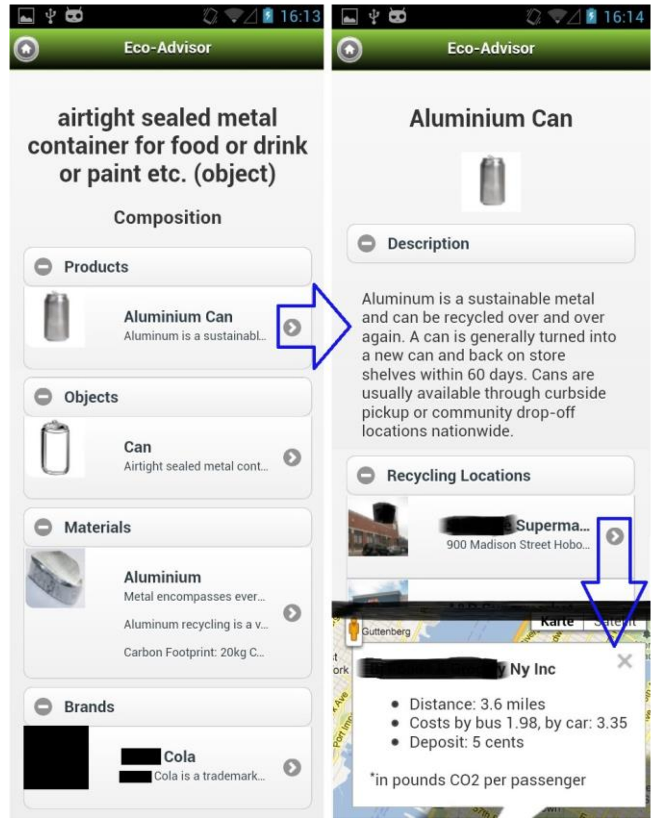


Figure 3. Composition, detail, and location view for a Thing/Object.

the things are recognized) that can be improved by training the image recognition algorithm. Search terms from all three ways of interaction are expanded by synonyms from the WordNet [21] dictionary to match additional entries in the database. The system architecture is kept modular to support adding and removing information sources. The system architecture, navigation, and interaction are described in detail in [15]. A screenshot from the mobile application is presented in Figure 3 and shows the search result for an aluminum can manufactured by a certain brand.

IV. CASE STUDY

In the following, a survey is presented that evaluates the mobile mashup and its underlying data mashup built on top of the domain modelin terms of usability and usefulness. First, the user interface is evaluated to check if the navigation and interaction method is easy to handle for the user. Second, the data mashup stored in the aforementioned domain model is evaluated to find out if the integrated information sources are helpful (1) in the way they are presented, (2) while the user has to solve different tasks from the recycling domain.

A. *Research Question and Experimental Design*

When the first running prototype of the Eco-Advisor mobile application was finished, feedback was gathered by involving a small probe of people in order to validate concept and basic design decisions. The main question the experiment sought to answer was the following one:

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-289-9

Do the mobile mashup and the domain model help a user to achieve recycling goals more efficiently compared to a stationary Web browser?

Here, "efficiency" comprises various facets of the original task, including quality of result (subjective measures such as user satisfaction, objective quality of recycling), efforts required to perform this kind of recycling, as well as efforts needed to deal with the application (time, interaction steps). In addition, the experiment aimed at gathering information concerning the preferred way of interaction with such a service. Acquiring information from such a service can be realized in quite different ways of interaction ranging from search by text, category, to image taken from the subject of interest.

In order to address these questions, three experimental tasks were defined, which had to be executed by participants of an experiment. These tasks had to be solved with the mobile application on a mobile device ("app variant"), and with a regular web browser ("browser variant"), respectively. The web browser was installed on a regular desktop PC in order to remove effects from potential issues specific to the interaction with mobile web browsers from the experiment (e.g., entering URLs, need for zooming gestures). Furthermore, the web browser was pre-configured in order to support participants in the requested tasks. This setup was chosen based on the assumption that users interested in recycling would have created bookmarks and other pointers to knowledge relevant for performing such tasks. Thus, the browser configuration seeks to reduce search for information sources as such, and instead to leverage search for information using these sources.

During **Task1 (Conventional Recycling)**, the participant is confronted with an object that has to be recycled in a conventional way in the vicinity. In the browser variant, the participant will find his or her location in an opened Google Maps tab and additional tabs with websites about recycling. The offer of opened websites on a workstation instead of an empty browser on the mobile phone makes the comparison between browser and app variant fairer and prevents the occurrence of a bias. During the study of results, the reader should keep in mind the difference between the two settings.

During **Task2 (Environmental Impact)**, the participant is confronted with a set of objects and is asked to choose the most environmental-friendly one among them. During task execution in the browser variant, the participant can continue his or her Web browser session from Task1.

During **Task3 (Creative Recycling)**, the participant is confronted with one of the objects from Task2. For this object, the participant should search a creative way of recycling which stands in contrast to conventional ways of recycling in Task1.

During the three tasks, the main factor is the Search for Information regarding the domain of sustainability. Every participant interacts on both levels Web browser and mobile application. Each task is related to one particular hypothesis:

H1: *The mobile application supports a more efficient search for conventional ways of recycling than a common stationary Web browser.*

H2: *The mobile application supports the user in judging an object's environmental impact more efficiently than a common stationary Web browser.*

H3: *The mobile application supports a more efficient search for creative recycling methods than a common stationary Web browser.*

For measuring support of these hypotheses in the respective tasks, the study relies on several parameters: one measurement is time. The time a participant takes to accomplish one task is measured and allows for comparing which kind of search method (stationary browser/mobile application) leads faster to results. Another measurement is the satisfaction of the user concerning search result and interaction comfort. The participants are asked to rank their opinion in both categories (satisfaction and comfort) on a five point Likert scale (ranging from 1 (disagree) over 3 (neutral) to 5 (fully agree). To check a user's preference, the participant has to select the preferred search variant per task (stationary browser/mobile application). To check if the domain model and the information it provided was helpful, each participant specified the criteria taken into consideration for the decision eventually made at the end of each task.

To receive feedback on usability related aspects, a user rating in the dimensions usefulness, readability, navigation, and visualization is gathered on a 5 point Likert scale, respectively.

Questions about the preferred search mechanism (by text / by category / by image) and ideas for improvement are meant to provide the developer some feedback for further improvements.

The (potential) persuasive nature of the mobile application is tested by asking about the influence of the mobile application on the participant's current recycling behavior: if the information offered by the mobile application would be available during decision making, would people expect a change in their behavior?

Finally, at the end of the study, an overall preference (stationary browser versus mobile application) is asked for.

B. Setup

The experiment was conducted in-lab under the supervision of one instructor. The participants sat at a table in front of a common PC workstation. On the workstation, participants filled out questionnaires and solved the tasks in the browser variant. The instructor guided through the experimental procedure, explained the tasks, and answered questions. For the mobile setting the mobile device Google Nexus S by Samsung was used.

The objects during task execution contain three objects from the category soda pop beverages. It was decided to use beverages from one well-known brand, to allow a brand specific search and to avoid that an unknown product will confuse a user. As questions of the survey are answered on the workstation, it can be profited by the advantage of fast result analysis and automated time measurements during the experiment. Most of the questions were of closed nature, while in some cases open questions were asked where the participant had to fill in an answer into the text field, for example the result of each task. All questions were mandatory, except the questions for problems during execution and ideas for improvement. During operations in the browser variant, the browser's history was used to log visited pages and used search terms. During operations on the mobile phone, search terms and navigation paths were logged on server-site.

C. Procedure

The experiment was divided into three phases: In the first phase, the participant had to answer a set of questions on his or her demographical background, the experience level concerning computer, mobile phone, and Internet usage, and the knowledge about recycling.

In the second phase, all participants had to solve three tasks. To solve these tasks two tools were provided: a Web browser on the workstation and a mobile phone with an application. For each task the participant had to use the Web browser in the first run and the mobile application in the second. After each run the participant had to answer a set of questions. In order to balance competition of mobile application and browser variant, in the latter one, 7 Web pages were already open in the browser's tabs once a session started. Those pages contained the same content that is integrated in the data mashup behind the mobile application. However, during task execution the participants were allowed to open new tabs and to start an own free search.

In the third phase, the study concluded with questions about the preferred search method, problems during task execution, and ideas for improvement. Additionally, it was asked if the presented mobile application could influence the participant's recycling behavior and if he or she preferred the mobile application over the stationary browser or not.

D. Result

The study lists 22 records, 2 experts and 20 non-experts. The average participant was 26 (median) years old. In the following presentation of the results percentages are rounded to integers. 13 female (59%) and 9 male people (41%) took part. Regarding the occupation, among the participants were 2 pupils (9%), 18 students (82%), and 2 professionals (9%, one software engineer and one researcher). Areas of work are wide spread and include linguistics and translation, computer science and IT, literature and culture, business administration and economics, and education.

The technical experience level regarding the usage of stationary and moveable computers was relatively high. 22 (100%) use a computer that is connected to the Internet, 16 (73%) use a mobile phone with Internet. On the stationary computer 8 (36%) surf more than 20 hours per week and 8 (36%) less or equal than 10 hours per week. On the mobile, only 4 (25%) spent more than 10 hours per week in the internet, while 8 (50%) are only between 0 and 2 hours online. While browsing the Web on the mobile, 4 out of 16 (25%) use predominantly applications. 4 (25%) additionally search for information about products during a shopping trip.

The participants' recycling knowledge was diverse. 19 (86.36%) are recycling their trash, 13 (68%) self-motivated, and 11 (58%) through regulation (multiple selections possible). 13 (68%) consider a product's environmental impact while coming to a decision during a shopping trip. Those who do, consider all different kinds of factors, energy consumption during operation as well as production and packaging. Those who do not, don't have time, are not informed enough, or have other reasons. Additionally, 8 (36%) knew what a carbon footprint is and were able to explain it, in most cases precisely.

Task1: Browser. All participants except one (the participant was not really motivated to spend some minutes on a location search) found a location for the glass bottle. The average distance to the user location was 0.71 miles. Two locations (9%) were subtracted out, one location was a container service and the other a junk hauling service. 4 (19%) identified trash cans, 5 (24%) chose supermarkets, and 10 (48%) identified a recycling center as point of disposal. Decision criteria were distance in most cases (15 / 71%), deposit value in 4 cases (19%), the "fastest result" in 2 cases (9%), and missing information on trash cans in 1 case.

Task1: Mobile application. All participants found a location for the glass bottle. The average distance to the user location was 0.36 miles, 0.35 miles lower compared to the results from the browser search. Distance was the most frequently mentioned decision criteria. Only one participant named carbon emissions associated with the trip as a decision criterion.

The preferred search method for Task1 was the mobile application (15 votes out of 22 / 68%).

Task2: Browser. All participants except one were able to identify one product out of three (glass bottle/plastic bottle/aluminum can) as the most environmental friendly one. 12 (57%) decided for the glass bottle, 6 (29%) for the plastic bottle, and 3 (14%) for the aluminum can. The decision criteria were carbon footprint (17 / 77%), the product's composition into materials (6 / 27%), and studies found through a search engine (1 / 5%). One participant said: "glass bottle is re-usable and I am safe from molecules from the plastic bottle entering my drink".

Task2: Mobile application. All participants were able to identify one product out of three (glass bottle/plastic bottle/aluminum can) as the most environmental friendly one. 9 (41%) decided for the glass bottle, 10 (45%) for the plastic bottle, and 3 (14%) for the aluminum can. While 43% of the participants changed their mind, 57% kept the decision from the browser variant.

The preferred search method for Task2 was the mobile application (16 votes out of 22 / 73%).

Task3: Browser. All participants except one (95%) found a creative way of recycling for the aluminum can. Several creative ways of recycling were discovered: potting plants, lanterns, aluminum boat, pen and pencil holder, build a children's telephone, tinker decorative items, sculptures, art, camping cooker, solar furnace, ashtray, money box, and so on. Asked, if the knowledge about reusing a product would influence the participant's buying decision was approved by 5 out of 21 / 24%).

 ISBN: 978-1-61208-289-9

Task3: Mobile application. All participants identified a creative way of recycling for the aluminum can. Additional results were a children's drum set, a candy box, a seed storage, a picture frame, gift wrapping, hooks, and film canisters. All participants except 3 (86%) found a new creative way of recycling different from the one they found in the browser variant. Knowledge about reusing the product could influence the participant's buying decision in 9 (41%) out of 22 cases, 17% more compared to the browser variant.

The preferred search method was the mobile application (14 votes out of 22 / 64 %).

Satisfaction and Comfort during the tasks is shown in Figure 4. The time measurement during the tasks resulted in the values that are presented in Table I.

TABLE I. AVERAGE EXECUTION TIME IN MINUTES

	Browser	*Application*
Task1	8:17 min.	7:10 min.
Task2	7:09 min.	5:17 min.
Task3	6:26 min.	5:25 min.

The concluding questions showed that most participants preferred the traditional search mechanisms "search by text" (13 / 59%) to the "search by category" (4 / 18%) and the uncommon "search by image" (5 / 23%). In the four categories usefulness, readability, navigation, and visualization the lowest average rating received the navigation (3.27) on a scale between 1 (worst) and 5 (best). Visualization was rated with 3.36, usefulness with 4.05, and readability with 4.14. Many participants experienced problems to find information placed at the leaf level of the navigation tree although a legend with hints on the underlying content was given on the screen. Room for improvement was seen in the navigation ("too complicated", "less clicking"). One participant suggested placing favorites on the home screen. Another one suggested integrating more pictures to improve the visualization, e.g., to visualize the creative ways of recycling. Asked if the mobile application could influence the participants recycling behavior, 73% responded with "yes". After all, the mobile application was mentioned as the preferred method of acquiring recycling information (15:7 / 68% : 32%).

E. Findings and Discussion

Feedback obtained in the categories navigation and visualization indicates that potential for improving the mobile application lies in the optimization of navigation concept and the presentation of content. For example, some participants had difficulties to find the content that was necessary to solve the task. Especially pieces of information on recycling locations which is provided in bubbles on the map, for example information on carbon emissions associated with a trip from the user location to the recycling location, are hard to discover. This information lays 5 navigation steps away from the start screen and hidden behind a 4 categories menu, which is too far. Especially users not familiar with mobile applications in general

Figure 4. Satisfaction (1=not satisfied, 2=satisfied in parts, 3=indifferent, 4=satisfied, 5=very satisfied) and comfort (1=not comfortable, 2=comfortable in parts, 3=indifferent, 4=comfortable, 5=very comfortable) during task execution.

became frustrated very fast, as they did not understand the mobile application's concept.

An interesting phenomenon is the development of time that was necessary to solve the tasks (cf. Table I). The first task took in average 7:10 minutes on the mobile application. For Task2 and 3 the duration lowered by about 2 minutes. This fact supports the statement of one participant who said, "after I was used to the mobile application I found it very helpful". However, since a mobile application might be installed right before a situation where its support is needed, it should be usable with little to no training. Therefore, this barrier has to be overcome. It has to be mentioned that in this experimental setting only a brief introduction to the mobile application was given. Usually, the user reads a description from the app store and may have a better understanding of the mobile application in advance. Thus, further experiments should start with an informing page about the mobile application as it is common in the big mobile application portals.

Nevertheless, having a look on the average task execution times in the stationary browser and the app variant, the app variant outperforms the browser variant in all three tasks. This result underlines that, after understanding the mobile application, the participants were able to find information faster using the mobile application than using the Web browser. Having a look at the level of satisfaction concerning the investigated result in Figure 4, the level of satisfaction was higher for the mobile application in all tasks. The perceived comfort during task execution was also higher when searching with the mobile application. The fact that the average distance to the identified recycling location during Task1 was about 0.35 miles lower in the app variant, while distance was the most important criterion for the participants shows that the implemented map visualization was easy to understand. These aforementioned results show that the three task-related hypotheses are supported in all categories, time, satisfaction, comfort, and user preference. Table II depicts the "delta" in all categories that were used to measure hypotheses support. Only some users used the uncommon search method "search by image". People with a great interest in technics found this search variant "very nice".

 ISBN: 978-1-61208-289-9

TABLE II. HYPOTHESES MEASUREMENTS APPLICATION VS BROWSER

Delta[a]	*Time*	*Satisfaction*	*Comfort*	*Preference*
H1	-1:07 min.	+0.71	+0.36	+36%
H2	-1:52 min.	+0.05	+0.53	+46%
H3	-1:01 min.	+0.45	+0.15	+28%
Avg.	***-1:20 min.***	***+0.40***	***+0.35***	***+37%***

a. Delta = Measurement(Browser) - Measurement(Application).

16 out of 22 (73%) participants reported that the mobile application could influence their recycling behavior. 15 (68%) participants reported that the mobile application is the preferred method of research for the tasks given. Both facts together support the appropriateness of the provided kind of support and indirectly of the employed domain model.

V. CONCLUSION AND FUTURE WORK

Sustainable behavior requires people to take a considerable amount of diverse information from distributed sources into account for decision making. This article reported on a domain model for a mobile mashup, which integrates such sources automatically. In order to gain feedback concerning the appropriateness of model and system architecture, a case study was conducted. In an experimental setup, participants had to perform recycling-related tasks with a mobile application implementing the mobile mashup approach, and with a browser-based solution on a desktop PC providing similar, but non-integrated features. Findings include that participants were able to find faster more accurate results when using the mobile application. Beyond, they were more satisfied with the mobile application's results and with the way of interaction provided by the mobile application.

Thus, the mobile mashup concept turned out to be of value for supporting people in making recycling-related decisions. However, this conclusion is limited in some ways. For instance, the user group shares certain demographic aspects, and the experiment did not involve true real-world interaction, where time pressure, interruption, and cognitive load might influence the results. Consequently, potential directions of future research should include a revision of the proposed interaction method in order to support new users in getting familiar with the mobile application. Furthermore, positive feedback obtained during the experiment indicates that persuasive technics might combine well with the mobile application concept. A context model could help to involve more user related constraints during decision support.

ACKNOWLEDGMENT

This research was funded in part by the German Federal Ministry of Education and Research under grant number 01IA11001 (project RES-COM) and 01IS12050 (Software Campus project OMM++).

REFERENCES

[1] M. Hornberger, "SustainHub project," URL: http://www.sustainhub-research.eu, accessed at 07/09/2013.

[2] WORLD Environmental Organization, "Recycling Database – Reduce, Reuse, Recycle," URL: http://www.world.org/weo/recycle, accessed at 02/07/2013.

[3] S. M. Burn and S. Oskamp, "Increasing Community Recycling with Persuasive Communication and Public Commitment," Journal of Applied Social Psychology, vol. 16, no. 1, Feb. 1986, pp. 29-41.

[4] E. M. Huang and K. N. Truong, "Sustainably Ours: Situated sustainability for mobile phones," ACM interactions, vol. 15, no. 2, Mar.-Apr. 2008, pp. 16-19.

[5] V. Margolin, "The Waste Manifesto," ACM interactions, vol. 16, no. 4, Jul.-Aug. 2009, pp. 12-15.

[6] Lennox Industries, "Lennox," URL: https://play.google.com/store/apps/details?id=com.lennox.lennoxmobile, accessed at 05/06/2013.

[7] This Is Green, Inc., "This Is Green," URL: https://itunes.apple.com/us/app/this-is-green/id337495391, accessed at 05/06/2013.

[8] Adbrownies Advertising Ltd., "Low Carbon Life," URL: http://www.appleapp.com/low-carbon-life.html, accessed at 05/06/2013.

[9] SampathG, "Green Footprint Calculator," URL: https://play.google.com/store/apps/details?id=com.virdea.mobile.android.carbonfp, accessed at 05/06/2013.

[10] M. Barton, "MeterRead," URL: http://www.macworld.com/product/71548/meterread-sale-go-green-save-money-save-earth-.html, accessed at 05/06/2013.

[11] PowerTutor.org, "PowerTutor," URL: https://play.google.com/store/apps/details?id=edu.umich.PowerTutor, accessed at 05/06/2013.

[12] The Apple Seed Store, "Green Gas Saver 1.0," URL: http://iphone-apps.toptenreviews.com/green/3015-green-gas-saver-screenshot4.html, accessed at 05/06/2013.

[13] GSH Group, "GSH ienergy," URL: https://itunes.apple.com/en/app/gsh-ienergy/id377909195, accessed at 05/06/2013.

[14] GreenWave Reality, Inc., "DONG Energy eFlex," URL: http://dong-energy-eflex.topapp.net/, accessed at 05/06/2013.

[15] S. Knoch and A. Kröner, "A mashup supporting sustainable decision making," Proc. 2013 ninth International Conference on Intelligent Environments (IE13), IEEE Press, Athens, Greece, 16-17 Jul. 2013, pp. 274-277.

[16] World Wide Web Consortium (W3C), "OWL Web Ontology Language," URL: http://www.w3.org/TR/owl-ref/, accessed at 08/12/2013.

[17] S. Rachuri et al., "A Model for Capturing Product Assembly Information," Journal of Computing and Information Science in Engineerning, vol. 6, Mar. 2006, pp. 11-21.

[18] S. Fenves, S. Foufou, C. Bock, N. Bouillon, and R. Sriram, "CPM2: A Revised Core Product Model for Representing Design Information," National Institute of Standards and Technology (NIST), Gaithersburg, MD, NIST Interagency/Internal Rep. (NISTIR) 7185, Oct. 2004.

[19] H. J. Pels, "Classification Hierarchies for Product Data Modelling," Production Planning Control, vol. 17, no. 4, June 2006, pp. 367–377.

[20] J. H. Lee et al., "A Semantic Product Modeling Framework and Its Application to Behavior Evaluation," IEEE Transactions on Automation Science and Engineering, vol. 9, no. 1, Jan. 2012, pp. 110-123.

[21] G. A. Miller, "WordNet: A Lexical Database for English," Communications of the ACM, vol. 38, no. 11, Nov. 1995, pp. 39-41.

Towards a Middleware to Infer the Risk Level of an Activity in Context-Aware Environments Using the SRK Model

Alfredo Del Fabro Neto, Rafael Boufleuer,
Bruno Romero de Azevedo, Iara Augustin
Informatics Graduation Program
Federal University of Santa Maria
Santa Maria/RS, Brazil
emails: {alfredodfn, rafaboufler}@redes.UFSM.br,
{brunodea, august}@inf.UFSM.br

João Carlos D. Lima
Department of Languages and
Computer Systems
Federal University of Santa Maria
Santa Maria/RS, Brazil
email: caio@inf.UFSM.br

Cristiano C. Rocha
MobilEasy Technologies Ltda.
Florianópolis/SC, Brazil
email: cristiano.rocha@MobilEasy.com.br

***Abstract*—It is inherent that every activity (human or non-human) has an associated risk level to the actors involved. This risk level may vary from no-damage risk to a very high risk level. Based on the regular actor's behavior, the risk level of an activity can be changed to a higher or lower level (i.e., how apt the actor is related to an activity). This paper proposes a novel approach to analyze the risk level of activities considering the behavior of the actors using the Skill, Rule and Knowledge (SRK) cognitive architecture proposed by Rasmussen in 1983. Taking advantage of the Internet of Things paradigm we developed a context-aware middleware that analyzes the actors' behavior based on the SRK model to infer the risk level associated with their activities.**

Keywords*—*Activity Recognition; Activity Risk Level; Context-Awareness; SRK Model.

I. Introduction

The usage of technologies to obtain information about the environment has been increasing. It provided the ground for research areas in pervasive and ubiquitous systems making possible the ideas proposed in the famous paper by Weiser in 1991 [1]. One of these research areas is the so-called Internet of Things (IoT), which can be seen as the integration between humans and application seamlessly "through the new dimension of "Things" communication and integration" [2]. This means that everyday life objects could be embedded within a sensor that would provide information about it to anyone who wants it (i.e., other objects, systems, etc.) [3].

To meet these requirements it is necessary for the systems to access the context data in which their actors are included. It is important to notice that actors may not only be humans as they can also be, e.g., the environment. There are many definitions of what a context is. Dey [4] shows some of them and proposes his own definition: "Context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves". A system that uses the context informations to act and react to environmental instigations in a transparent way to its users can be called context-aware system [5], term that was first introduced by Schilit et al. [6].

The IoT community has been using context-aware paradigm in order to develop solutions for different domains, which are classified as [2]: industrial (e.g., sales, enterprise services, etc), environmental (e.g., recycling, energy management, etc) or social (e.g., e-inclusion, healthcare, etc). A useful feature for providing these domains with knowledge about the context that they are inserted into is activity recognition [7][8]. These activities may be seen as human or non-human activities, e.g., environmental or system actions. They may contain certain risks that could cause damage to the envolved actors. In this sense, our paper focuses on activity recognition to determine the risk level associated with it. We developed a middleware based on the Skill, Rule and Knowledge (SRK) [9] cognitive architecture to model the actor's behavior in order to provide the risk level categorization.

To test our middleware we believe that the target scenarios would involve monitoring the activities of people who need special care, the environmental activities (e.g., fire detection) and intrusion detection based on environmental behavior.

This paper is structured as follows: In Section II, we present a model activity that was used in our research. In Section III, we define the components of our middleware architecture, their correlations and the influence of the priority system in the information flow. In Section IV, we present the work to be done, and, in Section V, we draw our final considerations.

II. Activity Model

Since the main objective of this work requires the recognition of activities, it is necessary to understand what an activity is and its relation to the actor's context.

In order to design our middleware, we needed a well defined approach that could characterize an activity. In Subsection II-A, we present the Kuutti's approach [10] to represent the relationships between the components of the Activity Theory.

A. Activity Theory

The basic notion behind the Activity Theory is that the subject is participating in an activity because he wants to achieve some specific goal. His interest is focused on an activity's object that he wants to use and/or modify in order to achieve an expected result. The interaction between the subject

 ISBN: 978-1-61208-289-9

and the object is mediated by tools. This way, a basic triangle of "subject", "object" and mediation by "artifact" is created.

With the addition of "community", the Activity Theory allowed the representation of social and cultural contexts as well as mediations between people and computing devices (smartphones and other artifacts developed for pervasive environments).

This way, the activity is turned into the basic unit of analysis and is the most important means to develop the object and the subject. Therefore, the simplest Activity Theory model already offers a solid approach to the task of comprehending the interaction between human beings and the world.

The Kuutti's approach [10] to the Activity Theory is based on three relationships:

- between the subject and the object mediated by tools and artifacts;
- between the subject and the community mediated by rules; and
- between the object and the community mediated by division of labor.

Since the subject is an active part of the community and it has social activities, the relations between the subject and the community as well as the community and the object are mediated by a set of rules and the division of labor.

The formalization of the relationships between the components of the Activity Theory allows the use of this theory in interactions (mediation) between individuals, objects and the community. The IoT is based on this mediation, since the objects are embedded with sensors that provide: i) information about the object (artifacts), ii) the presence of individuals (rules) and iii) information about the community (division of labor).

B. Context-aware model based on Activity Theory

The Activity Theory is an important theoretical reference to aid the components' explicitation that form the contextual knowledge to be incorporated in pervasive systems. The Activity Theory's components can be related with the contextual knowledge taxonomy [11], as shown in Table I.

TABLE I: BASIC ASPECTS OF AN ACTIVITY AND THEIR RELATION TO A TAXONOMY OF CONTEXTUAL KNOWLEDGE [11]

CHAT aspect	Category
Subject	Personal Context
Object	Task Context
Community	Spatio-Temporal Context
Mediating Artefact	Environmental Context
Mediating Rules	Task Context
Mediating Division of Labour	Social Context

The Cultural Historical Activity Theory (CHAT) aspects were mapped in a flexible way, allowing two or more aspects to participate in the same context category. The contextual knowledge vision is based on the premise that there are different interpretations, i.e., the contextual information in a setting can be considered as part of the knowledge model, in another setting as the own knowledge model. This flexibility allows the designers of pervasive systems to concentrate on the aspects of the knowledge level instead of modeling irrelevant details.

The taxonomy of context proposed by [12] has a pragmatic view of construction artifacts and incorporates to the context-aware systems the general concepts found in the Activity Theory (Fig. 1).

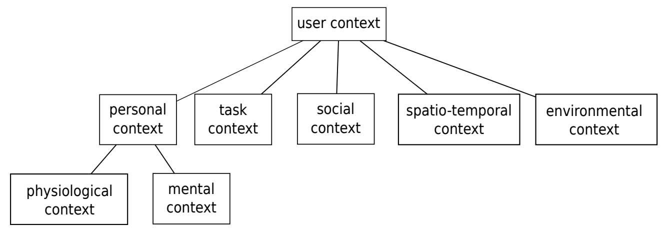


Figure 1. Context taxonomy [12]

C. Activities Recognition

The main objective of the activities recognition lies in recognizing the actors common activities in real situations interpreted from the retrieved sensor data.

The activity recognition process imposes some challenges related to interpretation because of the possibility of non-deterministic activities by the actor. Thus, probabilistic methods are apoppriated to activities recognition process [13][14]. Many researchers have been using algorithms based on probabilistic methods in order to construct activity models [15].

The comprehension of the human activities encompasses both the activities recognition and activities patterns discovery [15]. We extend this concept to not just humans, but different actors as a generic way to approach the subjects being analyzed. The activity recognition aims to the precision of the detection of activities based on a pre-defined activities model. On the other hand, the activities patterns discovery focus on the search for unknown patterns and it is directly performed over the data from low level context without any pre-defined model.

Although both techniques differs, they aim the perfectioning of the user-centric computing technology. This way, both of them are complementary, i.e., the discovery of activity patterns is capable of helping in the activity definitions that can be recognized and manipulated later.

III. PROPOSED ARCHITECTURE

To develop the proposed middleware architecture presented in Fig. 2, we relied on the layered conceptual framework for context-aware systems [5]. We chose this framework model because it separates in a well defined manner the components of sensoring, processing and management of the contextual data. It also provides the advantages of extensibility and reusability of systems. Therefore, our middleware extends this architecture and incorporates the SRK model in order to allow the actors' behavior analysis.

Our work's contribution relies in the introduction of a specialized layer (*SRK Classifier*) to create a well structured

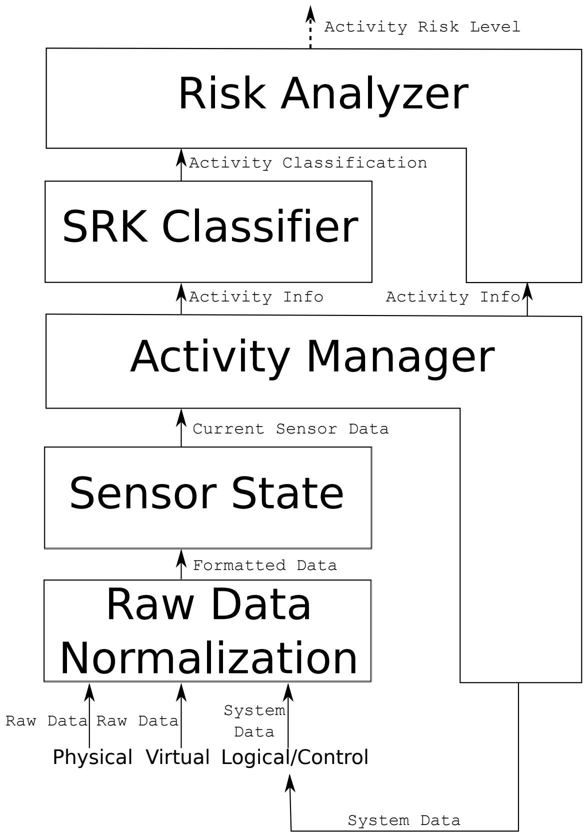


Figure 2. Middleware architecture proposed.

model of the user capabilities in performing some activity. It is important because each person has abilities that makes him more or less apt in doing a specific activity, which has a direct influence when the risk level analysis is made. The decisions of what should be done relies in this classification.

In the next subsections, we introduce the components of the proposed middleware (see Fig. 2).

A. Raw Data Normalization and Sensor State

The main goal of the *raw data normalization* component is to give meaning to the raw data retrieved from the different types of sensors (physical, virtual and logical) and format it to some defined pattern, e.g., the Message Queuing Telemetry Transport for Sensor Networks (MQTT-S) [16] standard [17]. It is from them that the context information is gathered. As it will be seen in section III-E, this layer gathers informations not only from the physical and virtual sensors, but from the logical sensors as well. These logical sensors are feeded by the system itself, incorporating informations that already have passed through some classification and semantic process. This aggregation of context information already treated allows the system to perform more precise analysis.

The *sensor state* component obtains the formatted sensor data and stores it to always provide to the above layers the latest relevant changes in the sensors values in a formatted way. This provisioning can be made by pushing the information accordingly to some priority criteria (the priority system is further explained in subsection III-E). There is also a subscription system that works like a blackboard model [5]. To make use of the subscription system, an activity has to subscribe to a certain event. An event is triggered when one or more sensores have a change in their state. In order to reduce the amount of events triggered (since a sensor may change their status with a really small variance) the system can define a certain minimum percentage necessary of changes in the state of each sensor for an event to be triggered.

B. Activity Manager

The *activity manager* component stores pre-defined activities registered manually and a log of activities that the system detected that happened. Each activity has associated sensors that describe the required values for the activity to be detected. With these informations and based on the sensors current states the *activity manager* can reason and then infer if an activity is happening, has happened or may happen. Using some learning techniques it might be possible for the system to correlate activities and sensors in order to identify new activities. It is interesting to notice that an activity may be used, e.g., to identify other activities thus acting as a logical sensor.

C. SRK Classifier

The *SRK classifier* component is proposed to define if an activity is a skill, rule or knowledge based on the work by Rasmussen [9], as follows:

- Skill: actions that the actor develops routinely without any conscious control or attention, e.g., walking with crutches;
- Rule: actions that need conscious control or attention, e.g., deviate from a hole while walking with crutches;
- Knowledge: actions where every option for solving the task through skill and rule based routines could not solve the problem, e.g., first attempt to walk with crutches.

According to Neal et al. [18], the more recent theories see habits as automated responses stimulated by aspects in the context or by the environment itself. The same way as humans, intelligent systems can control their actions and reactions based in an habitual behavior. In order to be a more generic middleware, we noted that skills should be characterized as habits, which makes more sense since actors can be anything (i.e., humans, environment, systems, etc).

The analyses performed by this component uses the historical data from the *activity manager* component and its own classification logs to determine which of the three classifications the activity belongs to. This classification also depends on the frequency and duration with which the activity is realized. Therefore, an activity may change levels, e.g., an activity that is classified as a rule may become an habit after some time practicing.

D. Risk Analyzer

The *risk analyzer* component receives a classified activity and determines the associated risk level. This is done by verifying the sensors' informations (which are requested to the *activity manager*) related to the activity being analyzed and its SRK classification. In order to clarify it, we present a simple example: An elderly person is walking in the kitchen of his house and its floor is wet. Our system would detect that the *walking* activity is complicated for this user (by managing its classification in the *SRK classifier*) and verify that the kitchen's floor is wet (by sensors spread in the environment). With this informations our system can calculate the risk level associated with that activity at that moment. This risk level can be used by other applications to aid some sort of decision-making (e.g., warn the user about the danger he might be in).

The priority system was developed because many activities may occur at the same time and that some of them may have related sensors with an abnormal data indicating that it needs to be processed first thus having higher priority. In Subsection III-E, the priority system will be discussed.

E. Data Flow and Priority System

The data flow is as follows (Fig. 2):

1) The system receives raw data from several sensors (physical, virtual and logical/control);
2) This data has to be normalized since it comes from different sources;
3) The latest formatted data from the sensors is kept in the *sensor state* layer, which can be queried from the layer above, *activity manager* (as presented in the Subsection III-A);
4) The *activity manager* gets the current sensors' data to detect activities;
5) Since activities can be created by the system itself, they can be used as logical sensors;
6) The *SRK classifier* gets activities and classify them according to the actor;
7) Lastly, the *risk analyzer* captures the classification of the activity being analyzed from the *SRK classifier* and its informations from the *activity manager* to infer its risk level.

However, the system data flow is affected by priorities, which have a major role in our middleware. The system first uses priority in the *raw data normalization* component to pass to the *sensor state* component the informations with higher priority. This way, the current state of the more relevant sensors will be updated first acting as a prioritized queue.

It is in the *sensor state* component where the priority of the sensors can be changed. As in the *raw data normalization* component, the information is passed according to its priority. The *activity manager* component is capable of creating a control sensor, which is an system activity that tells the *sensor state* to change the priority of one or more sensors. Besides that the priority is only used to determine which information goes first to the next component, which is the same and only function of the priorities in the *SRK classifier* component. The *risk analyzer* component has only the attribution of requesting the *activity manager* to change the priority of sensors.

IV. Research Agenda

We intend to verify and analyze which context model is more suitable to our middleware (object-oriented model or the use of ontologies) and then investigate standard protocols for the inter-layer communications (e.g., MQTT-S [16]).

After we have a complete and detailed model of our middleware and the flow of informations in it, we can specify different scenarios in a detailed way to evaluate the middleware. Since the collection of information for representing the behavior of people would be a time and resource consuming task, we intend to use public datasets of experiments made with this purpose to simulate the sensors states in our test scenarios.

V. Conclusion

This paper showed a novel idea for a middleware that analyzes the risk level of activities of some actor in a pervasive environment. In most studied researches, we could not observe the use a formal approach for activity recognition. So, this paper has two contributions: the use of the Activity Theory for modelling activities and the use of the SRK model for their classifications.

The context in which this idea was first conceived was to help people who need special care. However, we tried to design the middleware in a way that it could be used to more generic scenarios, including actors not only as people but as, e.g., objects or systems as well. Our aim was to provide the capability of verifying the risk level associated with an activity to applications with different domains. It could be applied even for non-human activities like a burning house or an opened door when everyone in the house is sleeping, etc.

Using the SRK cognitive architecture in our middleware to help analyze and infer the risk level of activities seems to be a promising technique, because it uses the aptitude of the actor to perform some activity based on his activity historical performance. The SRK model also allows the adjustment of the capability classification (skill/habit, rule or knowledge) of the actor to perform the same activity over time.

References

[1] M. Weiser, "The computer for the 21st century," vol. 265, no. 3. Nature Publishing Group, 1991, pp. 94–104.

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

[3] C. C. Aggarwal, N. Ashish, and A. Sheth, "The internet of things: A survey from the data-centric perspective," in *Managing and Mining Sensor Data*. Springer, 2013, pp. 383–428.

[4] A. K. Dey, "Providing architectural support for building context-aware applications," Ph.D. dissertation, Georgia Institute of Technology, 2000.

[5] M. Baldauf, S. Dustdar, and F. Rosenberg, "A survey on context-aware systems," vol. 2, no. 4. Inderscience, 2007, pp. 263–277.

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

[7] C. F. Crispim-Junior, F. Bremond, and V. Joumier, "A multi-sensor approach for activity recognition in older patients," in *The Second International Conference on Ambient Computing, Applications, Services and Technologies-AMBIENT*, 2012.

[8] C.-H. Lu and L.-C. Fu, "Robust location-aware activity recognition using wireless sensor network in an attentive home," vol. 6, no. 4. IEEE, 2009, pp. 598–609.

[9] J. Rasmussen, "Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models," no. 3. IEEE, 1983, pp. 257–266.

[10] K. Kuutti, “Activity theory as a potential framework for human-computer interaction research,” 1996, pp. 17–44.

[11] A. Kofod-Petersen and J. Cassens, “Using activity theory to model context awareness,” in *Modeling and Retrieval of Context*. Springer, 2006, pp. 1–17.

[12] M. Mikalsen and A. Kofod-Petersen, “Representing and reasoning about context in a mobile environment,” 2004, pp. 25–35.

[13] T. Gu, Z. Wu, X. Tao, H. K. Pung, and J. Lu, “epsicar: An emerging patterns based approach to sequential, interleaved and concurrent activity recognition,” in *Pervasive Computing and Communications, 2009. PerCom 2009. IEEE International Conference on*. IEEE, 2009, pp. 1–9.

[14] P. Srivastava and W.-C. Wong, “Hierarchical human activity recognition using gmm,” in *AMBIENT 2012, The Second International Conference on Ambient Computing, Applications, Services and Technologies*, 2012, pp. 32–37.

[15] E. Kim, S. Helal, and D. Cook, “Human activity recognition and pattern discovery,” vol. 9, no. 1. IEEE, 2010, pp. 48–53.

[16] U. Hunkeler, H. L. Truong, and A. Stanford-Clark, “Mqtt-sa publish/subscribe protocol for wireless sensor networks,” in *Communication Systems Software and Middleware and Workshops, 2008. COMSWARE 2008. 3rd International Conference on*. IEEE, 2008, pp. 791–798.

[17] J. Indulska, R. Robinson, A. Rakotonirainy, and K. Henricksen, “Experiences in using cc/pp in context-aware systems,” in *Mobile Data Management*. Springer, 2003, pp. 247–261.

[18] D. T. Neal, W. Wood, and J. M. Quinn, “Habits–a repeat performance,” vol. 15, no. 4. SAGE Publications, 2006, pp. 198–202.

T-Bookmarks: Providing TV-related Web Resources at Anywhere

Douglas Véras, Thiago Prota, Carlos Ferraz
Center of Informatics
Federal University of Pernambuco (UFPE)
Recife, Brazil
{dvs,tmp,cagf}@cin.ufpe.br

***Abstract*— In many parts of the world, the technologies of Digital TV and Connected TV are in a vast number of homes. Given this scenario, there are many sources of information in which the users may obtain access, especially with the union of "broadcast" and "broadband" technologies in a single TV set. Thus, this paper proposes a model (named T-Bookmarks) for convergent applications in Hybrid TV, responsible for retrieving Web resources of different categories (news, blogs, images, and others), related to programs watched by users. Also, our proposal takes into account the way resources are presented to the user, since she does not always want to visualize Web resources on the TV screen, as the experience of watching a TV program may be compromised. Therefore, that information is provided through a Web server so it may be accessed in any device connected to the Internet at any moment.**

Keywords-digital TV; connected TV; ubiquitous computing; distributed systems; recommender systems.

I. Introduction

In many parts of the world, Digital TV (DTV) is in a vast number of homes [11]. Regarding today's TV technology, two aspects may be considered: *a) transmission via radio broadcast characterized by the same information being sent to several TV receptors at the same time* and *b) the broadband technology that allows the distribution of information for the user on demand*. The latter one is known as "Connected TV", which is vastly present in the international market [5]. However, one technology does not exclude the other, and they may be complementary, according to Hybrid TV (Digital and "Connected") technology [4], increasing the interaction experience of the user and the quality of information she receives.

Nevertheless, the searching for information in these technologies is motivated by different causes. The traditional searches or recommendation for TV shows/sites on TV become linked, once the search for sites, for example, is generally oriented to tasks arising from the search for a service that attends a necessity, entertainment, curiosity, and others. Particularly, on a Connected TV, this search or recommendation for a certain content on the Web (a film, a song, Web pages, extra contents, etc.) may arise from the TV show being watched by the user who might have used a Web browser and found something of her interest, many times cluttering the continuous flow of the TV show. This fact may be noted, given the number of people that watch TV and surf the Web at the same time [7], sometimes even accessing the site of the TV show under exhibition. It reveals that TV and Web are complementary media.

Thus, there are some new challenges related to the recommendations made to the users of this new kind of hybrid TV, including: what information would be useful to the user when such information is in a broader context, not only related to a similar TV show, but also to any resource available in the Web? How are those recommendations exposed to the user, since many Web resources of different categories (videos, music, images, news, and others) may be suggested depending on the TV program the user watches? How must that information be arranged to make its access be easy and intuitive? In the last case, the displaying of results may be done in a Web page, so the user may access this information using his/her Personal Computer (PC), Smartphone, or even the Hybrid TV, through a Web browser. Therewith, access to available information on the Web, related to the TV shows watched by the user, would be done in a simpler and more direct way than if it was done searching for those resources in a conventional way through Web browsers, either on the "Connected" TV or PC.

Therefore, this paper presents a distributed system, acting as a model to convergent applications, based on Hybrid TVs, linking aspects from the DTV and Web fields. This system, called T-Bookmarks, provides recommendation of Web resources related to TV shows watched by the users. For that, T-Bookmarks architecture serves as a basis for the construction of applications responsible for recovering such Web resources, and organizing them in a Web server. Thus, the recommendation results may be accessed in any Internet device.

The remainder of this paper is structured as follows: In the next section, some related works are presented regarding to the access of additional information about TV shows, comparing them to proposed system in this paper. In Section III, the T-Bookmarks solution is presented, while in Section IV, a case study, as a proof of concept, of the system is presented under a real environment. In Section V, their implementation and profile algorithm are discussed. Section VI concludes the paper and gives hints on future work.

II. Related Works

A system that searches for Web contents-related to the TV shows and provides integration with other devices (smartphones, for example) is proposed by Patel et al. [8]. Although this system has provided many capabilities that make the search easier, such as: indication/correction of

 ISBN: 978-1-61208-289-9

terms inserted incorrectly and exhibition of contents related to the searched terms, including different categories of contents (images, videos, related news, and others), the user still need search for information or TV shows of his interest which requiring the insertion of text to determinate the searched item. This may result in a loss of the user's time when searching for contents of interest. On the other hand, a system that brings these contents in an automated manner (implicitly to the user) would make this process easier, with the option of searching the TV show that is currently in exhibition or based on the history of TV shows preferred/watched by the user. Thus, this system is proposed in this article (named T-Bookmarks).

Another work that aims the convergence between TV and Web is proposed by Dimitrova et al. [1]. In it, two recommendation systems of content related to TV shows were conceived. The first one is a recommendation system of news (MyInfo), while the second one recovers information about films (InfoSip). To access news (MyInfo), the users may select one of six categories: weather, traffic, sports, financial news, headline and local news. Besides, they may see the filmographies (InfoSip) based on the history of what is watched on the TV. By using a remote control, users may receive information related to the context of film as predefined questions such as: "Who is this actor?", "What is this song?" and "What city is this?", among others.

However, this information is not explored completely, since it is limited to certain TV shows types. Also, it does not allow the user to access resources available on the Internet, because it is restricted only to the TV device. Instead, T-Bookmarks system provides access to information (categorized Web resources) for any kind of TV show and from any device with Internet access.

Ma and Tanaka [6] present a method for dynamic integration between TV shows and Web-related contents, where a prototype system called WebTelop is described. Web pages related to the TV show's content are retrieved automatically in real time and they are used to add related extra content to TV show in the form of subtitles. During TV show transmission, a virtual agent is displayed on the screen to help the user to surf the related web pages. However, the work proposed in [6] disposes resources only based on the TV show being exhibited and to be accessed on the TV itself. Alternatively, T-Bookmarks system provides resources for the running TV show as well as for the TV shows most watched by the user. Furthermore, he may visualize the resources in a Web page, on the "Hybrid TV", tablet, smartphone, and other devices with access to the Internet. Besides that, the resources are persisted for access at any moment.

III. T-BOOKMARKS SOLUTION

Basically, the T-Bookmarks is a distributed system that aims to serve as base in the construction of applications responsible for capturing, through the usage history, which TV shows are "favorites" by the user, and after that, recovering Web resources of different categories (news, videos, images, and blogs), related to these "favorite" TV shows. Furthermore, the user have the option to recover Web resources related to the current TV show that he is watching, without requiring to generate a list of "favorite" TV shows. Figure 1 shows the general idea of the T-Bookmarks system.

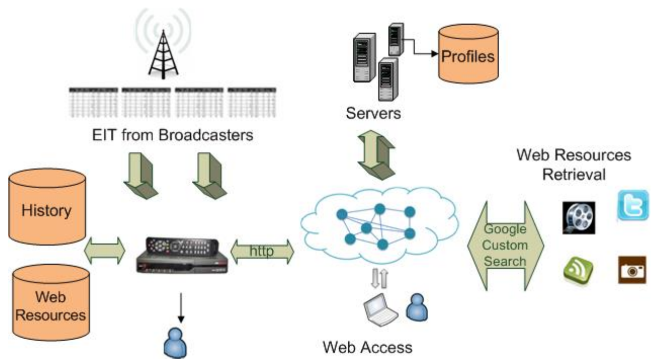


Figure 1. T-Bookmarks solution overview.

As can be seen in Figure 1, T-Bookmarks uses the EIT (Event Information Table), which are tables emitted by broadcasters together with the TV shows in order to obtain the exhibited TV show's data, such as: title, channel's name, TV show's time/date of begin/end, and others. More than the information of TV shows and programming guide, the receptor must be capable of storing the user's usage history, allowing the user profile's creation with his "favorite" TV shows being captured in an implicit manner.

From there, the system requires a search API (Web service) for Web resources according the user's profile (or search for the TV show in exhibition) since the reduced processing capability of TV devices as well as manipulation of a large mass of data and resources available on the Web. After the response of this search, based on the favorite TV shows or on the current TV show, the addresses (Web URLs) are saved in a categorized way according to their type (videos, news, images and blogs). These URLs are saved in the TV itself and, after that, they are sent to server, where the resources are provided to the user.

Finally, another important point to be highlighted is in the manner how these resources may be accessed by the user. Once these resources are in a server, they may be accessed by any platform with Internet access, including the "Connected TV" itself, at any moment.

A. Use Scenario

In this section, a T-Bookmarks use scenario is presented in order to illustrate the main functionalities provided as description of the system general idea. For this purpose, a short story is described from the point of view of the user, as follows.

"A viewer, turning on its TV, decides to watch a talk show, but in the meantime, he surfs the Internet (on PC) seeing some emails. During that TV program, he starts to search the Web for videos related to the talk show in exhibition, but he does not want to waste time searching for these videos, so he activates the T-Bookmarks by selecting the option to recommend resources related to the running TV show, and continues to read his emails. After reading some emails, he then decides to see what were the Web resources,

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-289-9

especially videos, related to the talk show. The application creates a list of Web resources related to the running TV show and makes it available on a Web server, thus it can be accessed anywhere (including his PC). By accessing the list, the user finds some interviews that had not watched yet and decides to access them on his computer. Among these TV shows, the viewer watches some ones for a short time, while he watches other ones completely, and so on. On the weekend, he decides to surf the Internet looking for varied content, nothing very objective, but related to the TV shows he more liked, which also was inferred implicitly by the T-Bookmarks. Then the user opens the application and requests resources related to his favorite TV shows. Hence, he accesses a resource list on a Web page by selecting blogs related to one of its favorite TV shows by visualizing them on its own TV via Web browser."

B. Architecture

T-Bookmarks is a distributed system based in the client-server model. The client model has an architecture divided in layers and makes requests to a Web service [3] in order to find Web resources (a list of Web addresses – URLs). In the following sub-topics, the T-Bookmarks Client and Server modules are presented, respectively.

1) Client Module

The Client module's architecture (see Figure 2) is divided in three layers: Model, View and Control (MVC). The inferior layers provide services for the adjacent superior layers through the facade present in each one. Below, the essential components of each layer are described.

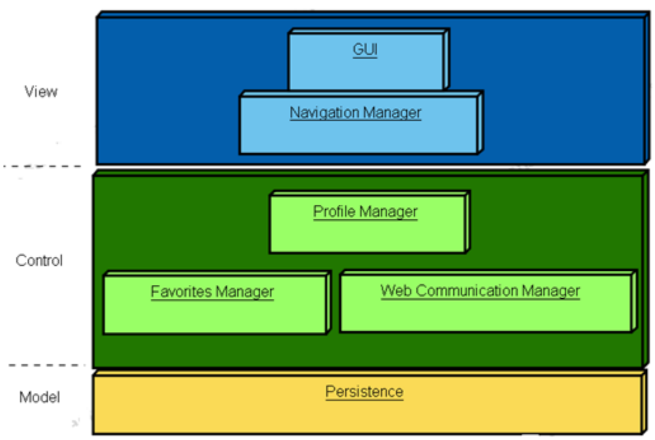


Figure 2. T-Bookmarks client module architecture.

The View layer is responsible by interface with the user. It is used to receive user's input data or requests and to present the result of the interaction with the system to the user. In the T-Bookmarks, there are two main components situated in this layer as described in the following:

- GUI – this component is responsible for exhibiting the recommendation options of Web resources ("Current" or "History" based) to the user as previously mentioned in the use scenario (Section III.A). Furthermore, it shows to the user, his unique identifier, that will be utilized to access the Web page containing the recommendation results;
- Navigation Manager – this component is responsible for providing navigation between recommendation options (exhibited by GUI) to the user through remote control. Besides it is responsible for passing the user's choices to the system. When the user selects an option (pressing the "ENTER" key from remote control), the Navigation Manager forwards this information to the Profile Manager component, in the Control layer, and then it sends a feedback message to the user through GUI, so he is aware that the desired option was selected.

The Control layer, situated between the View and Data (Model) layers, defines the core behavior of the system, interpreting the user's actions and determining what the system should do. Three main components are in this layer, as briefly described in the following:

- Profile Manager – this component interprets/receives the requests from Navigation Manager component. Furthermore, it is the Client main component, responsible for: *a) managing the user's profile through watched TV shows*, and accessing services in the Model layer; *b) requesting the Web resources to the Web Communication Manager component* according the selected option by the user ("Current" or History"); and *c) extracting the basic TV show's data such as*: title, channel name, duration time, among others, and transforming them in data structures represented by "business" objects of the application like: user profile, search criteria, and others;
- Favorite Manager – this component is required by the Profile Manager with objective of calculating the TV show "score" in order to classify it as a favorite TV show or not. The Favorite Manager component also creates the list of favorite TV shows, ordering the TV shows with highest score first. In this component, a profiling algorithm is implemented (proposed in [10]), based in the time that the TV show is watched; and
- Web Communication Manager – this component is responsible for accessing services that are out of the Client module (available on the Web). Its objectives are: *a) recover Web resources:* through an API (Google Custom Search API [3]) according each resource category - Videos, News, Images and Blogs. This Web API invokes functions of this component when a request is completed (callback). Thus, the Web Communication Manager accesses services from Model layer in order to persist the search results, distinguishing Web resources from "Current" or "History" recommendation; *b) communication with Server module:* sending the Web resource lists so that they can be accessed through the Web page provided by the Server module and retrieving the user identifier, generated in the Server from first time that the application is load. This identifier is sent to the GUI component in

order to the user can use it to access the Web resources.

The Data layer (Model) models the data behind the "business" layer. Thus, it is only responsible for storage, manipulation and generation of data. In the architecture of the T-Bookmarks, the Persistence component offers services for persistence of data from the user profile, with his watched and favorite TV shows as well as services for persistence of Web resources as result of user requests.

2) Server Module

The Server module architecture follows the MVC pattern (similarly as Client module architecture), as shown in Figure 3. This architecture is directed basically to receive Client module data and to save them in a database, as well as receiving requests from the users through Web browsers in order to show them the persisted data.

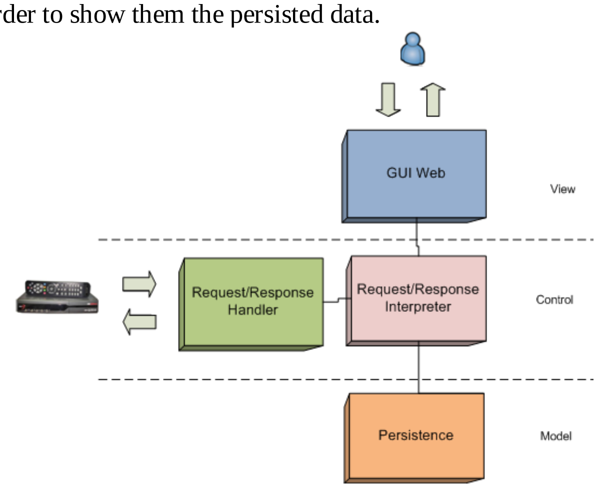


Figure 3. T-Bookmarks server module architecture.

As can be seen in Figure 3, the Server module has three layers composed by four main components according to the following descriptions:

- Request/Response Handler – this component is common to the majority of Web servers, being responsible to receive the requirements of services, returning the results through a response. In this case, servlets are utilized as entry port for requests from the Client module that communicates with the server for: *a) Obtaining identifier for posterior access of the recommended data*. This identifier is unique for each application (not user) and generated automatically by the server in the Requirement/Response Interpreter component; and *b) Sending the Web URL recovery through the API:* in order to make it available for access through Web page;
- Request/Response Interpreter – this is the main component of the server, responsible for receiving the requests from the Client module as well as the Web interface (GUI), interpreting them and taking the correspondent actions. After these actions are taken, responses are sent back to the Client module or user by accessing the Web page. According to the respective requests, the functionalities provided to the Client module are: *a) Generation of unique identifier (ID) for each application:* where once time loaded, the application requires the identifier and after server response, saves the ID locally, similar to a Web cookie; and *b) Interpreting data sent by the Client module:* according user's ID, kind of resource (images, videos, news or blogs) and Web recovery mode (based on the user's usage history or on the running TV show). Beyond interpreting actions from the Client module, this layer is also responsible for interpreting user actions through GUI Web. These actions are: requesting the exhibition of Web resources having as entry an identifier; selecting Web resources categories for exhibition in the browser itself. For this reason, the Requests/Response Interpreter accesses the data saved in Persistence component and returns the Web resources to GUI Web component;
- GUI Web – this component belongs to Presentation layer. It is responsible for showing to the user the Web resources sent from the Client module, according to the user ID. For each identifier, the GUI Web presents a list of resources solicited by the user in the Client module in a categorized way: by Web resources related to the running TV show, and, by Web resources related to the usage history, both cases are classified according to kind of resources: images, videos, news, and blogs. In addition, the user can type the ID, through a search box, in order to retrieve his desired Web resources. So, with a list of available resources he can select a resource which will be opened on same page or in other page, as the user wishes; and
- Persistence – this component, representing the Data Layer, provides storage and searching functions through files. Among these files are: files with the last ID of a registered application, serving for querying as well as storage (when it is necessary increase it). In addition, this layer offers to reading/writing in files with the user's Web resources, persisted separately according user's ID, their Web resource category (images, videos, news or twitters) and their recommendation option ("current" or "history").

IV. Case Study

As a proof of concept, in this section, an experiment performed in the Brazilian System of Digital TV (SBTVD) [2] under a real environment is described. For this, a Smart TV Samsung [9] was used, with Internet access, where the T-Bookmarks client application was installed. Moreover, the server application was executed in a computer from the same network (intranet) of the TV.

As it was previously described, every client application on the TV has an identifier generated automatically through a request to the server when the application is loaded at first time on TV. If this number is generated correctly, no alert is displayed on the screen and the user can select one of the options using the remote control (navigation by arrow keys

"UP" and "DOWN", pressing "ENTER" on the desired option). When one option is selected, a feedback message is sent to the user becoming him aware of the selected option, while he is waiting the system response. We recall that the application is executed in "background", where the user profile is implicitly created based on the TV shows which he watches.

Figure 4 shows the response obtained when the user chooses the "Current Recommendation" option, where resources are recovered and sent to the server application. Hereafter, according to this response, the data can be accessed on the server's website through the "4" identifier (generated and showed on TV). Thus, this page may be accessed by a browser on PC, smartphone, the Samsung Smart TV, or any other device with Internet access.

Figure 4. User selects the "Current Recommendation" option.

Likewise, the user can select the "History Recommendation" option. Then the T-Bookmarks retrieves Web resources related to his favorite TV shows and shows to the user the same response on the TV screen (confirmation of the data sent to the server, with his application ID).

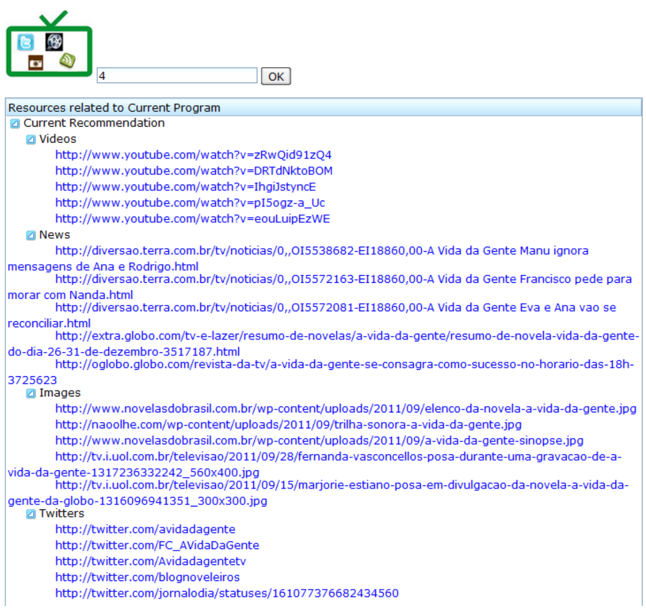


Figure 5. Recovery results of web resources on web page.

The result of these requests may be viewed through the Web page available on the server. In the first phase of this experiment, a user watched the "A Vida da Gente" TV show at the "Globo HD" channel and he selected the "Current Recommendation" option; thus, the T-Bookmarks has generated a list of resources available on the web page associated with the ID "4", as seen in Figure 5.

In the second phase of this case study, a user watched TV shows in the created environment during a period of one week (Sunday to Saturday). In this period, the T-Bookmarks has collected information about which TV shows has been watched and their respective scores according "watched" timing (in minutes) relative to TV show total duration, as shown in Table 1.

TABLE I. LISTING OF TV SHOWS WATCHED WITH THEIR RESPECTIVE SCORES.

TV Show	Channel	Duration	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Satuday	Total Score
Globo Esporte	Globo HD	30	0	0	0	15	0	5	0	8,365
Jornal Hoje	Globo HD	30	0	30	30	30	20	5	0	31,685
Video Show	Globo HD	35	0	0	0	0	20	5	0	8,454
Mulheres de Areia	Globo HD	100	0	90	90	90	90	55	0	37,679
Globo Notícia	Globo HD	3	0	3	3	3	0	3	0	28,2
Malhação	Globo HD	30	0	30	30	30	0	30	0	30
A Vida da gente	Globo HD	50	0	20	20	0	50	50	20	29,648
Fina Estampa	Globo HD	80	0	40	40	40	80	60	60	37,840
Jornal da Globo	Globo HD	30	0	0	0	0	30	30	0	15
Glee	Globo HD	50	0	0	0	0	50	0	0	7,833
Vidas em Jogo	Record HD	45	0	45	45	45	45	45	0	38,75
Rei Davi	Record HD	60	0	0	50	0	50	0	0	14,342
Dr. House – (Season 5)	Record HD	60	0	0	0	0	20	0	0	4,138
Fantástico	Globo HD	140	60	0	0	0	0	0	0	4,478
Altas Horas	Globo HD	120	0	0	0	0	0	0	30	2,968

Also, a list of favorite TV shows of this user was created according the first three TV shows "most favorite", in this order: "Vidas em Jogo" ("Record HD" channel) with score 38.75, "Fina Estampa" ("Globo HD" channel) with score 37.84 and "Mulheres de Areia" ("Globo HD" channel) with score 37.679. Thus, these three TV shows had related Web resources recovered and saved on the server, where the user could access them, similarly to what occurred in the recommendation from the "Current Recommendation" option (shown in Figure 5).

V. DISCUSSION

In this section, the main implementation aspects of the T-Bookmarks are discussed, as well as the use of the algorithm proposed in [10].

A. *Implementation*

The client-server architecture definition, for this novel design of recommender system cross-domain (TV and Web), is necessary due to computational limitations of TV sets and the mass of data to be fetched (any resource on the Web). Thus, T-Bookmarks uses a search service available in the Web cloud, removing the complexity of its implementation from client application.

As it was described previously in this paper, there are two options for retrieving Web resources provided to the

 ISBN: 978-1-61208-289-9

user: *a) related to the TV show in exhibition;* and *b) related to user's favorite TV show list (implicitly generated)*. For this, some data from EIT tables regarding to the TV shows are needed. However, few data are extracted from these tables compared to the amount of data available, so future implementations could utilize more data, such as the TV show "category", in order to bring more relevant results to the recommendation. Furthermore, the use of data from the broadcasters may not always be sufficient, even the possibility of errors by them, which directly affect the user profiling, implicitly made by the application. Thus, in order to alleviate this problem, consultation services for TV shows information on the Web may be used, for example.

Currently, T-Bookmarks searches for Web resources belonging to four categories: Video, News, Images and Blogs. However, new categories could be added to the client-server system, even using the same search API currently used, as well as through different search engines, which could also be included in the proposed architecture.

Another aspect that could be changed, according to the requirements, is the amount of recommended resources, since it is currently limited to five resources by category (Video, News, Images, and so). This was defined in order to provide the user only the most relevant Web resources by the Google Search API. Furthermore, the amount of favorite TV shows to be used in the search results is limited to only three (the first three programs in the favorite list). Multiplying these factors, the T- Bookmarks provides to the user sixty Web resources in the request for the list of favorites, while the search based on the running TV show retrieves twenty Web resources.

Furthermore, the recovery of Web resources was based on a single technique of recommender systems: based on user history, determining its favorites TV shows list implicitly. However, other techniques (ontologies, context-aware, collaborative filtering, among others) could be used to improve the relevance of the resources available to the user as well as other algorithms may be used to generate the user profile implicitly, different from the algorithm used currently [10].

Finally, the use of an identifier does not allow the installed application can be used by different users. For this, an authentication mechanism would be necessary, that it could hinder the T-Bookmarks use, because the user would need to register and login to each application use. Moreover, the application could be personalized in many ways such as: the choice of the quantity of resources searched; the choice of resource categories shown; among others.

B. Recommender Algorithm

According to the implementation of the algorithm proposed in [10] and the result of the case study presented in Section IV (see Table 1), the three TV shows with the highest score were used to search for Web resources when the user selected the recommendation based in favorite TV shows on the T-Bookmarks GUI.

The calculation of the TV show score takes into account its relative time watched and total duration. As we have seen, the "Globo Notícia" ("Globo HD" channel) has duration of three minutes while the "Malhação" ("Globo HD" channel) TV show has duration of thirty minutes. In both cases, the user watched the entire TV show on the same days and the same number of times each, but in the first case the score was 28.2, while in the second one the score was 30. This is a result of factors *Wf(d(c))* and *Se(d(c))* [10], which increase the score of programs with higher durations.

However, this algorithm does not take into account the amount of times a TV program is showed during the week, so these TV shows hardly enter into the top of the list of favorites, as it is the case of "Fantástico" and "Altas Horas" programs, which are showed only once a week; "Rei Davi", which is showed twice a week, among others. Furthermore, TV series like "Glee" and "Dr. House" showed only in one or two days a week should have more "weight" in their scores because they are TV shows with certain "inter-dependence" between each episode.

VI. Conclusion and Future Work

This paper has presented a distributed system, called T-Bookmarks, as a proof of concept, based on a client-server architecture, which captures which TV shows are the user's "favorites" (implicitly) and then retrieves Web resources from different categories related to these "favorite" TV shows. In T-Bookmarks, recommendation results are available in a Web page so the user can access them on any device with Internet access. Thus, users do not have to spend much time searching for Web resources related to the TV shows they are watching as this task is performed by the system. Finally, this way of viewing also allows the user to see the Web resources when he feels needed (any time), since the Web resources are saved in a Web server.

As future work, we plan to perform tests with multiple users to enhance the prototype according to the needs raised in the experimental tests. These tests aim to assess: *a) the prototype interface (GUI)*; *b) the algorithm for calculating the implicit user profile*; *c) the relevance of Web resources presented to users according to their watched TV shows.*

References

[1] N. Dimitrova et al., "Content augmentation aspects of personalized entertainment experience.", Procs. of the 3rd Workshop on Personalization in Future TV, 2003, pp. 42-51.

[2] Forum SBTVD, available at http://forumsbtvd.org.br, retrieved: July, 2013.

[3] Google Custom Search API, available at https://developers.google.com/custom-search/, retrieved: July, 2013.

[4] HbbTV, ETSI Technical Specification 102 796, "Hybrid Broadcast Broadband TV", available at http://www.hbbtv.org/pages/about_hbbtv/specification.php.

[5] R. Jacoby, "Why Connected TV Is Poised To Revolutionize Entertainment", available at http://mashable.com/2011/05/11/future-connected-tv/, retrieved: July, 2013.

[6] Q. Ma and K. Tanaka, "WebTelop: dynamic TV-content augmentation by using web pages.", Proc. of IEEE International Conference on Multimedia and Expo (ICME), 2003, Vol.2, pp. 173-176.

[7] N. Wire, available at http://blog.nielsen.com/nielsenwire/online_mobile/40-of-tablet-and-smartphone-owners-se-them-while-watching-tv/, retrieved: July, 2013.

[8] M. Patel et al., "Google TV search: dual-wielding search and discovery in a large-scale product.", Procs. of the 1st Intl. Conf. on Designing Interactive User Experiences for TV and Video, Publisher: ACM, 2008, pp. 95-104.

[9] Smart TV Samsung, available at http://www.samsung.com.br/smarttv/, retrieved: July, 2013.

[10] D. Weis et al., "A user profile-based personalization system for digital multimedia content.", Proceedings of the 3rd Intl. Conf. on Digital Interactive Media in Entertainment and Arts, 2008, pp. 281-288, publisher: ACM.

[11] World media televisions per capita (Factbook), available at http://www.nationmaster.com/graph/med_tel_percap-media-televisions-per-capita, retrieved: July, 2013.

Ubiquitous Learning – Treating the Context of Students Learning Styles

Andreia Rosangela Kessler Mühlbeier, Felipe Becker Nunes, Gleizer Bierhalz Voss,
Rafaela Ribeiro Jardim, Roseclea Duarte Medina

PPGI - Programa de Pós-Graduação em Informática
UFSM - Universidade Federal de Santa Maria
Santa Maria - Brasil
{dehiamuhlbeier, nunesfb, gleizer.voss, rafa.rjardim, roseclea.medina}@gmail.com

***Abstract*— The increasing use of mobile devices and the wide dissemination of learning environments encourage educators to find methods that involve mobile and ubiquitous computing. This scenario allows the development of environments geared to teaching, in which characteristics of the context information from the student should be treated apart. Thus, it arises a new paradigm, ubiquitous learning, and enabling differentiated instruction. This article presents a survey that has founded, by analyzing evidential theory and practice, the similarity between the different dimensions of learning styles, presenting an adaptation of the ubiquitous environment seamlessly with the style of the predominant user.**

Keywords-learning styles; ubiquitous learning; mobile computing.

I. INTRODUCTION

Technological development boosts the appearance of several tools that give students and teachers the ability to create various teaching methodologies in virtual learning environments, these in different ways to compose and accommodate formal education immersed in the educational environment.

However, by using new methodologies and aiming access everywhere at every moment, facilitates the unification of resources and experiences that are fundamental to identify the learning styles of the students, thus enabling the building of a customized environment, which can contribute to achieving good results in the educational process.

Mobile learning (of English Mobile Learning or m-Learning) allows learning to occur anytime and anywhere. However, despite providing mobility, it does not provide an apprenticeship able to consider the context-sensitive information of the student [1].

In this sense, through studies listed on the work of [2] and [3], there are some important questions: the use of various styles in the detection of student's profile can be an obstacle in the definition of his way to learn? Which style would be adopted in the adaptation of a virtual learning environment?

This article presents a survey that shows by evidential theory and practice analysis the similarity between the different dimensions of learning styles, presenting an adaptation of the ubiquitous environment seamlessly with the style of the predominant user.

This paper is organized as follows. In Section II, we introduce the theory base for ubiquitous learning. In Section III, we introduce the study of learning styles. In Section IV, we introduce the results and discussions. In Section V, there are partial considerations and at the last Section are presented the references.

II. UBIQUITOUS LEARNING

Ubiquitous learning or u-learning is viewed these days as a new learning paradigm. This expansion of previous learning paradigms goes from conventional learning to e-learning (e-learning), as well as from e-learning to mobile learning (m-learning) and currently, called u-learning [1].

From Yahya et al. [4], the definition for ubiquitous computing environment is "an area that incorporates a set of embedded systems (computers, sensors, user interfaces, and service infrastructure), which is enhanced by computing communication technologies."

However, to provide customized environments according to the profile of each student, it is necessary to collect information from the technological context in which each is located. According to Gasparini et al. [5], in order to be considered U-learning a system must meet at least two requirements:

- In must capture the context of the student without him noticing through devices or command lines;
- It must adapt the interface, content, presentation or navigation considering the user's profile.

In turn, Quinta [6] points out in his system for capture of context that, first, must be considered "for whom" will be held this adaptation, bringing the capture of the user's context through the standards of the World Wide Web Consortium (W3C) XML reading. In second place, it is seen "where" content will be adapted that could happen on the server, the client or intermediary agent (proxy). The next stage acts "when" this adjustment is made, before the requisition, or during the two cases, and thus can take better advantage of each adaptation depending on the media file.

 ISBN: 978-1-61208-289-9

Then, it is chosen "**what**" will be adapted and if indeed this adaptation is necessary and lastly "**how**" will be realized this conversion.

III. Learning Styles

Learning may be conceptualized as a natural process in the life of every human being and completed in adulthood. People of the same class, age, nationality, race and / or religion prefer to learn in different ways, in other words, they have their own preferences at the time of learning.

Is important to know how to identify the learning styles of students, since the more strategies he has developed, the chances of developing various forms of presentation of information in learning situations experienced by him will be greater [7]. Thus, the teacher is able to plan and encourage students to develop their skills according to their learning skills.

Learning styles are divided into dimensions in order to detect the prevalence of individual approaches of learning. A large number of them are available in the literature, with a tendency to merge styles presented by each author in their models.

According to Buther [8], "learning style is the consistent and personnel way in which people use their qualities and skills to define themselves, to find, evaluate and process information."

According to Honey and Mumford [9], "It is admitted also that each person is able to identify the characteristics of various learning styles, although generally, each person has a dominant style."

The author Felder and Soloman [10], understands the learning styles as preferences and dominant features in the way that people receive and process information. For him, "learning styles are skills that can and should be developed in the subject."

Already the author Kolb [11], "Notes that each individual develops a particular learning style, giving priority to certain skill."

According to Bariani [12], "examines that the forms are stable in relation to the characteristics of the cognitive structure of a person and which are modified from the direct or indirect influence of new events through culture and relate data of reality elaborating conclusions about them".

Table 1 shows the main characteristics associated with each of the styles.

The learning style refers to the preferences of the subject in the learning process. Based on the concepts of authors on learning styles, you can identify your own style.

IV. Discussion and Results

The study consisted of a videntiary-theoretical – practical analysis of the similarity between different dimensions of learning styles in order to prove that a person can have interwoven styles, as well as detect which dimensions are most suitable to use in a virtual learning environment.

The developed research integrated a master's dissertation entitled "Adaptation of Mobile Learning Engine Moodle (Moodle Mle) to different cognitive styles by using adaptive hypermedia" [2] in which the mobile MLE Moodle VLE it was adapted to different learning styles using adaptive hypermedia techniques. It was created and applied an Expert System for diagnosing styles, denominated SEDECA. The learning style was identified through the Instrument "Questionnaire" based on the research instruments proposed by [9], [12], [13] and [14]. The author Mozzaquatro [2], worked with four dimensions of learning styles: Reflective, serious, Holistic and Divergent in order to support the modeling process and implementation of the AVA to be adapted. The questionnaire was given to students of Undergraduate and Graduate distance learning mode.

The second work integrated to the research was the thesis entitled "Learning Styles and Strategies customized to students the methods and distance" [3] that addressed seventeen dimensions of learning styles, implementing a computer system that detects preferred styles and recommends different learning strategies to present educational materials and teaching resources according to the individual preferences of students. A system of objective questions was created: The SDLS (Learning Styles Detector System) consisting of a "Questionnaire", composed by sixty-one issues based in the research instruments proposed by [8], [9], [10], and [11]. The result of the test identified preferences regarding the following dimensions: Accommodating, Analytical, Assimilating, Active, Convergent, Divergent, Global, Intuitive, Personal, Pragmatic, Realistic, Reflective, Sensory, Sequential, Theoretical, Verbal, Visual.

The described researches have been validated by teachers and students of Graduate and Postgraduate modalities and distance education. After analysis of the characteristics, preferences and state of the different dimensions of learning styles, can confirm the similarity between of the aforementioned models.

The study shows that the styles are interwoven, and every person has more than one learning style, but each person has only one style of predominance, as shown in Figures 1 and 2.

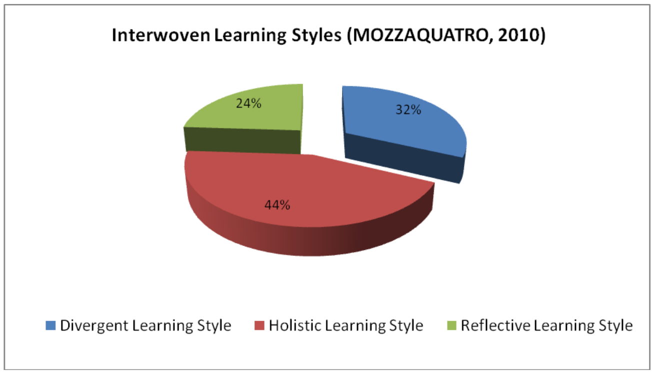


Figure 1 – Interwoven Learning Styles: Divergent – Holistic – Reflective.

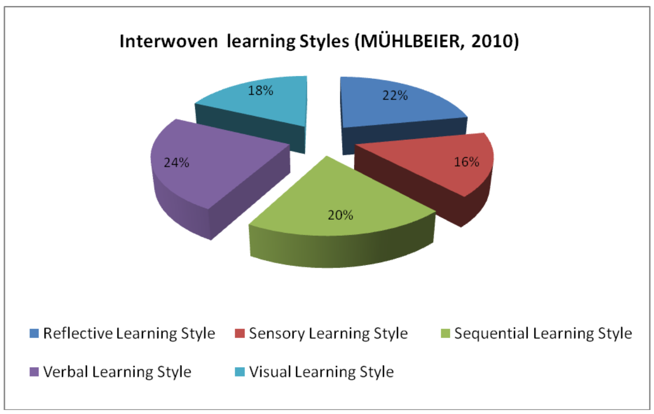


Figure 2 – Interwoven Learning Styles: Reflective- Sequential – Sensory – Verbal -Visual.

As shown in Figure 1, three styles have been interwoven, in other words, the results of the population on the survey showed by Mozzaquatro [2], demonstrate that subjects tended to predominate in the following styles: Holista (44%), Style Divergent (32 %) and reflective style (24%).

Figure 2 illustrates the analysis of the research developed by Mühlbeier [3], that presents the chance of the student to present learning style interwoven in which one style becomes predominant. It can be observed that the population interviewed presented five interwoven learning styles: Verbal (24%) Reflective (22%), Sequential (20%), Visual (18%) and Sensorial (16%).

As seen in Figure 3, the learning style Global approached by Felder and Silvermann [11], presents as the vision the whole approach. Compared to the learning Holistic style, integral of the Bariani Model [12], it was found that both have similar characteristics, with emphasis on the global context. Based on this assumption it is conclude that an individual with the Holistic Learning Style has the same preferences as a Global Style one.

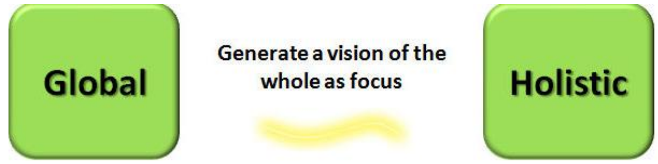


Figure 3 – Similarity Global and Holistic.

Figure 4 – Similarity Intuitive and Active.

Figure 4 illustrates the intuitive and active learning styles and members of the Felder and Silvermann Model [10], which mention individual characteristics motivated by new experiences and innovations. Concluding that people with such characteristics can suit both dimensions.

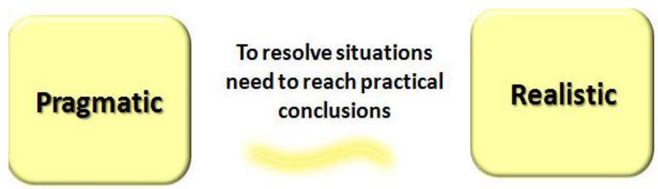


Figure 5 – Similarity Pragmatic and Realistic.

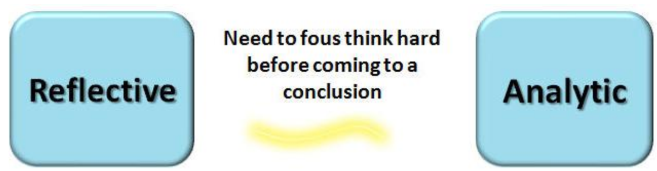


Figure 6 – Similarity Reflective and Analytic.

The realistic style of the Buther integral Model [8], presents itself as characteristic to resolve practical situations. The Honey and Mumford Model [9], shows the pragmatic dimension that has experience and theories applied to practice. It was noted that both dimensions have similarity between them, as shown in Figure 5. In Figure 6 we observe that Buther's Analytical Model dimensions and Reflective Model Honey and Mumford, have as a common characteristic "focus", meaning that reflective or analytical individuals think hard before reaching a conclusion.

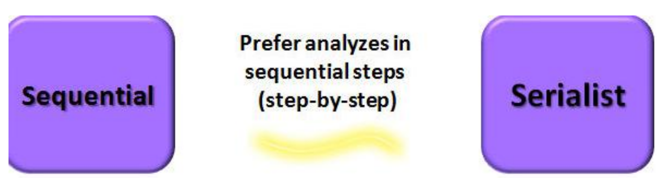


Figure 7 – Similarity between Sequencial and Serialistic.

As shown in Figure 7, the sequential and serialist learning styles integrates Felder - Silvermann and Bariani

 ISBN: 978-1-61208-289-9

models present as individual preferences analysis in sequential steps, in other words, a step by step. Characteristics prevalent in individuals who are also sequential serealists.

After analyzing the similarities between all surveyed dimensions, it was proven that to adapt a virtual learning environment, it is necessary to build a system that detects learning styles using a wide range of dimensions, since both have similarities, influencing positively the relation cost-benefit.

V. Conclusion

Virtual learning environments haves been lately focused to the functional aspects, looking to evolve through recommendations to become closer to classroom teaching with the use of mobile devices. A ubiquitous learning environment is able to meet the needs of students, exploring their preferences of learning styles.

It is believed that the presented research achieved its objectives, contributing to a development of the adaptation of virtual learning environments for different learning styles. The main contribution of this work was to prove the similarity between the different dimensions of learning styles addressed by the authors of the studied models, so there is no reason to use a wide range of dimensions of learning styles.

Based on the developed research, it was possible to verify the existence of interwoven styles, in other words, every person has predominance in one dimension, but integrates others in lesser extent. The study shows that diagnosing learning styles is a complex task, and requires analyzes in order to offer to the student an individualized learning experience, showing them the information in a personalized and targeted way. The identification of prevalent learning styles now has a key role on enabling more consistent educational practices with actions that prioritize autonomy and cooperation in the process of teaching and learning environment by adapting the student's predominant ubiquitous model.

References

[1] K. Mandula, S. R. Meda, and R. Kambham, "Implementation of Ubiquitous Learning System Using Sensor technologies , In: IEEE International Conference on Technology for Education, Hyderabad, India: IEEE Computer Society, pp. 142-148, 2011.

[2] P. M. Mozzaquatro, "Adaptation of *Mobile Learning Engine Moodle* (MLE *Moodle*) to different cognitive styles by using adaptive hypermedia", Dissertation - Federal University of Santa Maria (UFSM), 2010.

[3] A. R. K. Mühlbeier, "Personalized Styles and Strategies Learning to students from the classroom and distance teaching", Conclusion Work Course - University of Cruz Alta (UNICRUZ), 2010.

[4] S. Yahya, E. A. Ahmad and K. A. Jalil, University Teknologia MARA, "The definition and characteristics of ubiquitous learning: A discussion", International Journal of Education and Development using Information and Communication Technology (IJEDICT), vol. 6, Issue 1, 2010, pp. 117-127.

[5] I. Gasparini, A. P. Kemczinski and M. S. Oliveira, "User model sensitive to the cultural context in an adaptive e-learning system", Computers in Education: Theory and Practice, vol. 14, n. 1, UFRGS, 2011, Available: <http://seer.ufrgs.br/InfEducTeoriaPratica/article/view/21974> Retrieved: April, 2013.

[6] M. R. Quinta, "Adaptation of content to multiple contexts - ODIN system", Federal University of Goiás - Institute of Informatics, In: 2011, Available: <http://www.inf.ufg.br/mestrado/sites/www.inf.ufg.br.mestraof/files/uploads / dissertations / MarceloQuinta.pdf>. Retrieved: May, 2013.

[7] L. C. B. Cavellucci, "Learning Styles: in search of individual differences", 2006, p. 10-12. Available: <http://www.iar.unicamp.br/disciplinas/am540_2003/lia/estilos_de_aprendizagem.pdf>. Retrieved: April, 2013.

[8] K. A. Buther, "Learning Styles: the dimensions psychological, cognitive and affective", Translated by Renata Costa de Sá Bonotto and Jorge Alberto Reichert. Porto Alegre: Editora UFRGS, 2003.

[9] P. Honey, A. Mumford, "The Learning Styles helper's guide", Maldenhead Berks: Peter Honey Publications, 2000.

[10] R. M. Felder and B. A. Silvermann, "Index of Learning Styles Questionnaire". Department of Chemical Engineering North Carolina State University Raleigh, NC 27695-7905, 1993. Availeble:<http://www.engr.ncsu.edu/learningstyles/ilsweb.html>. Retrieved: April, 2013.

[11] D. A. Kolb, "Experiential Learning: experience as the source of learning and development". Englewood Cliffs: Prentice-Hall, 1984.

[12] I. C. Bariani, "Cognitive Styles of College and Initiation Scientific", Campinas UNICAMP. Thesis (Ph.D. in Education), Faculty of Education, University of Campinas, 1998.

[13] R. M. Felder, "Reaching the Second Tier: learning and teaching styles in college science education", Journal of College Science Teaching, vol. 23, n. 5, pp. 286-290, 1993, Available: <http://www.ncsu.edu/felderpublic/Papers/Secondtier.html>, Retrieved: April, 2013.

[14] M. Geller, "Distance Learning and Cognitive Styles: building a new look at virtual environments", Porto Alegre: UFRGS, Thesis (Ph.D. in Computer Education), Graduate Program in Computer Education, Federal University of Rio Grande do Sul, 2004.

TABLE I. DIMENSIONS OF LEARNING STYLES.

	Dimensions	Characteristics
Butler Model	Realistic	Members with learning style Realist see the world as a place orderly and predictable facts, actions and results. They are guided by experience; follow guidelines and make sure that others are aware of the actions that practiced; dislike change for change or fix what is not in trouble.
	Analytic	The members of the analytical learning style perceive the world as a logical system that can be understood through analysis and constant study, have a theory for almost everything, like the scientific method, technical information and evidence.
	Pragmatic	Pragmatists members see the whole world to the parts and the parts to the whole world, simultaneously, have the ability to adapt and adapt; attribute the same weight to facts and values, formulate strategies and tactics to make things happen; suffer less with the incompatibilities of styles.
	Personal	Members with personal style have the following characteristics: to understand the world as a place where harmony is essential and the prevalence is up to the well. These guys love to be seen as helpful, supportive, open and trustworthy; welcome multiple points of view and then assimilate.
	Divergent	The divergent perceives the world as a place of infinite possibilities, where imagination and experimentation combine themselves to test and find out what may prove to be, like change, dissent and novelty seeking.
Honey and Mumford Model	Active	They prefer new experiences, have an open mind, are enthusiastic about anything new, are sociable and engage constantly with others, try to be the center of all activities, are interested in challenges and problematic situations, manifest strong implication in action.
	Reflective	The prioritize members observe before action, like to observe the experiences of diverse perspectives, focus on reflection and meaning making, gather information both from their own experience as the experience of others, they prefer to think before coming to any conclusion.
	Theoretical	Tend to establish relationships, deduce, integrating the facts in coherent theories, tend to be perfectionists, they like to analyze and synthesize. Their approach to problems are consistent and logical. They seek rationality and objectivity, they are uncomfortable with subjective conclusions, lateral thinking or surface appearance.
Felder and Silvermann Model	Sensory	Sensory members enjoy learning facts; solve problems with established methods without complications and surprises, they are more finicky, and do well in practical work (laboratory, for example).
	Intuitive	Members of intuitive group prefer to discover possibilities and relationships; enjoy newness and get bored with repetition, they feel more comfortable to deal with new concepts, abstractions and mathematical formulas, are faster and more innovative work.
	Visual	Members of the visual group remember more of what they see - pictures, diagrams, flow charts, films and demonstrations.
	Verbal	Members of this group take greater advantage of verbal words - written or oral explanations and mathematical formulas.
	Active	Active members tend to comprehend and retain information more efficiently discussing, applying concepts and / or explaining to others, enjoy working in groups.
	Reflective	Reflective members need time to reflect on the information received; They prefer individual jobs.
	Sequential	They prefer to learn sequential linearly, logically sequenced steps in; tend to follow logical paths to find solutions.
	Global	Global members tend to learn at random, forming a vision of the whole, are able to solve complex problems quickly, but have difficulty explaining how they did.
Kolb Model	Divergent	Divergent members prefer to learn by concrete experience and reflective observation. They show up skilled in situations that require new and creative ideas, are able to analyze situations from different points of view and relate them in an organized whole, comprise people. Question feature WHY?
	Assimilating	Assimilators members learn by reflective observation and abstract conceptualization. They use inductive reasoning, respond to information in a logical way, when they have time to reflect, they have facility to create abstract models and theoretical and they do not worry about the practical. Question feature WHAT?
	Converging	Convergent members use deductive reasoning, with practical application of ideas, learn by trial and error, are skilled to solve problems and make decisions. Question feature HOW?
	Accommodating	Members of these group prefer this style based on learning and active experimentation in concrete experience, adapt to immediate circumstances; like challenges, act more by feeling than by logical analysis. Question feature THAT WAY?
Bariani Model	Reflective	They reflect before taking a particular course of action, have more organized thoughts. People whose thoughts are more organized, sequenced and that are weighted thoughts, previous response, reflexives are considered.
	Serialist	The serialist people are tighter focus on separate topics and logical sequences, searching later, patterns and relationships in process to confirm or not their hypothesis. They choose simple hypothesis and a logical-linear (a chance to another step-by-step). They are often good and skilled analysts on solving problems.
	Holistic	The holistic group gives greater emphasis to the global context, from the beginning of a task, examining preferences in a large amount of data looking for patterns and relations between them. They can solve complex problems quickly or put together things and are often good with synthesis.
	Divergent	The members of divergent group perceive the world as a place of infinite possibilities, where imagination and experimentation combine to test and find out what may prove to be, like change, dissent and novelty seeking.

Source: Adapted from [2].

Treating Context Information in a Ubiquitous Virtual Learning Environment (UVLE^{QoC})

Application of metrics for Quality of Context (QoC)

Felipe Becker Nunes, Gleizer Bierhalz Voss, Andreia Rosangela Kessler Mühlbeier, Fabricio Herpich, Roseclea Duarte Medina
PPGI - Post graduation Program in Informatics
Federal University of Santa Maria (UFSM)
Santa Maria, RS
(nunesfb, gleizer.voss, dehiamuhlbeier, fabricio.herpich, roseclea.medina) @gmail.com

***Abstract*—Context information captured by computer applications, assists in characterizing the momentary situation in which the user is located. However, during this construction process it must be taken into account not only the context that is used, but also the Quality of the Context (QOC) to be supplied in order to provide a real contribution to user interaction with the environment. This paper presents a study on the development of a ubiquitous virtual learning environment (UVLE^{QoC}), which performs the processing of context information by applying metrics of Quality of Context. The environment also integrates two systems: U-SEA to treat computational context and SEDECA to identify the cognitive profile of the user, adapting the contents and tools of the environment according to the formulated context. The environment will provide a context to guarantee a minimum level of quality, adapting their resources according to the preferences of the users.**

Keywords- context; quality of context; ubiquitous; moodle

I. Introduction

Given this global scenario that is constantly evolving, it is possible to notice changes in several areas; among them is the education, in which there was a paradigm shift from traditional learning, exclusively in face-to-face classrooms, with the addition of on line tasks. The use of Information and Communication Technologies (ICTs) provided the insertion of new methods and learning techniques, which aim to help users interact with computer applications.

Ubiquitous computing is an example of an emerging area, which, according to Yahya et al. [1], can be described as a learning paradigm that uses ubiquitous computing environments and allows learning to occur "anywhere and anytime". Thus, students have at their disposal in the most propitious time, content suitable for their learning, taking into account the location and context in which it is inserted.

One of the key aspects that are directly linked to ubiquitous computing is context information, which, according to Dey [2], can be defined as any information that can be used to characterize the situation of entities that are considered relevant to the interaction between a user and an application. The origin of this information may come from various sources, such as the analysis of users' cognitive styles on the environment and the speed of the network connection.

Examples of computer applications that use context information are the Virtual Learning Environments (VLE), which have been adapted to suit the needs of users, as seen in the work of [3] and [4]. However, the context used in these environments are subject to imperfections as explained in [5], wherein the context information have a high probability of defect or inconsistency caused by inaccuracies in the acquisition and maintenance of context.

This difficulty may reflect in providing a situation out of date with respect to the current moment the user is located, committing to provide a proper context for the user by the application, thus hindering its interaction with the tool to perform activities. Subarea Quality of Context aims to address this issue, which, according to Krause e Hochstatter [6], can be defined as any information describing the situation of an entity, i.e., its context, and that can be used to determine the value of information for a specific application. Thus, the QoC aids for the final result presented by the tool reflect more realistically as possible the situation where the user is located.

This paper presents a proposal for the development of a ubiquitous virtual learning environment called UVLE^{QoC}, which aims to treat the context information by applying metrics of Quality of Context. The environment will integrate two systems: U-SEA and SEDECA, thus will be captured the context information about connection speed and type of cognitive profile of the user. The information captured pass through an evaluation process, only selecting those who achieve a minimum percentage of confidence, in order to adapt the resources and tools of the environment according to the situation of the user.

The structure of the paper is organized as follows: in section one, an introduction about ubiquitous computing, context and quality of context, the problems in these areas and the proposed work is presented. Section two presents the related work in the area of QOC. Section three the functioning environment, its goals and advantages. Finally, section four presents a conclusion and future work.

 ISBN: 978-1-61208-289-9

II. Related Work

To assist the course of this study, some related works were evaluated, and are described following:

Yasar et al. [7] aimed to provide efficient communications for the established network; for this proposal, the communication was divided into two phases: the first is focused exclusively on QOC, while the second aims to determine the reputation of the nodes involved in the communication. Their contributions are the definitions of information acceptable limits, in order to ensure a minimum quality, eliminating those that are not within these standards.

In [8], a context-sensitive framework that supports management of QOC in several layers is presented. The authors compared the probabilities of the context to be valid using different analysis algorithms, such as the selection of the newest and largest service relativity. The framework allows to evaluate the context in its raw form, discarding those that are duplicated or inconsistent to provide context information with an acceptable level of quality for applications.

The work of [4] examines the conflicts that can be generated in gathering information to support decisions of adaptive applications. The authors explain that the QOC parameters can be used to perform these tasks, seeking to resolve internal and external conflicts based on two indicators: the probability of correctness and reliability.

The solutions presented above cover various application domains, such as networks of nodes and vehicular service selection, while this proposal is focused in education, with the use of virtual learning environments. The application of metrics of QoC in the information captured on the ubiquitous environment, seeking to enhance the quality of the context provided to users, is presented.

III. Proposal

This study presents a work in progress, which to design a Ubiquitous Virtual Learning Environment with Quality of Context ($UVLE^{QoC}$). In this environment, the following context information is collected: network connection speed and the cognitive profile of the users, applying over them quality indicators to ensure the attainment of an acceptable level of confidence, selecting the information to be used in the formulation of context of the user.

Once the context has been formulated, the virtual environment will modify the materials and tools to display them according to the user preferences. For example, it will be presented only those materials in video format and chat tool instead of long text and forums if the user has selected them in its preferences. The access may be accomplished either via desktop, or via mobile devices, where the environment adapts itself automatically to the device type of the user.

The objective is to create guarantees by using metrics such as precision and reliability, so that the environment use context information to the appropriate momentary situation in which the user is located. Thus, the $UVLE^{QoC}$ performs the necessary adjustments, as the selection of features and interface, to improve the process of interactive learning.

Its architecture can be seen in Figure 1, being divided into four parts. The first part deals with the information extracted from the selected context sources, which were based on the work of [3] and [9].

The data on the speed of the network connection will be extracted from the environment U-SEA [3], through an algorithm that performs the calculation and determines the connection speed of the user. At the same time, the information about the cognitive profile of the user is collected at SEDECA environment [9] by applying two questionnaires. The first one with 16 questions that define the type of cognitive style, and the second one with 13 questions that define the users preferences for the type of materials and tools that are included on the discipline.

Both environments were developed in Moodle [10], in different versions; so, for the development of work, they were integrated into a new environment Moodle in its version 2.4, serving both as a source of context and to make the adjustments made by them.

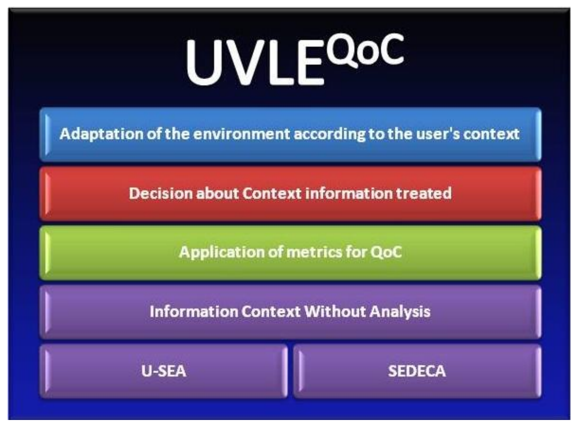


Figure 1 – Architecture of UVLEQoC

The second part is characterized by the application of quality metrics on information collected for the formulation of context. They will be inserted in the calculations and algorithms designed to assess the quality of the information, in which we defined the following quality indicators:

- **Accuracy:** is investigated how these data show the reality, so you can see if the data network connection, cognitive styles and expertise are in line with what is actually happening;
- **Probability of correctness:** it analyzes the probability that a context information is correct;
- **Resolution:** Sets the level of detail that has the context information, so that they can get more and more details about the user's context and thus create a suitable profile;
- **Reliability:** checks in percent values, how the information obtained from the context sources are correct. This check is performed based on pre-defined rules and calculations.
- **Time Accuracy:** are established minimum and maximum periods of time, to check if the collected information is valid and need to be catched again.

- **Integrity:** Checks if the data obtained from the context sources not suffered unauthorized modifications, which can generate incorrect information.

The preparation of these quality indicators were based on the creation of specific calculations and the use of algorithms based on the work of [7] and [11], for the definition of their operation and application.

The third part corresponds to the analysis of the results obtained with the calculations performed by checking that the verified information has a minimum percentage; so, they can be used in the formulation of the user's context. It is a decision process, where only those who are within acceptable limits will be used to build the user context for the environment.

In the fourth part, based on context information properly treated, are performed all necessary changes in the environment, as the types of materials (videos, slides, longer texts, images, etc.) and the types of tools that (chat, forum, task, etc.) that will be displayed, according to the connection speed of the user and their cognitive profile. If the network connection speed is less than the predetermined value, the environment will adapt itself in order to load resources and a lighter interface, thus seeking to promote a better user interaction with the application. And the materials and tools in the environment are loaded according to the preferences defined by the User in SEDECA. For this, an additional field will be created in the form of insertion of files in the environment where it will indicate the type of material that by entering, for example, if it is a video, image, long text, slides, etc.

As the technology used to develop this environment, the environment Moodle was chosen to be used. Besides the aforementioned environment, a tool called Wamp Server [12], which integrates three technologies (PHP, MySQL and Apache), was chosen to be used. Apache server will act as a venue for the hosting environment. For the database, MySQL will store all information from AVA. Finally, PHP is a programming language used for developing Moodle; therefore, it will be used in the development and implementation of quality metrics.

For adaptation in mobile devices, we will use the Bootstrap theme, which is integrated into Moodle environment, where the selection of the theme to be used is made by the administrator of the environment. In the case of access via workstations or laptops the default Moodle's common interface is exhibited. However for access via mobile devices, the interface is automatically adapted to user's device.

The environment is in development stage and it has been integrated into it the theme Bootstrap to make access from mobile devices and the module U-SEA, which verifies the connection speed of the User. Figure 2 shows the adaptation of the environment to mobile devices using Bootstrap, which presents only those files that are smaller than 400Kb, because the user's connection speed is less than 500Kb. For example, the file template.pdf was not made available to the user, because its size is larger than allowed.

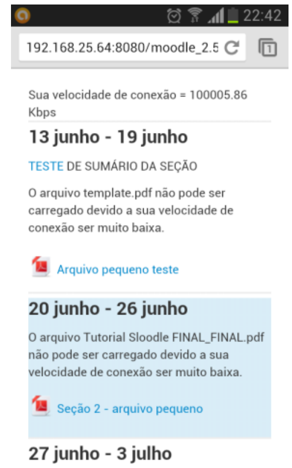


Figure 2 – UVLE^{QoC} accessed via mobile device with the U-SEA

UVLE^{QoC} adapts to the cognitive profile of users and their technological context, with a guaranteed minimum level of quality, with the goal of improving students' interaction with the environment. Thus, materials and resources best suited to their preferences and needs will be inserted by improving the process of teaching and learning, in which everyone will know your style, its technological limitations and be exempt from the difficulties explained by them, as a failure to access the archives of the discipline and the types of materials that are not consistent with their gender.

For example, if incorrect information is collected for the formulation of context and is not to go through an assessment of quality, there is no guarantee of this information has a desired level of accuracy. This may impair the use of the environment by the user, as in the case of adaptation according to their cognitive style and speed of network connection, if the collected information is incorrect, the environment will adapt erroneously damaging user interaction and achievement of their activities.

With the application of QoC indicators, the UVLE^{QoC} provides assurances that the information they are captured, will undergo an evaluation process in order to examine whether these holds the minimum required level of quality and then be used in the construction of the user context and more accurately reflect the reality in which it is inserted.

IV. Conclusion and Future Work

Context information is used to characterize the state of an object, reflecting the momentary situation in which it is located. One of the problems related to this is the inconsistency that these data are subject to, resulting in reflection of a state that is not consistent with reality. With this, the subarea of Quality Context aims to address this information through rules and algorithms verification in order to ensure that the information used in the construction of the context is correct.

This paper presented a proposal that is developing a virtual learning environment ubiquitous, which handles context information by applying metrics of QOC. In addition, it will integrate the modules U-SEA and SEDECA to perform the adaptation of resources and tools of the environment, according to cognitive style and connection speed of the User.

Thus, we seek to guarantee a minimum level of quality for information that are captured, by formulating an appropriate context the momentary situation of the user, performing a proper adaptation of the environment and facilitating their interaction.

As future work, we intend to finish this development environment and to test it with groups of undergraduate and graduate students to validate their implementation and fix the necessary aspects. In addition, we intend to create activities in the virtual world OpenSim to UVLE^{QoC} interconnected through technology Sloodle, so that users can interact with these two environments.

REFERENCES

[1] S. Yahya, E. Ahmad, and K. Jalil, “The definition and characteristics of ubiquitous learning: A discussion”, in International Journal of Education and Development using Information and Communication Technology. (IJEDICT), vol. 6, Issue 1, pp. 117-127, 2010.

[2] A. Dey, “Understanding and using context”, in Personal and Ubiquitous Computing, vol. 5, pp. 4–7, Feb. 2001.

[3] S. Piovesan, E. Amaral, C. Arenhardt, and R. Medina, “U-SEA: A learning environment ubiquitous using cloud computing”, in International Journal of Engineering and Technology (iJET) – Volume 7, Issue 1, 2012.

[4] J. Filho and N. Agoulmine, “A quality-aware approach for resolving context conflicts in context-aware systems”, in IEEE/IFIP 9th International Conference on Embedded and Ubiquitous Computing (EUC), Melbourne, Australia, pp. 229–236, Oct. 2011.

[5] V. Santos, “Context management in collaborative systems”. Master's Dissertation, Universidade Federal do Pernambuco (UFPE), May 2006.

[6] M. Krause and I. Hochstatter, “Challenges in Modeling and Using Quality of Context (QoC)”, in Lecture Notes in Computer Science, pp. 324-333, 2005.

[7] A. Yasar, , K. Paridel, D. Preuveneers, and Y. Berbers, “When efficiency matters: towards quality of context-aware peers for adaptive communication in VANETs”, in Intelligent Vehicles Symposium (IV), pp.1006-1012, Jun. 2011.

[8] D. Zheng and J. Wang, “Research of the QoC based Middleware for the service selection in pervasive environment”, in International Journal of Information Engineering and Electronic Business (IJIEEB), pp. 30-37, Feb. 2011.

[9] P. Mozzaquatro, F. Franciscato, P. Ribeiro, and R. Medina, “Modeling a Framework for adaptation of virtual learning environments moving to different cognitive styles”, in CINTED UFRGS, Porto Alegre, Brazil, v. 7 nº 3, Dec. 2009.

[10] Moodle´s Official Website. Last accessed Jul. 2013, at: https://moodle.org/ .

[11] A. Manzoor, H. Truong, A. Malik, and S. Dustdar, “On the Evaluation of Quality of Context”, in EuroSSC 2008, LNCS 5279, pp. 140–153, 2008.

[12] Wamp Server Official Website. Last accessed Jul. 2013, at: http://www.wampserver.com/.

Context Awareness for Service Desk Management

Jaziel Souza Lobo
Instituto Federal de Sergipe – Campus São Cristóvão
Aracaju(SE), Brazil
jaziel.lobo@gmail.com

Roseclea Duarte Medina
Centro de Tecnologia-Universidade Federal de Santa Maria
Santa Maria(RS), Brazil
roseclea.medina@gmail.com

***Abstract*—This paper presents a customization of a service desk tool inserting some context-aware computing elements based on location, temporal and expertise awareness. As main result, a context-aware service desk system, that also allows access from mobile devices and improves technical calls distribution based on professional expertise and geographical location, was concept.**

Keywords-Help Desk; Service Desk; Context-Aware; Location Awareness

I. Introduction

Modern organizations are becoming more dependent on Information Technology (IT) day by day, fact that turns necessary the implementation of effective IT management to justify high investments on this sector and to aggregate values for these companies [1]. Second [2], 88% of financial executives assert that the existing IT services operation efficiency are more worrisome than the attendance of new necessities. For Cusick and Ma [3], the "Impact on revenues is directly related to system availability" that can be proven with facts like the 10 million dollars Symantec company prejudice suffered after a sales systems fail [4]. Another example is of E-Bay company that had around U$ 3 and 5 million dollars decrement on its founds and a 26% decline on its actions after a 22 hours stop caused by a system fail [2]. Thus, when a problem arises that causes abnormal functioning of IT services, it is expected that the user has a quick response and clear support staff in order to minimize the damage that can be caused. The team responsible for resolving IT issues, was initially named Help Desk [5], but due to its importance and new services added to your area, now called Service Desk [6].

According to Zahedi et al. [7], time is an important parameter considered for the help desk so that the technicians of this activity, precisely because of time constraints to solve all kinds of problems, often refer to the help desk as "hell desk". To [2], one of the key factors for the success of a service desk is the allocation of human resources that have the appropriate profile for solving different types of problems. A service desk technicians whose work without any planning, attending calls disorderly, or whose technicians are assigned to calls that they do not have the expertise (and practical experience) to solve the problem, it may cause: for technical loss of time offset by unnecessary, for the user to idle due to lack of solution of the incident on the first call, and the company recorded losses by stopping the services.

This paper presents the adaptation of a service desk tool developed in the program Graduate of Federal University of Santa Maria (UFSM) in Brazil. The adjustments involve characteristics of computing context-sensitive location, adding the time and technical expertise to the context. As main results, we obtained a context-sensitive service desk system Context Aware Service Desk (CASD), which enables your access via mobile devices, with the optimization of the distribution of calls across the geographical location and expertise of the technician. To validate were inserted dummy data so that it was possible to test the behavior and operation of the system. The tests were conducted on the campus of UFSM with smartphone that had integrated GPS.

This article is organized as follows: Section 2 presents aspects of ubiquitous computing, emphasizing sensitivity to context. Section 3 discusses the related work. Section 4 presents the proposed project, and the changes made to the service desk tool. Experiments and discussion of the tests are presented in Section 5 and Section 6 presents the conclusions.

II. Ubiquitous Computing and Context Awareness

In 1991, the term Ubiquitous Computing was defined by Mark Weiser as a paradigm that enables integration and communication of multiple devices and resources (software and hardware) in a real environment that would enable the user to perform activities without awareness of this usage [8][9]. Other paradigms have emerged as Pervasive Computing, which provided access to information and computing resources for users anywhere, anytime and using any device. Currently, the terms Pervasive Computing and Ubiquitous Computing are used as synonyms by many researchers and so will be considered in this text.

One of the main research areas of ubiquitous computing is Context-Aware Computing in which you want to obtain inputs, called contexts, which are current user information, contextualizing the environment where it is located and the device computational used [10]. To Dey [11], context is any information that can be used to characterize the situation of an entity where the entity can be a person, place or object that is considered relevant to the interaction between a user and an application, including the user and the application itself. According to the author, a system is context-aware if it uses context to provide information and / or services relevant to the user, where relevancy depends on the user's task. According to Satyanarayanan [9], a system of pervasive computing that strives to be minimally intrusive has to be

 ISBN: 978-1-61208-289-9

context sensitive and be aware of the state and around its user to, based on this information, change their behavior.

A key component of ubiquitous computing is the location awareness that uses a positioning system for the location of the user [12]. To Loureiro et al. [10], "a positioning system is a tool used to determine and record the location of an object on Earth's surface." With the trying to add services and business models for ubiquitous devices in the network, the World Wide Web Consortium (W3C) established in March 2007, the working group Ubiquitous Web Applications that defined an API for Geolocation. This API has a high level interface to location information (latitude and longitude) and provides support for mobile browsers and applications of Location Based Services (LBS) [13] [14].

III. Related Work

The search by related work showed that many authors approach aspects of how to improve the control of the management of incidents or address issues of how best to implement the Information Technology Infrastructure Library (ITIL). As far where we investigated, there were not found studies with different approaches on the optimization of the work team to reduce the cost of diagnosis by technicians lacking the expertise ideal for the reported incident. Following some of the studies reviewed.

Zahedi, Rahimov and Soleymani [7], proposed a help desk that simulates a technical support center via a web site, where visitors ask questions and receive advice automatically. Already Jäntti [15], explored which are the basic functional requirements for an incident management system. Bartolini, Stefanelli and Tortonesi [16], present HANNIBAL software, a decision support tool for business impact analysis and incident management improvement. In Marcu et al. [17] it is proposed a method to correlate with user called open call by automatically opened monitoring systems to avoid loss of events for diagnostic time with the same problem. References [3] and [18] discuss what is the best way to implement the Information Technology Infrastructure Library (ITIL) processes.

IV. Context Aware Service Desk - CASD

This paper proposes to adapt a service desk system that minimizes the occurrence of a second call to the same calls that had not been closed for lack of expertise of the technician who performed the first service. The Context Aware Service Desk (CASD) was designed for smartphones and mobile phones with great computing that, in addition to the original functionality, have an integrated GPS. In addition to this feature, the system proposes to reduce the loss of time with unnecessary travel and enable access from anywhere at anytime and with any device. Amaral [19] presents a service desk system that was integrated into a corporate portal for the realization of incident management and access rights of the user (authentication, authorization, auditing) with collection of data for decision support, based on technical multivariate statistics. This was the system used for the adaptation in this paper.

The CASD is divided into two main tasks: Task 1 - responsible for identifying a context C and Task 2 - responsible for adapting the system to the context. The context C is defined by the time the coach makes the access to the system and it is composed of location data (latitude and longitude) and technical profile and the time of access to the system.

A. *Task 1 - Identification of Context*

The proposed work deals with the suitability of a Service Desk to provide services relevant to technical support by identifying a context C. The system consists of a finite technicians set defined as $t \subset \{t_1, t_2, \dots, t_n\}$ with $n > 1$; expertises defined as $e \subset \{e_1, e_2, \dots, e_n\}$ with $n > 1$; buildings defined as $P \subset \{P_1, P_2, \dots, P_n\}$ with $n > 1$; profiles defines as $pf \subset$ {Rating, Attendance, Administrator, Rating_and_Attendance} and priorities defines as $pr \subset \{1,2,3,4,5\}$.

1) *Getting the Geographic Location of the Technician*

The proposal with the W3C Geolocation API is a site that, when visited, can get the coordinates (latitude and longitude) of the device that is accessing it without any client application is installed on this device. This mechanism was implemented in the service desk so that when a technician logs in to the system, the script runs Javascript: navigator.geolocation.getCurrentPosition (position), which returns an object by positioning function position.

2) *Access Time*

The access time, to the system, is used to verify if the technician is at his work time. The temporal context entered here is to allow the support teams distribution per shift scales to meet on call 24x7, works like most IT teams.

3) *Technical Profile*

Based on division that occurs in the Central Appointments User (CAU) UFSM, technicians are mapped into four profiles: Technical Classifier, Service Technician, Technician and Administrator Meets and Sorts. The Technical Classifier is one that is allowed to sort the calls by setting the priority and the expertise necessary to solve the problem reported. The Technical Assistance will be allowed only to view and meet the tickets already sorted. The Technician Meets Sorts and builds the permissions of the two previous roles and the last technician is one who has administrator profile.

4) *Internal or External Access*

It's possible determine if a technician t_i will have access to the system when outside the building of the central services (external environment) or if it will only have access at building's vicinity (indoor). The location inside or outside of the coach is given by calculating the distance d, in miles, between the technical and building services center. For this work, it was found that at a distance of up to 300m ($d \leq 0{,}3$), which is almost the distance between a building and the other on campus, the system will consider that the technician is called the internal environment. Otherwise, it is considered that the technician is in the external environment.

5) *Distance Calculation between t_i and p_j*

When reporting an incident, the user has to enter the building and room where the problem occurred. The buildings must be first registered with their respective coordinates. The calculation of the distance between a technician t_i and a building P_j is used both to verify that the technician is in internal or external environment as to sort the list of calls that the technician will attend. Knowing the points formed by the t_i and P_j, coordinates, apply the formulas of spherical trigonometry to calculate the distance d in degrees formed by circular arcs between these points and these points at point A (Figure 1) [20].

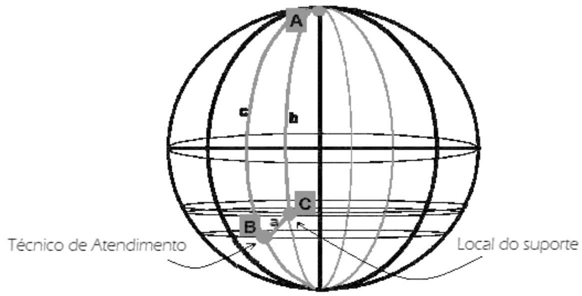


Figure 1. Arcs to calculate the length between two points using spherical trigonometry.

To calculate the distance will be used three equations. The first equation is given by the expression:

$$\cos(a) = \cos(b) * \cos(c) + \sin(b) * \sin(c) * \cos(ABC) \quad (1)$$

where,

- a = BC arc formed by the longitude difference of the two coordinates
- b = AC arc that is equal to 90 - latitude of the building support
- c = AB arc that is equal to 90 - latitude of technical assistance

The coordinates values in (1) must be reported in decimal values and the values received to invoke the W3C Geolocation method are in degrees, minutes and seconds, so it is necessary to convert to radians. Found the value of cos(a) in (1), applies (2) that calculates the arc cosine of the value found:

$$Arc = arc\cos(\cos(a)) \quad (2)$$

Unlike what happens in (1), the value found in (2), which represents the arc formed between the building and technical and which is in radians, must be converted to degrees to apply in (3). After obtaining the value of the arc in degrees, to calculate the distance d in km, you should apply a rule of three between this value and the value of the full arc of the circumference of the Earth. Knowing that the radius of the Earth is 6371 km, the Arco_T= 2 * π * radius = 40,030 k. Obtained the value of the full arc of the earth, applies (3) to calculate the distance in kilometers of the two points:

$$d = (40.030 * Arco)/360 \quad (3)$$

6) Algorithm of Task 1

The algorithm shown in Figure 2 represents the execution for task 1 and is subdivided in three sub-tasks, which are inExpedient, inPlace and length, as detailed below:

```
Task 1 – Identify Context
initialize lat ← 0; Lng ←0;
∀ authentication do
  if getLocate do
    if ti ∈ t then
      if ti.pf = sorter then
        if inExpedient(ti) and inPlace(lat, lng) then
          applyContextSorter();
        else
          write (“ user outside of working hours or without
                  permission to access this area outside the
                  place of service.”);
      else if ti.pf = attendant then
            if inExpedient(ti) then
              applyContextAttendant();
            else
              write (“ user outside of office hours”);
          else if ti.pf = sortAndAttends then
                if inExpedient(ti) then
                  if inPlace(lat,lng) then
                    ApplyContextSorter();
                  else
                    ApplyContextAttendant();
                else
                  write (“ user outside of office hours”);
              else if ti.pf = admin then
                    applyContextAdministrator();

Task 1.1 – inExpedient(ti)
  initialize now ← 0; Lng ←0;
  now ← accessTime;
  if (ti.inputShift1 ≤ now and now ≤ ti.outputShift1) or
     (ti.inputShift2 ≤ now and now ≤ ti.outputShift2) then
     return true;
  else
     return false;

Task 1.2 – inPlace(lat,lng)
  if length(lat, lng, P1.lat, P1.lng) ≤ 0,1 or ti.outside then
     return true;
  else
     return false;

Task 1.3 – length(plat, plng, q.lat, q.lng)
  initialize a←0; b←0; c←0; cosA←0; arcCosA←0;
  a = math.radians(plng – qlng);
  b = math.radians (90 – qlat);
  c = math.radians (90 – plat);
cosA=math.cos(b)*math.cos(c)+math.sin(b)*math.sin(c)*
       math.cos(a);
  arcCosA = math.degrees(math.acos(cosA);
  return (40030.00 * arcCosA)/360;
```

Figure 2. Algorithm for Identify Context

B. *Task 2 - Application of the Context*

Defining the profiles of the technician, also defined the contexts that should be applied to each of them. It is through the Task 2 that the system will suggest the actions the user can do.

1) Context for the Rating Technician

The rating technician is who has permission only to classify the tickets. This way the system must allow access only to view the calls that have been reported, but haven't yet been classified. In addition to this permission, the system will react differently if the technician is not in your expedient time or if some company considers relevant to this technician should not access the system outside of the building of center service and define him without permission to extern access.

2) Context for Attendance Technician

Unlike that what happens with the rating technician, even outside of the expedient time, the technician with profile to attendance should be able to access the system, provided that it is to complete a ticket or to inform the progress of the call and put it on availability to another technician, what is being called a shift change. This peculiarity was created for cases where the technical have not completed the service, must overcome their expedient time to fulfill the Service Level Agreement (SLA). This condition is only allowed if the technician has any call that is not finished yet.

3) Context for the Rating and Attendance Technician

The rating and attendance technician is who can classifier and attend the calls. It is during the identification of the context that the system will suggest, based on the location of the technician, the interface to be used. When it is identified that the technician is out of the building of center service, the system automatically will show a list of calls for attend, suggesting that he is in the external environment and probably is consulting new calls to attend. If the technician is in the building of center service, the system will direct you to the screen of the classification of the calls. Though the system suggests a feature, anytime the technician can switch between the two options.

4) Context for the Administrator Technician

The technical who has the administrator profile can view all calls classified or not, those in attendance and those already completed. Apart of observing, the administrator technician can classify calls and register the buildings and their locations. For this context, the system will show a screen with the not attended calls and provide other options.

V. EXPERIMENTS AND RESULTS

This study makes part of a major project and in this part of the project the objective was identification of context. In this way, the CASD was developed with the Django framework [21], which uses the Python programming language [22]. The implementation occurred in Eclipse [23], by installing the Pydev plugin that integrates the Python programming language and Django framework in the Eclipse development environment. For data persistence were used the MySQL database and web server like Apache HTTP Server. For the tests registered a set of expertises, with buildings and technical profile for classification, attendance, grading and attendance and administrator profile. A mass of data for the different incidents buildings with different priorities and expertise was fed.

The initial tests focused on the consolidation of the mechanism for obtaining the location of the technician via the mobile device. Figure 3 shows capture screens in this initial test phase.

Figure 3. Solicitation of permission to access the device localization.

In the figure, can be seen several devices that are displaying the message that prompts the user for permission to access your location. The tests proved that the W3C Geolocation method is effective and it works effectively across multiple devices. The second set of tests aimed to validate the formula for calculating the distance between the technician and the buildings with calls for service. At this stage it created a parallel application with only two functions: registering buildings with their coordinates and perform a comparison between the buildings registered. The register of buildings invoked the W3C Geolocation API and comparator used the formula of spherical trigonometry to display the registered buildings ordered according to the distance between them and a building used as a reference. Tests showed that the formula had always used the buildings in the correct order of distance from smallest to largest. The final stage was used to validate the operation and behavior of the system.

At this last phase, classification, customer attendance and administration functionalities have been joined. By admin functionality was possible to register the buildings and track instances of reported calls in the system. By the classification, the technicians defined the priority and expertise required for each ticket. About the costumer attendance, the tickets were presented to technicians according to their expertise and ordered according to the priority and the shortest distance between the building and location to meet the technician;

VI. CONCLUSION

The delay in diagnosis of incidents have generated higher losses for organizations and one of the factors that has generated this delay is the recurrence of care that is caused by the allocation of technicians who do not have the right expertise to solve the reported incident. To solve this problem were included characteristics of ubiquitous computing at the service desk system adaptation. The tests were done with six technicians registered at the system with different profiles and with five tickets, which were attended at the buildings marked on Figure 4, by different technicians with different profiles according to the test plan already created. The tests show that using a context that considers the location of technical expertise minimizes the recurrence of calls since the technicians were allocated according to their knowledge to the solution of the problem reported. During testing, 100% of calls were allocated to technicians with the proper expertise. Even if one has obtained the maximum percentage, it is noteworthy that in this case the data were fed into the simulation and behaved as expected. In a real environment, for this to occur, it is necessary that the technician, who will make the classification of tickets, is highly qualified for the correct classification of so-called reported.

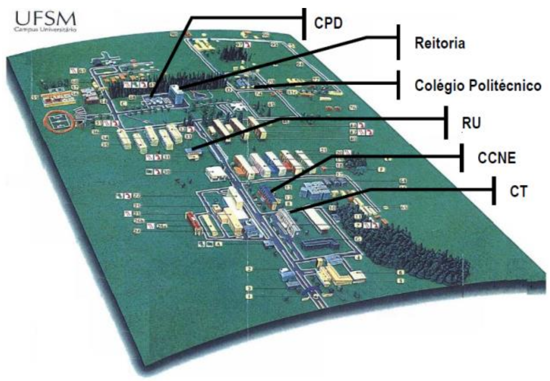


Figure 4. Identification of the buildings with registered tickets for attandance. Image adapted from the website of the Federal University of Santa Maria.

Another system's built-in capability is the ability to automatically direct the technician to answer calls geographically closer, reducing the waist of time with unnecessary travel by technician, leaving a building A to building B and then was informed that there was another call in building A. During testing nearby buildings, facing each other, were called were registered and reported in two buildings. By the time that a technician, with the called expertise, made access to the system and is called for the two calls on each side of the street and with equal distances for each building, it was identified that the CASD had called the building that stood on the same side of the street which was positioned the technician. It shows that the implementation of context awareness, to reduce unnecessary travel, gave greater flexibility in attendance, which enables the company to earn by having a lower downtime, and by reducing diagnosis time of the incidents.

Comparing with other similar systems available up to date, it is observed that the applications that have some mechanism for direct calls triggered by the user, present several problems, precisely the user's lack of technical support. Another feature present in these softwares is that the assignment of the priority of the call is also performed by the user. In this case, for the same reason of lack of technical knowledge or by having interest in the prompt resolution of your problem, users can sort it's calls always with the highest priority, impacting on the performance of the IT industry. In CASD case, the technical service center has the profile classifier that has the responsibility for determining the priority and expertise necessary to resolve the reported problem. The features like sorting the call according to the location of the technician and the definition of work time for the technical definition of scales, that are present in CASD, were not found in other softwares studied. Considering the point of view of implementation, tests have shown that the system is technically feasible and that the adjustments made in this paper can easily be implemented in different types of systems in the same category.

The results of this study bring expectations for new research and future work such as the creation of a knowledge base consisting of problems and solutions that help both accelerate the process of diagnosing problems as to "retain" the knowledge generated by the company. Another point that could be addressed in a future work concerns the allocation of expertise for calls. It would be interesting to evaluate the content of the text entered, in the system, in the description of the problem and that before that there were automatically determine the expertise required to solve this incident, conducting a self-rating of the tickets.

VII. References

[1] R. V. Ferreira, Impact of Investments in Information Technology on Firm Value Generation: multicase study with Baking Companies of the State of Minas Gerais, 2010. pp. p. 195. Dissertation (Administrator Master's degree) – Federation University of Paraná, Curitiba, PR, 2010.

[2] I. L. Magalhães and W. B. Pinheiro, IT Services Management: in Practice: An approuch based on ITIL, São Paulo: Novatec, 2007, p. 667.

[3] J. J. Cusick and G. Ma, "Creating an ITIL inspired Incident Management approach: Roots, response, and results," Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP, April 2010, pp. 142-148, doi:10.1109/NOMSW.2010.5486589.

[4] IDG Now, "Problem with antivirus ativation generates a loss of $ 10 million to Symantec," 2011. [Online]. Available: http://idgnow.uol.com.br/seguranca/2010/10/28/problema-de-ativacao-gera-prejuizo-de-us-10-milhoes-a-symantec/. [retrieved: Jul, 2013]

[5] G. O. T. Cavalari and H. A. X. Costa, "Model and Development of a Help-Desk System to Lavras' City Hall," doi: 10.5329/RESI, 2005. [Online]. Available: http://revistas.facecla.com.br/index.php/reinfo/article/view/158/0. [retrieved: Jul, 2013].

[6] M. Jäntti and J. Kalliokoski, "Identifying Knowledge Management Challenges in a Service Desk: A Case Study.," Proceedings of the 2010 Second International Conference on Information, Process, and Knowledge Management, Feb. 2010, pp. 100-105, doi:10.1109/eKNOW.2010.20.

[7] M. Zahedi, H. Rahimov and F. Soleymani, "A Two-Level Automatic Help Desk Based on a New Statistical Approach," Internet and Web Applications and Services, 2008. ICIW '08. Third International Conference on, June 2008, pp. 530 -534, doi:10.1109/ICIW.2008.46.

[8] M. Weiser, "The Computer for the 21st Century," Pervasive Computing, IEEE, vol. 1, no. 1, Jan.-March 2002, pp. 19-25, doi:10.1109/MPRV.2002.993141.

[9] M. Satyanarayanan, "Pervasive computing: vision and challenges," Personal Communications, IEEE, vol. 8, no. 4, Aug. 2001, pp. 10-17, doi:10.1109/98.943998.

[10] A. A. F. Loureiro, R. A. R. Oliveira, T. R. d. M. B. Silva, W. R. Pires Júnior, L. B. R. d. Oliveira, R. A. Moreira, R. G. Siqueira, B. P. S.

Rocha e L. B. Ruiz, “Context Awareness Computing: Challenges and Tendecies,” Brazilian Simposium of Network Computer and Distributeds Systems, May 2009, pp. 99-149.

[11] A. K. Dey, "Understanding and Using Context," Personal Ubiquitous Comput, vol. 5, no. 1, Jan. 2001, pp. 4-7, doi:10.1007/s007790170019.

[12] F. O. Akgul and K. Pahlavan, "Location awareness for everyday smart computing," Telecommunications, 2009. ICT '09. International Conference on, May 2009, pp. 2-7, doi: 10.1109/ICTEL.2009.5158609.

[13] W3CGeo, "W3C Geolocation Working Group," 2008. [Online]. Available: http://www.w3.org/2008/geolocation/. [retrieved: Jul, 2013].

[14] S. J. Vaughan-Nichols, "Will HTML 5 Restandardize the Web?," Computer, vol. 43, no. 4, April 2010, pp. 13-15, doi:10.1109/MC.2010.119.

[15] M. Jäntti, "Defining Requirements for an Incident Management System: A Case Study," Systems, 2009. ICONS '09. Fourth International Conference on, Mar. 2009, pp. 184 -189, doi:10.1109/ICONS.2009.17.

[16] C. Bartolini, C. Stefanelli and M. Tortonesi, "Business-impact analysis and simulation of critical incidents in IT service management," Integrated Network Management, 2009. IM '09. IFIP/IEEE International Symposium on, June 2009, pp. 9 -16, doi: 10.1109/INM.2009.5188781.

[17] P. Marcu, G. Grabarnik, L. Luan, D. Rosu, L. Shwartz and C. Ward, "Towards an optimized model of incident ticket correlation," Integrated Network Management, 2009. IM '09. IFIP/IEEE International Symposium on, June 2009, pp. 569 -576, doi:10.1109/INM.2009.5188863.

[18] R. Pereira and M. M. Silva, "ITIL maturity model," Information Systems and Technologies (CISTI), 2010 5th Iberian Conference on, June 2010, pp. 1-6.

[19] É. M. H. d. Amaral, Proactive management of security incidents through the quantification of data and the use of multivariate statistical methods, 2010. pp. p. 137. Dissertation (Master’s degree on Production Engineering) - Federal University of Santa Maria, Santa Maria, RS, 2010.

[20] V. L. Silva and M. P. Sucena, "Location of facilities: a case study applied to the proper choice of airport for minimizing logistics costs of distributing pharmaceutical products," Twenty-Third Congress on Transport Research and Education, Nov. 2009.

[21] Django, “Django Software Foundation,” 2010. [Online]. Available: http://www.djangoproject.com/. [retrieved: Jul, 2013]

[22] Python Software Foundation, “Python,” 2010. [Online]. Available: http://www.python.org/. [retrieved: Jul, 2013]

[23] Eclipse, “Eclipse - The Eclipse Foundation open source community website,” 2010. [Online]. Available: http://www.eclipse.org/. [retrieved: Jul, 2013]

Fostering Access to Data Collections in the Internet of Things

Alexander Kröner
Georg Simon Ohm University of Applied Sciences
Nuremberg, Germany
Alexander.Kroener@th-nuernberg.de

Jens Haupert, Christian Hauck, Matthieu Deru, Simon Bergweiler
German Research Center for Artificial Intelligence
Saarbrücken, Germany
{jens.haupert, christian.hauck, matthieu.deru, simon.bergweiler}@dfki.de

***Abstract*—Using label technology, a physical object may be employed to build a continuously growing data collection, which can, for instance, be exploited for product quality monitoring and lifecycle management. Along the objects lifecycle, queries to such a collection may stay quite similar, e.g., get unusual observations. However, expectations to a good answer may change, as with time different entities will come into contact with the object. This article reports on work in progress concerning a framework for collecting data concerning things, which aims at 1) decoupling logic employed for interpreting such a collection from processing hardware and 2) using the collection itself for transporting such logic. Main contributions include an approach to hardware-abstraction of processing logic at the object or remote, and an app store for retrieving interpretation and presentation logic.**

Keywords—Ubiquitous computing; RFID tags; Distributed information systems; Supply chain management

I. Introduction

Within the Internet of Things, physical objects may function as a focus for digital data and services concerning the artifact itself as well as associated things, people and processes. This function enables an object to take a new role as a data collector and provider in a broad range of scenarios, e.g., users may manually associate data with an object in order to socialize and foster discussion in a community [1], tools may automatically collect usage data in support of pay-by-use accounting [2], and products may steer and document their production [3] and transport rules that support reasoning of healthcare applications [4]. Existing applications of such technology are typically deployed in "closed" scenarios, i.e., requirements of users and applications are known before the collection process starts.

This reflects only to some extent a supply chain with continuously changing users and requirements. In order to facilitate communication between stake holders in such an "open" scenario, a uniform interaction behavior of the collection would be advantageous, e.g., a uniform way to "check integrity" of an object, i.e. compliance to criteria specified by a third party on an individual base for objects or kinds of objects.

In the following, Section II provides an example scenario, where one stakeholder has to employ logics provided by another stakeholder. Section III summarizes requirements that arise from this scenario. Then, Section IV wraps up work accomplished so far concerning so-called Active Digital Object Memories (ADOMe), a platform for processing logic in a way that allows for embracing a broad range of infrastructure approaches common to Internet of things applications. Section V extends this approach with a concept of an app store supporting distribution of the processing logic. Afterwards, work related to ADOMe and app store is briefly reviewed in Section VI, followed by a summary in Section VII.

II. Scenario

The following logistics scenario deals with integrity control during transportation of a heterogeneous set of goods (see Fig. 1). Each good is packaged in a way matching its nature (e.g., fragility) and value. All packages are tagged with some kind of label technology, which allows for automatically identifying the object. Depending on the respective kind of package, this technology may range from passive RFID (Radio Frequency IDentification) to embedded systems with integrated sensing and processing capabilities.

A retail chain advertises the quality of products sold in its stores, which is subject of the companys own, particular strong quality guideline. In order to leverage compliance to this guideline, the company provides business partners along the supply chain with constraints on parameters that need to be monitored. A supplier uses these parameters in order to configure an integrity test for each package destined for this particular retailer; for accuracy, input parameters should be sensed and processed by the package itself or by IT infrastructure near the object. Performing this configuration task is supported by a hardware-abstraction layer, which allows for assigning tests to packages independent from the kind of label technology provided by the respective package.

During loading a truck, by means of this layer a dialog between package, truck, and an app store hosting implementations of tests matching the retailers parameters. Result of this dialog is an assignment determining which technical component (truck or package) has to conduct the monitoring task, an assignment, which may differ for each package.

Finally, the packages arrive at the retailers receiving area. There, an employee uses a mobile device in order to identify the respective object and access its digital records values sensed during transport as well as testing methods and their results. The access is performed using an app, which downloads from the app store a method suited to visualize the test chosen by the logistics expert. This visual adaptation is performed automatically in the background; even for very different kinds of packages the employee experiences always the same way of interacting with the respective object.

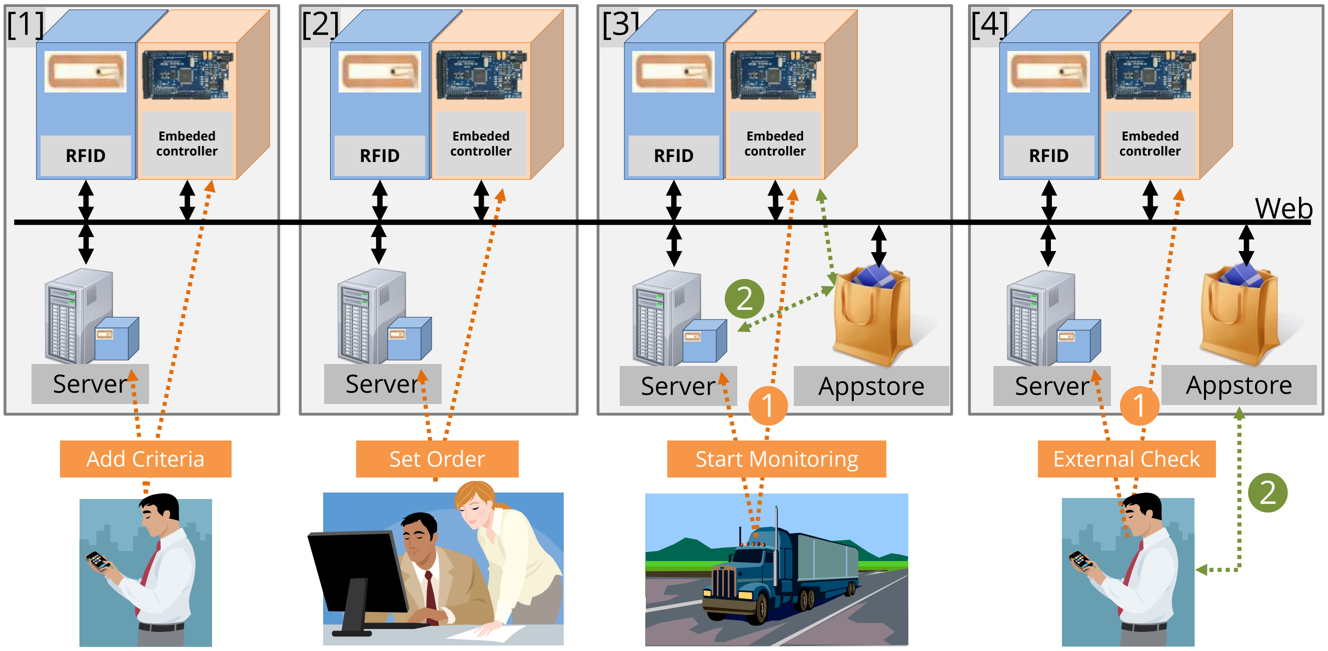


Fig. 1. Logistics scenario with on-product criteria storage [1], monitoring orders [2], active logic processing [3], and external checks [4].

III. REQUIREMENTS

The envisioned framework has to balance between the need for flexibility (to support different hardware platforms, open processes, new requirements in the future), for standardized structures that ease data retrieval and communication among different and cross-domain content providers and for normative description of object-related logic. Corresponding requirements to the architecture model can be divided in three parts that reflect a 3-tier approach to the realization of such a model, namely the collection model for data storage (1), support for active analysis functionality (2), and a common access architecture and infrastructure support for hardware abstraction with an app-store-like approach (3).

Data storage is the basic functionality of any ADOMe. In order to enable a party to analyze data added to the memory by another party, a common storage model is required, which is shared by all parties along the object's life cycle (Requirement 1, R1). In addition the model should be flexible enough to cover a large variety of data formats (including encodings and further data types) and should ease the process of data retrieval (R2). Due to the cross-domain usage of such memories involving several partners, a common infrastructure that provides an abstract memory access (including protocol and data exchange specifications) independent from any memory implementation or hardware platform is necessary to ease the task of memory access for existing and newly created applications (R3). Setting up activity and analysis support on top of the data storage can extend the functionalities of object memories, by allowing the memory to process data autonomously based on given algorithms. Based on this functionality we want to enable applications to ask common (but pre-defined) semantic questions (R4), rather than processing the entire memory data, which might be a complicated process due to possibly very large and capacious memories and in contrast a slow connection speed. In addition the memory should proactively process data with rules based on expert knowledge, and deploy results to memory storage or return the result on queries (R5). Finally, the system should provide a hardware abstraction layer, to enable a transparent access for clients (R6), which includes a compensation of missing object features by the environment as well as a mechanism to retrieve machine and platform-compatible logic code for the given use case (R7).

IV. THE ADOME FRAMEWORK

To fulfill the given requirements, we already implemented a framework for active object memories. The data within such an ADOMe are structured according to the Object Memory Model (OMM) [5]. This model partitions a memory in blocks with data of the same dedication and nature (R1). For the content of a block the model does not prescribe any data format or encoding. A block can be created and modified by external entities, which also define the content and encoding of the data of this specific block. The data of such a block can be either stored directly within the block as payload or are just linked to an external web-based source. In addition each block consists of a set of metadata, describing the content the block, creator and contributors, encoding and intention. Applications may use this metadata (e.g., attributes such as format, encoding, and namespace) to choose the best fitting information from the memory or utilize attributes like title and description to present memory blocks in a human readable form to end users (R2).

Accessing data collections for some object usually requires identifying this object as well as looking up the related collection infrastructure. The ADOMe framework does not impose particular constraints on the employed object identifier: in order to look up a memory, the framework resolves identifiers against the “IDs block” of hosted OMMs, which may contain arbitrary identifiers. Thus, access requires in the first place the address of the framework instance. This address is encoded as URIs (Unified Resource Identifier) suited to represent any

 ISBN: 978-1-61208-289-9

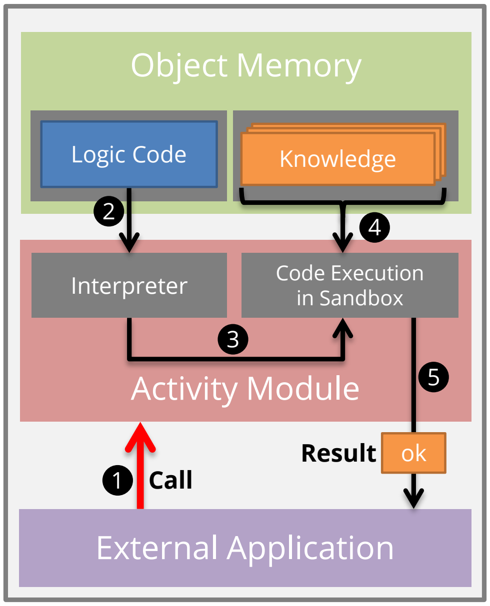


Fig. 2. Execution of ADOMe architecture triggered by external call (1-5).

installment. Based on the information about ID and URI, a root path to an object memory inside its container can be determined and directory and feature negotiation services can be accessed by a fixed REST-based (Representational State Transfer) interface that delivers URLs to pieces of the memory as well as functions implemented by the memory (R3).

Using the REST interface and OMM metadata, enables applications to employ their very own decision criteria to retrieve memory contents and process them outside of the framework. They even may store the employed decision criteria as contents within the memory, describe them through metadata, and thus share them with other applications (cf. [4]). However, this approach requires involved applications to implement the same command processor (e.g., a particular rule engine). Furthermore, the approach is little suited if a memory is required to autonomously prepare answers for future requests, as required by many applications focused on monitoring. Therefore, the ADOMe framework allows for extending its basic collection function with additional processing methods, which can then be explicitly selected applications to issue queries like check integrity (R4). Here, two concepts are mandatory for an ADOMe: a feature container providing means for statically implemented, optional, customer specific features, and a sandbox containing a logic processor. Both concepts can be triggered by external calls (see Fig. 2 and 3), by memory change events and by an internal timer (R5).

The feature container of an ADOMe provides a generic means to extend the functionality of object memories with so-called optional features. Installed by the owner of the framework during memory initialization, other parties may invoke and parameterize these features, for instance a producer may tailor its products memories to their respective application area. Optional features have to implement a REST-interface. For deployment, the framework provides a registration mechanism, which links the features services to the REST interface of the ADOMe. In addition, an ADOMe provides a processor for

Fig. 3. Performing built-in and on-product integrity check on a smart package equipped with ADOMe and accessed by a mobile handheld device.

software logic that allows for deploying custom evaluation functions for extracting meaning from data in the memory. If installed by the owner, a feature can be used by other parties to install their own compression, selection and accumulation methodologies, and to handle content types during the memories life cycle not foreseen during memory initialization. The logic processor runs in a sandbox environment that provides a generic interface integrated into the ADOMe's REST interface and a safe API to the memory storage of the ADOMe. The aforementioned framework has been implemented and has been used to realize several use cases utilizing active logic code. Hence, the framework can serve as the foundation for the following extension.

V. Hardware Abstraction & AppStore

Based on the given scenario we focus on a heterogeneous set of goods equipped with different techniques, ranging from simple barcodes to embedded systems with storage and processing capabilities. A common architecture to access different objects with the same approach and to access a similar functionality would be highly desirable from the user point of view. Currently we are developing a concept for hardware abstraction that allows accessing users and the objects themselves to complement their missing functionalities on their own, or at least to inform the outer environment about requested but not available functionalities.

This concept is based on the aforementioned ADOMe framework that is built on modular software components. This allows the framework to run on a large variety of platforms raging from high-end servers to small embedded systems (e.g., with storage capability only). The URI we use for memory identification and the RESTful interface for communication allows for transparent memory access independent of the concrete hardware and software implementation, and enables systems with missing functionalities (e.g., logic processing) to outsource these to server-based solutions by passing-through incoming requests (R6).

To combine software modules with object memory hardware for logic processing several approaches are possible. Applications can install and use software modules located in the memory storage (see Fig. 4). The processing and execution

 ISBN: 978-1-61208-289-9

Fig. 4. Client application displaying ADOMe-based measurements created by a software module downloaded from an external appstore.

location (either in embedded systems on-product or on server-based solutions off-product) is transparent for the access application. In case of no fitting software module available in memory, an access to a so called app store is possible based on a semantically defined logic description. This store contains a large set of logic modules for different applications and platforms, ranging from fixed domain-specific code to generic modules capable of parameterization. If available the framework downloads, installs and executes the downloaded module from the store. Client applications performing operations based on memory data can be extended with in-memory or app store modules the same way (R7).

VI. Related Work

This work is related to research and development concerning frameworks that leverage collecting and processing data related to physical objects, and as such related to the Internet of Things. Related research comprises embedded systems as well as web-based data stores. So-called Collaborative Business Items (CoBIs) ilustrate the benefits of delegeting small parts of a well-defined business process - e.g., monitoring and self-assessment tasks - to objects with embedded sensing and processing capabilities [6]. In order to decouple such a service from the employed hardware, SmartProducts [7] seek to dynamically integrate resources - including web-based structures - in the object's environment into the service realization. Complementary to our proposal, this work puts particular emphasis on semantic device and data descriptions for products with embedded technology. An example of collecting object-related data in a web-based data store is the Tales of Things electronic Memories (TOTeM) system. It seeks to foster communication between humans via personal stories digitally linked with things [1]. Its infrastructure shares aspects of an ADOMe, in particular a unified approach (which actually inspired parts of OMM) for structuring data concerning a thing, and open web-based information storage. EVRYTHNG [8] follows a similar approach in order to enable information sharing concerning objects; it extends it with Active Digital Identities for objects, where services linked with an object employ information collections (concerning the object, or objects of the same kind) in order to adapt to the user. The question of how implementation and provision of such services can be supported is addressed by Xively [9]. The web-based service supports not only hosting and sharing object data, but also software products and descriptions concerning devices, which we propose extending in a way that supports exchanging such components across devices using unified data structures and semantic descriptions.

VII. Conclusion and Outlook

This article summarized work in progress concerning a framework for setting up so-called Active Digital Object Memories. Its contribution is twofold: In order to leverage the application of such data collections in open scenarios, this framework seeks to 1.) embrace different ways of deploying answering logic in a collection, and 2.) provide abstraction from the technical diversity of existing infrastructures for collecting object-related data, which employ technology embedded into physical objects, virtual data stores located in the Web, and combinations of both approaches. Future work will address in the very first place the proposed method of distributing processing logic within this framework: the app store implementation. A first prototype illustrates the feasibility of this approach in a manufacturing scenario involving passive RFID, Android tablets, and embedded devices; however, more efforts are needed to verify that concept for broader range of embedded system platforms as well as logic hosted for deployment.

Acknowledgment

This research was funded in part by the German Federal Ministry of Education and Research under grant number 01IA11001 (project RES-COM). The responsibility for this publication lies with the authors.

References

[1] R. Barthel, K. Leder Mackley, A. Hudson-Smith, A. Karpovich, M. de Jode, and C. Speed, "An internet of old things as an augmented memory system," in *Personal and Ubiquitous Computing*, vol. 17. Springer London, 2011, pp. 321–333.

[2] D. Fitton, F. Kawsar, and G. Kortuem, "Exploring the design of a memory model for smart objects," in *Ambient Intelligence and Smart Environments*, vol. 4: Workshops Proceedings of the 5th International Conference on Intelligent Environments. IOS Press, 2009, pp. 33–38.

[3] P. Stephan, G. Meixner, H. Kling, F. Flrchinger, and L. Ollinger, "Product-mediated communication through digital object memories in heterogeneous value chains," in *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications*. IEEE, 2010, pp. 199–207.

[4] M. Schneider, M. Velten, and J. Haupert, "The objectrules framework - providing ad hoc context-dependent assistance in dynamic environments," in *Proceedings of the Sixth International Conference on Intelligent Environments. July 19-21, Kuala Lumpur, Malaysia*. IEEE Computer Society CPS, 2010, pp. 122–127.

[5] A. Kröner, J. Haupert, M. Seiler, B. Kiesel, B. Schennerlein, S. Horn, D. Schreiber, and R. Barthel, "Object Memory Modeling - W3C Incubator Group Report," http://www.w3.org/2005/Incubator/omm/XGR-omm-20111026/, 2011, [retrieved: June 2013].

[6] C. Decker, T. Riedel, M. Beigl, L. Moreira, S. Souza, P. Spiess, and S. Haller, "Collaborative business items," in *Proceedings of IE 07: 3rd International Conference on Intelligent Environments, Ulm, Germany*, 2007, pp. 40–47.

[7] M. Mühlhäuser, "Smart Products: An introduction," in *Constructing Ambient Intelligence: AmI 2007 Workshops*. Springer, 2007.

[8] EVRYTHNG Ltd., "EVRYTHNG Every Thing Connected," http://evrythng.com/, [retrieved: June 2013].

[9] Xively (by LogMein Inc.), "Xively - Internet of Things Platform Connecting Devices and Apps for Real-Time Control and Data Storage," http://xively.com/, [retrieved: June 2013].

 ISBN: 978-1-61208-289-9

The Effect of Feedback on Chord Typing

Adrian Tarniceriu, Pierre Dillenbourg, Bixio Rimoldi
School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
adrian.tarniceriu@epfl.ch, pierre.dillenbourg@epfl.ch, bixio.rimoldi@epfl.ch

***Abstract*—The amount of visual feedback when using a mobile device in a busy context is often limited. For example, while texting and walking in a crowded place, we need to focus on the environment and not on the phone. A way to type fast, accurately and with limited visual feedback is represented by chording keyboards. We present a study on such a chording keyboard prototype and analyze the influence of having visual, audio or no feedback at all on the typing process. The typing rates are the same under all three conditions, with an average of 20 words per minute, after approximately 350 minutes of practice. The average error rates are the lowest in the absence of feedback (2.41%) and the highest when the users can see what has been typed (4.03%). Considering these results, the proposed text input method is a viable option in situations where visual attention is already committed to other tasks.**

Keywords-chording keyboard; text entry; blind typing; feedback

I. Introduction

As mobile computing devices become increasingly popular, people want to be able to access them at all times. However, in a mobile environment, the amount of visual attention that can be devoted to a smartphone is often limited. This happens because the vision is already focused on the environment, and cannot be committed at the same time to the device or to the display. Many people initiate phone-calls while walking, but other functions such as text messaging or e-mail writing are less accessible. Even so, more than 40% of people write text messages while walking in public places [1]. This is potentially dangerous, as the visual attention is committed to typing and not to the surrounding environment. Therefore, to increase security and reduce the risk of accidents, it is important to find a method for entering text with limited visual feedback and without the need to look at the keys.

Using a chording keyboard [2] for text input will reduce the visual constraints. These keyboards enable users to generate a character by simultaneously pressing a combination of keys, similarly to playing a note on a musical instrument. With five keys, there are 31 combinations in which at least one key is pressed, enough for the 26 letters of the English alphabet plus five other characters. If the keys are placed in a position that is naturally under the fingertips, one can type with only one hand and without looking at them. The vision (or auditive feedback) is still needed occasionally to verify the output and to correct errors, but this requires considerably less commitment than continuously looking at the input device.

The likely reason chording devices are not very popular is that users require some training before being able to type, to learn the correspondence between key combinations and characters. A previous study [3] shows that people can learn to type with a five-key chording keyboard in less than 45 minutes if the key-to-character mapping is conveniently chosen.

We will analyze how different types of feedback affect the ability to type with a chording keyboard. To obtain the experimental data, users were asked to type under three different conditions: (1) with visual feedback, when they can see what has been typed; (2) with audio feedback, when they cannot see, but they hear the character that has been typed; (3) with neither visual nor audio feedback. We analyze the typing rates, the accuracy and the distribution of errors. We also evaluate the effect of the input device form factor on the typing process.

The paper is organized as follows. In Section II, we overview existing typing studies that evaluate different types of feedback. In Section III, we describe the experimental setup used in the study and in Section IV, we present the results. In Section V, we conclude the paper and discuss future directions.

II. Related Work

The condition when users type without looking at the text-input device and/or the display mostly occurs in mobile environments and is usually denoted as "blind" or "eyes-free" typing [4]. This explains why most blind-typing studies are performed using mobile keyboards such as 4×3 multi-tap keypads, mini-QWERTY, touchscreen keyboards, Twiddler [5], or other chording keyboards.

Silfverberg examined the effect of both tactile and visual feedback when using mobile phone keypads [4] and found that reduced tactile feedback increases the typing error rate. In addition, low visual feedback also leads to more errors, decreasing accuracy. A similar study made by Clawson et al. [6], concerning typing with mini-QWERTY keyboards, demonstrates the importance of seeing the keys while typing, but does not show any significant differences in typing

 ISBN: 978-1-61208-289-9

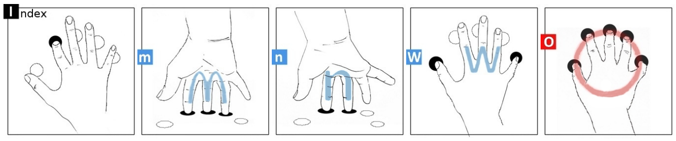


Figure 1. Examples of letter mappings for "**i**", "**m**", "**n**", "**w**" and "**o**".

speeds and error rates when users can or cannot see the typed text.

The above studies stress the importance of seeing the input device in the case of 4×3 multi-tap keypads and mini-QWERTY keyboards. However, this should not be an issue for most chording keyboards, which are specifically designed to be operated without looking at the keys. Typing experiments with limited visual feedback for the Twiddler chording keyboard were performed by Lyons et al. [7], and show that, surprisingly, typing and error rates actually improve with reduced visual feedback. Mascetti et al. propose and evaluate a Braille typing system for smartphones [8]. As it is intended for visually impaired persons, there is no visual, but only audio feedback.

Other studies, where participants do not look at the typing device or are involved in dynamic activities that require vision commitment, include the chording glove [9], a two-handed chorded software keyboard for PDA's [10], half-QWERTY touch typing [11], or the keyboard proposed by Gopher and Raij [12].

The chording keyboard used in this study has five keys, placed directly under the natural position of the fingertips. As users do not have to move their fingers from one key to another, it should make no difference if they are able to see the keys or not.

In a previous study [3], we proposed and evaluated a key-to-character mapping for a five-key chording keyboard. It is designed to minimize the learning time by assigning intuitive mnemonics to each character. Five examples of mnemonics are shown in Figure 1: "**i**" is given by the initial of the finger pressing the key (index); "**m**" and"**n**" are given by the shape of the fingers pressing the keys; "**w**" is given by the shape of the fingers not pressing the keys; for "**o**", we imagine five dots spread around a circle, and we obtain it by pressing all the keys. The complete mapping is given in the Appendix. The first part of the study evaluated the learnability of the mapping. We found that it can be completely learned in less than 45 minutes.

In the second part of the study, we analyzed the achievable text-entry rates and accuracy. We also assessed the difficulty of different key combinations by measuring the associated composition times. After 250 minutes of typing, the mean typing rate was 15.2 words per minute (wpm), with a maximum of 19.2 wpm. As a reference, after the same training time, the typing rates for multi-tap mobile phones are 12.4 wpm [13], for Twiddler 20.6 wpm [5], and for half-QWERTY approximately 24 wpm [4]. Rates of 20.3 wpm were reached by expert T9 users [14]. Note that the experimental conditions were not the same for all devices, and the given values are only indicative. The mean error rate during the experiment, accounting for both corrected and uncorrected errors, was 7.42%.

The chording keyboard was simulated on a regular desktop QWERTY keyboard. It only allowed for the use of five keys, each representing a key of the chording keyboard.

III. Experimental Setup

The input method that we present is designed to be used in situations where the visual attention is partially or totally unavailable for the typing process. In these conditions, audio feedback is often suggested as an alternative. This is indeed useful in some environments, but could be difficult to use in noisy areas. Considering this, we designed a 3×10 within-subjects experiment, where we analyzed three different typing conditions. Under the first condition, subjects are able to see the outcome of what they have typed, under the second condition they receive voice output for each typed letter (without visual feedback), and under the third condition they receive no feedback at all about the typing. From here on, we will refer to these conditions as visual, auditive, and no-feedback, respectively.

We asked ten PhD students from our university (eight male and two female, between 24 and 31 years old) to type using a chording keyboard. All of them had participated in the learnability part of the previous study, so they already knew the mapping and had had 45 minutes of training. They used a five-key chording keyboard prototype with the keys placed around a computer mouse, presented in Figure 2. We chose this design for the prototype because we wanted the subjects to see a practical application of a chording device that allows for typing and screen navigation at the same time. The keyboard is designed using an Arduino Pro Mini microcontroller board and communicates with the computer by Bluetooth. The buttons are placed in a position that is naturally under the fingertips when a user holds the palm on the mouse.

The participants were asked to type for 10 sessions of 30 minutes. Each session consisted of three rounds of 10 minutes separated by breaks of 2 minutes, and each round corresponded to a different typing condition. The order of the typing conditions was random for each session, but the same for all subjects. For each user, the typing sessions took place on consecutive days, with the exception of weekends.

The first session enabled the subjects to remember the mapping between keys and characters, and a help image showing the key combination for the letter to be typed was displayed. Starting with the second session, this image was no longer displayed. In the beginning of each round, the participants warmed up by typing each letter of the alphabet.

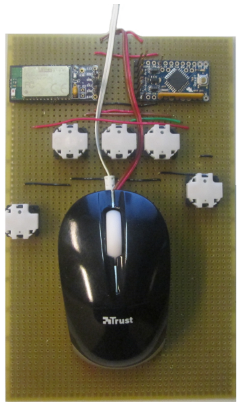


Figure 2. Chording keyboard prototype

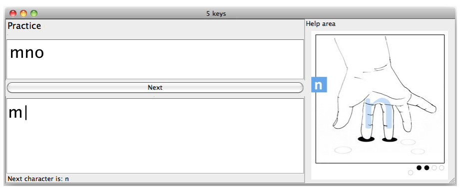


Figure 3. Typing application screenshot for the visual feedback condition

Afterwards, they typed phrases from a set considered representative of the English language [15]. These phrases were pre-prepared before the experiment to contain only small letters and no punctuation signs.

A Java application was designed to display the text to be typed, to monitor the pressed keys and, in case of errors, to check what character was typed in lieu of the correct one. A screenshot of the application for the visual feedback condition is shown in Figure 3. The top-left window contains the text to be typed, and the bottom-left window represents the typing area. The help image is displayed on the right.

The subjects were instructed to type as quickly and accurately as possible. They were told to not correct eventual mistakes and keep typing, but this was not enforced and they could delete typed text. As a reward for the time commitment during the experiment, they received a fixed monetary compensation for the first nine sessions. For additional motivation, for the last session, the reward was proportional to the number of typed words and to the accuracy.

The total amount of data gathered during the experiment consists of $40\,345$ words, out of which 4052 (10.17%) contain errors. The total number of characters is $219\,308$, from which 5889 (2.69%) are errors.

IV. Results

The main purpose of the experiment is to analyze how different typing conditions affect a person's typing rate and accuracy. Even if the mapping is the same, the input devices are different between this study and our previous work. This enables us to also verify the influence of the form factor of the device on the text entry process.

A. Typing Speed

We use the words-per-minute measure to describe the text entry speed. This is defined as

$$wpm = \frac{60L}{t}\frac{1}{5} \qquad (1)$$

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. Because the average word length for the typed text differs from one session to another, the use of the above formula provides a more reliable estimate than actually counting the words.

In Figure 4, we show the average typing rates for each session and for each condition. For the first three sessions, the rates are higher for the no-feedback condition, and the analysis of variance tests showed that the differences are statistically significant ($F = 10.85$, $p < 0.001$). From the fourth session onward, there are no more visible differences between the typing rates. Moreover, the effect of the feedback type is no longer significant ($F = 0.28$, $p = 0.75$). This probably happens because in the beginning subjects pause typing to check the provided feedback, visual or audio. As they gain experience, they become more confident and do not analyze the feedback so often, therefore reducing the differences between conditions.

In Figure 5, we show the typing rates for each user and for each session, during the no-feedback condition. As it can be noticed, the fastest subject can type three times faster than the slowest subject, the differences being statistically significant ($F = 53.8$, $p < 0.001$).

The average typing rates at the end of the experiment are 19.77, 20.16 and 20.00 wpm for the visual, audio and no-feedback conditions, with maximums of 31.24, 30.48 and 31.78 wpm, respectively. Considering the participants' experience from the previous experiment, these values correspond to approximately 350 minutes of practice. Because the text entry rates will probably still improve, we use

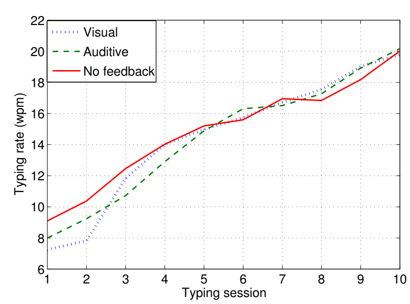


Figure 4. Average typing rates for each condition and for each typing session

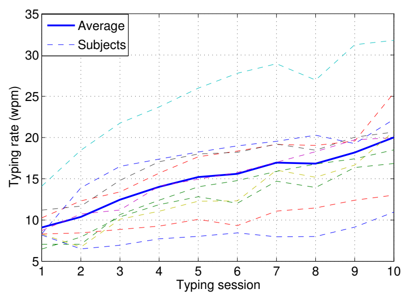


Figure 5. Average typing rates for each subject and for each typing session, for the no-feedback condition

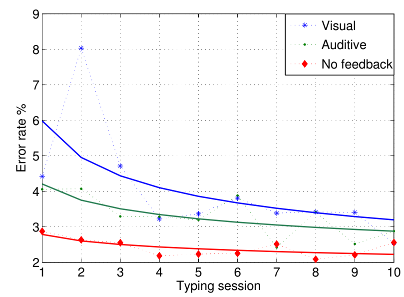


Figure 6. Average error rates for each condition and for each typing session

exponential regressions to estimate how fast people will be able to type after longer training periods. Based on these calculations, after 20 sessions (300 more minutes of practice), the average could be 26 wpm, and the fastest typist could reach 42 wpm.

B. Error Rates

The error rates are computed as the number of errors divided by the number of typed characters. Errors include substitutions (when one character is typed for another), insertions (when an extra character is typed) and deletions (when a character is missing). Each substitution, insertion and deletion counts as one error.

In Figure 6, we display the average error rates for each session, accounting for both uncorrected and corrected errors, and the corresponding exponential regressions. All of the error rates are below 5%, except one. The averages for all sessions and for all users under the visual/auditive/no-feedback conditions are 4.03%, 3.30% and 2.41%, respectively. Analysis of variance tests show that feedback plays a relevant role in the error rates ($F = 25.57$, $p < 0.001$).

Initially, it might seem surprising that the error rates are the lowest for the no-feedback condition and the highest for the visual condition. However, this is explained by the fact that increased cognitive loads generally lead to more errors [16]. For our study, the cognitive load is the highest in the visual condition: users can check the whole typed phrase; it is reduced by the audio condition when users only hear the last typed character, and minimum in the absence of feedback. Noticing an error could cause someone to become less focused, thus favoring new mistakes.

In Figure 6, we notice the high error rate for the second typing session, visual condition. The reason for this could be the fact that in the first session the help image was always displayed, whereas in the second session it was hidden. Moreover, the first typing condition in session 2 was the visual one, giving subjects more practice time for the audio and no-feedback conditions, which do not have much of an increase in the error rates.

The error rates decrease during the first four sessions (with the exception mentioned in the previous paragraph), but afterwards they remain stationary or even increase. Similar effects, when after a certain point the error rates do not decrease anymore as users gain experience, were also noticed by Matias et al. [11] and Lyons et al. [5].

Error Patterns: In Figure 7, we present the error rates for each character and for each typing condition. We notice that the character errors respect the pattern of the overall error rates: the highest for the visual condition and the lowest for the no-feedback condition, this being the case for 20 of the 27 analyzed characters.

For all three conditions, the error rates are higher for characters that are less frequent in the English language, such as “**q**” and “**j**”. The character error rates are similar between the three conditions, up to a scaling factor: if a character has an error rate lower than other characters for a specific condition, it usually also has a lower error rate relative to the same other characters for the other conditions. This is confirmed by the correlation coefficients between the error vectors, all above 0.9.

To analyze in more detail the error patterns, we determine the confusion matrices for each typing condition. They are square matrices 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 of the total number of occurrences for character i. The three matrices are similar, with correlation coefficients higher than 0.99. If we consider only the erroneously typed characters (by setting the diagonal values to zero), the correlation coefficients are above 0.9, still showing a strong similarity: if one character is frequently typed instead of another under one condition, the same will happen under the other two conditions; if the probability for one character to be typed instead of another is low under one condition, it is also low under the other two conditions.

C. Typing-Device Form Factor

When designing a keyboard, an important aspect is the form factor: where the keys are placed, how hard they have to be pressed, and what the provided tactile feedback is.

 ISBN: 978-1-61208-289-9

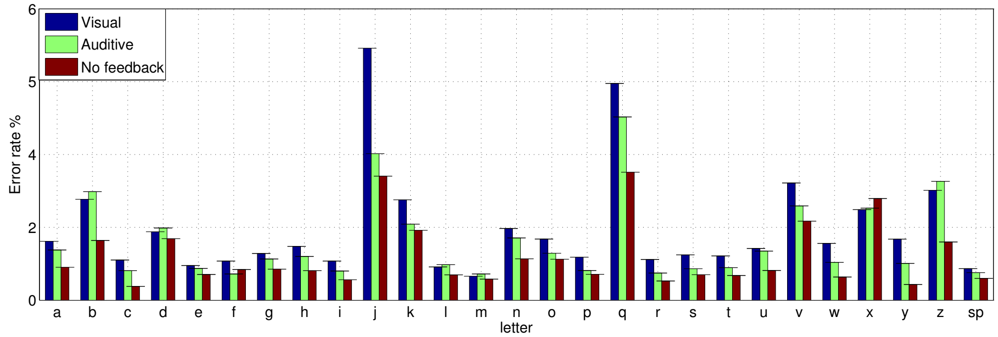


Figure 7. Error rates for each character for each typing condition

Modifying these characteristics will most likely influence the typing process, because a keyboard can be perceived based on them as difficult or comfortable to use. In the absence of visual/auditive feedback, tactile feedback is important because it could be the only way to let the user know that a character has been typed. This is why we used buttons that provide a strong passive feedback, and not touch or pressure sensors.

We evaluate the influence of the typing device form factor on the typing process by comparing the typing rates and the error rates under the visual condition with those from our previous work. Even if the experiments took place at different times, this is motivated by the fact that the main difference is the input device, whereas the other experimental conditions are similar. This time, the subjects used a real prototype, not a chording keyboard simulated on a regular QWERTY desktop keyboard; the keys are not the same and also placed in a different position. For both studies, the participants had the same education level (all master's or PhD students at our university), the gender distribution was similar (two females out of ten vs. one female out of six), the typed phrases were from the same set, the Java application used to see the text to be typed was the same, and the participants received similar financial compensations.

As can be observed in Table I, after the same amounts of training time, the typing rates are higher than those in the previous work (15.73 wpm, compared to 12.16 wpm). The time values start from 1.5 hours and not from 30 minutes (the duration of one session), thus including the previous typing experience of the subjects. The differences in the typing speed are statistically significant ($F = 12.16$, $p < 0.001$), thus confirming the importance of the input-device form factor.

TABLE I. AVERAGE TYPING RATES FOR THE VISUAL CONDITION AND FOR THE PREVIOUS EXPERIMENT

Training time (hours)	1.5	2	2.5	3	3.5	4
Visual condition (wpm)	7.23	7.82	11.83	14.01	14.97	15.73
Previous study (wpm)	6.79	7.83	8.87	9.40	10.37	12.16

The average error rate at the end of the previous study was 6.43%. After the same typing time, the average error rate for the visual condition was 3.91%. The difference could also be explained by different input devices and is statistically significant ($F = 55.2$, $p < 0.001$). Even if the average error rates are not the same, the error patterns are similar for the two studies, therefore we can conclude that they depend on the mapping more than on the typing condition or on the form factor.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented a study evaluating the effect of different types of feedback for a chording keyboard. The subjects were asked to type under three conditions: with visual feedback, with auditive feedback and with no feedback at all. Because of the keyboard design, whether the user can see the keys or not should not make any difference on the typing process — at the end of the experiment, participants confirmed that they did not look at the keys. Similarly, someone playing a saxophone does not look at the keys to be pressed.

After approximately 350 minutes of typing (taking into consideration the previous typing experience of the subjects), the average entry rates are approximately 20 wpm under all three conditions, with the maximums above 30 wpm. We conclude, therefore, that having visual, audio or no feedback has no influence on the typing speed.

The average error rates are 2.41% under the no-feedback, 3.30% under the auditive and 4.03% under the visual conditions. This is explained by the fact that the cognitive loads are different under the three typing conditions: the highest under the visual and the lowest under the no-feedback condition. Hence, not seeing the typed text actually represents an advantage. The error patterns are similar between conditions, the characters with the highest error rates and the most common substitutions being the same.

This study shows that the lack of visual or audio feedback does not impede the typing process, therefore the chording keyboard can be used reliably in situations where a person is not able to continuously check the output. Besides this,

 ISBN: 978-1-61208-289-9

the keyboard can be used with only one hand. The small number of keys also represents an advantage from the size and design flexibility point of view. As the study took place in an office, to go one step further, we should set up the experiment in a dynamic environment (for example, have the participants walking or jogging).

In addition to the effect of different typing conditions, the experiment enables us to evaluate the importance of different form factors for the input device: using the keyboard prototype and not a simulated keyboard as in our previous work leads to higher typing rates and lower error rates. However, no attempt was made to optimize the form factor, and other designs might further improve these measures.

So far, we have envisaged the chording keyboard as a means of typing in dynamic or busy environments. Due to its advantages, it can also be successfully used in other areas: For example, it can facilitate text input for disabled users who can only use one hand, or for persons who are visually impaired.

Appendix

In Table II, we present the key combinations corresponding to the characters used in our study.

TABLE II. FIVE-BIT CODES FOR THE USED CHARACTERS

Character	5-bit code	Character	5-bit code
a	00110	q	01101
b	10111	r	00010
c	10100	s	10101
d	11101	t	10000
e	11000	u	01001
f	01010	v	10011
g	11100	w	10001
h	11001	x	11011
i	01000	y	10110
j	01011	z	10010
k	11010	*space*	11110
l	00111	*backspace*	01111
m	01110	*enter*	00011
n	01100	*period*	00100
o	11111	*comma*	00101
p	00001		

Each key combination can be represented by a five-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, 10111 (the codeword for "**b**") means that all fingers except the index press the keys.

References

[1] H. Isaac, R.C. Nickerson and P. Tarasewich, "Cell phone use in social settings : Preliminary results from a study in the United States and France," in Decision Sciences Institute Conference, November 2004.

[2] J. Noyes, "Chord keyboards," Applied Ergonomics, vol. 14, no. 1, 1983, pp. 55 – 59.

[3] A. Tarniceriu, P. Dillenbourg, and B. Rimoldi, "Single-handed typing with minimal eye commitment: A text-entry study," in The Sixth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, September 2012, pp. 117 – 122.

[4] M. Silfverberg, "Using mobile keypads with limited visual feedback: Implications to handheld and wearable devices," in Human-Computer Interaction with Mobile Devices and Services (L. Chittaro, ed.), vol. 2795 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2003, pp. 76–90.

[5] K. Lyons, et al., "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), ACM, 2004, pp. 671–678.

[6] J. Clawson, K. Lyons, T. Starner, and E. Clarkson, "The impacts of limited visual feedback on mobile text entry for the twiddler and mini-qwerty keyboards," in Wearable Computers, 2005. Proceedings. Ninth IEEE International Symposium on, Oct. 2005, pp. 170 – 177.

[7] K. Lyons, D. Plaisted, and T. Starner, "Expert chording text entry on the twiddler one-handed keyboard," in Proceedings of the Eighth International Symposium on Wearable Computers, ISWC '04, (Washington, DC, USA), 2004, pp. 94–101.

[8] S. Mascetti, C. Bernareggi, and M. Belotti, "Typeinbraille: Quick eyes-free typing on smartphones," in Computers Helping People with Special Needs (K. Miesenberger, A. Karshmer, P. Penaz, and W. Zagler, eds.), vol. 7383 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 615–622.

[9] R. Rosenberg and M. Slater, "The chording glove: a glove-based text input device," Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, vol. 29, no. 2, 1999, pp. 186 –191.

[10] K. Yatani and K. N. Truong, "An evaluation of stylus-based text entry methods on handheld devices studied in different user mobility states," Pervasive and Mobile Computing, vol. 5, no. 5, 2009, pp. 496 – 508.

[11] E. Matias, I. S. MacKenzie, and W. Buxton, "One-handed touch typing on a qwerty keyboard," Hum.-Comput. Interact., vol. 11, Mar. 1996, pp. 1–27.

[12] D. Gopher and D. Raij, "Typing with a two-hand chord keyboard: will the qwerty become obsolete?," Systems, Man and Cybernetics, IEEE Transactions on, vol. 18, July-Aug. 1988, pp. 601 –609.

[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), 2001, pp. 111–120.

[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), ACM, 2001, pp. 365–371.

[15] I. S. Mackenzie and 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), ACM, 2003, pp. 766–767.

[16] A. Baddeley, Working Memory. Oxford Psychology Series, No 11, Clarendon Press, 1986.

Exploring the Tradeoffs of Configurability and Heterogeneity in Multicore Embedded Systems

Tosiron Adegbija and Ann Gordon-Ross*
Department of Electrical and Computer Engineering
University of Florida, Gainesville, Florida, USA
*Also with the Center for High Performance Reconfigurable Computing (CHREC) at the University of Florida
e-mail: tosironkbd@ufl.edu, ann@ece.ufl.edu

***Abstract*—Embedded systems, such as smartphones, have become ubiquitous due to the versatility of these devices for various applications, which have varying application resource requirements. Due to these variances, system resources should be specialized to the executing applications' resource requirements in order to adhere to design/optimization goals (e.g., reduced energy consumption, improved performance, etc.). In multicore systems, core heterogeneity and/or configurability affords specialization, however, this design choice exacerbates design challenges and complexity due to an embedded system's stringent design constraints. We evaluate the benefits and tradeoffs of heterogeneous and configurable cores as compared to homogeneous cores for design-constrained multicore embedded systems. Our studies provide valuable insights and guidelines on design choices and show that combining heterogeneity and configurability provides unique opportunities for fine-grained specialization.**

Keywords-heterogeneous cores; homogeneous cores; configurable architecture; low-power architecture; multicore

I. Introduction and Motivation

Multicore architectures are becoming prevalent in ubiquitous embedded systems (e.g., automotive systems, consumer electronics, smartphones, etc.) as an alternative to single-core architectures for achieving design/optimization goals, such as reducing cost, energy consumption, time to market, and/or increasing performance. However, this single- to multicore architecture shift significantly increases design challenges and complexity when coupled with an embedded system's stringent design constraints and resource availability (e.g., energy, power, area, real-time deadlines, size, etc.), which affords challenging design decisions. Additionally, designers must consider the applications' varying resource requirements during execution [12], thus necessitating *specialization* to the applications' unique requirements in order to adhere to design goals, which is becoming increasingly difficult to achieve using traditional homogeneous cores due to widely disparate application requirements.

One method to achieve multicore system specialization is by using disparate—heterogeneous—cores with varying characteristics (e.g., processor family/version, performance, die area, etc.). For example, the Open Multimedia Applications Platform (OMAP) chip contains a microprocessor core (ARM926) and a digital signal processor (DSP) coprocessor core (TMS320C55X) [20]. Even though the designer can select different cores to meet the varying applications' requirements, the design space is limited to the number of core combinations.

To increase adherence to design goals, configurable cores have configurable parameters (e.g., cache size, core frequency and/or voltage, etc.), whose values/configurations can be determined statically at design time or dynamically during runtime. In a configurable homogeneous core system [8], the cores have identical characteristics, but the cores' configurations are specialized to the applications' requirements. Depending on the level of configurability, the cores can have either the same or different configurations. Configurable core systems have large design spaces that consist of all combinations of parameter values, and thus provide more fine-grained specialization.

However, fine-grained specialization exacerbates design complexity, challenges, and decisions. Selecting between a heterogeneous or configurable homogeneous core system affects the level of design goal adherence, but this selection affects competing design goals, including design complexity, energy consumption, performance, runtime overhead, time to market, etc. For example, in heterogeneous core systems, designers have limited configuration options—lower design complexity—but must carefully select the most appropriate cores, and thus adherence to design goals may be limited due to the coarse-grained design space.

Alternatively, configurable homogenous cores may adhere more closely to design goals, but significantly increase design complexity, and time to market, since the cores' configurations must be tuned. Tuning evaluates an application's requirements and determines the best configuration for design goal adherence, but incurs overhead in terms of time, performance and/or energy overhead.

Concomitant to system specialization is application scheduling, which determines the most appropriate core to execute the application on [16]. Scheduling decisions, whether made a priori or at runtime, must consider the cores' characteristics and configurations since this information can significantly affect the system's adherence to design goals. Heterogeneous cores offer less specialization, thus designers must carefully select the cores to maximize the potential for design goal adherence. Configurable homogenous cores alleviate the core selection challenge and increase design goal adherence potential, but complicate scheduling decisions due to a larger design space.

Previous research for general purpose systems showed that heterogeneous and configurable homogeneous cores improve

 ISBN: 978-1-61208-289-9

energy consumption and performance as compared to homogeneous cores [11][12], however, there is little research with respect to the unique, and highly constrained, embedded system design goals. Prior scheduling and design space exploration methods [10] for embedded systems did not compare heterogeneity versus homogeneity. Furthermore, to the best of our knowledge, no prior work studied the tradeoffs (with respect to energy consumption, time to market, runtime overhead, etc.) between using heterogeneous and configurable homogeneous cores for achieving specialization, or whether configurable homogeneous cores provide an appreciable increase in design goal adherence to offset the increase in design and scheduling complexity and tuning overhead (e.g., energy, power, performance, etc.).

In this paper, we present an empirical comparison of the tradeoffs between heterogeneous and configurable homogeneous core embedded systems with respect to the energy delay product (EDP) and cache configuration and core frequency specialization. We also evaluate the EDP savings attained by using configurable heterogeneous cores, which leverage the advantages of both heterogeneity and configurability. Our evaluations provide valuable insights and guidelines to assist designers with design challenges and decisions.

The remainder of this paper is organized as follows. Section II presents the related work, Section III identifies the design challenges and studied architectures, Section IV discusses our experimental methodology, and our results are presented in Section V. Section VI posits multicore design suggestions and Section VII concludes the paper and discusses future work directions.

II. Related Work

Kumar et al. [12] proposed a single-instruction set architecture (ISA) heterogeneous multicore system to reduce power in general purpose computers, where each core provided different performance versus power tradeoffs. In [11], Kumar et al. showed that heterogeneous core systems provided power and throughput advantages for applications with varying execution requirements. Balakrishnan et al. [3] investigated the effects of data input size for recurring applications verses core scheduling, and concluded that heterogeneous core systems were beneficial for performance when core scheduling decisions considered the input size/characteristics. Grochowski et al. [9] studied heterogeneous cores with respect to energy and throughput improvements. However, all of these works focused on evaluating heterogeneity in general purpose computers, where throughput typically outweighs energy consumption.

Configurable core systems can be composed of any core with any configurable parameter(s). Zhang et al. [28] showed that applications have varying cache requirements and proposed a configurable cache architecture that determined Pareto optimal cache parameter values trading off energy consumption and performance, showing average energy savings of 40% as compared to a conventional, non-configurable cache. Gordon-Ross et al. [8] showed that configuring the cache to a particular application's requirements reduced memory access energy by 62% with performance improvements in most cases. Semeraro et al. [25] showed that dynamically scaling core voltage using multiple clock domains resulted in EDP savings of 20%. Albonesi [2] presented complexity-adaptive processors where the instruction per cycle (IPC)/clock rate tradeoff could be dynamically altered to match the application's changing requirements, and reduced time per instruction by an average of 9%.

Whereas prior works clearly motivate the benefits of heterogeneous cores and configurable cores, to the best of our knowledge, we are the first to investigate the tradeoffs between heterogeneous and configurable core systems for fine-grained specialization in embedded systems. Our studies and outcomes provide designers with valuable insights into design decisions when choosing an appropriate system configuration for specialization to the applications' requirements.

III. Design Challenges and Architectures

Since incorporating specialization into embedded system design imposes many daunting design challenges, this section details some of the challenges introduced when considering heterogeneity and configurability, and illustrates our evaluated system architectures.

A. Heterogeneity

In heterogeneous core systems, the cores' non-uniformity enables designers to statically select different cores that are suitable for different application requirements. If the applications are scheduled to the most suitable cores, performance and energy improvements are possible as compared to a homogenous core system. To provide a wide variety of suitability for diverse application requirements, most traditional heterogeneous core systems contain disparate cores [20]. However, large core disparity can introduce additional overheads and design challenges, especially if the cores have different ISAs, necessitating additional design time, area overheads, more complex scheduling, multiple binaries for each application, etc. Since leveraging cores with the same ISA, but different characteristics, eases system design while still offering specialization, we evaluate single-ISA systems, however, our fundamental tradeoff analysis is applicable to any heterogeneous core system with diverse core ISAs.

Since embedded systems typically have a large design space and several options for heterogeneity (e.g., ISA, core interconnect, memory hierarchy, etc.) one of the major challenges of heterogeneity is determining the key core characteristics that should differ in order to most closely adhere to design goals. The designer must evaluate the system and anticipated applications to select the appropriate cores, which places pressure on the time to market. Oftentimes this process is not straightforward at design time for general purpose embedded systems that execute a wide variety of applications (e.g., smartphones, tablets, etc.), and may require lengthy application evaluation/pre-analysis and design space exploration when the application(s) (or application domain(s)) is/are known a priori. Ideally, the cores would have enough diversity such that there exists a core that would be suitable for any application that could potentially execute on the system.

Scheduling further compounds the core selection challenge since the benefits of core diversity can only be exploited if the scheduler is aware of the cores' tradeoffs with respect to the applications' requirements. We evaluated the scheduling

 ISBN: 978-1-61208-289-9

policy's criticality on EDP (Section V) and observed that naïve scheduling decisions (e.g., the scheduler does not consider core tradeoffs, and randomly schedules applications to arbitrary cores) can severely degrade EDP, even resulting in higher EDP than a simple homogeneous core system. Whereas the scheduler must effectively analyze application versus core tradeoffs, this analysis and scheduling must not impose excessive overhead [4]. Prior works showed that effective scheduling can be integrated into the operating system, thus avoiding hardware overhead [7].

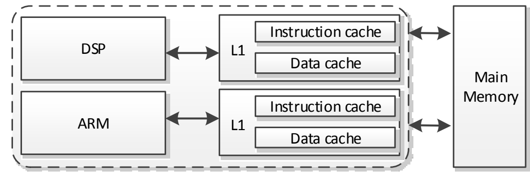
A simple, yet effective, approach to scheduling is the sampling-based method [4][12], which samples different application-to-core mappings at runtime and selects the best schedule based on design goals. This method introduces performance overhead due to periodic application migration across cores for application-to-core mapping evaluation, especially for systems with a large number of cores, and can incur significant overhead when executing an application on an unsuitable core. This overhead is less significant for heterogeneous core systems with replicated cores and in systems with persistent applications (e.g., smartphones), since the best schedule only needs to be determined once and can be reused for each subsequent application execution. Other prior works proposed more complex scheduling methods [4][7][10], however, due to sampling's simplicity, effectiveness, and ease of implementation, and thus appropriateness for embedded systems, we leverage sampling-based scheduling in our experiments, however, our fundamental tradeoff analysis could evaluate any scheduling method.

B. Configurability

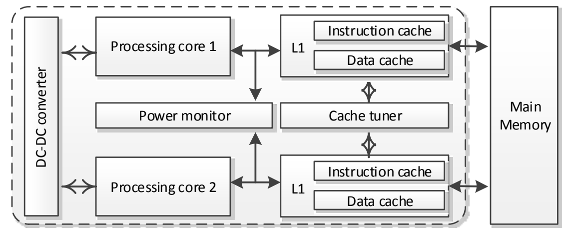
System configurability is key to adhering to design goals, thus much research has explored configurability with respect to instruction set extensions [14], core voltage [25], issue queue [6], reorder buffer [21], etc. Since research shows large potential EDP savings when combining dynamically configurable caches and core frequency [18], in this work we focus on these parameters, however our fundamental tradeoff analysis is applicable to any configurable parameters.

When using multiple configurable parameters, the design space increases rapidly, especially for interdependent parameters, thus exacerbating design challenges due to potentially intractable design spaces and large tuning overhead. One major challenge when leveraging configurability is specialization granularity, which determines how often the configuration changes. Application-based tuning [28] uses a single configuration that represents the best configuration for the average run of the entire application. Phase-based tuning [8] achieves finer-grained specialization by changing the configuration during application execution based on the application's varying runtime requirements. Whereas phase-based tuning increases design goal adherence potential, phase-based tuning requires identifying phase changes (changes in requirements) and determining the best configuration for each phase, thus increasing tuning overhead.

Dynamically determining the best configurations without incurring significant tuning overhead is especially challenging for large design spaces and fine-grained specialization. Heuristics [8] significantly prune the design space and analytical models/methods [1] can directly determine the best configuration sans design space exploration, thus significantly reducing tuning overhead. Despite the overheads, dynamic configurability alleviates costly system/application pre-analysis by the designer, thus resulting in a shorter time to market as compared to heterogeneous core systems [24].



(a)



(b)

Figure 1. Layout of a sample (a) heterogeneous dual-core system and (b) configurable homogeneous dual-core system

C. Illustrative System Architectures

Fig. 1 illustrates our evaluated dual-core system architectures, however, our evaluation methodology is applicable to any system with any arbitrary number of cores and any configurable parameters. Figure 1 (a) depicts a heterogeneous dual-core system with the following on-chip components: two processing cores with different clock frequencies and private level one (L1) instruction and data caches (iCache and dCache, respectively) with different cache configurations for each core. The clock frequencies and cache configurations are tuned at design time and remain static throughout the system's lifetime. Figure 1 (b) depicts a configurable homogeneous dual-core system with the following on-chip components: two identical processing cores with private configurable L1 instruction and data caches, and lightweight, low-overhead tuning hardware—a cache tuner and a DC-DC converter [18][28]. The cache tuner orchestrates dynamic cache tuning by changing the caches' configurations, evaluating and determining the best configurations, and fixing the system to run in those configurations. The DC-DC converter dynamically tunes the core frequency based on power measurements from the power monitor.

IV. EVALUATION METHODOLOGY AND EXPERIMENTAL SETUP

To quantify the EDP variances for heterogeneous, configurable homogeneous, and homogenous core systems, we modeled cache and core frequency configurations common to consumer embedded systems [10] (e.g., the Tegra 2 [26]) and evaluated dual-core systems, which are common in general purpose consumer embedded systems. Even though our experiments in this paper represent state-of-the-art embedded systems [26], our fundamental tradeoff analysis is applicable to future and/or more complex systems (e.g., n-core systems with multi-level caches) because the fundamental design goals and challenges are independent of these characteristics.

TABLE 1. DUAL-CORE SYSTEMS AND CONFIGURATIONS. THE CONFIGURABLE SYSTEM'S PARAMETER VALUES REPRESENT RANGES. THE HETEROGENEOUS SYSTEMS' PARAMETER VALUES REPRESENT CORES' VALUES AS CORE1/CORE2.

System	Cache size	Associativity	Line size	Clock frequency
Homogeneous	32 Kbyte	4 way	64 byte	2 GHz
Configurable	16 – 32 Kbyte	1 – 4 way	16 – 64 byte	1 – 2 GHz
Heterogeneous-1	16/32 Kbyte	4 way	64 byte	1/2 GHz
Heterogeneous-2	8/16 Kbyte	4 way	64 byte	800 MHz/1 GHz
Heterogeneous-3	8/32 Kbyte	4 way	64 byte	800 MHz/2 GHz

TABLE 2. TEST SCENARIOS

Name	Core descriptions
Test scenario 1	Naively-scheduled Heterogeneous-1
Test scenario 2	Optimally-scheduled Heterogeneous-1
Test scenario 3	Configurable homogeneous
Test scenario 4	Configurable heterogeneous

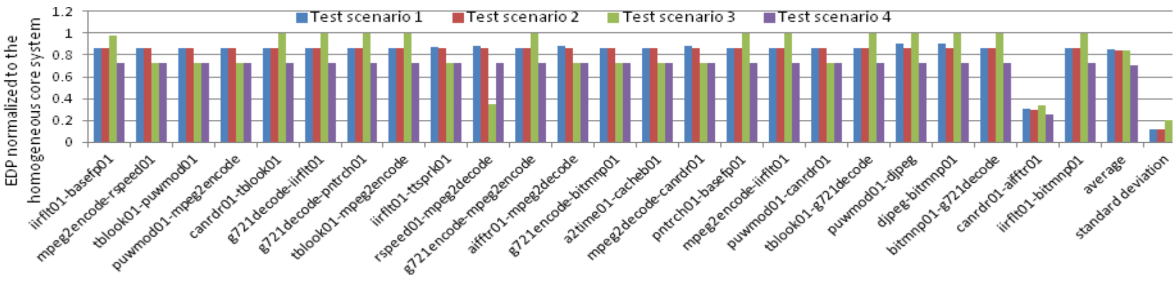


Figure 2. EDP normalized to the homogeneous core system (baseline of one) for the test scenarios in Table 2

Table 1 depicts our experimental dual-core systems and the systems' configurations, which represent actual embedded systems (e.g., Nokia Lumina 620 [19] and Motorola's Atrix 4G smartphones [17]). All cores had separate, private L1 instruction and data caches connected directly to off-chip main memory. Even though our evaluated systems could include a level two (L2) cache, our work evaluates the effects of L1 cache specialization, therefore we do not need to model the L2 cache, however, our work could be easily extended to include an L2 cache. The homogeneous core system served as our *base system* for comparison purposes. The configurable system represented the configurable homogeneous and configurable heterogeneous core systems, and offered cores with varying cache parameter values and operating frequencies (denoted as ranges). The configurable homogeneous core system's cores were tuned simultaneously to a single, homogenous configuration (the cores had the same configurations), resulting in a design space of 108 core configurations. The configurable heterogeneous core system's cores were tuned independently to heterogeneous configurations (the cores could have different configurations), resulting in a design space of 108^n configurations, where n is the number of cores.

In order to compare to static designer-selected cores, we evaluated three different heterogeneous core systems, denoted as Heterogeneous-1, -2, and -3. Based on empirical analysis, Heterogeneous-1 represented the best average configuration for all workloads (Section V), and thus served as the base heterogeneous core system. Heterogeneous-2 and -3 offer different core selection options for situations where designers cannot perform extensive design time analysis, and therefore the designer must make a "best guess" of the applications' requirements.

We exhaustively modeled and evaluated all configurations using GEM5 [5], which we modified to support heterogeneous cores, and McPAT [15] to calculate the systems' cores' EDPs in Joule seconds:

$$\begin{aligned} EDP &= core\_power * running\_time^2 \\ &= core\_power * (total\_cycles/core\_frequency)^2 \qquad (1) \end{aligned}$$

where *core_power* includes the core's components, such as the network interface units (NIUs), peripheral component interconnect (PCI) controllers, etc., and the cache, and *total_cycles* is the number of cycles for a single workload execution.

In order to reduce the sensitivity of the results to a particular set of simulated workloads and model real-world embedded system applications, we created twenty-four multiprogrammed workload sets by selecting two random single-threaded benchmarks from seventeen EEMBC [22] Automotive benchmarks (the entire EEMBC Automotive benchmark suite could not be evaluated due to compilation errors) and six random MediaBench [13] benchmarks for image, video, and audio processing. To ensure that both cores executed the same number of cycles and both benchmarks executed at least once to completion, we looped the faster benchmark in each set. The benchmarks were always scheduled to a separate core during execution, such that both cores were active throughout execution. Since embedded systems typically execute small applications with relatively stable characteristics throughout execution, we leveraged application-based tuning in our experiments.

Table 2 depicts our test scenarios. We determined optimal scheduling using sampling such that the lowest EDP schedule represented the best case. Since naïve scheduling randomly schedules applications to cores, we used the highest EDP schedule in order to compare with the worst case schedule. In practice, naïve scheduling would not necessarily achieve the worst case EDP, but comparing to the worse case provides a clearer picture of core selection tradeoffs. We determined the best configurations for the configurable homogeneous and configurable heterogeneous cores using exhaustive search for all of the workloads.

V. RESULTS

Fig. 2 depicts the EDP for the test scenarios in Table 2 normalized to the homogeneous core system (baseline of one) for all experimental workloads, denoted as the x-axis benchmark combinations, and the average and standard

 ISBN: 978-1-61208-289-9

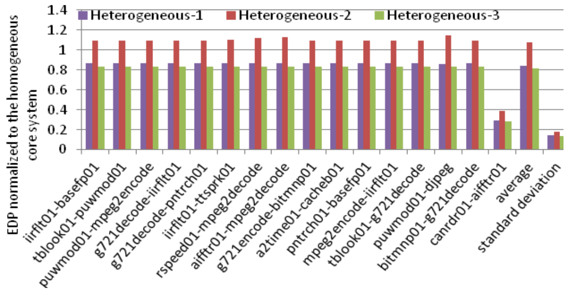


Figure 3. EDP of the heterogeneous systems in Table 1 normalized to the homogeneous core system (baseline of one)

deviation across all workloads. For brevity, we show comparisons with only the base heterogeneous core system, Heterogeneous-1.

Test scenarios 1 and 2: Compared to the homogeneous core system, the naively-scheduled Heterogeneous-1 system achieved similar EDP savings for most workloads, averaging 15% over all workloads, ranging from 9% for *djpeg-bitmnp01* and 69% for *canrdr01-aifftr01*, with a standard deviation of 12%.

The optimally-scheduled Heterogeneous-1 system achieved average EDP savings of 16% over the homogeneous core system, with savings increasing to 13% and 70% for *djpeg-bitmap* and *canrdr01-aifftr01,* respectively. This optimally-scheduled system performed only marginally better than the naïvely-scheduled system, 1% on average over all of the workloads, with differences as high as 4%, but *all* workloads benefited from heterogeneity regardless of scheduling. The EDP savings for these systems were similar because both cores' configurations represented good average configurations for all of the benchmarks (i.e., the cores were averagely-suited for all of the benchmarks), thus neither core resulted in prohibitively high EDP. However, we point out that we exhaustively evaluated our benchmark suite to determine the best average heterogeneous cores' configurations, and in practice, a designer would have to expend considerable design time effort to replicate these results. This comparison shows that scheduling decisions have little effect if the cores' configurations are carefully selected to match the applications' requirements.

Therefore, given limited time to market and potentially unknown applications, we evaluated heterogeneous core systems more representative of actual designer-selected cores. Figure 3 depicts the EDP of the optimally-scheduled Heterogeneous-1, -2, and -3 systems (Table 1) normalized to the homogeneous core system. Heterogeneous-2 *increased* the EDP by 7% on average as compared to the homogeneous core system, and Heterogeneous-3 and -1 decreased the EDP by 19% and 16% on average, respectively. Additional results, not shown for brevity, also showed that scheduling had a significant impact on EDP for these systems, with Heterogeneous-2 and -3 showing EDP *increases* of 65% and 26% on average as compared to the homogeneous core system, respectively, as compared to Heterogeneous-1's 15% decrease in EDP.

Test scenario 3: The configurable homogeneous core system revealed similar EDP savings as the optimally-scheduled Heterogeneous-1 system, with average EDP savings of 16%, ranging from no savings for some workloads to 66% for *canrdr01-aifftr01*, as compared to the homogeneous core system. Even though the configurable homogeneous cores afforded a larger design space, and potentially finer-grained specialization than Heterogeneous-1, since the cores could only select a *single* (same) average best configuration for all the benchmarks, the specialization granularity and EDP savings were reduced.

Test scenario 4: The configurable heterogeneous core system achieved the highest overall average EDP savings of 29% over the homogeneous core system, ranging from 27% to 75% for *djpeg-bitmnp01* and *canrdr01-aifftr01*, respectively, with a standard deviation of 10%. Compared to the configurable homogeneous core system and the optimally-scheduled Heterogeneous-1, -2, and -3 systems, the configurable heterogeneous core system achieved average EDP savings of 11%, 16%, 34%, and 13%, respectively. The configurable heterogeneous core system achieved the maximum EDP savings because each core was tuned independently for each application's requirements, thus affording the finest-grained specialization and the largest EDP savings. We note that these results do not include runtime tuning overhead, but prior work showed that these overheads can be minimal [23].

VI. MULTICORE SPECIALIZATION GUIDELINES

Previous work established that two cores introduce sufficient heterogeneity to attain appreciable EDP reductions [12]. In this paper, we significantly extend prior evaluations by studying the tradeoffs between core diversity and scheduling, and the benefits of incorporating configurability into heterogeneous core systems. Based on our findings, we suggest practical multicore specialization design guidelines intended to aid embedded system designers in selecting appropriate core characteristic and configuration diversity to meet design goals.

Our findings revealed a clear tradeoff between scheduling efficiency/effort and core diversity. Increased core diversity with a good balance of characteristics/configurations (e.g., Heterogeneous-2, as opposed to Heterogeneous-1 or Heterogeneous-3) enhances the benefits of heterogeneity by reducing the EDP, however, to realize these benefits, the scheduling policy must be very effective. With a less effective (i.e., naïve) scheduling policy, less core diversity is advantageous, such as using Heterogeneous-1, which contained good average configurations for the executing applications, such that either core will reveal good average savings, but neither core will reveal the largest savings. Therefore, designers can adjust their core selection efforts based on the system's scheduling policy

Maximum EDP savings is achieved using configurable heterogeneous cores where the cores can be independently tuned and the larger design space reveals greater EDP savings potential. Based on previous work [27], we conjecture that the design space can be significantly reduced with similar EDP savings by using cores with different configuration subsets. Each subset would be specialized to a different application domain's requirements, and applications from a particular domain would be scheduled to the appropriate core. Tuning would therefore only evaluate the reduced design space available on that core, which may be a fraction of the total,

 ISBN: 978-1-61208-289-9

unsubsetted design space (only three/four in [27]), thus significantly reducing tuning overhead, however, the tradeoff is extensive design time analysis.

VII. Conclusions and Future Work

Configurable and/or heterogeneous cores specialize a system's configurations to the applications' execution resource requirements to adhere to design goals, however, this specialization exacerbates design complexity, challenges, and decisions. System core selection (e.g., homogenous, heterogeneous, or configurable homogeneous/heterogeneous cores) affects the level of design goal adherence and affects competing design goals. In order to assist designers in core selection, we evaluated and empirically quantified the benefits and tradeoffs of heterogeneous and configurable core systems as compared to homogeneous core systems with respect to cache configuration and core frequency specialization. We provided insights and guidelines for designers and showed that the best energy delay product (EDP) savings can be achieved by using configurable heterogeneous cores, which leverage the advantages of both configurability and heterogeneity. However, since configurable heterogeneous cores result in exponentially large design spaces, our future work will explore and evaluate the impact of reducing the configurable heterogeneous cores' design space by designing heterogeneous core systems with different configuration subsets in each core, where each subset is specialized to a different application domain.

Acknowledgment

This work was supported by the National Science Foundation (CNS-0953447). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] T. Adegbija, A. Gordon-Ross, and A. Munir, "Dynamic phase-based tuning for embedded systems using phase distance mapping," International Conference on Computer Design, October 2012, pp. 284-290.

[2] D. Albonesi, "Dynamic IPC/clock rate optimization," Internation Symposium on Computer Architecture, July 1998, pp. 282-292.

[3] S. Balakrishnan, R. Rajwar, M. Upton, and K. Lai, "The impact of performance asymmetry in emerging multicore architectures," International Symposium on Computer Architecture, June 2005, pp. 506-517.

[4] M. Becchi and P. Crowley, "Dynamic thread assignment on heterogeneous multiprocessor architectures," Computing Frontiers, 2006, pp. 29-40.

[5] N. Binkert, B. Beckmann, G. Black, S. Reinhardt, A. Saidi, A. Basu, J. Hestness, D. Hower, T. Krishna, S. Sardashti, R. Sen, K. Sewell, M. Shoaib, N. Vaish, M. Hill, and D. Wood, "The gem5 simulator," Computer Architecture News, May 2011, pp. 1-7.

[6] D. Folegnani and A. Gonzalez, "Energy-efficient issue logic," Internation Symposium on Computer Architecture, July 2001, pp. 230-239.

[7] S. Ghiasi, T. Keller, and F. Rawsom, "Scheduling for heterogeneous processors in server systems," Proc. Computing Frontiers, ACM Press, 2005, pp. 199-210.

[8] A. Gordon-Ross and F. Vahid, "A self-tuning configurable cache," Design Automation Conference, July 2007, pp. 234-237.

[9] E. Grochowski, R. Ronen, J. Shen, and W. Wang, "Best of both latency and throughput," International Conference on Computer Design, October 2004, pp. 236-243.

[10] K. Kim, D. Kim, C. Park, "Real-time scheduling in heterogeneous dual-core architectures," International Conference on Parallel and Distributed Systems, 2006, pp. 91-96.

[11] R. Kumar, D. Tullsen, N. Jouppi, P. Ranganathan, "Heterogeneous chip multiprocessors," Computer, November 2005, pp. 32-38.

[12] R. Kumar, D. Tullsen, P. Ranganathan, N. Jouppi, and K. Farkas., "Single-ISA heterogeneous multi-core architectures: the potential for processor power reduction," International Symposium on Microarchitecture, December 2003, pp. 81-92.

[13] C. Lee, M. Potkonjak, and W. Mangione-Smith, "MediaBench: a tool for evaluating and synthesizing multimedia and communications systems," International Symposium on Microarchitecture, December 1997, pp. 330-335.

[14] R. Leupers, K. Karuri, S. Kraemer, and M. Pandey, "A design flow for configurable embedded processors based on optimized instruction set extension synthesis," Design, Automation and Test in Europe, 2006.

[15] S. Li, J. Ahn, R. Strong, J. Brockman, and D. Tullsen, N. Jouppi, "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," International Symposium on Microarchitecture, December 2009, pp. 469-480.

[16] J. Luo and N. Jha, "Battery-aware static scheduling for distributed real-time embedded systems," Design Automation Conference, 2001, pp. 444-449.

[17] Motorola Atrix 4G. http://www.motorola.com/us/consumers/Motorola-ATRIX-4G/72112,en_US,pd.html [retrieved: April, 2013]

[18] A. Nacul and T. Givargis, "Dynamic voltage and cache reconfiguration for low power," Design, Automation, and Test in Europe, February 2004, pp. 1376-1377.

[19] Nokia Lumina 620. http://www.nokia.com/global/products/phone/lumia620/ [retrieved: April, 2013]

[20] OMAP1710 processor architecture and features: http://focus.ti.com/general/docs/wtbu/wtbuproductcontent.tsp?templateId=6123&navigationId=11991&path=templatedata/cm/product/data/omap_1710 [retrieved: April, 2013]

[21] D Ponomarev, G. Kucuk, and K. Ghose, "Reducing power requirements of instruction scheduling through dynamic allocation of multiple datapath resources," International Symposium on Microarchitecture, December 2001, pp. 90-101.

[22] J. Poovey, M. Levy, and S. Gal-On, "A benchmark characterization of the EEMBC benchmark suite," International Symposium on Microarchitecture, December 2009, pp. 18-29.

[23] M. Rawlins and A. Gordon-Ross, "An application classification guided cache tuning heuristic for multi-core architectures," Asia South Pacific Design Automation Cconference, 2012, pp. 23-28.

[24] B. Ristau, T. Limberg, and G. Fettweis, "A mapping framework for guided design space exploration of heterogeneous MP-SoCs," Design, Automation and Test in Europe, March 2008, pp. 780-783.

[25] G. Semeraro, G. Magklis, R. Balasubramonian, D. Albonesi, S. Dwarkadas, H. Dwarkadas, and M. Scott, "Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling," High Performance Computer Architecture, February 2002, pp. 29-40.

[26] Tegra 2 super chip processors. http://www.nvidia.com/object/tegra-superchip.html [retrieved: April, 2013]

[27] P. Viana, A. Gordon-Ross, E. Keogh, E. Barros, and F. Vahid, "Configurable cache subsetting for fast cache tuning," Design Automation Conference, July 2006, pp. 695-700.

[28] C. Zhang, F. Vahid and W. Najjar, "A highly-configurable cache architecture for embedded systems," 30th Annual International Symposium on Computer Architecture, May 2003, pp. 136-146.

Smart Chair: What Can Simple Pressure Sensors under the Chairs' Legs Tell Us about User Activity?

Jingyuan Cheng, Bo Zhou, Mathias Sundholm, Paul Lukowicz
Group Embedded Intelligence
German Research Center for Artificial Intelligence
Kaiserslautern, Germany
{jingyuan.cheng, bo.zhou, mathias.sundholm, paul.lukowicz }@ dfki.de

***Abstract*—In this paper, we investigate how much information about user activity can be extracted from simple pressure sensors mounted under the legs of a chair. We show that it is possible to detect not only different postures (0.826 accuracy for 5 subjects and 7 classes), but also subtle hand and head related actions like typing and nodding (0.880 accuracy for 5 subjects and 5 classes). Combining features related to postures and such simple actions, we can detect high-level activities such as working on a PC, watching a movie or eating in a continuous, real-life data stream. In a dataset of 105.6 hours recorded from 4 subjects, we achieved 0.783 accuracy for 7 classes.**

***Keywords*—ubiquitous computing; smart chair; resistive pressure sensor; activity recognition**

I. Introduction

In general, people spend a lot of time sitting: when working in the office, in a meeting, in the theater, when having a meal, when playing games or watching TV. In [1], it has been shown that on average, people with sedentary jobs sit 9.95 hours at working days and 8.07 hours when not working. From the above, instrumented chairs may seem like an obvious approach to activity recognition. On the other hand, there is the question that how much information can be extracted from chair-mounted sensors. Per definition, when sitting, people tend to engage in activities that involve little motion of the body trunk and are mostly determined by hand actions and cognitive processes. At the same time, the body trunk is the main physical interface between the chair and the user, and thus the main potential source of information. As a consequence, the vast majority of work on chair based activity monitoring has focused on posture analysis and the measurement of physiological parameters that can be extracted from the trunk by electrodes integrated in the seat (see related work section). Such work has so far relied on sensors integrated in the seat itself or the backrest (often using smart textiles), which implies significant effort in terms of unobtrusive instrumentation in every day environments.

A. Paper Contribution

Building on such work, the core ideas that this paper explores, are two fold:

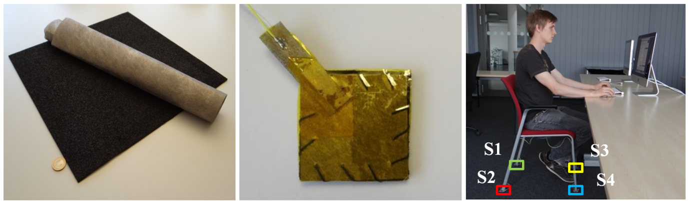
1) Many shoulder, arm, hand and head related actions leave a subtle but characteristic signature in terms of weight distribution changes on the chair. Such signatures can be combined with the correlation between posture and activity for the recognition of complex activities, which are not directly related to trunk motions.
2) The weight distribution and posture changes can be detected using simple, cheap pressure sensors mounted under the chair, without the need to instrument the seating surface or the backrest. This means that existing chairs can be easily retrofitted, allowing for the instrumentation of entire offices, meeting rooms, or theaters.

From the above ideas, the paper makes the following concrete contributions:

- We describe the measurement system, which is based on our own resistive foam sensor, which is simple and cheap; while at the same time providing the required resolution and measurement range.
- We validate the hypothesis that posture changes can be detected using pressure sensors integrated under the chairs' legs (as opposed to pressure sensors matrices on the seat itself as investigated by related work). This is done in experiments with 5 subjects and 7 postures (e.g. sit straight, lean left / right / forward / backward, raise one hand, cross one leg over the other knee), giving a recognition rate of 0.826 on average.
- We validate the assumption that actions related to arm, hand and head motion produce detectable signatures in the pressure signals. This is done in experiments with 5 subjects and 5 actions (e.g. typing keyboard, clicking mouse, nodding, clapping hands and sitting still), giving a recognition rate of 0.880 on average.
- We demonstrate that such information can be used to discriminate high-level activities such as playing a computer game, working on the computer and eating a snack. From 4 test subjects, 105.6 hours of data recording, we are able to divide high-level activities into 7 classes and achieve an overall accuracy of 0.783.

B. Related Work

As early as 1997, Tan, etc. studied the possibility of using a chair as a control interface [2], the idea led to a series of papers with on-chair sensor matrix, used for real-time posture tracking or game interface [3]. By using a near-optimal sensor placement strategy, Multu, etc. managed to reduce the sensor number to 19 while still having 78% accuracy from 10 postures [4]. Beside posture, pressure matrix can also recognize who is sitting on a car seat [5]. There are already



(a) (b) (c)

Fig. 1. System Design; a) raw materials for the pressure sensor; b) a sensor prototype; c) a chair equipped with sensors

a wide range of pressure sensors/matrices commercially available with different shapes, sizes, pressure ranges [6].

Another approach has been to use the contact area between the torso and the chair to measure various physiological parameters. Examples are the measurement of vital signs in aircraft seats [7], car seats [8] and regular chairs [9].

C. Paper Structure

In this paper, we first introduced the system setup in Section II. Then the evaluation in Section III is done by recognizing postures, minor activities and daily activities from experiment results. At last, current work is concluded, and future plans proposed.

II. System Design

Our prototype is composed of a normal office chair with 4 pressure pads, one under each leg. The hardware is completely hidden under the chair (as shown in Fig. 1). With a 5500mAh battery, the system runs continuously for 36 hours.

A. Physical background

The above system provides two types of information: (1) the total weight (vertical force component) applied to the chair; (2) the distribution of the weight/force applied to the four legs. Obviously, different body postures lead to different weight distributions (sitting straight or leaning aside). Thus, the low frequency "baseline" of the signal is essentially determined by posture and posture changes. On top of this baseline, there are higher frequency fluctuations related to the fact that body parts do not move in isolation. Instead, the human body is a coupled mechanical system where even a subtle movement of one part influences many others. For example, nodding is not only the head moving up and down, but a complex action concerning the whole vertebral column, the buttocks and sometimes the feet. Even subtle limb motions can propagate all the way to the chair's legs. As a consequence, many activities, that one may not obviously associate with weight distribution changes (e.g. typing), have a weak but characteristic signature in the high frequency signals from pressure sensors in the chair's legs. In addition, different high-level activities are associated with certain posture types (e.g. leaning forward when typing).

B. System Design

The core requirements resulting from the above considerations are high dynamic range and high precision. In our sensor, the baseline weight of a person sitting in a chair produces a signal $\sim 2V$ while the smallest labeled signal for "typing", is in the range of single mV. At the same time, since we aim to eventually be able to easily retrofit whole meeting rooms or theaters, a simple, low cost system is needed.

Our solution is based on polyethylene foam that changes its resistance as a function of mechanical pressure. The resistance of a single $3mm$ layer varies from several hundred $k\Omega$ to several $M\Omega$. Preliminary tests indicated that the force from an occupied chair is large enough to saturate a single layer. As a consequent state, to improve the stiffness of the whole sensor pad, we have compressed 4 layers of $1 \times 1cm^2$ PE foam into the thickness of one layer and embedded it into the center of a single layer $4 \times 4cm^2$ pad using normal thread. Two electrodes made of metal textile are then secured by normal thread and isolating tape, which turns the whole pad into a resistive pressure sensor.

Each sensor is connected in series to a resistor; a low noise 2.5V DC voltage is applied across the two, converting resistance change to voltage change, which is then fed into a 4 channel 24-bit ADC sampling at $25Hz$. The overall noise level of the analog circuit is $\sim 0.22mV(RMS)$.

To enable the user to freely move the chair and to isolate the noise from mains power, we use a battery to power the system and a 2.4GHz Zigbee module to transfer data.

III. Evaluation Strategy and Results

We evaluate the system in two steps. First we demonstrate the ability of our system to recognize a set of predefined postures and simple actions in controlled lab experiments. We then proceed to investigate the usefulness of our concept for the discrimination of more complex, high-level activities in real life data streams.

A. Predefined Postures and Minor Activities

The subject is seated on the chair in front of a desk with a computer and asked to perform the activity displayed on the screen, in total 12 postures and actions $\times$ 20 times each. The postures are sitting straight, leaning forward / backward / left / right, sitting with one leg cross the other knee, and sitting with one hand raised in the air. The actions are, nodding, clapping hands, typing on the keyboard, moving and clicking the mouse. We also include the class "vacant chair' in the experiment. Each activity lasts for 15 seconds, followed by a small pause, where the subject returns to the default posture (sitting straight). The activities' order is randomized in each round except "vacant chair", which is always at the round's end so that the test subject seat him/herself differently in the next round.

Overall 5 healthy subjects (1 female, 4 males, aged 23-34 years) participated in the data recording.

The first and last 2 seconds of each activity are left out, where the subject reacts to the instructions or returns

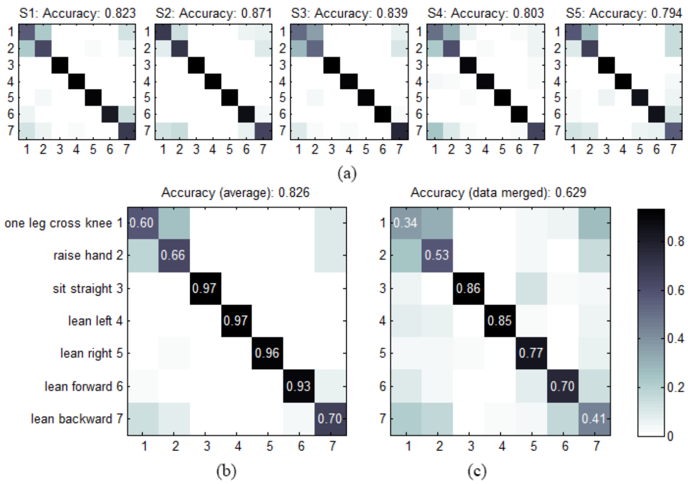


Fig. 2. Confusion matrix of predefined postures: a) each test subject b) average of all subjects c) when merged into a single dataset

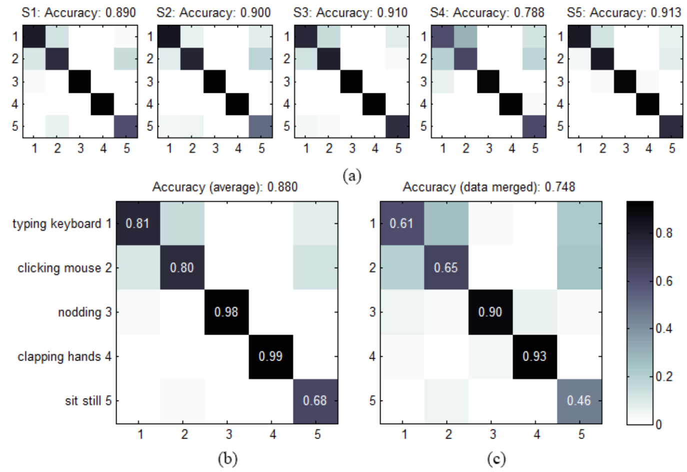


Fig. 3. Confusion matrix of predefined activities: a) each test subject b) average of all subjects c) when merged into a single dataset

to the default posture. The following features are used for classification:

1) Cross-channel features (8 total):

- Mean and RMS of 4 sensors combined.
- Center of weight in two directions: calculated as the differences between sensors on the left and right side of the chair, and between the sensors on the front and back side of the chair, both divided by the sum of all 4 sensors.
- Activity level features: calculated as mean, median of the summed 1st derivative's absolute values in time domain of 4 sensors.
- Median and mean of absolute values sum of 4 sensors after removing DC using a high-pass filter ($f_c = 0.5Hz$).

2) Single channel features (7× 4 total):

- The mean magnitude, the central frequency, and the magnitude of the central frequency in 5 uniform frequency bands between 0 and $12.5Hz$.

We first evaluate 7 sitting postures. The classification is performed using stratified 10-fold cross-validation with an LDA classifier [10]. The average accuracy with subject dependent training is 0.826 (e.g. if the chair knows who is sitting on it, balanced F-score in Fig. 2 b). If data from all subjects is merged into one (e.g. the chair doesn't know the user), then the accuracy drops to 0.629 (Fig. 2 c)).

We then evaluated the 4 simple actions of the hands and the head: typing on the keyboard, clicking the mouse, clapping hands and nodding. The first two are tiny activities with hand and little arm movement; in the latter two, the vertebral column also moves a little. A 5th class is added as sitting still, where we randomly pick 20 out of 100 postures without movement (viz. sitting straight and leaning aside). The average accuracy is 0.880 for the user independent and 0.748 for the across all users case (details in Fig. 3).

B. High-level Daily Activities

Next we consider 7 high-level activities that are routinely performed while sitting: working on PC, eating, playing video games, watching movie, talking with others, browsing the Internet and vacant seat. We record data from 4 healthy subjects (1 female, 3 males, aged 24–34 years). Thus, the subjects are asked to sit on the equipped chair and perform their normal work routine (which mostly takes place in front of a computer in the student room) for at least 8 hours × 3 days. They are also asked to play games and watch movies (which we assume is not part of a daily work routine). For privacy and practicability issues, the experiment is not video recorded and not precisely labeled. Instead, the subjects keep a log of their major activities throughout the day in an experience sampling like approach. Overall, the data encompasses 105.6 hours, out of which 79.2 hours the chair was occupied (details in Table. I).

One key feature for distinguishing the above high-level activities is the ”activity level”, defined as mean and median of the 1st derivative' absolute value in time domain. It increases starting ”watch movies”, through ”play games”,”browse the Internet”,”work on PC”,”eat”,”talk with others”. Even though the exact way of performing the specific high-level activities differs across subjects, their activity levels are still comparable.

Another important feature is the weight center, which is related to the attention on the PC screen for certain activities. It is further in the back when watching movie and browsing website where overview of the whole screen is important. When talking or eating, no screen is needed, so occupants tend to sit comfortably in the middle; while when it comes to computer work and game, where close attention needs to be paid to the screen and frequent keyboard/mouse operation is required, the weight center is pushed forward.

In a 5 min window jumping in steps of 30 sec over the daily data stream, the same set of features were calculated as in Section III-A. When the windows covers multiple high level activities,a majority decision is made. Classification is performed using 10-fold cross-validation method with an LDA classifier. The confusion matrix is given in Fig. 4.

The accuracy is 0.783 for the user dependent case and 0.645

 ISBN: 978-1-61208-289-9

TABLE I
STATUS AND DURATION FOR EACH SUBJECT (HOURS)

Status	Center of weight	Typical activities	S1(F)	S2(M)	S3(M)	S4(M)	Sum hours
watch movie	back/middle	seldom any movement	3.7	2.4	2.8	1.6	10.5
play game	front/middle	frequent mouse movement, seldom body movement	4.4	1.6	4.8	1.6	12.4
browsing the Internet	/	some mouse and body movement	0.4	1.9	2.4	0.5	5.2
computer work	much in the front	frequent typing on keyboard, some mouse and body movement	12.5	9.9	8.9	10.5	42
eat	front/middle	some arm/body movement	0.5	0.1	1.7	0.4	2.7
talk with others	middle/back/aside	bursting arm/head/body movement	0.8	2.7	1.5	1.4	6.4
nobody on chair	/	much less pressure, no movement	10.2	6.7	1.5	8.5	26.4
Sum hours			32.6	25.3	23.7	24.0	105.6

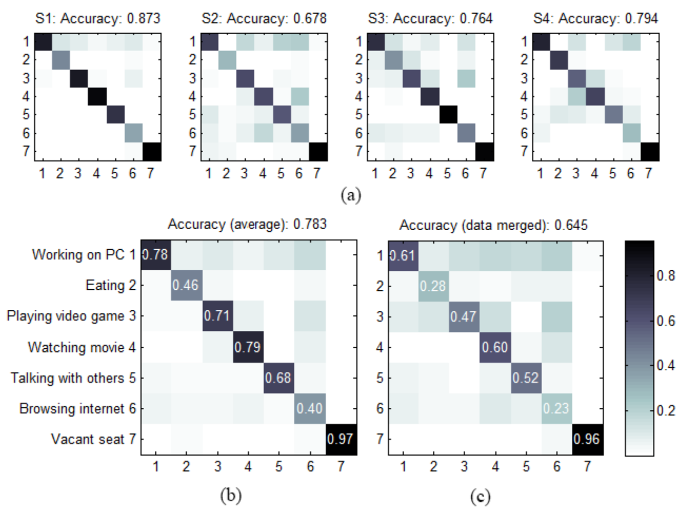


Fig. 4. Confusion matrix of high-level daily activities: a) each test subject b) average of all subjects c) when merged into a single dataset

accross all users. This is far from perfect, but also far above random, making it possible for anonymous implementations. Note that errors are not only due to misclassification, but also related to labeling inaccuracies and the complexity of the activities. For example, working on computer might well include browsing the Internet for a short while (to seek for information). Some subject ate snacks when watching movie, or browsed the Internet while having main meal.

IV. CONCLUSION

While the classification results shown in the previous section are far from perfect; they are also far above random. Overall, we believe that the fact that such results can be achieved using signals from just four simple pressure sensors mounted under the chairs legs is surprising and relevant for a wide range of applications. In particular the ability to easily retrofit large rooms (e.g. a conference room, theater) opens up the interesting research opportunities in the area of social interaction. Another possibility is to equip other furniture (table, bed, couch and etc.) for more complex activity recognition at home or in public spaces; yet the physical sensors need to be improved to support larger force. We will also investigate the possibilities of putting the sensors under furniture with different support, such as castors. In such an approach, visual and audio information is not required, therefore implementation could be anonymous, protecting the occupants' privacy. Finally, we will investigate the combination of furniture integrated sensors with information from users' smart-phones as a way of tying activities to specific persons.

Note that although our validation is done in the office and based on office activities; the aim of the paper is not to develop a practical system for the tracking of office activities and/or related applications. Instead, the main contribution is to show that high-level activities not necessarily related to body torso trunk motions can be detected from simple, easily retrofitted sensors integrated under the chairs' legs and the office activities are merely an easily obtainable dataset.

REFERENCES

[1] C. G. Ryan, P. M. Dall, M. H. Granat, and P. M. Grant, "Sitting patterns at work: objective measurement of adherence to current recommendations," Ergonomics, vol. 54, no. 6, 2011, pp. 531–538.
[2] H. Z. Tan, A. Pentland, and I. Lu, "The chair as a novel haptic user interface," in Proceedings of the Workshop on Perceptual User Interfaces, Banff, Alberta, Canada, 1997, pp. 56–57.
[3] H. Z. Tan, L. A. Slivovsky, and A. Pentland, "A sensing chair using pressure distribution sensors," Mechatronics, IEEE/ASME Transactions on, vol. 6, no. 3, 2001, pp. 261–268.
[4] B. Mutlu, A. Krause, J. Forlizzi, C. Guestrin, and J. Hodgins, "Robust, low-cost, non-intrusive sensing and recognition of seated postures," in Proceedings of the 20th annual ACM symposium on User interface software and technology, ser. UIST '07. New York, NY, USA: ACM, 2007, pp. 149–158. [Online]. Available: http://doi.acm.org/10.1145/1294211.1294237
[5] X. Xie, B. Zheng, and W. Xue, "Object identification on car seat based on rough sets," in Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on, 2011, pp. 157–159.
[6] I. Tekscan, "Tekscan pressure sensors," http://www.tekscan.com/pressure-sensors, accessed: 2013-05-07.
[7] J. Schumm et al., "Unobtrusive physiological monitoring in an airplane seat," Personal and Ubiquitous Computing, vol. 14, no. 6, 2010, pp. 541–550.
[8] M. Walter, B. Eilebrecht, T. Wartzek, and S. Leonhardt, "The smart car seat: personalized monitoring of vital signs in automotive applications," Personal and Ubiquitous Computing, vol. 15, no. 7, 2011, pp. 707–715.
[9] Y. G. Lim, K. K. Kim, and S. Park, "Ecg measurement on a chair without conductive contact," Biomedical Engineering, IEEE Transactions on, vol. 53, no. 5, 2006, pp. 956–959.
[10] W. J. Krzanowski, "Principles of multivariate analysis: A user's perspective," 1988.

VANET Route Selection in Urban/Rural Areas using Metric Base Traffic Analysis

Syed Fakhar Abbas
Communication Network Research Center (CNRC)
Department of Computer Science, COMSATS Institute of Information Technology
Defense Road, Jinnah Campus, Lahore, Pakistan
syedfakhar15@gmail.com

S. R. Chaudhry
Communication Network Research Center (CNRC)
Department of Computer Science, COMSATS Institute of Information Technology
Defense Road, Jinnah Campus, Lahore, Pakistan
saqib.chaudhry@ieee.org

Ghulam Yasin
Division of Science and Technology
University of Education
Lahore Pakistan
ghulam_yaseen@hotmail.com

***Abstract*— Vehicular Ad hoc Network (VANET) is a subordinate of Mobile Ad hoc Network (MANET). It helps for the provision of wireless communication between the vehicles in the road side equipment and the vehicle. Nowadays, Intelligent Transportation System (ITS) has a major impact on improving the quality and efficiency of the transportation system. When we talk about luxury and secure implementation of VANET, vehicle will query data from any other vehicle through multi hop infrastructure. While data is moving in VANET it will suffers from recurrent interruptions due to frequent mobility and sporadically linked network system. In this paper, studies and comparison of the following three routing protocols AODV, DSR and DSDV have been done. The analysis has been measured and evaluated in both urban and rural environments. Contrary to the pessimistic conclusion of previous works, by incorporating the classification accuracy is improved in the routing algorithms used. The proposed system analyzes the vehicles density, data drop, and throughput and end-to-end delay. It was observed from the results that in form of high through put and low packet drop DSR shows better performance compared to DSDV and AODV in rural environment and urban high density area, while AODV shows better performance in comparison to DSR in a VANET environment that is of low density.**

Keywords- ITS; VANET; AODV; DSR; DSDV

I. INTRODUCTION

Wireless technology in the field of has been growing fast previous more than two decades. Developments have played a vital role to explore new horizons doors for researchers to dig out unique cost-effective solutions for different applications. Cellular networks, Ad hoc networks, ,wireless sensor networks, Visible Light Communication, Wi-Fi, Wi-Max are some of the illustrations of new wireless network technologies that have been used in security, telecommunication, engineering, location tracking, network monitoring, remote sensing, medical, education, and tracking systems.

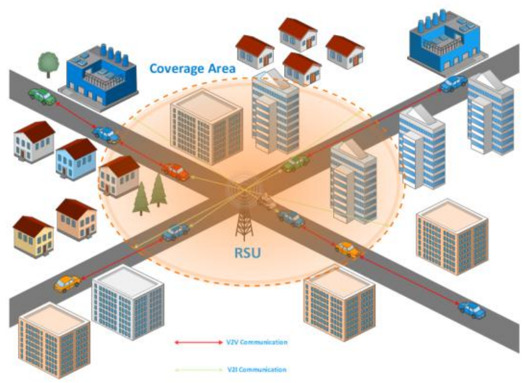


Figure 1. Communication architecture of VANET.

Fig. 1 shows the architecture of VANET in which different components are shown. This is a generic architecture in which data is transferred among different On-Board Units (OBU). Intelligent Transportation Systems (ITS) is a wide-ranging technology system applicable to transportation to make system safer, more effective, and more reliable and more environment friendly, without altering the existing infrastructure. Technologies ranges include sensor network, control technologies, communications system, computer informatics, transportation, engineering, telecommunications, computer science, finance, electronic commerce and automobile manufacturing.

Intelligent Transportation Systems (ITS) have been developed to improve the safety, security and efficacy of the transportation systems. The field of Inter Vehicular Communications (IVC), including both Vehicle-to-Vehicle communication (V2V) and Vehicle-to-Roadside communication (V2R), also known as VANET, is recognize as an important component of ITS in various national plans [1].

Figure 2. Inter Vehicular Communications (IVC).

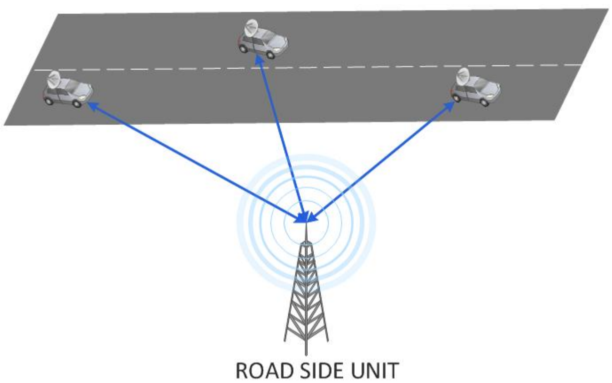


Figure 3. Vehicle to Infrastructre Communication (V2I).

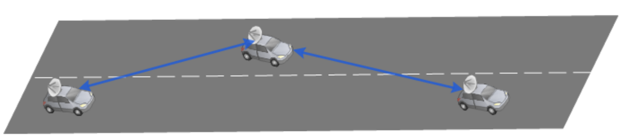


Figure 4. Vehicle to Vehicle Communication (V2V).

Vehicular Ad-Hoc Network (VANET) communication has become a progressively important research topic in the area of wireless networking as well as the automotive industries. The goal of VANET research is to develop a vehicular communication system to enable cost-effective and fastest communication of data for the benefit of passengers' safety [2] and comfort.

A vehicular ad hoc network (VANET) is an emerging research area for the communications industry and academician. Researchers proposed an entirely new wireless networking concept, i.e., vehicular ad hoc network which can increase passenger safety and reduces vehicle collisions on the road. Wireless communication among moving vehicles is unique and innovative research era in the academics and in the corporate sector, driven by the vision to communicate information among vehicles to ensure the safety and comfort of the users [3–4].

Nowadays, automobile industry have equipped their new vehicles with Global Positioning Systems (GPS), digital maps and even wireless interfaces, e.g., Honda-ASV3.The network architecture of VANET can be classified into three categories: Cellular network/WLAN, ad hoc network, and hybrid network [5].

This paper summarizes the impact of topology-based routing protocols in urban and rural area scenarios. The remainder of the paper is organized as follows: Section 2 is about description of problem statement. Section 3 describes the wireless access methods for VANET. Section 4 discusses the different routing protocols in VANET. Section 5 presents the experimental and model scenarios. Section 6 analyzes the results of urban and rural area scenarios. We conclude in Section 7.

II. PROBLEM STATEMENT

With the emergence of vigorous wireless network technologies and wireless applications intensive research have been conducted in VANET. Vehicles are needed to equip with Global Positioning System (GPS) and IEEE 802.11 wireless adaptors to create ad-hoc network for data sharing. Path is pre-established for data transfer and network configuration changes frequently.

Challenges faced by ad hoc network include security, bandwidth limitation, energy utilization, scalability, network performance etc. while measuring different parameters mentioned above we use different protocols to show the network performance and quality of service. The proposed system classifies the changing vehicular density, data drop, and throughput and end-to-end delay of vehicles in the network.

To guarantee the performance of network multiple protocols have been analyzed in the simulation scenario like AODV [13], DSR [13] and DSDV. Performance has been measured in rural and urban areas using following parameters: changing vehicular density, data drop, and throughput and end-to-end delay of vehicles in the network.

 ISBN: 978-1-61208-289-9

III. Wireless Access Standards for Vehicular Technology

Dedicated Short Range Communications (DSRC)/Wireless Access in a Vehicular Environment (WAVE) refer to a set of developing standards for mobile wireless communications. DSRC/WAVE is part of the Federal Highway Authority's Vehicle Infrastructure Integration (VII) initiative and supports Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) communications for emerging Intelligent Transportation Systems (ITS). DSRC/WAVE systems fill a role in the wireless infrastructure by facilitating low delay, expansion geographical area, high data rate, and high mobility communications.

A. DSRC Dedicated Short Range Communications

5.9 GHz DSRC (Dedicated Short Range Communications) is communications service which ranges between short to medium to supports vehicle and vehicle to vehicle communication environments. DSRC is meant to be an accompaniment to mobile communications by providing fast data transfer rates in surrounding environment where less delay during the communication and separating small communication zones are important [7]. DSRC modes of operations are Ad hoc mode characterized by distributed multi-hop networking (vehicle–vehicle), and Infrastructure mode characterized by a centralized mobile single hop network (vehicle-gateway).

B. IEEE 802.11P/WAVE SYSTEM

The IEEE has accomplished the standards IEEE P1609.1, P1609.2, and P1609.4 for vehicular networks and just released them for experimental use to the network users [7]. A fourth standard, P1609.3 is still under further development. P1609.1 is the standard for Wireless Access for Vehicular Environments (WAVE) Resource Manager. It describes the services and interfaces of the WAVE resource manager application as well as the message data format. It delivers instant access for different applications to the other network architecture. The WAVE stack usages is an altered version of the IEEE 802.11a, known as IEEE 802.11p [8], for its Medium Access Control (MAC) layer protocol. It uses (Carrier Sense Multiple Access/Collision Avoidance) CSMA/CA as the basic medium access protocol for link sharing and usages one control channel to set up transmissions, which then should be done over some transmission channels. By using the OFDM (Orthogonal Frequency Division Multiplexing) system, it delivers both (Vehicle to Vehicle) V2V and V2R (Vehicle to Road-Side) unit wireless communications over long distances up to 1000m,while taking into account the environment, that is, high absolute and relative velocities (up to 200 km/h), fast multipath fading and dissimilar scenarios (rural, highway, and city).

IV. VANET Routing Protocols

In VANET, the routing protocols are classified into five major categories:

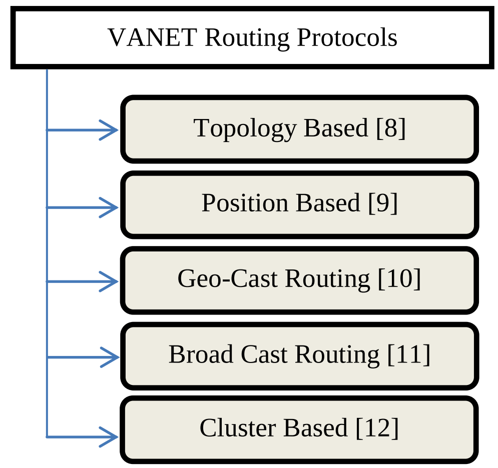


Figure 5. VANET routing protocols Categories

VANETs consist of vehicles which have to follow the traffic rules of movement unlike MANETs in which nodes move randomly without any movement restrictions. VANETs have very flexible and complex topology due to multiple routes on which drivers move at different speeds, with different driving behavior, whereas in MANET's topology changes are less frequent. Due to these distinguished differences between MANETs and VANETs, MANETs routing protocols have to be studied first and verified for their compatibility in VANET environments. The routing protocols, which are selected for this study, belongs to a special branch of MANET routing protocols namely Topology Based Routing Protocols. The main reason for such a selection is the dynamic topology aspect of VANETs, which has a straight implication on routing protocol analysis.

A. Topology Based Routing Protocols

In this paper, we have considered proactive and reactive protocols. Topology based routing protocols use link's information, which exists in the network and use as packet forwarding. These protocols can be categorized into:

- Proactive (Table-Driven) routing protocols.
- Reactive (On-Demand) routing protocols.
- Hybrid routing protocols.

B. Proactive-Routing Protocols

Proactive routing protocol is also named as "table-driven" routing protocol. In proactive routing protocol, nodes in a mobile ad-hoc network incessantly assess routes to all accessible nodes and try to uphold reliable, up-to-the-minute routing information in their routing tables. FSLS, DSDV, OLSR etc. are examples of proactive-routing protocols [14].

1) Destination Sequenced Distance Vector (DSDV)

It is Table Driven routing protocol which is used in VANET; is grounded on classical Bellman-Ford algorithm [15]. Primarily every vehicle broadcasts its own route information tables to its neighbor vehicles. The neighbor vehicles keep up-to-date routing table by two type of packets- Full Dump Packet and Incremental Packet. Full Dump Packet comprises information about every contributing vehicle in the VANET. These packets are communicated intermittently after a long time intermission. Incremental Packet covers latest change in vehicle position since last Full Dump Packet. Routes are nominated with the up-to-date entry in the table. DSDV is better option for networks where location of nodes is less changeable.

C. Reactive Routing Protocols

Reactive routing protocols for mobile ad hoc networks are similarly known as “on-demand" routing protocols. In a reactive destination node but no route is accessible, it initiates a route detection process. It is initiated with RREQ (route request) packet, response is with route reply (RREP) and while link is not available it is received route error (RERR) packet. Reactive routing protocols has less overhead, a unique feature, while reactive routing protocols scalable than proactive routing protocols. But while using reactive routing protocols, source nodes may undergo routing protocol, routing paths are look for only while it is desirable. Henceforth these protocols are not appropriate for real–time applications. The Dynamic Source Routing (DSR) [13] and Ad hoc On-demand Distance Vector routing (AODV) [13] are examples for reactive routing protocols.

1) Ad Hoc on-Demand Distance Vector Routing (AODV)

It is a Source Initiated on Demand routing protocols used in VANET. In this protocol, every vehicle retains its route information of every vehicle. Sequence number is used to accept an acknowledgment of update for table entry. If a table entry is not used within a certain time limit, it will be erased from table and if there is any route is disconnected from vehicle to another vehicle, route error (RERR) packet is generated so that vehicle route is efficiently updated in its routing table.

2) Dynamic Source Routing (DSR)

It is Source Initiated on Demand routing protocol used in VANET and is grounded on link state routing algorithm. When a vehicle needs to transfer data to another vehicle, first it initiates route discoveries request up to that vehicle. For route finding, source vehicle recruits a route request (RREQ) packet in the network and other nodes forward the RREQ by changing their name as sender. Lastly when RREQ packet spreads to the destination vehicle or to a vehicle having path to the destination vehicle, a route reply (RREP) packet is unicasted to the sender node. If the reply packet is not received, the source vehicle resumes violent discovery of route up to the destination vehicle.

V. EXPERIMENTAL MODEL AND SCENARIOS

In the vehicular networks, basically, two different types of simulator are required; first, is network simulators, and second, is traffic simulator. But, in proposed solution, a hybrid simulator is used which offers a mixing services of both network and traffic simulator. The hybrid simulator EstiNet 7.0 is the commercial version of NCTUns network simulator and emulator. EstiNet 7.0 is a world-renowned tool and has been used by more than 20,000 listed users from 144 countries all over the world. EstiNet 7.0 is latest version and whose central technology is based on the novel kernel re-entering methodology invented by Prof. S.Y. Wang [10]. The various features of VANET supported by EstiNet 7.0 make it a clear choice for proposed research work.

A. Simulation Model and Parameters

TABLE I. GENERIC SIMULATION PARAMETERS

Routing Protocols in both scenarios	AODV, DSR and DSDV
Channel type	Wireless channel
MAC protocol	802.11p
Mobility model	Random Way Point(RWP)
Theoretical Channel Model	Two Ray Ground

B. *Urban Area Scenario*

Fig. 6 shows the urban area grid scenario where 80 vehicles devices identified as On-Board Units (OBUs) communicate with each other as well as with RSU (Road Side Unit) .Vehicles show the network behavior as the OBUs move within the network to analyze the performance of each protocol. While assessing the performance of a given scenario in the Vehicle-to-Vehicle communication (V2V) and Vehicle-to-Roadside communication (V2R) vehicles move within network and establish VANET. In this mobility model we used Random waypoint. By using this mobility model, vehicles are free to move to reach at random destination. Movement of the vehicles is calculated by the algorithm.

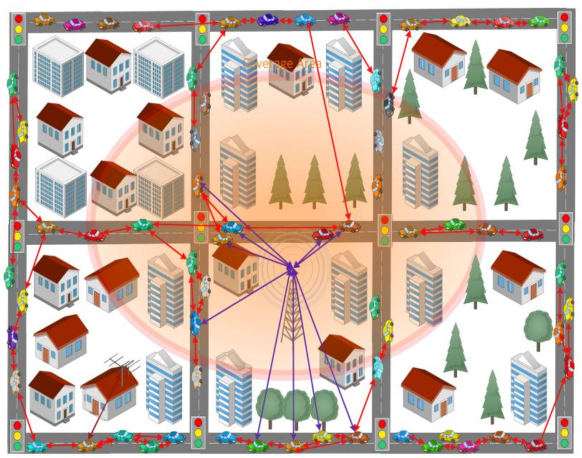


Figure 6. Urban Area Grid Scenario.

Table II describes the simulation parameter used in the urban area scenario.

TABLE II. URBAN AREA SIMULATION PARAMETERS

No. of Vehicles	80
Simulation area	2X2Km
Simulatio Time	90 minutes
Vehicle Speed	10 m/sec
Channel bandwidth	6 Mbps
Transport protocol	UDP
Transmitted power	28.8dbm

C. *Rural Area Scenario*

Figure 7. Rural Area Scenario.

Fig. 7 shows rural area grid scenario where 15 On-Board Units (OBUs) communicate with each other as well as with RSU (Road Side Unit) .Vehicles show the network behavior as the OBUs move within the network to analyze the performance of each protocol.

Table III describes the simulation parameter used in the rural area scenario.

TABLE III. URBAN AREA SIMULATION PARAMETERS

No. of Vehicles	80
Simulation area	2X2Km
Simulatio Time	50 minutes
Vehicle Speed	10 m/sec to 18 m/sec
Channel bandwidth	6 Mbps
Transport protocol	UDP
Transmitted power	28.8dbm

VI. PERFORMANCE RESULTS OF AODV, DSR, DSDV

The following graph shows the performance of the routing protocol using different metric considered above. The x axis represents the time in seconds and y axis represents the metric considered.

A. *Throughput*

From Fig. 8, it can be realized that in urban area scenario throughput with DSR protocol is better than with AODV and DSDV with throughput peak reaching up to 300 kB/s. It is also observed that, when speed is increased and more vehicles connected to the RSU than AODV performance suddenly degrades from about 270 KB/s to 40 KB/s, while DSDV performance slightly decrease remains moderately the same.

 ISBN: 978-1-61208-289-9

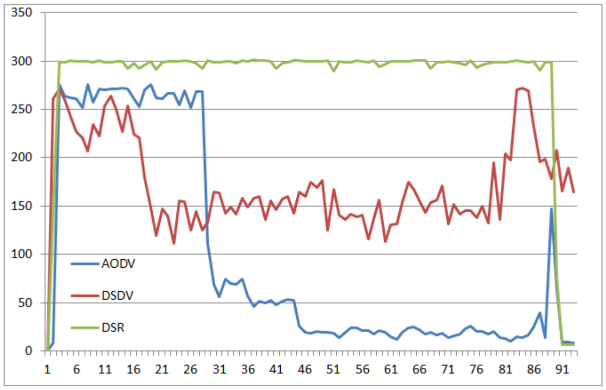


Figure 8. Throughput Performance in Urban Scenario.

Fig. 9 shows the Throughput in rural scenario. Clear implication from graph DSR Throughput is uppermost. AODV Throughput remains in between other two and DSDV Throughput is lowest among all three.

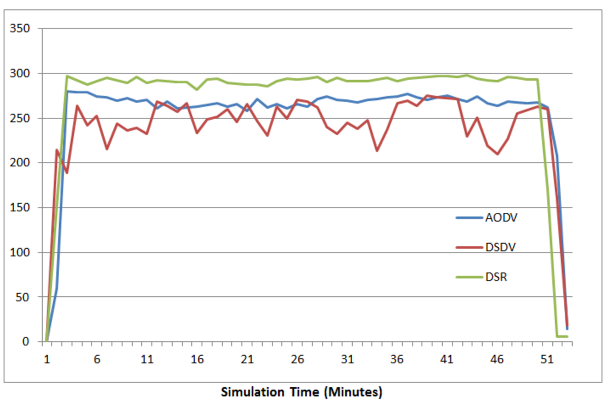


Figure 9. Throughput in Rural Scenario.

B. *Packet Dropped*

The packet drop performance of DSR protocol is much better than AODV and DSDV in urban and rural scenarios, as is seen from Fig. 10. About 10 to 20 packets dropped in urban area while less than 50 packets drop in rural scenario while this drop of packets better as time passes for the DSR protocol. As the speed is increased the packet drop rate for AODV protocol increases from 200 to 1100 drop packets in urban area whereas in rural area packet drop for AODV remain between 100 to 140 packets. In terms of dropped packets for DSDV's performance is the worst in both scenarios. The performance degrades with the increase in the number of nodes and speed. The performance degrades with the increase in the number of nodes and speed.

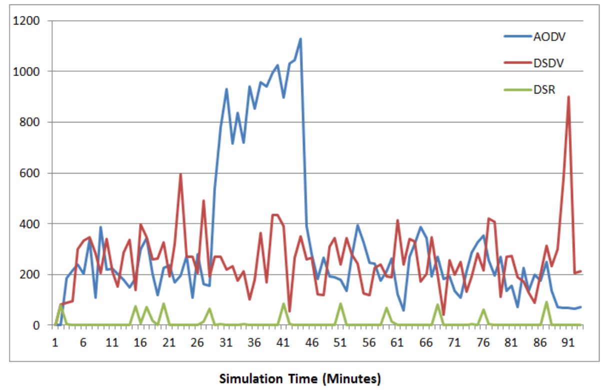


Figure 10. Packets Dropped in Urban Scenario.

Fig. 11 shows the drop packet performance of DSR protocol is better than for both AODV and DSDV in urban scenario. As the number of vehicles increased, data dropped ratio of DSDV suddenly increased from 29 to 480 packets that much worst in entire simulation.

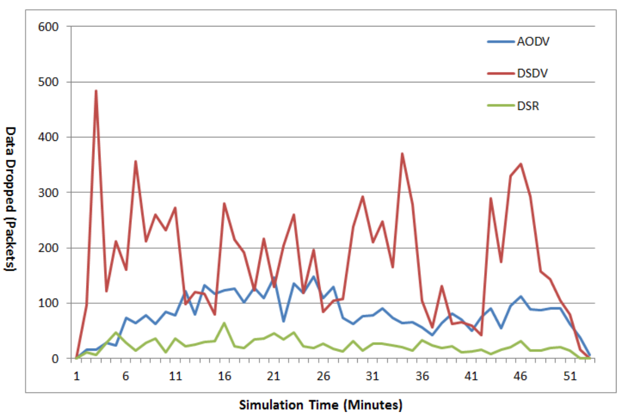


Figure 11. Dropped Packets in rural Scenario.

VII. CONCLUSION

In this research, analysis the simulations are performed to compare the performance of On-Demand (DSR and AODV) and Table-Driven (DSDV) routing protocols by using the two different scenarios. Also measured the metrics like data collide, data dropped, etc. The results presented in Table 5 for different traffic scenario and for three VANET routing protocols, performance of DSR has been found to be better than that of AODV and DSDV. From the parameter values characterizing the two traffic scenarios, DSR is found suitable for both rural and urban traffics scenarios. Thus it can be concluded that DSR outperforms from other routing protocols AODV and DSDV in both urban and rural area scenarios.

 ISBN: 978-1-61208-289-9

TABLE IV. COMPARISION OF ROUTING PROTCOLS

Parameters	**Scenario**	**Routing Protocols**		
		AODV	DSDV	DSR
Throughput	Urban	L	M	H
Throughput	Rural	M	L	H
Data Dropped	Urban	H	M	L
Data Dropped	Rural	M	H	L
Data Collide	Urban	H	M	L
Data Collide	Rural	M	H	L

H = High, L=Low, M = Medium

VIII. REFERENCES

[1] U.S. Department of Transportation, Intelligent Transportation Systems (ITS) Home [July, 2013] http://www.its.dot.gov/index.htm .

[2] Rylander, D.Volvo Constr. Equip., Eskilstuna, and Sweden Axelsson, "Using wireless communication to improve road safety and quality of service at road construction work sites (Poster)" Vehicular Networking Conference (VNC) IEEE Nov. 2012, pp. 14-16.

[3] Jedrzej Rybicki, Björn Scheuermann, Wolfgang Kiess, Christian Lochert, Pezhman Fallahi, and Martin Mauve. "Challenge: peers on wheels - a road to new traffic information systems" in Proceedings of the 13th Annual ACM International Conference on Mobile Computing and Networking 2007, pp. 215–221.

[4] Jakob Eriksson, HariBalakrishnan, and Samuel Madden, "Cabernet: Vehicular contentdelivery using WiFi" in Proceedings of the 14th ACM International Conference on Mobile Computing and Networking 2008, pp. 199–210.

[5] Kevin C. Lee, Uichin Lee, and Mario Gerla, "Survey of Routing Protocols in Vehicular Ad Hoc Networks," advances in Vehicular Ad-Hoc Networks: Developments and Challenges, IGI Global, Oct, 2009.

[6] DSRC (Dedicated Short-Range Communications), http://www.standards.its.dot.gov/Documents/advisories/dsrc_advisory .htm.

[7] Claudia Campolo and Antonella Molinaro, "On vehicle-to-roadside communications in 802.11p/WAVE VANETs" in Wireless Communications and Networking Conference (WCNC) IEEE, March 2011, Cancun, Mexico, pp. 28-31.

[8] Rani, P.,Sharma, N, Singh, and P.K. . "Performance Comparison of VANET Routing Protocols" in Wireless Communications, Networking and Mobile Computing (WiCOM), 7th International Conference Wuhan 23-25 Sept. 2011, pp .1 - 4.

[9] Daxin Tian,Kaveh Shafiee, and Victor C. M. Leung "Position-based directional vehicular routing" in proceeding GLOBECOM'09 of the 28th IEEE conference on Global telecommunications 2009 USA, pp. 3255-3260.

[10] Guoqing Zhang, Wu Chen, Zhong Xu, Hong Liang, Dejun Mu, and Li Gao "Geocast Routing in Urban Vehicular Ad Hoc Networks" in Computational Intelligence springer link , 2009, Volume 208/2009.

[11] Ahmed Soua, Walid Ben-Ameur, and Hossam Afifi "Broadcast-based Directional Routing in Vehicular Ad-Hoc Networks" in Networking and Internet Architecture (cs.NI) arXiv:1206.1358v1 [cs.NI] Wed, 6 Jun 2012.

[12] Yuyi Luo,Wei Zhang, and Yangqing Hu "A cluster-based directional routing protocol in VANET" in Networks Security Wireless Communications and Trusted Computing (NSWCTC), 2010 Second International Conference on , 24-25 April 2010.

[13] Hadi Sargolzaey, Ayyoub Akbari Moghanjoughi, and Sabira Khatun, "A Review and Comparison of Reliable Unicast Routing Protocols for Mobile Ad Hoc Networks", in IJCSNS International Journal of Computer Science and Network Security, vol .9, no.1, pp. 186-196, January 2009.

[14] Ramandeep Kaur and Chandan Sharma, "Review paper on performance analysis of AODV, DSDV, OLSR on the basis of packet delivery" in IOSR Journal of Computer Engineering (IOSR-JCE) vol. 11, Issue 1 (May. - Jun. 2013), pp. 51-55.

[15] Baruch Awerbuch, Amotz Bar-Noy, and Madan Gopal, "Approximate distributed Bellman-Ford algorithms" in Communications, IEEE Transactions on vol. 42 , Issue: 8, pp. 2515 – 2517, published year 1994.

Spaces-Based Communication: an Extension to Support Infinite Spatial Models

Diego Bernini, Francesco Fiamberti, Daniela Micucci, Francesco Tisato, Alessio Vertemati
Department of Informatics, Systems and Communication
University of Milano-Bicocca
Milano, Italy
Email: {diego.bernini, francesco.fiamberti, daniela.micucci, francesco.tisato, alessio.vertemati}@disco.unimib.it

***Abstract*—Space Integration Services (SIS) is a software communication platform that enables the seamless integration of sensors, actuators, and application-logic components through a multi-space model and a spaces-based publish/subscribe mechanism. The underlying model is based on finite spaces only. SIS has been extended in order to add support for spaces with an infinite number of locations, (e.g., spaces described by continuous coordinates, such as geodetic or Cartesian spaces). The existing conceptual model, valid for finite spaces, is reviewed and generalized to infinite spaces. The obtained model is tested using a prototype implementation realized by means of an additional layer on top of the finite-space version. Finally, the performance of the prototype is then compared to the one of the finite-space version in a series of experimental tests.**

Keywords*—*Infinite spatial models; spaces-based communication; software architecture; component-based architecture.

I. Introduction

Pervasive computing [1] aims at simplifying the everyday life through digital environments that are sensitive, adaptive, and responsive to the user's needs. A pervasive computing system requires the perception of the context in which the user operates to provide a richer and expanded mode of interaction, in addition to an intelligence for performing actions on the environment. From a technological point of view, pervasive computing relies on responsive environments. The term responsive environment [2] refers to physical environments enhanced by input devices (e.g., sensors or cameras) and output devices (e.g., displays, lights, motors). Input devices capture stimuli from the environment, whereas output devices execute actions on the environment given a predefined set of commands.

Responsive environments are, therefore, able to perceive and respond to users thanks to the presence of a computer system that receives data from the sensors (input stream) and sends commands to the actuators (output stream). For example, an application may locate the user onto the cartographic representation of the city and may also receive data from light sensors (input stream). On the basis of established rules, it could send commands to the street lights (output stream). This simple example emphasizes how responsive environments require establishing information flows from the devices to the applications and vice versa.

Space Integration Services (SIS) [3] is a software communication platform that enables the seamless integration of sensors, actuators, and application-logic components through a multi-space model and a spaces-based publish/subscribe mechanism. It provides various spatial models that can be used by applications to represent location-related information in order to support complex representations of the environment with a focus on location-aware systems. In this regard, different spatial representations can be put in appropriate correspondence to describe the localization according to different visions of the environment. The spatial representations supported by SIS are symbolic models (graphs and name spaces) and grid models (bi- and tri-dimensional). With the increase of accuracy and precision of localization sensors [4] and with the need to include geographical-related spatial models, a revision of the platform was required in order to deal with spatial representations that contain a potentially infinite number of locations, for example those described by continuous coordinates, such as geodetic and Cartesian spaces, but also unbounded grids described by discrete coordinates.

The paper is organized as follows. Section II presents the SIS conceptual model and the proposed extension for the inclusion of infinite spaces. Section III discusses the application of the model to a simplified case study. Section IV describes a prototype implementation used to test the extension. Section V presents the results of several tests aimed at estimating the performance of the proposed extension with respect to the existing SIS implementation. Section VI reviews related works. Finally, conclusions and future developments are presented in Section VII.

II. Conceptual model

The previous conceptual model was based on the assumption that *spaces* are finite sets of *locations* built from *spatial models* (e.g., graph spatial model). Non- empty sets of locations belonging to a space are named *spatial contexts*. In such a model, locations play a crucial role since the existence of both a space and a context is subject to the existence of the set of locations that constitutes them (space and context). In order to handle infinite spaces, that is, spaces that contain a potentially infinite number of locations, the conceptual model has been revised around the concept of spatial context.

A. Space

A *space* is a set of *potential location*s, that are all the locations that could be theoretically considered in that space, not only the ones actually used explicitly. For example, in a Cartesian space only a finite number of points (i.e., locations) can be actually used explicitly by a real application, but theoretically every point in $\mathbb{R}^2$ could be used by the application.

 ISBN: 978-1-61208-289-9

A *spatial model* defines 1) the types of locations that may belong to the spaces created by that model, 2) the way in which locations are arranged, and 3) at least one premetric that can be applied to a pair of locations. The premetric defines the distance between two locations as a positive, non-zero number if the two locations are distinct, and zero if the locations are the same. A single location might be, for example, a node of a graph, a symbolic label or a numeric coordinate. Two examples of spatial models for infinite spaces are (1) the Geodetic spatial model, defining locations as geodetic coordinates (i.e., latitude and longitude) in a geodetic coordinate reference system [5]; (2) the n-dimensional Cartesian spatial model, that models an Euclidean space $\mathbb{R}^n$. In this model, locations are represented by ordered tuples of real numbers in a orthogonal Cartesian reference system. These examples also highlight that a location in an infinite space can be identified with a potentially infinite precision, as in the nature of the coordinate system based on real numbers.

B. Spatial Context

In order to handle the selection of subsets of locations in an infinite space, the concept of spatial context (context in the following), already present in the finite space-based SIS model, has been refined. A *spatial context* C_S is a subset of potential locations of a space S. It is defined by a set of *characteristic locations* in S and by a membership function that states if a given location of S belongs to the context. Essentially, the membership function is a boolean function that is true when a location is in the spatial context. According to the membership function used, the following kinds of contexts have been identified: *enumerative*, *premetric declarative*, *polygonal*, and *pure functional*.

In a *enumerative context* the set of characteristic locations is non-empty and the membership function is based on the standard belonging relationship defined in set theory. For example, given a space S, an enumerative context is defined by the set $C_S = \{l_1, l_2, ..., l_5\}$ of characteristic locations.

In a *premetric declarative context* the members are all the locations within a given distance from a given characteristic location in terms of a premetric function (see Figure 1a).

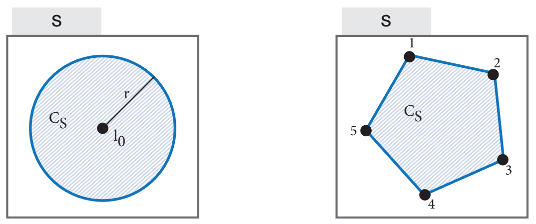
In a *polygonal context* the characteristic locations are the vertices of a polygon, and the membership function indicates the inclusion of a location in the region corresponding to the polygon itself (see Figure 1b).

Finally, in a *pure functional context* the set of characteristic locations is empty and the membership function is defined by using mathematical expressions defined in terms of the space coordinate system. For example, consider a Cartesian bi-dimensional space S. A context can be defined by the following membership function: for all locations (x, y) in S and for any given location (x_0, y_0) and (x_1, y_1),

$$f(x,y) = \begin{cases} \text{true} & \text{if } x_0 < x < x_1 \text{ and } y_0 < y < y_1 \\ \text{false} & \text{otherwise} \end{cases} \quad (1)$$

C. Projection

Related to the concept of space, is the concept of *projection*, which is widely used in geometry and in algebra.



(a) A premetric context (b) A polygonal context

Fig. 1: Different types of spatial context

Given a source spatial context C_S defined in a source space S and given a target space T, the result of a projection is a spatial context C_T on the target space T containing the locations that are obtained by applying a projection function f to all the locations in the source context C_S. The target space can be defined according to the same spatial model of the source, or to a different one. For example, a planimetry of a building is the projection of the tri-dimensional environment onto a bi-dimensional surface. Another typical example of the application of a projection is the transition from a geodesic representation of the Earth to any cartographic representation.

D. Mapping

Mappings relate different spaces. For example, a mapping can relate an area of a Cartesian space (representing the plant of a building) to a node of a graph (representing a synoptic view of the same building). Mappings are key concepts because, as it will be explained afterward, enable the communication among components, even if they rely on different spaces. Three kind of mappings have been defined: *explicit*, *projective* and *implicit*.

An *explicit mapping* is an ordered pair of contexts defined in different spaces (possibly based on different spatial models): given the two spaces S_1 and S_2 with $S_1 \neq S_2$, the ordered pair $\langle SC_1, SC_2 \rangle$ is an explicit mapping between the contexts $SC_1 \subseteq S_1$ (source) and $SC_2 \subseteq S_2$ (target).

Figure 2 shows two examples of explicit mappings between the context SC_m of the `Geo` space (a Geodetic spatial model) and the context SC_k of the `LocationNames` space (a graph spatial model), and between the context SC_i of the `PeopleIds` space (a name spatial model) and the context SC_m of `Geo`.

The target context may be defined independently of the source context. But when the target context is the result of the application of a projection to the source context, the mapping is termed *projective mapping* and is fully determined by the source context and the projection function.

Let SM be the set of all the defined explicit and projective mappings, and let SC_a and SC_b be contexts defined in different spaces. The *implicit mapping* $\langle SC_a, SC_b \rangle$ is derived if there exist k contexts $SC_1, \ldots, SC_k$ such that $\langle SC_a, SC_1 \rangle, \langle SC_1, SC_2 \rangle, \ldots, \langle SC_k, SC_b \rangle \in SM$ for $k > 1$.

In Figure 2 the dotted arrow represents an implicit mapping between the contexts SC_i of `PeopleIds` and SC_k of `LocationNames`.

 ISBN: 978-1-61208-289-9

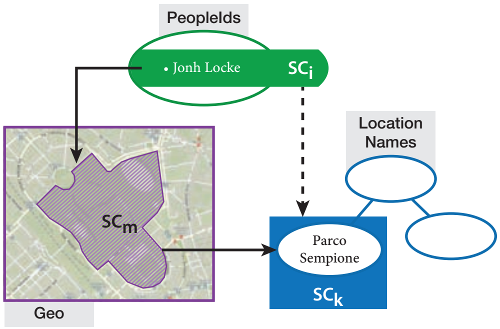


Fig. 2: Explicit and implicit mappings

E. Matching

A *direct matching* occurs when the intersection between two spatial contexts defined in the same space is not empty.

Let SC_1 and SC_2 be spatial contexts defined in different spaces. An *indirect match* between SC_1 and SC_2 occurs when there exists a mapping (explicit or implicit) $\langle SC_a, SC_b \rangle$ such that the intersections between SC_1 and SC_a and between SC_2 and SC_b are both not empty (see Figure 3).

Fig. 3: Indirect matching

F. Publication and subscription

As previously introduced, SIS enables information flows relying on the publish and subscribe mechanism. A *publication* includes a thematic information (whose semantics is up to the application domain) and one or more spatial contexts; a *subscription* includes one or more spatial contexts.

When an application performs a publication, the enclosed thematic information is received by all the applications that previously performed a subscription such that at least one of its contexts *matches*, either directly or indirectly, a context of the publication. The thematic information is enriched with all the contexts that contributed to the matching.

III. Applicative Scenario

This section will apply the concepts introduced in Section II considering an exemplified scenario. Consider a parcel distribution company with warehouses distributed in Europe (for the sake of simplicity, we consider only six warehouse distributed in Italy and Spain). The company exploits vehicles to distribute parcels. Each vehicle is equipped with a device (mounting a GPS) that periodically notifies its position. Before reaching the final warehouse, vehicles can pass through intermediate warehouses. Each time a vehicle enters a warehouse, different operations have to be performed according to the country's rules, including the decision of the next warehouse the vehicle has to reach.

As depicted in Figure 4, the required spaces are: S_1, a Geodetic spatial model covering the involved countries; S_2, a graph spatial model where each node corresponds to a specific warehouse (identifiers $wh_1, wh_2, ..., wh_6$) and each arc connects the warehouses that can be reached without any intermediate stop; finally, S_3, a name spatial model containing the identifiers of the two countries $(Italy, Spain)$.

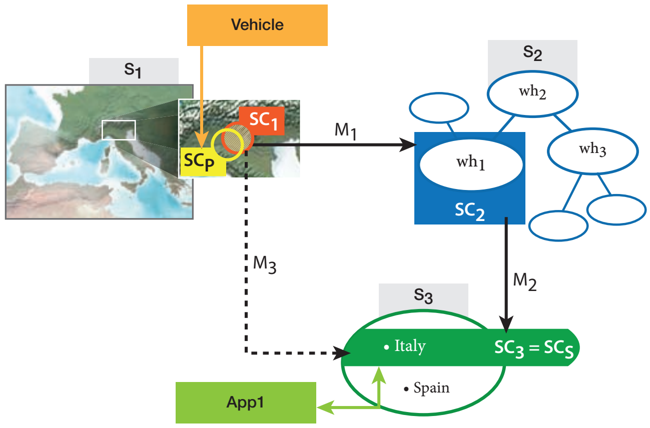


Fig. 4: Example scenario

Explicit mappings are required from S_1 to S_2 with the aim of localizing each warehouse in the Geodetic space. Mappings are in the form $M = \langle (latitude, longitude, radius), \{wh_i\} \rangle$, where the target is an enumerative context containing a node of the S_2 space (i.e., a warehouse wh_i) and the source is a premetric declarative context specifying the area in S_1 where the warehouse wh_i is located. For example, the mapping $M_1 = \langle SC_1, SC_2 \rangle$, where $SC_1 = (45.523653, 9.219436, 50)$ and $SC_2 = \{wh_1\}$. In Figure 4 spatial contexts on S_1 have been hugely enlarged for visualization purposes. Moreover, explicit mappings are required from S_2 to S_3 with the aim of localizing each warehouse in its country. Mappings are in the form $M = \langle \{wh_i\}, \{country\} \rangle$, where the source is an enumerative context containing a node of the graph (i.e., the identifier of a warehouse) and the target is an enumerative context containing the identifier of the country. For example, the mapping $M_2 = \langle SC_2, SC_3 \rangle$, where $SC_2 = \{wh_1\}$ and $SC_3 = \{Italy\}$. Finally, six indirect mappings are derived: their source contexts are the source contexts of the explicit mappings defined between S_1 and S_2, and their target contexts are the target contexts of the explicit mappings defined between S_2 and S_3. For example, $M_3 = \langle SC_1, SC_3 \rangle$.

Two applications are required (App_1 for Italy and App_2 for Spain), each implementing the local rules. Each application subscribes to the appropriate country to be notified when a vehicle enters a warehouse in the country of competence. For example, App_1 performs a subscription to the enumerative context SC_S in S_3 containing the location $Italy$ (i.e., $SC_S = \{Italy\}$). Periodically, the vehicles make publications in the S_1 space, thus sharing their position with all the interested applications. Publications are in the form $[vehicleID, \{(latitude, longitude, radius)\}]$, where $vehicleID$ is the thematic information that identifies the vehicle, and $(latitude, longitude, radius)$ is a premetric declarative context specifying the position of the vehicle iden-

 ISBN: 978-1-61208-289-9

tified by $vehicleID$ in the Geodetic space. For example, publication $Pub_i = [12345, \{SC_P\}]$, where 12345 is the identifier of the vehicle that performs the publication and $SC_P = (45.51788, 9.214071, 20)$ is the location in which it has been localized.

When the vehicle 12345 makes the publication Pub_i, App_1 is notified because an indirect match occurs. Indeed, the following conditions result true: SC_P intersects SC_1, SC_1 is indirectly mapped to SC_3 (M_3 mapping), and SC_S intersects SC_3. When notified, App_1 receives the thematic information 12345 enriched with all the contexts that enabled the matching, that is, SC_P, SC_1, SC_2, SC_3, SC_S. This way, App_1 is aware of the warehouse in which the vehicle 12345 is, and, if it is able to manage S_2, it can inspect the graph in order to decide the next stop for the vehicle.

Form the above example, it is clear that spaces can be used by applications not only to enable information flows, but also to reason about spatial configurations.

IV. Implementation

A. Prototype

The developed prototype is based on the implementation of the finite-space version of Space Integration Services.

The infinite-space extension of SIS is organized according to two different software layers [6] on top of the finite-space SIS core layer. Figure 5 shows the resulting structure.

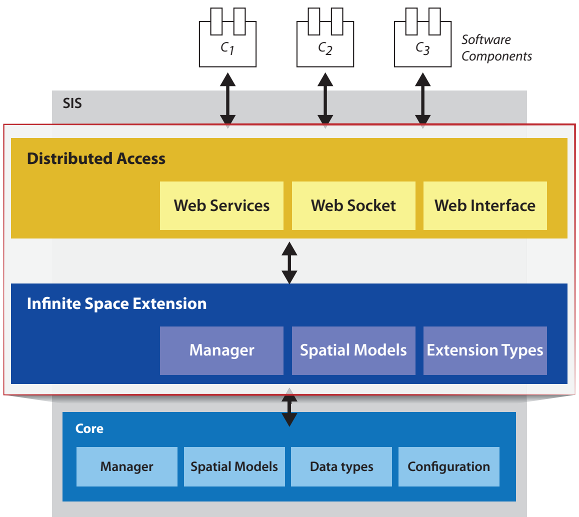


Fig. 5: The extended SIS Structure

The *Distributed Access* layer exhibits three different mechanisms allowing applications to interact with the platform. The Web Services interface provides access to the platform features by means of the Web Service Definition Language (WSDL) and of the Representational State Transfer (REST). Web Sockets [7] are a recent W3C standard for two-way asynchronous communication in the context of web applications; this technology makes available asynchronous communication of notifications to the applications. Finally, a Web interface visually exposes all the primitives of the platform and allows for the configuration of the users with their access permissions.

The *Infinite Space Extension* layer encloses the management of the new spatial models, the new data structures and the business logic to handle publications and subscriptions, the definition of contexts and the creation of mappings between contexts of different spaces.

The *Core* layer contains the finite spatial models (i.e., graph space, name space and grid space) with related primitive types and the manager in charge of monitoring instances of SIS itself. The Core layer uses the JESS rule engine in order to handle the operations of transitive closure and matching. This layer grants the full compatibility of the SIS extension with the previous versions of SIS.

The prototype has been developed in Java because the current Core layer is implemented in this language.

V. Performance evaluation

Several performance tests have been performed on the SIS prototype. For such tests, a SIS instance has been used running on an Intel Core i5 2.8 GHz PC with 4GB of RAM, running Windows(R) 7 64bit with the 64bit version of the Java Runtime Environment 1.6.33. Publications were generated by a mobile client and sent to the SIS server. The tests measured the mean reasoning time and the maximum RAM occupation. The mean reasoning time is the mean time between the reception of a publication and the moment at which notifications are made available to interested applications.

A. Mean reasoning time vs. Mappings

The first experimental setup allows studying the dependence of the mean reasoning time on the number of mappings. In this test, the mappings are created in the space configuration phase and do not change dynamically.

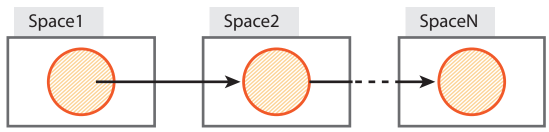


Fig. 6: Space configuration for the mapping test

The space configuration for this test consists of n Cartesian spaces, each containing a single premetric declarative context with a radius of 2 units. Every context is directly mapped onto a context in the next space, thus realizing a chain of $n \;\; 1$ explicit mappings, as shown in Figure 6. Considering the implicit mappings, the total number of mappings is therefore equal to $n(n \;\; 1)/2$. Publications occur on the context in the first space with a frequency of 50 Hz (i.e., every 20 ms), and a subscription is made on the context of the last space. With this generic configuration, the mean reasoning time for a publication can be measured as a function of the number of spaces n.

Figure 7 shows the measured mean reasoning time, expressed in milliseconds, for n varying from 2 to 300. Two different behaviors can be identified in the graph: for n between 2 and 90, the mean reasoning time is mostly constant (about 0.9 ms), whereas for $n > 90$ the reasoning time depends

 ISBN: 978-1-61208-289-9

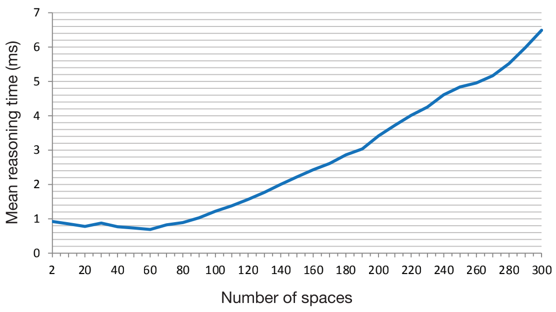


Fig. 7: Mean reasoning time vs number of spaces

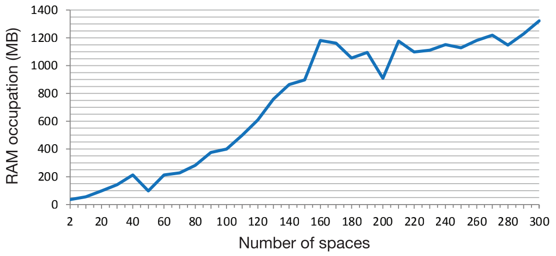


Fig. 8: Memory footprint vs number of spaces

more or less linearly on n. Figure 8 shows the RAM footprint, showing a constant increase for about half of the test and next an oscillatory behavior until the maximum value around 1.3 GB for 300 spaces. Such an amount of memory is probably due to the current implementation, which is not optimized and thus exploits a huge number of objects. The oscillatory behavior is probably due to the Java memory management.

Given these experimental results, a comparative test has been conducted against the finite-space implementation. In particular, when using the finite-space version of SIS (SIS 1), the Cartesian space has been approximated with a grid of regular cells.

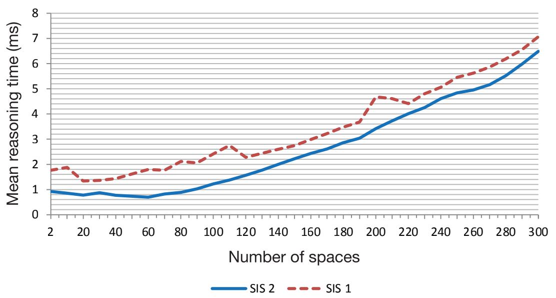


Fig. 9: SIS 2 vs SIS 1: mean reasoning time comparison

As pointed out by the graph in Figure 9, the prototype (SIS 2) is 0.7 ms faster than SIS 1 using a grid approximation. Considering the prototypical stage, the memory footprint is at acceptable levels when compared to SIS 1. Overall, an average gain of 50 MB was observed on the use of grid spaces as approximation of a Cartesian space.

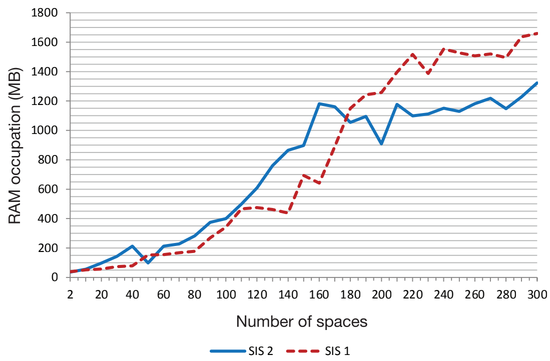


Fig. 10: SIS 2 vs SIS 1: memory footprint comparison

B. *Mean reasoning time vs. context size*

The second test aims to investigate the dependence of the mean reasoning time on the size of the contexts. The space configuration in this case includes two Cartesian spaces. One premetric declarative context was defined in each space, the first one for publication and the second one for subscription. During the test, the number of mapped spaces has been maintained fixed at two, whereas the radius of the premetric declarative contexts was varied according to the values 1, 5, 10, 15, 20 and 25 units.

To get a better idea of the results, a comparative test has been conducted between the SIS 2 prototype and the SIS 1 implementation with Cartesian spaces approximated by regular sized grid. As pointed out by the graph in Figure 11, the SIS 1 prototype has a constant mean reasoning time about 2 ms, while the SIS 1 implementation behaves like $O(n^2)$.

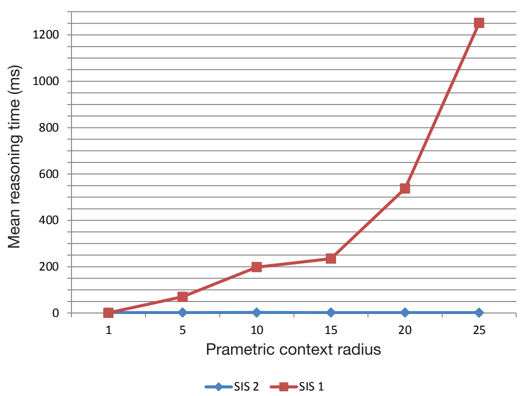


Fig. 11: Mean reasoning time vs. size of context

C. *Discussion*

The experimental tests show how the implemented techniques allow for a better management of different context sizes in infinite spaces, maintaining a constant time of reasoning together with a moderate use of the available RAM memory.

Moreover, the prototype has a better scaling behavior with respect to the finite-space implementation, both for increasing number of mappings and for increasing size of publication and subscription contexts. Such improved scalability can be understood in terms of the number of facts that form the knowledge base of the rule engine managing the operations of transitive closure and matching; the infinite-space implementation reduces the number of facts to one per context, and thus to two facts for a single mapping, whereas the finite-space implementation creates a fact for each location involved in the mapping.

The memory footprint globally resembles the one characterizing the finite-space implementation, but without the peak that has been evidenced during the tests. This memory occupation is caused by the supplementary data structures needed by the infinite-space implementation to correctly handle situations which involve both finite and infinite spaces. It is noticeable that these structures have a relevant weight when a reduced number of spaces and mappings is involved but become negligible as the number of spaces increases.

VI. Related work

Location-aware computing has been an active area of research. Different platforms at the state of art enable location-aware applications focusing on sensor fusion and reasoning with the help of a multi-spatial model, or hybrid model (as called by Becker et al. [8]).

Location Stack [9] defines a layered modeled for fusing location information from multiple sensors and reasoning about an object's location. It, however, does not incorporate a spatial model of the physical world and does not support representations of immobile objects. This leads to a lack of support for spatial reasoning relative to stationary entities such as rooms or corridors.

Loc8 [10], on the other hand, extends the Location Stack layered architecture by considering only high level position and data instead of low level sensor data. Reasoning is applied to that position data, enriched by the knowledge given by a base ontology, to infer additional spatial relationships.

The *Aura Space Service* [11] combines coordinate and hierarchical location models into a single hybrid model, which supports spatial queries. The focus of the Aura Space Service is only on modeling the physical space and supporting spatial queries. It does not address location inferencing and does not provide a framework for spatial reasoning.

MiddleWhere [12] uses the hybrid location model introduced by the Aura Space Service and enables the fusion of different location sensing technologies. MiddleWhere introduces also probabilistic reasoning techniques to resolve conflicts and deduce the location of people given different sensor data. The model of the world is stored in a spatial-enabled database.

Semantic Spaces from Microsoft Research decomposes the physical environment into a hierarchy of spaces. The locations of moving users or devices are correlated to actual physical spaces, thus it is capable of answering ”containment” queries. However, because of its inherent lack of metric attributes and precision, it is unable to compute distance accurately or represent locations precisely, which are requirements for some ubiquitous computing applications [13].

Semantic Spaces and Location Stack lack any support for infinite spaces and in general spaces with a coordinate system, while Loc8 and MiddleWhere have at least one spatial model with a Cartesian coordinate system and can handle different levels of precision on that space model. These two platforms substantially treat infinite spaces by using different granularities for location representation on a local and a global coordinate system.

VII. Conclusion and Future Work

The paper proposed an extension to the conceptual model of the SIS platform. This refinement comes in order to enable the use of infinite spatial models like the geodetic or the Cartesian ones. The approach that has been presented has involved the extension of the concept of spatial context and the use of that concept as the elementary unit at the basis of all the spatial operations enabled by the platform itself (i.e., mapping and matching). With the help of a prototypal implementation the revised conceptual model has been tested for performance evaluation. The tests that have been conducted have shown an overall increase of performance and capacity to handle spatial contexts with large extent as needed when using the geodetic space. The main future work consists in the deep integration of the Infinite Space Extension layer in the Core layer. This will enable a more efficient use of the rule engine.

References

[1] D. Saha and A. Mukherjee, “Pervasive computing: a paradigm for the 21st century,” *Computer*, vol. 36, no. 3, pp. 25–31, 2003.

[2] L. Bullivant, *Responsive environments: architecture, art and design*. Victoria & Albert Museum, 2006.

[3] D. Bernini, F. Fiamberti, D. Micucci, and F. Tisato, “Architectural abstractions for spaces-based communication in smart environments,” *Journal of Ambient Intelligence and Smart Environments*, vol. 4, no. 3, pp. 253–277, 2012.

[4] C. A. Patterson, R. R. Muntz, and C. M. Pancake, “Challenges in location-aware computing,” *Pervasive Computing, IEEE*, vol. 2, no. 2, pp. 80–89, 2003.

[5] ISO, *Spatial referencing by geographic identifiers*. International Organization for Standardization, Geneva, Switzerland, 2003, no. ISO 19112.

[6] D. Garlan and M. Shaw, “An introduction to software architecture,” *Advances in software engineering and knowledge engineering*, vol. 1, pp. 1–40, 1993.

[7] IETF. The WebSocket Protocol - RFC 6455. [retrieved: July, 2013]. [Online]. Available: http://datatracker.ietf.org/doc/rfc6455/?include_text=1

[8] C. Becker and F. Dürr, “On location models for ubiquitous computing,” *Personal and Ubiquitous Computing*, vol. 9, no. 1, pp. 20–31, 2005.

[9] D. Graumann, W. Lara, J. Hightower, and G. Borriello, “Real-world implementation of the location stack: The universal location framework,” in *Mobile Computing Systems and Applications, 2003. Proceedings. Fifth IEEE Workshop on*. IEEE, 2003, pp. 122–128.

[10] G. Stevenson, J. Ye, S. Dobson, and P. Nixon, “Loc8: a location model and extensible framework for programming with location,” *Pervasive Computing, IEEE*, vol. 9, no. 1, pp. 28–37, 2010.

[11] C. Jiang and P. Steenkiste, “A hybrid location model with a computable location identifier for ubiquitous computing.” Springer, 2002, pp. 307–313.

[12] A. Ranganathan, J. Al-Muhtadi, S. Chetan, R. Campbell, and M. Mickunas, “Middlewhere: a middleware for location awareness in ubiquitous computing applications,” in *Proceedings of the 5th ACM/IFIP/USENIX international conference on Middleware*. Springer-Verlag New York, Inc., 2004, pp. 397–416.

[13] B. Brumitt and S. Shafer, “Topological world modeling using semantic spaces,” in *Proceedings of the Workshop on Location Modeling for Ubiquitous Computing, UbiComp*, vol. 2001, 2001, pp. 55–62.

Evaluating Student Attitudes on Ubiquitous e-Learning

Timothy Arndt
Dept. of Computer and Information Science
Cleveland State University
Cleveland, OH, USA
arndt@cis.csuohio.edu

Angela Guercio
Dept. of Computer Science
Kent State University – Stark
North Canton, OH, USA
aguercio@kent.edu

Abstract—**This paper describes our preliminary work in progress on ubiquitous e-learning. Ubiquitous e-learning is learning which can take place anywhere, anytime. Following this paradigm, ubiquitous e-learners use mobile devices such as smart phones, tablets, and laptops to learn wherever they are. Furthermore, ubiquitous e-learning implies context-sensitivity so that the style of learning as well as the material is adapted to the e-learner's immediate surroundings. Ubiquitous e-learning has been identified by researchers as an increasingly important paradigm for the future, for both non-traditional learners as well as for today's generation of students who are increasingly comfortable with mobile devices as their primary computing platforms. In spite of this, evidence about students' attitudes towards ubiquitous e-learning is scarce. In order to guide and inform our future research in ubiquitous e-learning research, we have performed a survey of our computer science students. In this paper we present the results of this survey, our evaluation of the results, and our reflection on how these results will inform our future research.**

Keywords - e-learning; ubiquitous learning; mobile learning; instuctional technology.

I. INTRODUCTION

Several emerging trends point towards the growing importance of ubiquitous learning – learning which takes place at any time and at any place. Among these trends is a growing population of non-traditional learners. Many of these learners have full or part time jobs which require them to fit the learning into a crowded schedule. Learning must take place wherever and whenever possible and it is not possible to fit this learning into a fixed, rigid schedule. Older non-traditional learners often have family obligations which render them non-mobile as well – the learning must come to them rather than the other way around. Younger, more traditional, learners also bring new demands to the learning environment. This generation is used to being entertained when and where they want, and may find traditional learning methods to be too constraining. In order to meet the needs of this generation, a more flexible, adaptive approach to learning is required.

In the following section, we briefly describe experimental prototypes which we have previously developed which illustrate different aspects of adaptability for ubiquitous e-learning. Section 3 surveys research in ubiquitous e-learning. Section 4 presents the results of a survey on the attitudes of students towards ubiquitous e-Learning. These results are meant to guide the development of our research. Section 5 presents our analysis of the student survey. Section 6 gives conclusions and discusses future research.

II. ADAPTABILITY FOR UBIQUITOUS E-LEARNING

This section briefly reviews our previous research in adaptability for ubiquitous e-learning.

In [8], we described our research in multimedia software engineering applied to distance learning – in particular the Growing Book project, a multinational research effort supporting multi-lingual, multi-modal and multi-level learning. The metadata for courseware was described using an XML (extensible markup language) language called TAOML (teleaction object markup language) whose definition was given. We also described a dataflow transformer, based on XSLT (extensible style sheets language for transformations), for transforming the courseware from one desired output format to another. A prototype data transformer was developed in Java and demonstrated.

In [9], we further developed this approach, concentrating on ubiquitous e-learning and showing how the dataflow transformation approach could be used to support e-learning on different types of devices as well as diverse learning styles, described by user profiles. We moved towards standards-compliant metadata for learning objects and we developed a prototype system capable of generating learning scenarios for several different types of devices.

In the future, we wish to explore ubiquitous e-learning by incorporating context awareness, social networking and support for new mobile devices and networks into our previous research. The exact direction we take will be guided by the survey whose results are reported in this paper.

III. OVERVIEW OF RESEARCH IN UBIQUITOUS E-LEARNING

Ubiquitous e-learning builds on the increasing popularity of smart phones and tablets and the widespread availability of wireless networks in order to adaptively recommend learning resources which are appropriate for a user's given context, and to supply them

 ISBN: 978-1-61208-289-9

in a timely manner. In this section, we review previous research in ubiquitous e-learning.

Wang and Wu [1] developed a prototype system for ubiquitous e-learning for lifelong learners. Their system takes into account the learner's characteristics, behaviors and preferences to make appropriate courseware recommendations. Context-aware functionality was enabled through the use of RFID technology.

Chen, Chan and Wang [2] developed a website for ubiquitous e-learning with various devices (laptop, smart phone, etc.). A log of learning behaviors was kept at the website. SMS (short message system) messages were used to push information (e.g., reminders of homework and tests) to users. Information correlating device types to tasks performed was collected. The push and pull of course information to and from learners' phones seemed to increase student awareness.

Baird and Fisher [11] note the reliance on and expertise in social software of the rising generation of students and proposes the use of social networking media to foster the building of learning communities as well as to facilitate self-paced and customized learning experiences in synchronous and asynchronous learning environments. This work reviews the literature in Social Learning Theory and lists various social networking media with hints of how they may be exploited in e-learning.

Chen and Huang [3] designed a context-aware ubiquitous learning system based on radio frequency identification (RFID) in order to enable students to learn in an environment containing both traditional learning resources and resources from the real world. An experimental result is given showing the effectiveness of the approach.

Tzouveli, Mlyonas, and Kollias [4] aim at providing adaptability for web-based learning systems based on user profiling (similar to our approach in [9]). Electronic questionnaires were used to develop the user profiles and automatic assessment of the questionnaires is used to assign a particular profile to each user. The profile serves the proper learning material user to the learner. An experiment in five European countries shows the approach is promising.

Muntean and Muntean [5] combined the adaptability of [4] with a ubiquitous learning approach. They produced a system which provides support for the selection of learning materials, in the ubiquitous context, from one of a number of open corpus resources. Their system serves rich media content to the users which will best meet their needs while keeping in mind the limitations of the devices and networks currently being used.

Hwang and Tsai [6] give a survey of research in mobile and ubiquitous learning published in selected journals between 2001 and 2010. They find that the amount of research in this area has increased greatly over this time frame and that researchers have stressed the importance of situating students to learn from a real world environment as well as emphasizing the mobile and communication technologies which are key to supporting effective learning in the real world.

Agarwal and Nath [7] present a systematic study of various issues and challenges in ubiquitous e-learning. They identify the characteristics of ubiquitous learning as: Permanency; Accessibility; Immediacy; Interactivity; Situating of instructional activities; and Adaptability. Further, they identify six parameters which may be used to explore ubiquitous e-learning: Connectivity; Flexibility; Interactivity; Collaboration; Extended opportunities; and Motivation.

Liu and Hwang [8] explore the paradigm shifts in e-learning leading to the development of context-aware ubiquitous learning, proposing a set of significant values based on the relevant literature to guide the development of context-aware ubiquitous e-learning applications. They also present a context-aware ubiquitous e-learning case study. In addition to the RFID technology as identified in [3], they also mention Global Positioning System (GPS) as another technology which is expected to impact this area.

IV. SURVEY OF STUDENT OPINIONS ON UBIQUITOUS E-LEARNING

In order to guide our future research in ubiquitous e-learning, we performed a survey of our students. The students surveyed are all computer science majors. They are both undergraduate (bachelors degree) and graduate (masters degree) students, both domestic (U.S.A.) and international students. The survey (see Table I.) is composed of three parts: the first part (questions 1 – 3) concerns the use of mobile devices for study; the second part (questions 4 – 8) concerns the students' study environment; the third part (questions 9 – 12) concerns the type of ubiquitous interested in using. The survey employs a standard five-point Likert scale.

 ISBN: 978-1-61208-289-9

TABLE I. RESULTS OF STUDENT SURVEY

	Items addressed	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree	Score
1.	I use a laptop computer for coursework (studying online slides, videotaped lectures, doing assignments, etc.).	34	3	1	0	0	4.87
2.	I use a tablet device for coursework (studying online slides, videotaped lectures, doing assignments, etc.).	6	10	10	6	5	3.16
3.	I use a smartphone for coursework (studying online slides, videotaped lectures, doing assignments, etc.).	6	10	5	9	8	2.92
4.	I do coursework while commuting (listening to taped lectures in a car, studying on public transportation, etc.).	4	6	4	15	9	2.50
5.	I do coursework in "slack times" (waiting for appointments, between classes, etc.).	9	18	2	7	2	3.66
6.	I do coursework out of doors.	5	10	8	11	4	3.03
7.	I do coursework while travelling (weekend outings, vacations, business trips, etc.).	5	8	7	15	2	2.97
8.	I do coursework for one class while in another class.	0	7	5	7	19	2.00
9.	I would be interested in viewing videotaped lectures with seamless scaling and remembered last locations across all of my mobile devices.	12	16	6	3	1	3.92
10.	I would be interested in a mobile app which would automatically highlight relevant material from related courses when I am following a lecture in a different course.	16	14	2	4	2	4.00
11.	I would be interested in a mobile app which would alert me when classmates are in close proximity so that we could have impromptu study sessions or collaboration.	6	12	7	10	3	3.21
12.	I would be interested in a social networking-type app which would let me know when my instructor and/or classmates are online for chat/help.	19	13	5	1	0	4.32

V. ANALYSIS OF STUDENT SURVEY

In order to guide our research in ubiquitous e-learning, we have analyzed the results of the student survey. We briefly summarize our analysis here.

The first part of the survey attempts to determine the extent to which the students are using mobile devices for coursework. It is clear that the use of mobile devices – laptops, tablets, and smart phones – is almost universal among students, but the question we ask is more specific – how much are these devices used for coursework?

 ISBN: 978-1-61208-289-9

What we see from the survey is (as would be expected) that the use of laptops for coursework is almost universal. We also see that tablets score fairly highly, and smart phones just a bit less, with more than 40% of respondents agreeing (or agreeing strongly) that they use these devices for coursework. We take this number as a base level from which to start – if we can develop more useful apps for ubiquitous e-learning for tablet and smart phone devices, we would expect the number of users to increase. Overall, the results of this section are very favorable for the potential of research in ubiquitous e-learning.

The second part of the survey examines how much students currently work in situations which might be considered to part of the "ubiquitous" environment (out of doors, while between appointments, while commuting, etc.). While the responses show that many students do coursework while between classes or appointments (we conjecture that many of them are working on laptops between classes), the other possible learning situations have lower responses. This indicates a potential opportunity for more learning if we can enable this via appropriate applications.

The third and final part of the survey attempts to determine which type of mobile apps would be most appreciated by the students surveyed. The results of this section are very positive, with several of the proposed apps having very high scores. This indicates a willingness of the students to use such apps if they are available.

VI. Conclusion and Future Research

This work presents our initial research in ubiquitous e-learning. The results of our survey indicate both an opportunity for new applications (students have mobile devices and a significant number use them for coursework), an opportunity for significance (there are slack times, which are not currently regularly exploited for coursework), and a willingness to use the results of research in this area (indicated by the positive responses in the third part of the survey). We intend to use the results of this survey to direct our research direction in this area.

References

[1] S. L. Wang and C. Y. Wu, "Application of Context-Aware and Personalized Recommendation to Implement an Adaptive Ubiquitous Learning System", Expert Systems with Applications, vol. 38, 2011, pp. 10831-10838.

[2] G. D. Chen, C. K. Chang, and C. Y. Wang, "Ubiquitous Learning Website: Scaffold Learners by Mobile Devices with Information-Aware Techniques", Computers & Education, vol. 50, 2008, pp. 77-90.

[3] C. C. Chen and T. C. Huang, "Learning in a u-Museum: Developing a Context-Aware Ubiquitous Learning Environment", Computers & Education, vol. 59, 2012, pp. 873-883.

[4] P. Tzouveli, P. Mylonas, and S. Kollias, "An Intelligent e-Learning System Based on Learner Profiling and Learning Resource Adaptation", Computers & Education, vol. 51, 2008, pp. 224-238.

[5] C.H. Muvean and G.M. Muntean, "Open Corpus Architecture for Personalized Ubiquitous e-Learning", Pers. and Ubiquitous Comput., vol. 13, 2009, pp. 197-205.

[6] G. J. Hwang and C. C. Tsai, "Research Trends in Mobile and Ubiquitous Learning: a Review of Publications in Selected Journals from 2001 to 2010", British Journal of Educational Technology, vol. 42, no. 4, 2011, pp. E65-E70.

[7] S. Agarwal and A. Nath, "Some Challenges in Designing and Implementing Learning Material for Ubiquitous e-Learning Environment", Journal of Global Research in Computer Science, vol. 2, no. 2, 2011, pp. 29-32.

[8] T. Arndt, S.K. Chang, A. Guercio, and P. Maresca, "An XML-Based Approach to Multimedia Software Engineering for Distance Learning", in Future Directions in Distance Learning and Communication Technologies, T. Shih and J. Hung, Eds. London: Information Science Publishing, 2007, pp. 106-134.

[9] T. Arndt and A. Guercio, "Course Personalization for Ubiquitous e-Learning", ISAST Journal of Computers and Intelligent Systems, vol. 2, no. 2, 2010, pp. 1-11.

[10] G. Z. Liu and G. J. Hwang, "A Key Step to Understanding Paradigm Shifts in e-Learning: Towards Context-Aware Ubiquitous Learning", British Journal of Educational Technology, vol. 41, no. 10, 2010, pp. E1-E9.

[11] D. E. Baird and M. Fisher, "Neomillennial User Experience Design Strategies: Utilizing Social Networking Media To Support "Always On" Learning Styles", Journal of Educational Technology Systems, vol. 34, 2005-2006, pp. 5-32.

Self-Organizing Localization for Wireless Sensor Networks Based on Neighbor Topology

Range-free localization with low dependence on anchor node

Yasuhisa Takizawa
Faculty of Environmental and Urban Engineering, Kansai University
Osaka Japan
e-mail: takizawa@kansai-u.ac.jp

Yuto Takashima
Graduate School of Science and Engineering, Kansai University
Osaka Japan

Naotoshi Adachi
Faculty of Environmental and Urban Engineering, Kansai University
Osaka Japan

***Abstract*—The localization of sensor nodes is one of the key issues for sensor network systems. Therefore, to obtain precise location information, several localization systems have been researched. However, they require an arranged space using a large number of anchor nodes whose locations are well known, or they need advanced information such as radio conditions in the space. Otherwise, the systems cannot be used for a space that cannot be arranged in advance with suitable conditions for these systems operation. Furthermore, some localizations assume the use of advanced distance measurements, such as TOA and TDOA, to achieve high accuracy in estimating locations, but these advanced distance measurement schemes cannot be used for ordinary sensor node systems. To resolve these problems, we propose Self-Organizing Localization for wireless sensor networks. Self-Organizing Localization requires no distance measurement scheme and no advanced information on a space; even then it reproduces a geometry nearly similar to the network's original geometry without anchor nodes, and it reproduces a geometry with two anchor nodes that is nearly congruent with the original. Furthermore, using just three anchor nodes, it estimates node absolute localization with high accuracy. Therefore, it can be applied to any space and any sensor node. In this paper, the algorithm of Self-Organizing Localization is described, and its accuracy based on simulation evaluation is shown.**

Keywords-localization; wireless sensor networks; self-organizing maps;

I. INTRODUCTION

To achieve advanced sensing services, technology that senses the environment with precise location information is indispensable. Therefore, several localization systems that obtain accurate location information have been researched. They are classified into range-based localization [1-6] and range-free localization [8-10]. Range-based localizations assume the use of advanced distance measurement schemes between sensor nodes, such as Time Difference Of Arrival (TDOA) schemes and Time Of Arrival (TOA) schemes. However, such distance measurement schemes are not used in ordinary sensor node systems because they are not suitable for sensor nodes whose functions should be minimized. Some localizations use a Received Signal Strength Indicator (RSSI), which can be used in ordinary sensor node systems; however, these systems require advanced information, and they must have radio condition information sets in the space so they can estimate location with high accuracy. On the other hand, range-free localizations do not need distance measurement schemes; however, to estimate location with high accuracy, they require an arranged space where a large number of anchor nodes are deployed. Some range-free localizations do not need anchor nodes, but they need advanced information on the probable network topology of the space. Therefore, range-free localization cannot be used for a space that cannot be sufficiently arranged in advance.

To resolve these problems, we propose Self-Organizing Localization (SOL) for wireless sensor networks. SOL needs no distance measurement schemes, no advanced information on the space, and its dependence on anchor nodes is very low. SOL achieves the following results by using Self-Organizing Maps (SOM) [14, 15]:

- Without anchor nodes, it reproduces a geometry nearly similar to the network's original geometry.
- With two anchor nodes, it reproduces a geometry nearly congruent with the original, that is, it derives relative node locations on the network.
- With just three anchor nodes, it derives absolute node locations with high accuracy.

According to the above properties, SOL can be applied to any space and any sensor node.

SOL is based on our original localization [13], which assumed an ad hoc network that consisted of many nodes whose locations were unknown and a few anchor nodes whose locations were well known. The localization also assumed a distance measurement scheme that uses an RSSI. SOL eliminates distance measurement schemes because of its application to any sensor node. Then, SOL controls SOM localization based on neighbor topology, which is expressed by hop count between nodes.

In this paper, the algorithm on SOL using SOM is described, and its accuracy based on simulation evaluation is shown. In the rest of the paper, Section 2 describes related work, and Section 3 presents the algorithm of SOL based on our original localization. Then, Section 4 presents the algorithm of SOL based on neighbor topology. Furthermore, Section 5 shows evaluation results for the SOL and discusses its characteristics regarding accuracy of the estimated location.

II. Related Work

Node localizations are classified into range-based and range-free localizations. The typical range-based localizations are shown as follows. RADAR [1], Active Badge [2], and SpotON [3] have been proposed as location estimation methods that use an RSSI. RADAR requires space where radio wave propagation has been measured in advance. Since Active Badge and SpotON use an RSSI for sensing proximity to anchor nodes, these estimated locations have low resolution. Active Bat [4] and Cricket [7] have been proposed as location estimation methods using TDOA. Both estimate locations with high precision based on triangulation; however, they need a space arranged with a large number of TDOA devices. Iterative Multilateration [5] has been proposed as a location estimation method with a small number of anchor nodes; Dolphin [6] is a system that uses Iterative Multilateration. This method and system use triangulation to estimate location and propagate the estimated location to neighbor nodes. The method and system require highly precise distance measurements such as TDOA and suffer from location error that increases as estimated location propagation progresses.

On the other hand, typical range-free localizations are shown as follows. Centroid [8] estimates node location based on the centroid on three anchor nodes that the target node can communicate with directly. The centroid needs an arranged space in which a large number of anchor nodes are deployed. DV-Hop [9] proposed for location estimation using network topology, calculates average distance in 1 hop using communication between anchor nodes, and it estimates node location with the calculated average distance and the number of hops from the anchor node. It also requires a minimum of three anchor nodes. APIT [10] estimates node location based on the geometrical condition that a node can be inside or outside for multiple triangulation. The construction of APIT is based on a three-anchor-node unit, and thus it needs a large number of anchor nodes. These studies [11, 12] apply SOM to wireless localization and provide relative location without anchor nodes. However, they need a training set that leads SOMs to the proper map, and the training set is prepared with information on the space in advance. The accuracy of range-free localizations is very inferior to that of range-based localizations and is insufficient for many sensing services.

III. Our Original Localization

Our original localization [13] reproduces network geometry using SOM. In SOM, the number and range of neighbor nodes are important metrics, and SOM converges when the number and range of neighbor nodes are reduced in accordance with a convex decreasing function [14]. In accordance with the above characteristics of SOMs, our original localization has the following two strategies to effectively use the measured distances between nodes.

- In the early phase, the algorithm uses the locations and distances of both 1- and 2-hop nodes and reproduces an inaccurate but characteristic network geometry by emphasizing the distance relation between nodes.
- In the next phase, the algorithm uses the locations and distances of 1-hop nodes and shapes the geometry to minimize distance errors between neighbor nodes.

Therefore, the number and range of neighbor nodes correspond to the hop count, and, in the early phase, the algorithm actively and widely accepts the neighbor location. Then, in the next phase, the algorithm selectively accepts the neighbor location. Furthermore, in order to need no advanced information on the space, the original applies SOM in the following way:

- The SOM input vector is dynamically generated by the location and distance of neighbor nodes.
- The SOM winner is the node that receives the input vector from a neighbor node.

The following explains the algorithm based on the above SOM strategies and applications.

[step 1] Each node generates a random location as its estimated location $w_i(t)$ and then broadcasts its location $w_i(t)$ to neighbor nodes, where t is the number of estimation steps.

[step 2] The node i receives the estimated location information from a neighbor node j; that is, node i, which is the SOM winner, modifies its estimated location $w_i(t)$ to draw near the input vector $m_i(t)$, which is the location estimated from location $w_j(t)$ of node j. The distance $d_{ij}(t)$ between nodes i and j is provided by the node distance measurement function. Therefore, a modified vector $V_i^{\{1\}}(t)$ that reduces $|m_i(t)-w_i(t)|$ is generated (see Fig.1(a)):

$$V_i^{\{1\}}(t) = \frac{d_{ij}(t) - |w_i(t) - w_j(t)|}{|w_i(t) - w_j(t)|}(w_i(t) - w_j(t)) \tag{1}$$

Furthermore, when the estimation is in the early phase, the input vector $m'_i(t)$ is generated using location estimates $w_k(t)$ from a 2-hop node k in a set of 1-hop nodes from neighbor nodes j and the sum of distances $d_{ij}(t)$ and $d_{jk}(t)$. Therefore, a modified vector $V_i^{\{2\}}(t)$ is generated in which the relation of the 2-hop node k is the following (see Fig.1(b)):

$$V_i^{\{2\}}(t) = \frac{d_{ij}(t) + d_{jk}(t) - |w_i(t) - w_k(t)|}{|w_i(t) - w_k(t)|}(w_i(t) - w_k(t)) \tag{2}$$

If the modified location $w_i(t)$ of node i by (1) and (2) is the location nearer to the 2-hop node k than to the 1-hop node j, that is, if $| w_i(t) - w_j(t)| > | w_i(t) - w_k(t)|$, then the input vector $m'_i(t)$ is the relocation estimated with locations $w_k(t)$ and $w_j(t)$. $d_{jk}(t)$, which is the distance between node i and node k, becomes larger than $d_{ij}(t)$ when node i is on the broken circular line in Fig.1(c), and the modified vector derived as node i is relocated to the center of the range on the broken circular line. Therefore, the modified vector $V_i^{\{2\}}(t)$ is the following (see Fig.1(d)):

$$V_i^{\{2\}}(t) = w_j(t) - w_i(t) + \frac{d_{ij}(t)}{d_{jk}(t)}(w_j(t) - w_k(t)) \tag{3}$$

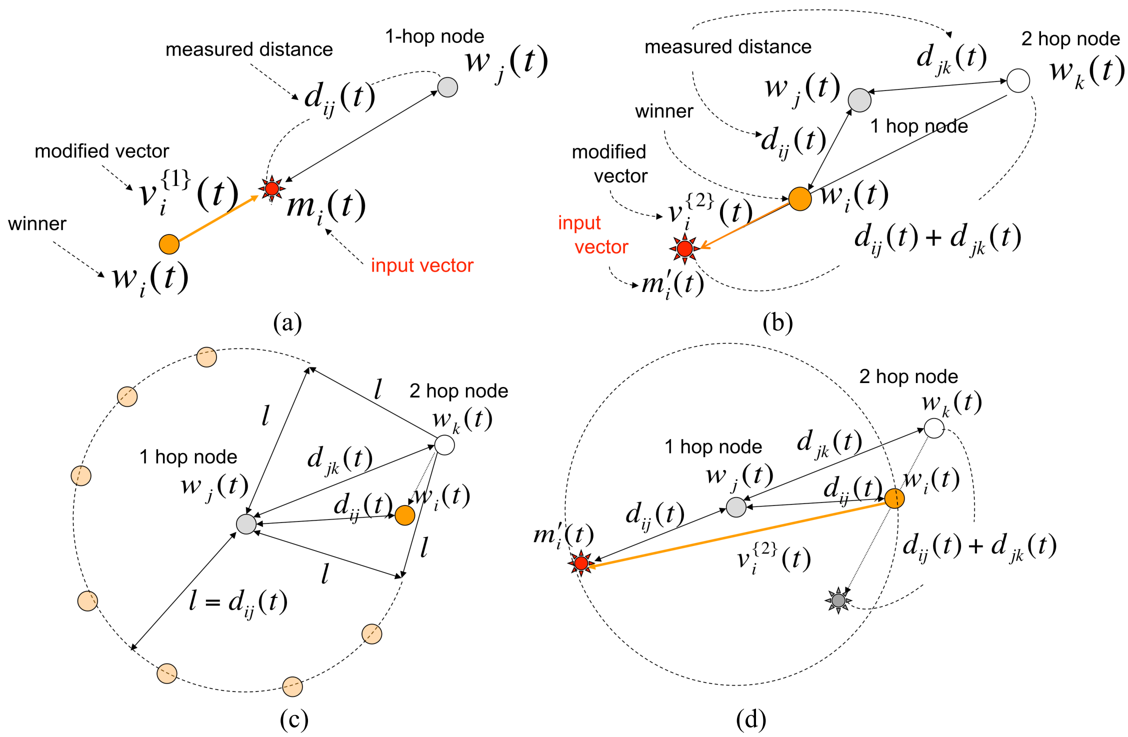


Figure 1. (a) Input vector, winner, and modified vector by 1-hop. (b), (c) and (d) Input and modified vectors by 2-hop.

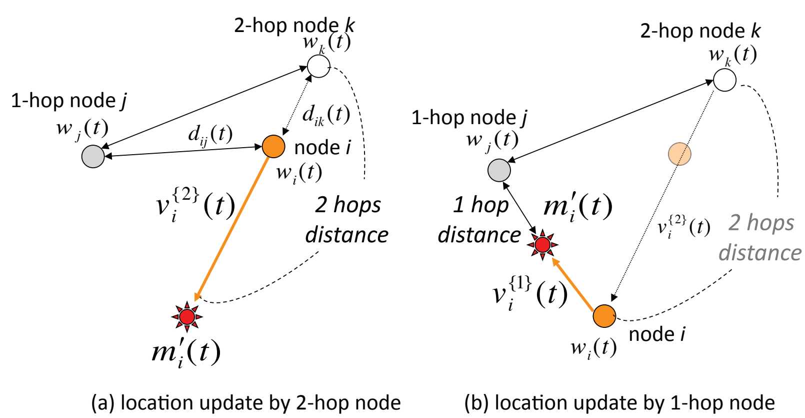


Figure 2. Input and modified vectors by 1-hop and 2-hop node on SOL.

Consequently, using $V_i^{\{1\}}(t)$ and $V_i^{\{2\}}(t)$, each node modifies and updates its estimated location as the following:

$$w_i(t+1) = \begin{cases} w_i(t) + \alpha_i \cdot (V_i^{\{1\}}(t) + V_i^{\{2\}}(t)) & t \le \tau \\ w_i(t) + \alpha_i \cdot V_i^{\{1\}}(t) & t > \tau \end{cases} \quad (4)$$

where τ is a phase threshold and $\alpha_i(t)$ is the learning rate of node i at step t. $\alpha_i(t)$ is defined as follows:

$$\alpha_i(t) = \eta \cdot \alpha_i(t-1) \quad (0 < \eta < 1) \quad (5)$$

where η is a positive constant of attenuation.

[step 3] The current estimated location is periodically broadcast to neighbor nodes in a period. The node that received the estimated location executes [step 2].

As above, each node repeatedly executes [step 2] and [step 3], and as a result, the network's original geometry is reproduced.

IV. SELF-ORGANIZING LOCALIZATION BASED ON NEIGHBOR TOPOLOGY

The algorithm of SOL works on each node autonomously and is composed as follows.

- Node location estimation function: this function reproduces the similarity to the network's original geometry based on SOMs without anchor nodes.

 ISBN: 978-1-61208-289-9

- Node location adjustment function: this function adjusts the reproduced geometry to the congruence with the network's original geometry with two anchor nodes and adjusts it to node absolute location with three anchor nodes.

In this section, each function as the algorithm of SOL is described.

A. Node location estimation based on neighbor topology

As described in section III, based on Euclidean distance between nodes, the original localization emphasizes graphical features of the topology by 2-hop nodes and aims to reproduce the topology. After that, it minimizes the difference of Euclidean distance between 1-hop nodes. That is, assuming that the Euclidean distance between nodes is accurate, the strategy of the original localization inputs a large amount of displacement by 2-hop nodes and the accurate distance by 1-hop nodes to SOM, and dynamically operates SOM. On the other hand, the SOL cannot use Euclidean distance as a relation between nodes because of the elimination of the distance measurement scheme from the original localization. Therefore, the SOL uses hop count as the relation between nodes, and its strategy is based on the neighbor topology, whose 1-hop neighbor node is nearer than the 2-hop neighbor node, and aims to reproduce the geometry that meets the neighbor topology between nodes. The neighbor topology is much rougher than the Euclidean distance provided by the distance measurement scheme; therefore, the SOL cannot dynamically operate SOM, and must operate SOM gradually. Accordingly, to eliminate the inconsistency with the neighbor topology from the reproduced geometry, the SOL carefully controls the location estimation by the 1-hop and 2-hop neighbor node as follows.

- SOL lets $d_{ij}(t)$, which is the distance between 1-hop neighbor nodes, be constantly 1, which is the number of hops.
- SOL estimates the location by 2-hop neighbor nodes only when the relative location to 2-hop neighbor nodes is inconsistent with the neighbor topology as in Fig.1(d) (that is, $d_{ij}(t)>d_{ik}(t)$). Because the number of hops is inaccurate as a distance between nodes, the modified vector $V_i^{\{2\}}(t)$ in Fig.1(d) and (3) is a large amount of displacement, and is very inaccurate. When such a modified vector by 2-hop neighbor node is frequently inputted to SOM, SOM oscillates, becomes unstable, and then converges to a state far from the optimal state. Therefore, SOL sets the modified vector as shown in Fig.2(a) only when the relative location to 2-hop neighbor nodes is inconsistent with the neighbor topology, and lets node keep away from 2-hop neighbor nodes
- Next SOL brings node close to 1-hop nodes as shown in Fig.2(b).

From the above, SOL aims to eliminate the inconsistency with the neighbor topology, and reproduces network topology. Summarizing, on the SOL, each node modifies and updates its estimated location as the following:

$$V_i^{\{1\}}(t) = \frac{1-|w_i(t)-w_j(t)|}{|w_i(t)-w_j(t)|}(w_i(t)-w_j(t)) \tag{6}$$

$$V_i^{\{2\}}(t) = \frac{1+1-|w_i(t)-w_k(t)|}{|w_i(t)-w_k(t)|}(w_i(t)-w_k(t)) \tag{7}$$

$$w_i(t+1) = \begin{cases} w_i(t)+\alpha_i\cdot(V_i^{\{1\}}(t)+V_i^{\{2\}}(t)) \\ \quad |w_i(t)-w_j(t)| > |w_i(t)-w_k(t)| \\ w_i(t)+\alpha_i\cdot V_i^{\{1\}}(t) \quad \textit{otherwise} \end{cases} \tag{8}$$

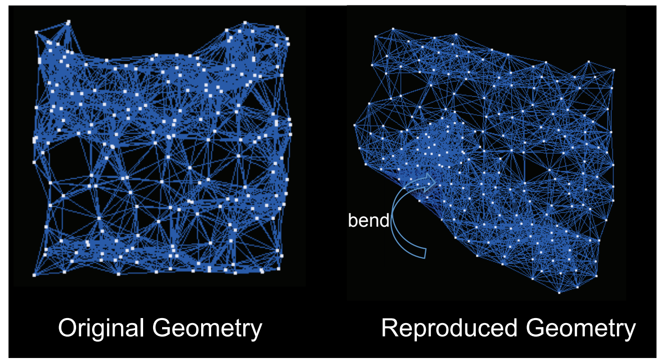


Figure 3. Example of mis-reproduced network geometry.

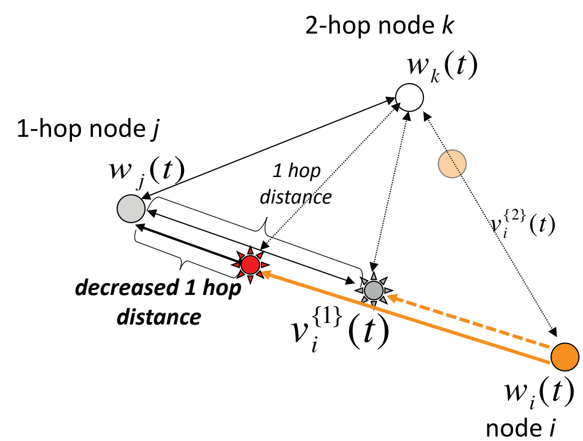


Figure 4. Location update based on decreased 1-hop distance.

In (6), (7) and (8), node j is a 1-hop node from node i, and node k is a 2-hop node from node i.

B. Node location re-estimation

When the number of neighbor nodes selected as input vectors is small, or when the range of neighbor nodes selected as input vectors is narrow, the reproduced geometry is correct locally, but is inconsistent with the entire geometry (mis-reproduction, see Fig.3).

SOL does not dynamically operate SOM, but aims to gradually reproduce network geometry which is narrow, using the 1-hop and 2-hop range nodes. Then, it may suffer from mis-reproducibility. Consequently, in SOL, each node confirms the inconsistency with neighbor topology at the end of iteration to measure mis-reproduced geometry as follows.

$$\frac{I_i^{\{2\}}}{N_i^{\{2\}}} < \theta \tag{9}$$

$I_i^{\{2\}}$ is the number of inconsistent 2-hop neighbor nodes on node i, $N_i^{\{2\}}$ is the number of 2-hop nodes on node i, and θ is the threshold of inconsistency. If (9) is not met, the node aims to dispel the mis-reproduction as follows.

- Reset the learning rate $\alpha_i(t)$ to 1, and re-estimate from the current estimated location to correct the inconsistency in 2-hop geometry.
- Broadcast the message of resetting the learning rate based on the number of message forwardings to neighbor nodes.
- The nodes that receive the message reset$\alpha_i(t)$ to *1/(the number-of-message-forwardings)*, and re-estimate from the current estimated location to correct the inconsistency in 2-hop geometry. And they also broadcast the message of resetting the learning rate based on the number of message forwardings to their neighbor nodes.

Thus, nodes reset smaller learning rate according as the number of message forwardings, and the re-estimation works in local range of mis-reproduced node. On the re-estimation, SOL decreases 1-hop distance that is used by the location update based on 1-hop neighbor node, because SOL brings a node closer to 1-hop neighbor node and raises the probability which the inconsistency with neighbor node topology is eliminated (see Fig.4). Furthermore, SOL makes smaller 1-hop distance as the number of re-estimations increases, and more strongly aims to eliminate the inconsistency with neighbor topology.

C. Node location Adjustment Function

It is expected that the reproduced network geometry has the geometric property of the network's original geometry; therefore, we assume that the reproduced geometry is nearly similar to the network's original geometry, and the reproduced network geometry is defined and adjusted as follows.

- Without an anchor node
  The reproduced geometry is a figure similar to the network's original geometry.
- With two anchor nodes
  Leaving the location of the anchor node unknown, the network geometry is reproduced. Then r, which is a similar scale, is derived using the estimated location and true location of two anchor nodes as follows.

$$r = \frac{d_{ab}}{D_{ab}} \tag{10}$$

$$d_{ab} = |w_a - w_b| \tag{11}$$

d_{ab} is an estimated distance between anchor nodes a and b according to (11), D_{ab} is the true distance between anchor nodes a and b, w_a is the estimated location of anchor node a, and w_b is the estimated location of anchor node b. Consequently, the adjustment for the reproduced network geometry from similarity to congruence is shown as follows:

$$w_i^A = \frac{w_i}{r} \quad i \in NW \tag{12}$$

w_i^A is the adjusted location of node i and NW is a set of nodes on the network. That is, each node can derive its relative location on the network using the estimation and the adjustment.

- With three anchor nodes
  As with the case of two anchor nodes, leaving the location of the anchor nodes unknown, the network geometry is reproduced, and then the three anchor nodes flood their true location and estimated location. The true location $W_A=(X_A,Y_A)$ of an anchor node is expressed as follows using its estimated location $w_A=(x_A,y_A)$.

$$\begin{aligned} X_A &= ax_A + by_A + t_x \\ Y_A &= cx_A + dy_A + t_y \end{aligned} \tag{13}$$

On each node, using simultaneous equations composed by (13) of three anchor nodes, these six coefficients (i.e., a, b, t_x, c, d, t_y) are gained. Also, its estimated location $w_i=(x_i,y_i)$ is transformed to the absolute location $w_i^A=(x_i^A,y_i^A)$ as follows by using affine transformation.

$$\begin{pmatrix} x_i^A \\ y_i^A \\ 1 \end{pmatrix} = \begin{pmatrix} a & b & t_x \\ c & d & t_y \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_i \\ y_i \\ 1 \end{pmatrix} \tag{14}$$

Summarizing, based on the assumption that the reproduced geometry is similar to the network's original geometry, SOL reproduces a similar geometry for the network original geometry without anchor nodes, and it reproduces a congruent geometry for the original geometry with two anchor nodes; that is, it derives relative node location on the network. Using three anchor nodes, the SOL reproduces the network geometry with absolute node location.

V. EVALUATION OF ACCURACY

A. Evaluation Method

TABLE I. SIMULATION PARAMETERS FOR PROPOSED METHOD

Maximum communication range on wireless media	0.2
Wireless media access control	CSMA/CA
Initial estimated location	random
Constant of attenuation η	0.99
Threshold of Inconsistency θ	0.05
Number of iterations for update	600
Maximum number of re-estimation	2
Decreased 1-hop distance	1/(number-of-estimations)

The space in which nodes are deployed is defined as a 1.0×1.0 plane. Table 1 shows the summary of simulation parameters used in the evaluation.

The similarity and congruence with the network's original geometry is evaluated based on (10) and (11) as follows.

$$r_{ij}^M = Mean[r_{ij}] \quad r_{ij} = \frac{d_{ij}}{D_{ij}} \quad i,j \in NW \tag{15}$$

$$r_{ij}^V = 1 - \frac{(Mean[r_{ij}])^2}{Mean[r_{ij}^2]} \tag{16}$$

Mean[x] is the average of set *x*, and *NW* is the set of nodes on the network. The accuracy of absolute location is evaluated based on (12) as follows.

$$Err_{ave} = \frac{1}{N}\sum_{i=1}^{N} \left| W_i - w_i^A \right| \tag{17}$$

N is the number of nodes, W_i is the true location of node i and w_i^A is the estimated and adjusted location of node i .

B. Evaluation of Similarity and Congruence

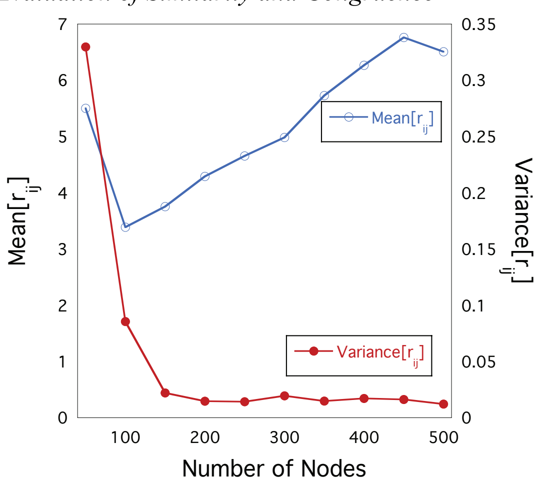


Figure 5. Dependence of r_{ij}^M and r_{ij}^V on number of nodes without anchor node.

Figure 5 shows the dependence of r_{ij}^M (broken line) and r_{ij}^V (solid line) on the number of nodes (50~500 nodes) without an anchor node. The r_{ij}^M and r_{ij}^V shown are averaged over 20 network topologies generated with randomly deployed nodes. When the number of nodes is smaller than 100, r_{ij}^V is a large value. In that case, the original geometry is sparse or fragmented because the density of nodes is low. Therefore, SOL cannot effectively estimate a network geometry in which the density of nodes is low. When the number of nodes exceeds 150, r_{ij}^V becomes small, and the estimated geometry is very near to being similar to the original geometry. When the number of nodes exceeds 250, r_{ij}^V approaches 0.01, and the estimated geometry is very similar to the original geometry. r_{ij}^M is not 1 for any case, regardless of number of nodes. Therefore, the scale of reproduced geometry is different from the original geometry, but is graphically similar to the original geometry.

Figure 6 shows r_{ij}^M and r_{ij}^V on the number of nodes (50~500 nodes) with two anchor nodes. The two anchor nodes are respectively the nearest node to the origin and the farthest node from the origin. When the number of nodes exceeds 150, r_{ij}^M approaches 1 and r_{ij}^V approaches 0.01. Then, the reproduced and adjusted geometry is nearly congruent with the network's original geometry.

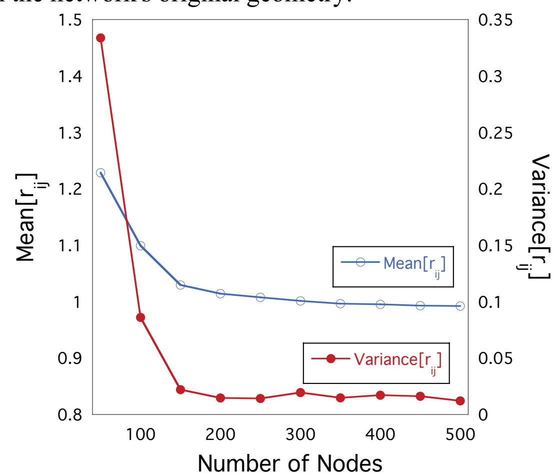


Figure 6. Dependence of r_{ij}^M and \$ r_{ij}^V on number of nodes with two anchor nodes.

C. Evaluation of absolute location

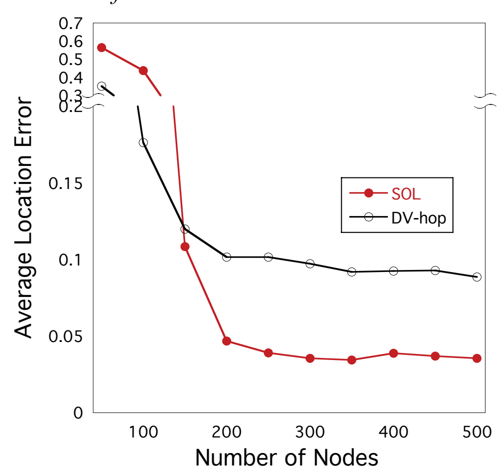


Figure 7. Dependence of average location error on number of nodes with three anchor nodes.

To evaluate accuracy of the absolute node location, compare its accuracy with that of DV-Hop in accordance with (17). DV-Hop can estimate node location with just three anchor nodes and without special distance measurement and previous information on the space in advance. The others do not meet the above restriction. DV-Hop calculates the average distance in 1 hop using the hop count between anchor nodes based on the minimum hop route and the distance between anchor nodes, and it estimates node

 ISBN: 978-1-61208-289-9

location with triangulation that uses the location of each anchor node and the calculated distance to each anchor node. Figure 7 shows the comparison of SOL with DV-Hop on average location error in accordance with (17). The three anchor nodes are respectively the nearest node to the origin, the farthest node from the origin, and the farthest node from the above two anchor nodes.

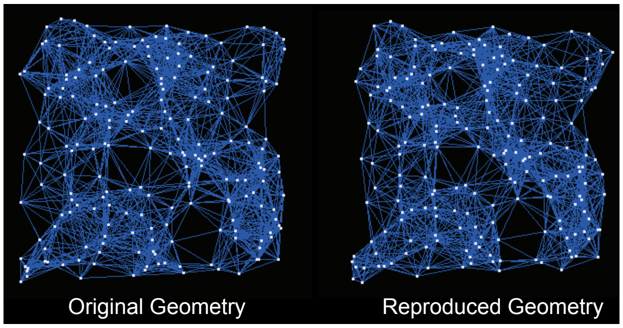


Figure 8. Comparison of reproduced geometry with original geometry on 200 nodes.

When the number of nodes is smaller than 100, SOL location accuracy and DV-Hop location accuracy are low. On a low-density network, SOL cannot effectively reproduce network geometry, and DV-Hop cannot estimate node location. Furthermore, the route based on the minimum hops becomes a zigzag or bent path rather than a straight-line, and, thus, the calculated average distance on 1 hop is inaccurate. Therefore, the accuracy on the estimated node location of DV-Hop becomes low. Any localization that depends on network topology has the problem that accuracy decreases in low-density networks.

When the number of nodes exceeds 200, the average location error of DV-Hop approaches approximately 0.1, and that of SOL approaches approximately 0.04. Therefore, the accuracy of SOL is much superior to that of DV-Hop (see Fig.8).

VI. CONCLUSION

In this paper, Self-Organizing Localization for wireless sensor networks was proposed. SOL requires no distance measurement schemes and no advanced information on the space, and its dependence on anchor nodes is very low. On the suitable density of nodes, SOL achieves the following results by using SOM.

- Without anchor nodes, it reproduces a geometry very similar to the network's original geometry.
- With two anchor nodes, it reproduces a geometry nearly congruent with the original, that is, it derives relative node locations on the network.
- With just three anchor nodes, it derives absolute node locations with high accuracy

Given the above properties, SOL can be applied to any space and any sensor node.

REFERENCES

[1] P. Bahl and V. N. Padmanabhan, “IRADAR: An In-Building RF-based User Location and Tracking System”, In Proc. IEEE InfoCom 2000, vol. 2, 2000, pp. 775–784.

[2] R. Want, A. Hopper, V. Falcao, and J. Gibbons, “The Active Badge Location System”, ACM Trans. on IS, vol.10, no. 1, 1992, pp. 91–102.

[3] J. Hightower, G. Boriello, and R. Wat, “SpotOn: An indoor 3D Location Sensing Technology Based on RF Signal Strength”, University of Washington CSE Report 200-02-02, 2000.

[4] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster, “The Anatomy of a Context-Aware Application”, MobiCom1999, 1999.

[5] A. Savvides, C. Han, and M. Srivastava,“Dynamic Fine-Grained Localization in Ad-Hoc Networks of Sensors”, In Proc. IEEE MobiCom 2000, 2000, pp. 1–14.

[6] M. Minami, Y. Fukuji, K. Hirasawa, S. Yokoyama, M. Mizumachi, H. Morikawa, and T. Aoyama, “'DOLPHIN: A Practical Approach for Implementing a Fully Distributed Indoor Ultrasonic Positioning System”, In Proc. ACM Ubicomp 2004, 2004, pp. 166–179.

[7] N. Priyantha, A. Miu, H. Balakrishman, and S. Teller,”The Cricket Compass for Context-aware Mobile Applications”, MobiCom2001, 2001.

[8] N. Beutel, J. Heidermann, and D. Estrin, “GPS-less Low Cost Outdoor Localization for Very Small Devices”, IEEE Personal Comm. Magazine, vol. 7, no. 5, 2000, pp. 28–34.

[9] D. Niculescu and B. Nath, “DV-based Positioning in Ad Hoc Networks”, Telecommunication Systems, vol. 22, 2003, pp. 267-280

[10] T. He, C. Chung, B. M. Blum, J. A. Stankovic, and T. F. Abdelzaher, “Range-Free Localization and Its Impact on Large Scale Sensor Networks”, ACM Trans. Embedded, vol. 4, no. 4, 2005, pp. 877-906.

[11] E. Ertin and K. Priddy, “Self-localization of wireless sensor networks using self-organizing maps”, Proceedings of SPIE, 2005.

[12] G. Giorgetti, S. K. S. Gupta, and G. Manes, “Wireless localization using self-organizing maps”, In Proceedings of the 6th international conference on Information processing in sensor networks (IPSN '07), 2007, pp. 293-302.

[13] Y. Takizawa, P. Davis, M. Kawai, H. Iwai, A. Yamaguchi, and S. Obana, “Self-Organizing Location Estimation Method Using Received Signal Strength”, IEICE Trans. on Comm., vol. E89-B, no. 10, 2006, pp. 2687-2695.

[14] T. Kohonen, “Self-Organizing Maps 3rd ed”, Springer, 2000.

[15] E. Bonabeau and F. Henaux, “Self-organizing maps for drawing large graphs”, Information Processing Letters, vol. 67, no. 4, 1998, pp. 177-184.

 ISBN: 978-1-61208-289-9

Application Driven Environment Representation

André Dietrich, Jörg Kaiser, Sebastian Zug, and Sasanka Potluri
Department of Distributed Systems
Otto-von-Guericke-Universität Magdeburg
Universitätsplatz 2, 39106 Magdeburg, Germany
Email: {dietrich, kaiser, zug, sasanka}@ivs.cs.uni-magdeburg.de

***Abstract*—Autonomous systems have to sustain in environments of growing complexity and under dynamically changing conditions. In addition, the number and the complexity of the tasks these systems have to fulfill increases constantly. To cope with these emerging problems, we present a general concept of constructing environment models. From these models, we can derive any kind of information in an application specific abstraction, which we call views. Furthermore, we describe the main problem in the generation of views as well as a possible solution to this.**

Keywords*—*Smart Environments, Environment Model, Robotics

I. Introduction

One of the main characteristics of smart environments is the spontaneous use of ambient information, obtained from remote sensors and servers. These sources of information (mobile or stationary) can be a part of the infrastructure of a smart space itself or gathered from other systems that exploit these location-dependent information and services. The ultimate goal is to perceive the environment with the best possible quality, including as many aspects as desired. For autonomous systems, such as mobile robots or even for the emerging field of cooperating cars, the benefits of exploiting environment knowledge and available sensory information are obvious; reaching from a substantial increase of sensing range to the integration of more and more detailed aspects into the environment perception. Literally, it is possible to look around the corner or to recognize obstacles and to plan to avoid them, long way before they come into sight of a local sensor.

Perception requires two things: the input from sensors that provide some basic physical data and a model that allows to interpret this information correctly. Designers of such applications have to deal with a couple of problems. Firstly, the heterogeneity of sensors may require intimate knowledge of sensor characteristics and complex adaptation procedures of the application. Secondly, they have to solve the problem of interpreting data coming from complex mobile remote directed sensors like Kinects, laser scanners or cameras. Solving both problems requires a huge amount of specific context information, as depicted in Figure 1.

On the lowest level, sensor data interpretation needs some form of knowledge about the sensor type and the physical nature of observed real world entities, which can be provided by metadata. We therefore had developed expressive description formats for sensors, to support their spontaneous use [1]. While these sensor descriptions cover low level sensor interpretation and fusion, this paper focuses on an application-oriented abstraction of perception on a higher level. We propose the

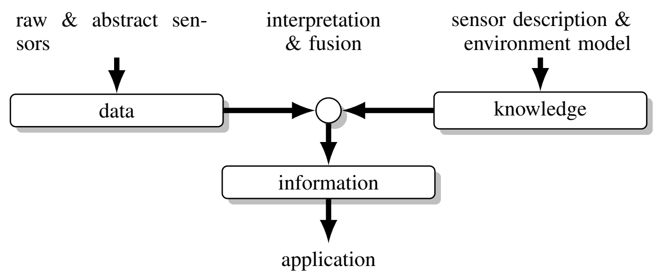


Fig. 1: Relation between sensor data and application specific information.

use of environment models as explicit knowledge for putting remote sensor data into an adequate context. We argue that cooperative systems, which exploit information from remote sensors, will need an explicit environment model to cope with the requirements of mobility. The information, generated with the help of an environment model and available data, can be represented in different flavors, such as an occupancy grid map, a relative position, or a set of objects or entities that fulfill some certain conditions.

In the related work, an environment model is rarely understood as a powerful filter and fusion mechanism for the heterogeneous and distributed sensor system, providing a well-defined and stable view of the physical surroundings. Additionally, as a major difference, the presented work does not consider cooperative functions of different systems based on an easy and efficient exchange of environment knowledge. In this paper we strive for a more general approach, which will support sensor interpretation, serve multiple applications, and also allows easily to exchange environment information between different mobile or immobile entities. This will be crucial for generating large scale and global environment perception. In contrast to our approach, other solutions are to a large extent rather application-specific and just tailored to a single dedicated control task. Furthermore, in most cases the environment models interpreting information coming from (local) sensors are intimately integrated into the application.

Before describing our approach and the problems we have to deal with, we will give detailed overview on related research areas within the next section and finish with a conclusion and an outlook on future work.

II. Related Work

There is a huge amount of proposals to model the environment from various areas in mobile robotics. An early overview

 ISBN: 978-1-61208-289-9

on state-of-the-art world models is given in [2] but as stated by the authors themselves, most of these models are designed for specific (robotic) applications only. Examples are “Constructive Solid Geometry” [3], a geometric world model based on primitive geometrical objects, occupancy grid maps [4], octrees [4], and many more. Nearly all of these examples for models are also directly linked to the available sensor equipment and the respective data formats. In contrast to the pure geometrical representations, there are also application specific models, like in [5] that shows an example for trajectory-planning of car like vehicles, which require an abstract 2D model of the environment.

Recently there is a shift to more general and complex environment models, which results from the fact that also the robotic applications and the environments itself become more and more complex. Models should serve different purposes like surveillance, decision making, trajectory planning, obstacle avoidance, etc. Also, the authors of [6] tackle the problem of specified and customized world models. They present a system that incorporates a 3D occupancy grid map, other abstractions with more than just spatial semantics are not considered (but it allows to adapt update rates and accuracy to serve different needs). A more general approach, with more information about the environment and the objects within, is presented by Hsiao [7]. It bases on the Open Dynamics Engine (ODE) [8], a high performance library for simulating rigid body dynamics. Next to the geometric description of the environment, it also allows to label included 3D objects with additional information, such as mass, velocity, color, etc.

The idea that an environment model can be constructed from multiple sources is presented in [9]. It describes a vehicle’s road environment conceptually as an (very specialized) object-oriented model. It uses a-priori information and information obtained from on-board sensors or through car2car communication. This means that, for example, the awareness of another car in front can be obtained from the local sensor system or from the communicated position of the front car itself. It makes no difference for the application, because all information is taken from the model representation only. This type of modeling is ideal for situation assessment, because all required sensor data of the environment is already translated into a simplified data-structure. This type of modeling is specialized on road environments, which restricts the type of entities (vehicles, pedestrians, and traffic signs), their representations (2D points), and also the application. Nevertheless, it shows that the application can be separated from the sensor interpretation and work only with information derived from a general model.

We discussed this separation of concerns firstly in [10] and presented a step by step approach of extracting functionality from an application that deals with environmental perception in [11]. Furthermore, we demonstrated the applicability of our concept in [12], see also our YouTube-channel [13]. Comparable to our approach, which is presented in the next section, Belkin and Kuwertz discuss their concept for a holistic environment model in [14] and [15]. It bases also on a scene-graph, representing the spatial relations between different geometrical entities/concepts. But in contrast to their solution, we do not believe that it is desirable in general to build a global environment model. Instead, every application should be able to generate its own local environment model for its own purposes (based on the available data within a smart environment) and abstract any kind of information from it. We had developed an architecture, based on a distributed database, which allows each entity to construct and update its own local environment model. Furthermore, all required data whether it is actual sensor data (using our Cassandra_ROS package [16]) or metadata describing sensors, actuators, or other objects of the environment, can be stored with arbitrary complexity as required.

Models based on logic predicates can be seen as another type of environment modeling, which also separates sensor interpretation and the use of the derived information. In a first step, all data has to be transformed into logic predicates, which subsequently can be queried easily according to various aspects. This approach is mainly used to determine complex action sequences as well as to describe complex situations and formally bases on the situation calculus [17]. Examples are “alGOl in LOGic” better known as GOLOG [18] with its specialized dialects, such as Con(current)GOLOG [19], which includes concurrency and the influence of external events. KnowRob [20] is also a knowledge processing system already integrated in the Robot Operating System (ROS) [21]. Logic-based approaches require huge knowledge bases and are slow in response, but due to the used concepts of predicates and rules, queries and situations can be expressed very simple (mostly in Prolog-syntax).

III. Environment and Views

It is a challenging task for a system to interpret varying and heterogeneous sets of networked remote sensors and also information obtained from other systems. It has to have the ability to “understand” the environment. On the highest level all information about the environment is captured in an environment model. The concept of an environment model can be described as a simplified co-simulation of the surrounding, which is continuously updated by local and remote sensors. It consists of elements and relations that represent all relevant contexts for perception and control.

Similar to the concept described in [14], [15], a simple scene graph may be sufficient enough, if only spatial information (e. g., distances, geometries, relative positions, etc.) is required. More sophisticated scenarios like cooperating cars require next to a geometric representation also data about masses, friction, velocities and forces. This type of data can be easily incorporated into the model by extending the prior scene graph with capabilities of the ODE (equal to the approach of by Hsiao [7]). The model as we use it for a robotic applications are even more complex, including dedicated information about the surroundings, its inhabitants (other robots and sensors), their capabilities, as well as on different objects and tools. To incorporate also such kind of information, we apply the OpenRAVE [22] (Open Robotics Automation Virtual Environment), an environment for testing, developing, and deploying real-world robotics applications.

As depicted in Figure 2, the environment model fulfills three tasks. First of all, sensor data is interpreted and filtered using its information. Secondly, it stores a history of states of the environment and the present state that is continuously

 ISBN: 978-1-61208-289-9

updated by sensor data. Thirdly, the model provides access to specific aspects of the environment. We call the respective information, depending on the application requirements, a "view". Figure 2 illustrates these relations.

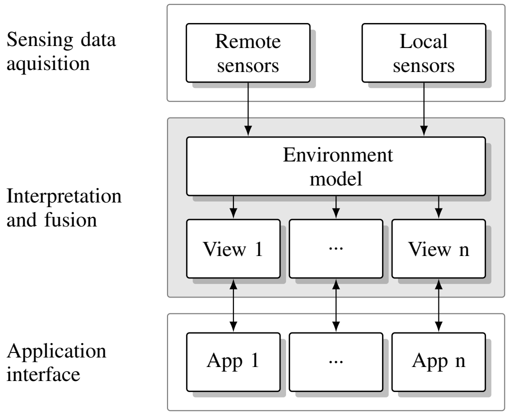


Fig. 2: View generation related to application requirements and based on a common environment representation.

A view is a well-defined and specialized abstraction of the external environment, while the environment model is a general representation and will serve multiple applications. This separation of concerns will ease the design of the application and support adaptation, because all the assumptions about the environment are made explicit in the description and can be changed. A specialized representation may be partly based on a-priori knowledge like a map but it is updated by the information derived from the sensors to include elements that are dynamically perceived, for instance, other vehicles and obstacles. As already mentioned, a 3D geometric environment model will comprise a coordinate system, geometric elements, the relations between them and the own position of the respective entity. If other elements were classified as mobile, their positions and relations will continuously be updated by sensor information.

Figure 3 shows the dataflow of a simple robotic scenario. In this example, a mobile platform operates in the aisle of our laboratory. It is equipped with a Kinect sensor and has access to virtual remote sensors that provide the static geometry of the building. It should be noted that the correct interpretation of sensor data from the Kinect already needs the environment model, i. e., the description about the position and the field of view. If we enrich the geometric model by the description of motion, we will be able to observe trajectories and use the environment model to detect and reason about jams and collisions. However, for our example we assume two functions, collision avoidance and short range path planning. Both applications utilize different aspects of the 3D environment model obtained from the view generator. The path planning algorithm receives an occupancy grid map of the vicinity. The collision avoidance algorithm does not need the sophisticated representation of the general environment model. The view defines a needed subset of the environment model, describing this as metadata to interpret sensor data. The view thus provides just the information needed by the application functions. A further important benefit of a view is that it shields the application from the details of the sensor system. In the example, the view would provide the unit "distance" independently of what sensors are used to derive this information. As indicated earlier, a varying set of sensors can participate in generating a view, without the need that the application must be aware of this.

IV. The Problem of Arbitrary Views

As shortly described in Section II, our previous work is based on developing an infrastructure for distributed environment modeling. According to the required view complexity, different models can be applied. All information about the surrounding are obtained from the environment model. But currently we have to apply specialized functions and filters to create views, specialized in terms of the type of environment model. Values such as the distance between two objects, their velocities, colors, or their relative positions can be easily extracted from the environment, while the generation of an occupancy grid map might be more difficult. And the definition of functions can vary according to the applied scene graph, physics simulation, etc.

It is easy to see that this solution is less satisfactory because complex queries or the definition of situations (based on the scene graph) have to be defined within the program code. Although, we started out with the undertaking of reducing source code complexity. Furthermore, calling such functions from the program code does not simplify the programmation and it does not meet the requirements of a dynamic access and changing demands. Queries and situations are hard coded and we have to parameterize our code to include dynamics. As also stated in [14], it is essential to have some kind of symbolic representation of the environment to solve these problems. Currently there is no system or solution, which can convert geometric models into logic-based and vice versa.

If we interpret environment models as an implicit and dynamically changing knowledge base, it should also be possible to query it in the same way as we query databases. In contrast to logic-based systems, we do not want to query the environment with predicates and rules, instead we want to query it with simple SELECT-queries, similar to ordinary SQL. In fact, querying with SQL or with Prolog is quite the same, since SQL is already an implementation of the Relational Calculus (form of Predicate Logic). While SQL is primarily intended to derive facts and relations (as we want to use it), Prolog is primarily a rules and inference engine. The result of an SQL-query, as we intend it, can be a single distance between two objects, a list of object having a special property, a specialized map of the surrounding, containing only objects bigger than a certain value, or a simplified model of the environment model itself. Everything can be put into SELECT statement using the WHERE clause to define additional conditions. Furthermore, it allows also to define situations, which occur, if the SELECT query returns a non-empty result. Therefore, we are currently developing a new kind of programming semantics/language that should allow to mix SELECT statements, running on our environment models and normal programming languages.

V. Conclusion and Future Work

As stated in [2], an environment model is the key component of any intelligent system, which must be able to describe and represent the environment as well as incorporated entities

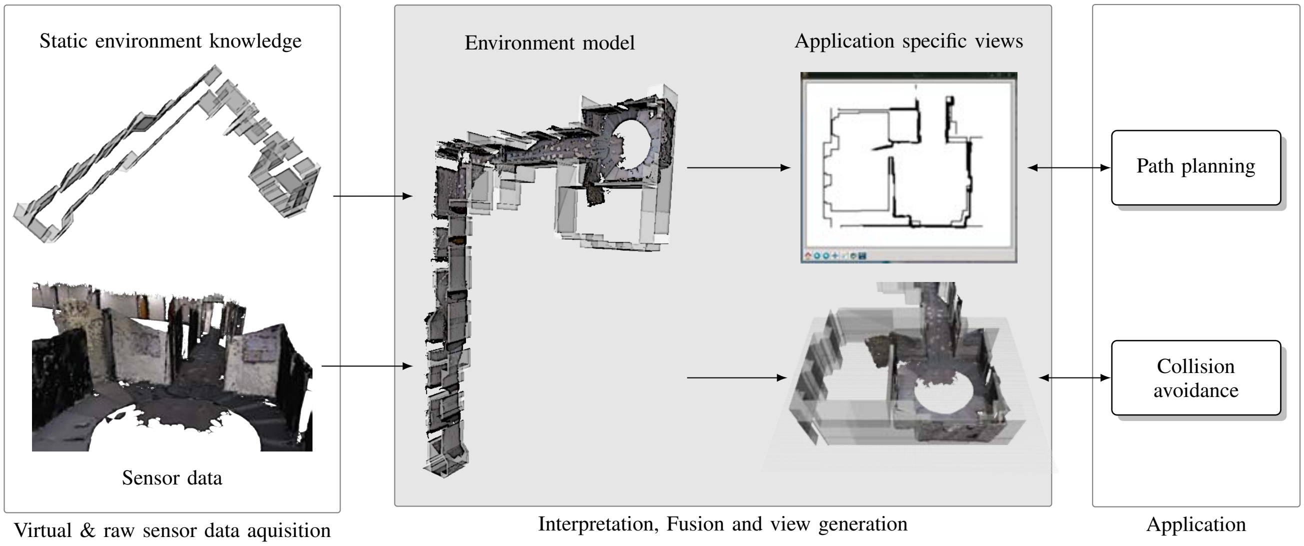


Fig. 3: Composing an environment representation from static & dynamic data and generate of application specific views from it.

in an adequate structure and complexity. We presented our general approach of building environment models and the concept of deducing views from it. To be able to cope with the complexity and diversity of application specific requirements, we will have to develop also new programming concepts and paradigms. Or at least combine well known concepts in a new ways, which we will do by applying SQL on environment models.

ACKNOWLEDGMENT

This work is funded and supported by the German Ministry of Education and research within the project ViERforES-II (grant no. 01IM10002B) and by the EU FP7-ICT program under the contract number 288195 "Kernel-based ARchitecture for safetY-critical cONtrol" (KARYON).

REFERENCES

[1] S. Zug, M. Schulze, A. Dietrich, and J. Kaiser, "Programming abstractions and middleware for building control systems as networks of smart sensors and actuators," in *Proceedings of Emerging Technologies in Factory Automation (ETFA '10)*, Bilbao, Spain, 9 2010.

[2] E. Angelopoulou, T.-H. Hong, and A. Y. Wu, "World model representations for mobile robots," in *Proc. of the Intelligent Vehicles Symposium*. IEEE, 1992, pp. 293–297.

[3] A. A. G. Requicha and R. B. Tilove, "Mathematical foundations of constructive solid geometry: General topology of closed regular sets," Production Automation Project, Univ. Rochester, Tech. Rep., 1978.

[4] S. Thrun, "Learning occupancy grid maps with forward sensor models," *Autonomous robots*, vol. 15, no. 2, pp. 111–127, 2003.

[5] T. Fraichard, "Smooth trajectory planning for a car in a structured world," in *Proc. of the International Conference Robotics and Automation*. IEEE, 1991, pp. 318–323.

[6] R. K. Harle and A. Hopper, "Dynamic world models from ray-tracing," in *Proc. of the 2. Annual Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 2004, pp. 55–64.

[7] K. Hsiao, N. Mavridis, and D. Roy, "Coupling perception and simulation: Steps towards conversational robotics," in *Proc. to the International Conference on Intelligent Robots and Systems (IROS)*, vol. 1. IEEE/RSJ, 2003, pp. 928–933.

[8] R. Smith, "Open Dynamics Engine project website," http://www.ode.org, [(online), as at: 29.03.2011].

[9] A. Furda and L. Vlacic, "An object-oriented design of a world model for autonomous city vehicles," in *Proc. of the Intelligent Vehicles Symposium (IV)*. IEEE, 2010.

[10] A. Dietrich, S. Zug, and J. Kaiser, "Model based Decoupling of Perception and Processing," in *ERCIM/EWICS/Cyberphysical Systems Workshop, Resilient Systems, Robotics, Systems-of-Systems Challenges in Design, Validation & Verification and Certification*, Naples, Italy, 9 2011.

[11] ——, "Towards Artificial Perception," in *SAFECOMP 2012 Workshops*, F. Ortmeier and P. Daniel, Eds. Springer-Verlag Berlin, 2012, pp. 466–476.

[12] ——, "Geometric Environment Modeling System," in *IFAC Conference on Manufacturing Modelling, Management and Control*, Saint Petersburg, 6 2013, pp. 1445–1450.

[13] A. Dietrich, "Youtube-channel: Ivs magdeburg," http://www.youtube.com/ivsmagdeburg, [(online), as at: 29.09.2013].

[14] A. Belkin, A. Kuwertz, Y. Fischer, and J. Beyerer, "World modeling for autonomous systems," *Innovative Information Systems Modelling Techniques*, vol. 1, pp. 135–158, 2012.

[15] A. Kuwertz, "Towards adaptive open-world modeling," Vision and Fusion Laboratory, Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT), Tech. Rep., 2012.

[16] A. Dietrich and S. Zug, "Cassandra_ros project website," http://www.ros.org/wiki/cassandra_ros, [(online), as at: 29.09.2013].

[17] J. McCarthy, "Situations, Actions, and Causal Laws," Stanford University Artificial Intelligence Project, Tech. Rep., 1963.

[18] H. Levesque, R. Reiter, Y. Lesperance, F. Lin, and R. Scherl, "Golog: A logic programming language for dynamic domains," *The Journal of Logic Programming*, vol. 19, no. 1-3, pp. 59–83, 1994.

[19] G. De Giacomo, Y. Lespérance, and H. Levesque, "Congolog, a concurrent programming language based on the situation calculus," *Artificial Intelligence*, vol. 121, no. 1, pp. 109–169, 2000.

[20] M. Tenorth and M. Beetz, "Knowrob – knowledge processing for autonomous personal robots," in *Proc. of the International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2009, pp. 4261–4266.

[21] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Ng, "Ros: an open-source robot operating system," in *ICRA Workshop on Open Source Software*, vol. 3, no. 3.2, 2009.

[22] R. Diankov, "Automated construction of robotic manipulation programs," Ph.D. dissertation, Carnegie Mellon University, Robotics Institute, 8 2010.

WSN Trends: Sensor Infrastructure Virtualization as a Driver Towards the Evolution of the Internet of Things

Andreas Merentitis, Florian Zeiger, Marco Huber, Nick Frangiadakis, Kostas Mathioudakis,
Konstantinos Sasloglou, Georgios Mazarakis, Vangelis Gazis, Zachos Boufidis
AGT International
Darmstadt, Germany
{amerentitis, fzeiger, mhuber, nfrangiadakis, kmathioudakis, ksasloglou, gmazarakis, vgazis, zboufidis}@agtinternational.com

Abstract— **Fueled by advances in microelectronics, wireless communications and the availability of affordable mobile connectivity, the last decade has seen a rapid increase in the number of devices connected to the Internet. This evolution is part of the transition to the Internet of Things (IoT), which envisions connecting anything at any time and place. While it can be argued that the IoT is already here, the next paradigm shift is already emerging on the horizon, targeting yet another order of magnitude increase in the number of interconnected devices and promising to bring people and processes in the equation. However, before such a paradigm shift can be realized, significant challenges with respect to scalability, cooperative communications, energy consumption, as well as convergence of sensor and analytics trends have to be resolved. Sensor Virtualization Technology capturing both the Virtual Sensors and Virtual Sensors Networks aspects, promises to alleviate or resolve these challenges, and pave the way towards the evolution of the Internet of Things.**

Keywords-Sensor Networks, Sensor Virtualization; Machine to Machine Communications; Internet of Things; Future Internet.

I. INTRODUCTION

Technological advances in the fields of sensor technology, low power microelectronics, and low energy wireless communications paved the way for the emergence of Wireless Sensor Networks (WSNs). These networks are currently used in a wide range of industrial, civilian and military applications, including healthcare applications, home automation, earthquake warning, traffic control and industrial process monitoring. A WSN is a system composed of small, wireless nodes that cooperate on a common distributed application under strict energy, cost, noise and maintenance constraints [1], [2]. Although many interesting applications have been implemented/developed for WSNs, further work is required for realizing their full potential as “the next big thing” that will revolutionize the way we interact with our environment.

As a promising step in this direction, during the last decade there has been a growing research interest in the Internet of Things (IoT), ranked as a disruptive technology, according to the US National Intelligence Council [3]. An early definition for the IoT envisioned a world where computers would relieve humans of the Sisyphean burden of data entry, by automatically recording, storing and processing all the information relevant to the things involved in human activities, while also providing “anytime, anyplace [...] connectivity for anything” [4].

Henceforth, and depending on the viewpoint, different understandings and definitions have been reported in the literature [5]-[7] regarding what the Internet of Things is about. However, while it is possible to argue that the IoT is already here [8], the next (r)evolutions are already on the horizon, ranging from the open effort to the Future Internet up to industry driven initiatives such as the National Instruments Data Acquisition Technology Outlook [9], the General Electric concept of “Industrial Internet” [10], and the CISCO initiated “Internet of Everything” [8], [11]. Such initiatives have differences in flavor and focus; yet, it is possible to distil the general trends and enablers that need to be in place for successfully realizing the shift to the next networking paradigm, whichever form it might take. In this paper, we argue that, among these enablers, Sensor Network Virtualization is a technology that has the potential to augment and unlock advances in several other fronts (e.g., scalability, cooperation, low energy solutions and convergence of Sensor Network and Data Analytics trends) that will pave the way towards this paradigm shift.

The rest of the paper is organized as follows: Section II presents some of the key networking trends that are commonly captured in several independent views for the next networking paradigm evolution. It finishes with a selection of four core areas where significant challenges remain unresolved. The selected areas and the nature of the challenges in each of them are then discussed in more detail in Sections III-VI. Section VII elaborates on the different aspects of sensor infrastructure virtualization. Their advantages are captured and the potential of using different virtualization flavors to address the challenges described earlier is explained. Finally, Section VIII concludes the paper.

II. IDENTIFICATION OF NETWORKING TRENDS

In parallel with the efforts towards interconnecting a myriad of smart devices according to the IoT vision, the Future Internet stands as a general term for research activities and communication paradigms towards a more up to date and efficient Internet architecture. Approaches towards the “Future Internet” cover the full range from small, incremental evolutionary steps up to complete redesigns (clean slate) of the core architecture and the underlying mechanisms, where the applied technologies are not to be limited by existing standards or paradigms (e.g., the client server networking model might evolve into co-operative peer structures). In general, most of the work in this area is summarized by the

 ISBN: 978-1-61208-289-9

Future Internet Assembly (FIA) [12], where it is underlined that whatever form the Future Internet may take, a set of core principles need to be preserved:

- *Heterogeneity support principle*, refers to supporting a plethora of devices and nodes, scheduling algorithms and queue management mechanisms, routing protocols, levels of multiplexing, protocol versions, underlying link layers or even administrative domains and pricing structures.
- *Scalability and Amplification principle*, describing the ability of a computational system to continue operating under well specified bounds when its input is increased in size or volume.
- *Robustness principle*, ensuring that each protocol implementation must transparently interoperate with other implementations.
- *Loose Coupling principle*, describing a method of interconnecting architectural components of a system so that those components depend on each other to the least extent practicable.
- *Locality principle*, which in the computer science domain focuses on the design of thrashing-proof, self-regulating, and robust logical systems.

However, apart from these principles that should only undergo small incremental changes (if any) a list of additional principles that need to be significantly adapted/relaxed or augmented is also provided. Here, we focus on a subset of this list that is related or overlapping to the IoT evolution:

- *Keep it simple, but not "stupid" principle* [12], which refers to the fact that in current Internet design, the complexity belongs always at the edges, while in a more flexible architecture inherently supporting heterogeneous "Things" this might not always be the case.
- *Polymorphism principle,* which refers to the ability to manipulate objects of various classes, and invoke methods on an object without knowing that object's type. The idea is to extend this principle to allow the same abstract components exhibiting different functional and non-functional behavior in case of changing environments or circumstances [12].
- *Unambiguous naming and addressing principle*, establishing that protocols are independent of the hardware medium and hardware addressing scheme. The proposal of the FIA initiative is to extend this principle in order to also capture the data and services.

Even more recently than the FIA initiative, CISCO has evangelized the Internet of Everything (IoE) as the next wave in the evolution of the networking paradigms [8]. With a clear all-IP focus, building on the same principles as Machine to Machine Communications (M2M) and the Internet of Things but extending them, the IoE envisions to increase the number of connections by yet another order of magnitude (from ~10 billion currently connected "Things"). However, arguably the biggest innovation is that it targets to include processes and people in the loop, facilitating and enabling communications that are more relevant in order to offer new capabilities, richer experiences and unprecedented economic opportunities.

In all the previous activities, as well as in various independent research efforts, it has already being identified that in future large-scale heterogeneous networks, the adoption of mechanisms achieving scalable, predictable and self-adaptive network behavior ("more relevant" in CISCO IoE terminology, "pushing the boundaries" in the GE Industrial Internet notion) will be a key enabler [8], [10], [11], [13], [14]. At the same time, with systems becoming continuously more complex in terms of scale and functionality, reliability and interoperability are getting increasingly important. Therefore, techniques for achieving dependable system operation under cost and energy constraints will be an important evolutionary step [2], [13], [14].

Current wireless network development is guided by horizontal mass-markets ("one size fits all"). More often than not, different verticals and niche markets require dedicated applications [14]. Consequently, the deployment or evolution of a wireless network in these areas often demands for expensive infrastructure replacement. Moreover, extending system and network capabilities, switching services or adopting the purpose of an operational network consisting of heterogeneous "Things" usually calls for costly (manual) reconfigurations and upgrades, while it often results in temporary unavailability of system services. On the other hand, dynamic changes during operation typically allow for only a limited subset or scope of updates. Solutions for such problems require capabilities for spontaneous ad-hoc cooperation between objects, self-adaptive behavior, exploitation of dynamic information, predictability of non-functional properties (e.g., energy consumption), and on-the-fly reconfiguration [13], [14], [15].

Summarizing, first and foremost, **scalability** is the key enabler for facilitating the (r)evolution of the Future Internet as the number of interconnected devices is expected to rise by yet another order of magnitude. The vast majority of these devices will be smart sensors with relatively limited computation resources. Thus, key challenges lie in efficient **cooperation** of heterogeneous network elements in order to realize advanced capabilities and services. Furthermore, innovations to **low energy solutions** create an attractive business case by offering benefits in terms of operational cost, long-term product reliability and increased lifetime of wireless and mobile elements (especially relevant for a significant portion of the myriad of electronic "Things" that will be battery powered). Last but not least, as the number of interconnected devices will increase a **convergence of the Sensor Network and Data Analytics trends** is required for effectively bringing processes and people into the equation. An overview of the respective trends, including some of the main open issues, is provided in the sequel of this section.

III. SCALABILITY OF COMMUNICATION AND MANAGEMENT

In order to realize the vision of ~50 billion devices connected to the Internet by 2020 [8], several scalability enablers need to be in place. One can argue that some of them are already here and they have driven the evolution towards the estimated ~10 billion interconnected devices that we have currently reached [8], [11]. Hardware node miniaturization, node capability enrichment and cost reduction, all fueled by Moore's law, are a good example of such enablers. Processing and storage availability are also improving thanks to the cloud computing paradigm. On the network protocol naming and addressing part, the transition to IPv6 has to take place sooner than later in order to facilitate the next jump in number of interconnected devices.

However, apart from the hardware node and protocol/ communication part, efficient management of this huge number of heterogeneous devices is also a big challenge. The concept of network management traditionally captures the methods and tools that are related to the operation, administration, maintenance, and provisioning of networked systems. In this context, operation is related to keeping the network working according to the specifications; administration is dealing with resource tracking and utilization; maintenance is concerned with changes and upgrades to the network infrastructure; and finally provisioning addresses dynamic, service-based resource allocation. However, catering for heterogeneous sensors and actuators with different requirements and operational properties calls for a paradigm shift; higher layers need to efficiently capture the changing dynamics of the systems and the lower layers need to transform this information into appropriate action, in an autonomous and scalable fashion.

In recent years, several extensions have been proposed to the traditional definition of network management that are specifically designed to address the topic of ever increasing network management complexity. The Self-Organizing Network (SON) notion was introduced by the 3rd Generation Partnership Project (3GPP) and targets to constitute future radio access networks easier to plan, configure, manage, optimize and heal compared to current state of the art. In similar direction, Autonomic Networking, inspired by the IBM initiated vision for Autonomic Computing [16], has been proposed as a means to create self-managing networks able to address the rapidly growing complexity of modern large scale networks and to enable their further growth, far beyond the size of today. The four main pillars of Autonomic Networking are self-configuration, self-healing, self-optimization, and self-protection, known also as self-CHOP features. However, the related technologies have so far found their way mostly in cellular networks or in smaller scale ah-hoc sensor networks. Frameworks for configurable and, to some extent, reusable deployment of SON functionality would be an important evolutionary step in the direction of scalable network management and lower maintenance cost.

IV. Cooperative Communications and Networking

Close cooperation between network elements is increasingly seen as an important driver for further evolution. In the FIA recommendations, it is referenced, for example, that the traditional client-server model will at least partially evolve into co-operative structures between peer entities. Cooperation frameworks cover the full range from information exchange, actions coordination and decision making. Moreover, such aspects are expected to be utilized in different context, thus spanning different communication layers and capabilities. A taxonomy of cooperative and collaborative frameworks was presented in [13].

In order to achieve cooperation between networks in multi-stakeholder networking environments, proper incentives need to be in place. Such incentives formulate the expected networking benefits that a single network can derive from its cooperation with another. Networks are only motivated to cooperate with other networks when this cooperation improves their performance according to such incentives [13]. However, in order to be effective and support generalization in a large scale dynamic environment, the incentives should not express low-level performance metrics, but instead indicate high level functional and network requirements. An incentive formulates a reason for cooperation between networks (i.e., if cooperation with another network can improve this high level objective, cooperation might be viable). Example incentives are (i) increasing coverage (to reach more clients), (ii) reduce energy consumption (to increase battery life), and (iii) increasing QoS guarantees (higher throughput, higher reliability, lower delay, etc.), among others [13].

Deciding, however, on the most beneficial cooperation settings requires mechanisms such as negotiation [13]. During negotiations, independent devices or complete networks with the required capabilities are identified and the utility of the cooperation is derived also as part of the cooperation incentive [17]. While significant research efforts have been invested in this area, large scale commercial application is still limited. Variations in the realization of the cooperation mechanisms and compatibility problems between the early products of different vendors are among the more important inhibitors, therefore ways to alleviate them will be particularly beneficial.

V. Low Energy Solutions

Energy efficiency is commonly perceived as one of the most important design and performance factors of a Wireless Sensor Network (WSN). This fact is only expected to increase in relevance as a myriad of additional mobile and portable devices will be connected to the Future Internet. The desired low energy behavior can be achieved by optimizing the sensor node as well as the communication protocol [18]. The goal is to reduce energy consumption and, consequently, increase the lifetime of the system.

At the level of the independent nodes, the fundamental limit of the energy requirements is calculated by taking into account the energy consumption of every hardware (HW) component on a WSN node like sensors and conditioning electronic circuitry, processing and storage, radio, etc. The components selected in the final node architecture will have a significant impact on the nodes' capabilities and lifetime. Thus, a holistic low-power system design should be pursued from the very beginning, creating the correct HW infrastructure base for further network, protocol, software and algorithmic energy efficiency optimization.

This holistic low-power system approach can further incorporate methods for energy harvesting from the

 ISBN: 978-1-61208-289-9

environment in order to utilize ambient energy sources (e.g., mechanical, thermal, radiant and chemical) that will allow extending lifetime and minimizing or possibly removing the need for battery replacement. Such a scenario would enable the development of autonomous wireless sensor networks with theoretically unlimited lifetime. Still focused on the sensor node level, but on the algorithmic part, ongoing efforts are made to design the sensor nodes in an inherent power-aware approach. The goal is to develop an adaptable system that is able to prioritize either system lifetime or output quality at the user's request. The radio communication and network protocol part is the other major source of energy consumption that can be targeted for optimization. Optimizing the network protocol is typically done with respect to a specific application domain, usually to favor bursts of transmission followed by cycles of low or no activity. As the range of transmission is also a very important parameter, low energy operation of a specific protocol version is often achieved only for a selected range, whereas other protocols are more efficient beyond that range.

Optimizations for low energy are a relatively mature field that has been (in different forms) around for a long time. However, most of the available solutions are customized for specific applications and are not directly transferable across different verticals and application domains. For example, a low energy protocol is typically "optimal" only with respect to a specific communication range and bandwidth, while other solutions might be preferable outside of this area. This implies that making the best selection usually requires a thorough understanding of the specific requirements and peculiarities of the targeted application domain and environment, so that the energy optimization can be appropriately tailored to these parameters. Therefore, a more transparent on-the-fly mechanism for node reconfigurations between different Pareto-optimal states is required to enhance sensor node reusability in the context of different vertical applications.

VI. CONVERGENCE OF THE SENSOR NETWORK AND DATA ANALYTICS TRENDS

In order to efficiently integrate processes and people in the IoT (as envisioned by the Internet of Everything), connected "Things" will need to share higher-level information with distributed peer entities, as well as with centralized processing units or people for further evaluation and decision making. This transformation from data sharing to information sharing is considered as particularly important in the IoE notion because it will facilitate faster, more intelligent decisions, as well as more effective control of our environment [8]. Similarly, in the field of industrial automation, there is clear movement towards keeping the pace with the rapidly increasing data footprint by a paradigm shift in data acquisition and processing [9].

In parallel with these activities, a significant evolution is taking place in the data analytics domain. In this case, the trend is to evolve from "descriptive analytics" that capture what is happening to "predictive analytics" that describe what is likely to happen. Similarly, a little further down the road is the progress from "diagnostic analytics" that describe why something is happening to "prescriptive analytics" that describe what should happen. Fusion of "hard" data coming from sensors with "soft" data from e.g., social networks is another important trend in this domain, which is already going in the direction of bringing humans into the equation. "Pervasive analytics" (in some cases even referenced as "butler analytics") are envisioning to bring the power of analytics in an ever increasing range of day-to-day applications and make them available to non-experts. The relation between sensor and analytic trends is depicted in Figure 1.

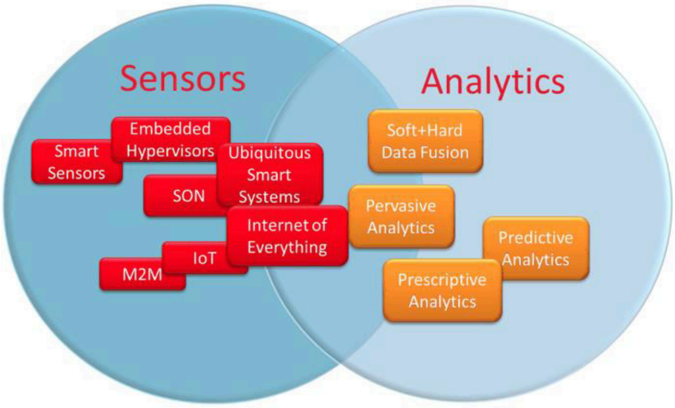


Figure 1. Convergence between sensor and analytic trends

The desired destination in this convergence is a framework of abundant sensor information taping at the "anytime, anyplace […] connectivity for anything" notion of the IoT combined with advanced analytic models that can provide real insight (in the form of human-consumable prediction and recommendation) for any situation and usable by everyone.

However, significant steps need to be taken before this vision is realized. "Analytics" is a very broad and varying field, and while wrapping them in a user friendly package is easy, using them in an irresponsive way without knowledge or respect for possible limitations or model constraints, to be the recipe for disaster [19]. Frameworks that can provide different tradeoffs of accuracy and execution time or easiness to interpret and (even more importantly) can at least make the users aware of model limitations and constraints would be an important driver towards approaching this vision.

VII. SENSOR INFRASTRUCTURE VIRTUALIZATION AS A DRIVER TOWARDS THE FUTURE INTERNET

Achieving a significant progress in the four open challenges identified in the previous sections calls for frameworks that either facilitate innovation or minimize the cost/risk for each of the four pillars identified previously (scalability, cooperation, low energy solutions and convergence of Sensor Network and Data Analytics trends). It is also important to underline that these pillars are not completely autonomous, but are mutually dependent. For example, one of the objectives of cooperation might be low energy operation, while the cooperation process by itself has to be scalable. Therefore, an important constraint is that possible solutions for each challenge are as transparent as possible to the other topics, to avoid setbacks in other fronts. A promising paradigm for addressing challenges in terms of decreasing the cost/risk as well as facilitating innovation in some of the topics identified previously is virtualization. Several types of virtualization can be distinguished, including Virtual Machines and OS Virtualization [20], Sensor Virtualization [21], and Sensor Network Virtualization [22].

While the first two types of virtualization have found their way into mainstream applications and are arguably the driving forces behind the cloud computing paradigm, the other two types are still not so widely used and there is some ambiguity in their definition in related literature. In this paper, we use the term Virtual Sensor (VS) to refer to a software entity that can serve as an aggregation point for multiple sensors, using physical sensor entities and a computational model to combine their measurements. There are several points that need to be elaborated in this definition. The Virtual Sensor can exist either in-field as a thin layer of virtualization software that is executed on physical sensors or it can be a mathematical model for aggregating information residing in a sensor management platform similar to [23]. Moreover, in both cases, the virtual sensor is doing more than interpolating values of physical sensors measuring the same phenomenon, as translation between different types of physical sensors is a far more interesting topic when models for the relations between the underlying phenomena are available. For example, one can estimate car pollution based on a model that combines car counting and weather conditions, while possibly utilizing on also the information from the few available pollution sensors.

Beyond independent Virtual Sensors, Virtual Sensor Networks (VSNs) are emerging as a novel form of collaborative wireless sensor networks [24] that can provide the common layer over which the evolution from connecting "Things" to the efficient interaction of the "Things" with processes and people can be realized. To some extent, VSNs are an evolution of the overlay network principle, which describes a network built on top of another network. The main differentiator of VSNs to overlays is that the latter are typically realized in the application layer only and constitute a temporary solution of true virtualization of the sensor network.

A VSN can be formed by providing logical connectivity among collaborative sensors [22], [25]. Nodes are grouped into different VSNs based on the phenomenon they track (e.g., number of cars vs. NO_2 concentration) or the task they perform (e.g., environmental monitoring vs. traffic control). VSNs are expected to provide the protocol support for formation, usage, adaptation, and maintenance of the subset of sensors collaborating on a specific task(s). Even nodes that do not sense the particular event/phenomenon (directly or indirectly by the notion of Virtual Sensor) could be part of a VSN if they allow sensing nodes to communicate through them. Thus, VSNs make use of intermediate nodes, networks, or other VSNs to deliver messages across VSN members [22].

However, the main goal of VSNs is to enable and promote sensor reusability and facilitate resource efficient, collaborative WSNs. By collaboration, nodes achieve application objectives of different use cases in a more resource efficient way. These networks may also evolve into a dynamically varying subset of sensor nodes (e.g., when a phenomenon develops in the spatial domain, the sensors that can detect it change over time). Similarly, the subset of the users having access to different subsets of the VSN can vary (e.g., the people that have access to the network change with time or specific operations on a sensor network subset are only available to specific groups of people based on their access rights, etc.). Finally, combined with the Virtual Sensor notion described earlier, VSNs can enable the same physical sensor, i.e., an induction loop or an LPR (License Plate Recognition) camera to support its primary application of traffic control, but also secondary applications such as environmental monitoring.

With respect to the challenges identified in previous sections of this paper, different flavors of sensor virtualization can provide answers to different facets of the open problems. For example, the Sensor Virtualization aspect that is based on introducing a thin abstraction middleware is a promising way to address sensor configurability and deployment issues related to network scalability, while it can also reduce the risk/cost of supporting a closed manufacturer solution (identified as an inhibitor for early adoption of related efforts). Moreover, to the extent that the sensors can execute different virtualization middleware versions with different configurations, performing on-the-fly reconfigurations between different Pareto-optimal states of energy/performance or accuracy/execution time will be easier. Finally, the virtual sensor aspect that aggregates sensor measurements to compensate for missing physical sensors of a given type is a significant step towards the vision of converged sensor and analytics in the IoE and the pervasive analytics notions.

On the VSNs side, a Virtual Sensor Network is inherently a collaborative networking paradigm that promotes node reusability in a resource efficient way. Thus, they are a particularly promising base for cooperative communications but also as a way to streamline sensor operations related to the management part of network scalability. Taking this property of VSNs (see first paragraph of Section IV) into account, an updated model of the 3D cooperative methods taxonomy introduced in [13] that also captures the different virtualization aspects described above and is provided in Figure 2.

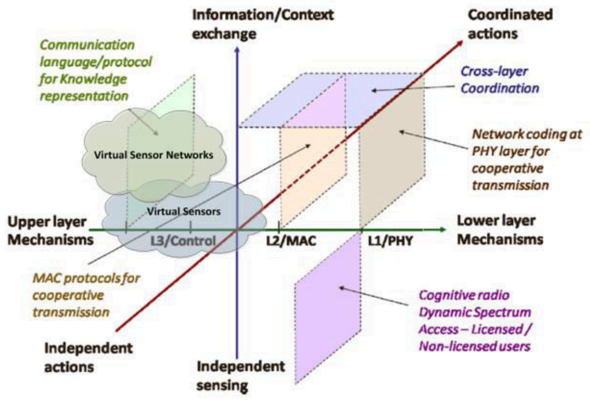


Figure 2. Sensor Infrastructure Virtulization depicted over the various dimensions of cooperative decision making and control.

The figure captures the scope of the cooperation as planes in a 3D space (information exchange, decision and configuration control, and layer mechanisms), where each dimension is associated with a set of enablers and technical areas.

Finally, with respect to the convergence of sensor and analytic trends, it is the most promising but arguably the most challenging of the four pillars. Bringing together in a dependable way an ever increasing number of sensor networks, owned from different stakeholders and forming on top of this infrastructure subsets of sensor nodes based on various criteria (e.g., temporal, spatial, thematic, etc.) is a good basis to generate an abundance of diverse data. Analytics can be applied in this ocean of data to make predictions by aggregating or excluding WSN subsets, fusing the data or the decisions of different subsets (according to the "analytics at the edge" notion [26], etc.). Nevertheless, any model can be misused if its constraints are ignored; making such limitations visible to the user is feasible (and virtualization can help), it is however his final responsibility to adhere to them.

VIII. CONCLUSION

The rapid proliferation in the number of devices connected to the Internet that occurred during the last decade is expected to continue, targeting yet another order of magnitude increase and promising to bring people and processes in the equation. However, in order to realize this paradigm shift, important challenges with respect to scalability, cooperative communications, energy consumption, as well convergence of sensor and analytics trends need to be addressed. In this paper, we have elaborated on the different flavors of Sensor Infrastructure Virtualization as a powerful enabler that can pave the way towards the Future Internet.

REFERENCES

[1] R. Zhang, Z. Zilic, and K. Radecka, "Energy Efficient Software-Based Self-Test for Wireless Sensor Network Nodes", VLSI Test Symposium (VTS), 2005, pp. 186-191.

[2] A. Merentitis, N. Kranitis, A. Paschalis, and D. Gizopoulos, "Low Energy On-Line Self-Test of Embedded Processors in Dependable WSN Nodes", IEEE Transactions on Dependable and Secure Computing, 2011, pp-86,100

[3] National Intelligence Council, "Disruptive Civil Technologies: Six Technologies with Potential Impacts on US Interests out to 2025," Apr. 2008. [Online]. http://www.dni.gov/nic/confreports disruptive tech.html, last access: July 2013.

[4] International Telecommunication Union "The Internet of Things" 2005.

[5] L. Atzori, A. Iera, and G. Morabito, "The internet of things: A survey," Computer Networks, vol. 54, no. 15, 2010, pp. 2787–2805. [Online]. http://www.sciencedirect.com/science/article/pii/S1389128610001568, last access: July 2013.

[6] D. Miorandi, S. Sicari, F. D. Pellegrini, and I. Chlamtac, "Internet of things: Vision, applications and research challenges," Ad Hoc Networks, vol. 10, no. 7, 2012, pp. 1497–1516. [Online]. http://www.sciencedirect.com/science/article/pii/S1570870512000674, last access: July 2013.

[7] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems, 2013, pp. 1645-1660 [Online]. http://www.sciencedirect.com/science/article/pii/S0167739X13000241, last access: July 2013.

[8] D. Evans, "The Internet of Everything: How More Relevant and Valuable Connections Will Change the World", CISCO White Paper [Online]. http://www.cisco.com/web/about/ac79/docs/innov/IoE.pdf, last access: July 2013.

[9] National Instrumnets, Data Acquisition Technology Outlook 2013, [Online]. http://www.ni.com/daq-trends/, last access: July 2013

[10] P. C. Evans and M. Annunziata, "Industrial Internet: Pushing the Boundaries of Minds and Machines", GE White Paper, Nov. 26, 2012, [Online]. http://www.ge.com/docs/chapters/Industrial_Internet.pdf, last access: July 2013.

[11] J. Bradley, J. Barbier, and D. Handler, CISCO White Paper, "Embracing the Internet of Everything To Capture Your Share of $14.4 Trillion", [Online].http://www.cisco.com/web/about/ac79/docs/innov/IoE_Economy.pdf, last access: July 2013.

[12] Future Internet Architecture (FIArch) Group, "Future Internet Design Principles", January 2012, [Online]. http://www.future-internet.eu/uploads/media/FIArch_Design_Principles_V1.0.pdf, last access: July 2013.

[13] G. Koudouridis, et al., "Enablers for Energy-Aware Cooperative Decision and Control in Wireless Networks", Vehicular Technology Conference C2POWER Workshop, May 2011, pp.1-5, Budapest, Hungary.

[14] N. Alonistioti, et al., "Towards Self-Adaptable, Scalable, Dependable and Energy Efficient Networks: The Self-Growing Concept", UBICOMM, 25-30 October 2010, pp. 324-327, Florence, Italy.

[15] E. Patouni, A. Lilis, A. Merentitis, N. Alonistioti, C. Beaujean, D. Bourse, and E. Nicollet, "Protocol Reconfiguration Schemes for Policy-based Equipment Management", Vehicular Technology Conference (VTC), September 2006, pp.1-5, Montréal, Canada.

[16] B. Jacob, R. Lanyon-Hogg, D.K. Nadgir, and A.F. Yassin, "A practical guide to the IBM autonomic computing toolkit" IBM, International Technical Support Organization, 2004.

[17] K. Chatzikokolakis, R. Arapoglou, A. Merentitis, and N. Alonistioti, "Fair Power Control in Cooperative Systems based on Evolutionary Techniques", UBICOMM, September 2012, pp. 111-116 Barcelona, Spain.

[18] G. Anastasi, M. Conti, M. D. Francesco, and A. Passarella, "Energy Conservation in Wireless Sensor Networks: a Survey" Journal of Ad Hoc Networks, Volume 7 Issue 3, May, 2009, pp. 537-568.

[19] N.N. Taleb, "The Black Swan: The Impact of the Highly Improbable", Random House Digital, Inc., 2010, ISBN 978-1400063512, doi=10.1007/s00362-009-0226-8.

[20] S. Crosby and D. Brown. "The Virtualization Reality." Queue 4, n.10, December 2006, pp. 34-41, doi=10.1145/1189276.1189289.

[21] S. Kabadayi, A. Pridgen, and C. Julien, "Virtual sensors: abstracting data from physical sensors", WoWMoM 2006, pp. 587-592, Buffalo-Niagara Falls, NY

[22] H. M. N. D. Bandara, A. P. Jayasumana, and T. H. Illangasekare, "Cluster tree based self organization of virtual sensor networks," GLOBECOM Workshops Nov. 2008, pp.c 1-6, New Orleans USA.

[23] V. Gazis, K. Sasloglou, N. Frangiadakis, P. Kikiras, A. Merentitis, K. Mathioudakis, and G. Mazarakis, "Architectural Blueprints of a Unified Sensing Platform for the Internet of Things" – International Conference on Computer Communications and Networks (ICCCN), 30 July – 2 August, 2013, Nassau, Bahamas.

[24] L. Sarakis, T. Zahariadis, H. Leligou, and M. Dohler, "A framework for service provisioning in virtual sensor networks" EURASIP Journal on Wireless Communications and Networking 2012, pp. 1-19.

[25] N. M. M. K. Chowdhury and R. Boutaba, "A survey of network virtualization," Computer Networks, vol. 54, no 5, Apr. 2010, pp. 862–876.

[26] Distributed Data Mining and Big Data, Intel White Paper [Online]. http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/distributed-data-mining-paper.pdf, last access: July 2013.

An Efficient Cooperative Parking Slot Assignment Solution

Naourez Mejri, Mouna Ayari, Farouk Kamoun
RAMSIS, CRISTAL Lab.
University of Manouba
Manouba, Tunisia
{naourez.mejri; mouna.ayari}@cristal.rnu.tn; frk.kamoun@planet.tn

***Abstract*—Finding a vacant parking place is one of the major concerns of drivers on the road. Hence, good parking management policies are required to efficiently assign solicited parking places to drivers especially in highly solicited urban environments. In this context, we propose and study an efficient semi-centralized parking slot assignment approach with two variants: without/with complete knowledge. In our proposal, each parking lot, in a given urban zone, is monitored by a local authority entity called Parking Coordinator (PC). Its main task is to process received requests from vehicles and offer them accurate parking slots. This selection takes into account the preferences of each vehicle as specified in its request. Through comparing the two variants, we investigate the impact of communication (or lack of it) between these parking coordinators on the good distribution of assigned parking places and the requests rate satisfaction. We study the efficiency of the proposed schemas in various contexts using the mathematical programming solver for linear programming CPLEX and compare them with the centralized approach. Results show, on one hand, that the centralized approach provides the highest rate of request satisfaction. However, as known, this approach heavily suffers from the scalability problem. On the other hand, experiments show that the solution with complete information outperforms the semi-centralized one and its performances matches those obtained in the centralized solution.**

Keywords-coordination; parking slot assignment; vehicles; service-based systems

I. INTRODUCTION

In densely populated urban areas, parking is one of the non negligible causes of congestion and travel delays. In fact, searching for a vacant parking place can become a time consuming and frustrating task for drivers in a hurry. Hence, good parking management policies are required and efficiently assigning solicited parking places to drivers is a priority. In this paper, we propose an new efficient and practical approach to guide drivers to parking. We aim to guarantee drivers satisfaction with the parking assignment and to improve the fairness among parking zones by balancing their occupancy-load. The idea behind our proposal is to use Parking Coordinators (PC) for each lot of parking places. These authorities are responsible for assigning the empty spots based on the vehicles' demands and drivers' preferences. Drivers looking for parking spots send requests over the Vehicular Ad hoc NETwork (VANET). Each request specifies among others a parking place location at a preferred distance from the driver's final destination. The PCs collect the requests of the drivers over a certain time window and assign a free parking slot to each one. We distinguish two variants of our proposed semi-centralized strategy for parking slot assignment: without and with complete knowledge. In the first variant, each PC acts independently and makes its decision without caring about other authorities decisions. However, in the second variant, each PC considers its surrounding PCs possible alternatives.

We compare our proposal with the centralized strategy and we investigate the impact of communication between PCs on the accuracy and efficiency of the parking assignment. We notice that the main contribution of our paper is a scalable localized solution that assigns parking places based on drivers' preferences and ensures a good distribution of available parking places. This task can be fulfilled before vehicles reach the parking zone. Hence, this considerably reduces waiting and search time. The remainder of this paper is organized as follows: Section II discusses related work addressing parking slots assignment. In Section III, we introduce the system model and outline the main assumptions and goals on which our solution is designed. We describe the two variants of our proposal in Section IV and compare their performances with the centralized solution in Section V. Finally, the last section concludes the paper.

II. RELATED WORK

Last years, the parking slot assignment problem has received particular attention in both industrial and academic levels. Different smart applications have been developed in order to assist drivers in their parking search. Proposed solutions can be classified in two big categories: parking solutions with infrastructure assistance and parking solutions relying on estimated/predicted information. In the first category, existing or added facilities gather accurate data about parking occupancy and capacities in order for drivers to find efficiently their parking places. Whereas, in the second category, such privilege no longer exists and the status of the parking places is either predicted or estimated with other methods. [1][2][3][4] and [5] are examples of

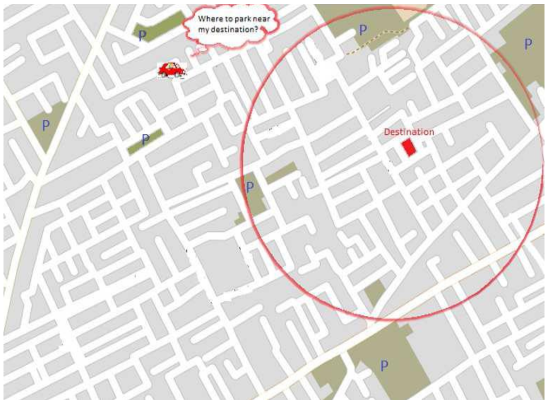


Figure 1. A vehicle looking for a parking within its zone of interest

solutions with infrastructure assistance. In SF-Park [1] and SmartParking [2], each parking spot is equipped with a fixed sensor to determine its occupancy/freeness. The infrastructure then advertises the available spots and manages their reservation. Moreover, a penalty mechanism is proposed in [2] to ensure that vehicles respect their assigned spots. However, these solutions require a large cost in order to adequately monitor the parking spaces even at the level of a downtown area. Hence, for more scalability, some parking garages manage their set of parking places with in/out counters at the entry and exit points to count the number of additional vehicles they can accommodate at any given time.

Such is the case of IrisNet [3] where web cameras are used to monitor individual parking spaces thus allowing users to query the system for vacant spaces on a web front-end. The same concept is used in the SPARK scheme [4]. Road Side Units (RSUs) were installed across a parking lot in a manner they could supervise the whole parking lot. They provide drivers with services such as real-time parking navigation, intelligent anti-theft protection and friendly parking information dissemination. Nevertheless, since such solutions are only valid on closed parking spaces and cannot be applied to curbside parking places, authors in ParkNet [5] proposed reducing the number of infrastructure/sensors required by equipping some special vehicles (such as cabs), instead of the parking places, with ultrasonic sensors to determine and reserve vacant places even in isolated areas of the road. However, the precision of these ultrasonic devices isn't very accurate and the solution concept requires the designated vehicles to permanently monitor the road state to check for parking availability. The solutions presented in [6] and [7] are examples of solutions based on predictability. In [6], although Caliskan et al. use the parking lot information disseminated by the parking automats, they rely more on the inter-vehicle broadcasts to allow the drivers find their preferred free parking lot, whereas in [7], each vehicle which leaves its parking place, becomes its coordinator. After collecting information among interested neighbors, it decides with which vehicle to share the parking coordinates. This process aims to reduce competition between vehicles in search for parking where only the elected vehicle knows the parking place exact location. However, this solution suffers from scalability since this process need to be iterated repeatedly for each freed parking place. Besides, it doesn't address how free parking places are being assigned at the initial process. We note here that other parking assignment solutions exist, however they assume that the information about parking occupancy is already available or that all the parking places are initially vacant and open for competition. They just worry about how to efficiently assign the parking places. An example of these solutions can be found in [8], where a parking slot assignment game was proposed with the vehicles being the rival players. Our proposal belongs to the second category since we aimed for a scalable solution without depending on existing infrastructure to ensure its compliancy with several topologies.

In our proposal, we aim at providing vehicles with parking spots in advance. In fact, in order to gain time, a vehicle can emit its request for a parking place in a specific region before reaching it. By the time it arrives, it will find a spot already reserved to it. Nevertheless, we note here that a vehicle can ignore this in advance request-answer mechanism and specify its desire for a parking place only when it reaches the parking garage (which can be accomplished through an online application).

III. SYSTEM MODEL AND DESIGN GOAL

A. *System Model*

Our network is composed of:

- A set $V = \{v_1, v_2, .. , v_N\}$ of N vehicles. Each vehicle v_i is looking for a free parking spot within its zone of interest ZI_i. The zone of interest ZI_i is defined as a circle of center the point of interest PI_i of the vehicle (i.e., its final destination) and of radius r_i. The radius value can be dynamically adjusted according to the day time, the network congestion, the geographic location and the willingness of the driver to walk once parked. This zone delimits the location of the parking places that can be assigned to the driver. It aims for him to be close enough of his destination and hence do not incur further time loss due to extra walking (see Fig. 1).
- A set $S = \{s_1, s_2, .. , s_M\}$ of M vacant parking slots.
- A set $PC = \{PC_1, PC_2, .. ,PC_L\}$ of L parking coordinators (PC). Each PC_k is a local authority managing a set SPC_k of size M_k. It consists of non overlapping slots of M such as:

$$\sum_{k=1}^{L} M_k = M \qquad (1)$$

These PCs are responsible of allotting the empty parking places under them to the vehicles while optimizing their own social welfare i.e., the cost/the benefit induced by this parking slot assignment.

We assume that all received vehicles requests are collected over a periodic time window and processed periodically in the same time. Moreover, we assume that each PC knows about the locations of the other PCs. Hence, it can determine the requests simultaneously processed by it and them. Such information can be easily obtained through an up-to-date topology map or dynamically by periodic broadcast through the existing infrastructure.

B. Design Goal

Our main goal is to provide drivers with available parking spots in advance. Hence, each vehicle can express its desire for a parking place in a request issued beforehand. This request contains information about the preferences of the driver. It includes its point of interest coordinates and the radius of its zone of interest. Moreover, the driver specifies its current location at the moment of the request preparation. Such information is used when processing requests to give priorities to approaching vehicles. While assigning slots to vehicles, each parking coordinator has to make sure that this assignment meets the following criteria:

- A spot can be assigned only once to a vehicle. It stays unavailable until the parked vehicle leaves it. Each spot offered to a vehicle v_i has to lay within its zone of interest ZI_i.
- Each PC aims to satisfy the maximum number of requests for parking places in the limit of the available ones.

IV. SEMI CENTRALIZED PARKING ASSIGNMENT SOLUTION

In the following section, we present both variants of our proposal and explain how both solutions work.

A. Semi-centralized solution without complete knowledge

Periodically, upon reception of requests, each parking coordinator processes the requests that are relevant to it and discards the rest. A request of a vehicle v_i is relevant for the parking coordinator PC_k if it satisfies (2).

$$\exists s_j \in SPC_k \mid dist(s_j, PI_i) \leq r_i \qquad (2)$$

The measure dist refers here to the Manhattan distance between two points.

Each parking coordinator PC_k periodically updates an association matrix called A_k where $A(i, j)_k$ is set to one if the request of vehicle v_i can be satisfied by the slot s_j (element of the set SPC_k). $A(i, j)_k$ is set to zero otherwise. Then, the parking coordinator computes for each vacant slot s_j, its solicitation factor. This latter displays how many requests the slot can possibly satisfy. This factor is obtained from (3).

$$F(v_i, s_j) = \alpha \frac{dist(v_i, s_j)}{\sum_{k=1}^{M_k} dist(v_i, s_k)} + \beta \frac{dist(PI_i, s_j)}{\sum_{k=1}^{M_k} dist(PI_i, s_k)} + \gamma \frac{\sum_{i=1}^{N_k} A(i,j)_k}{N_k} \qquad (3)$$

where:

- The factors α,β and γ are variables, for which the values are selected according to whether the distance to the parking spot, the distance to the destination or the less solicited slots are to be privileged.
- The distances $dist(v_i,s_j)$ and $dist(PI_i,s_j)$ are respectively the Manhattan distances between the vehicle v_i and the slot s_j and between the vehicle's destination PI_i and the slot s_j . Both distances are normalized by the sum of distances among the whole slots belonging to PC_k.
- N_k is the total number of requests relevant to PC_k.

We stress here that although in real life, the Manhattan distance can be different from the real driving distance; it is not easy to determine the latter in advance unless the vehicle's trajectory is predetermined in advance. Besides, it is obvious that the driver can change his mind anytime depending on traffic/road and weather conditions. Finally, in our simulations, we used a grid model where the Manhattan distance matches well the real distance. Our objective can then be modeled as an integer linear program optimization for each PC_k as shown in (4).

$$\text{minimize} \sum_{i=1}^{N_k} \sum_{j=1}^{M_k} y_{i,j} \, F(v_i, s_j) \qquad (4)$$

subject to

- $dist(PI_i,s_j) \leq R_i$
- $y(i,j) = \begin{cases} 1 & \text{if vehicle } v_i \text{ was assigned to spot } s_j \\ 0 & \text{otherwise} \end{cases}$
- α,β and α ≥0 | α+β+γ=1

As proven in [8], a system optimal solution can be computed for such problem in a (strongly) polynomial time. In this variant, each PC's decision is made independently of the other PCs decisions. This can obviously result in one request being answered more than once by different PCs while other vehicles can be left with no response to their requests. In order to remedy these possible solution defects, we opted in the second variant toward a solution where PCs are aware of each other.

The following section explains how this is accomplished.

B. Semi-centralized solution with complete knowledge

In this variant, the parking coordinators will share with each other information about the availability/occupancy of their parking places. Moreover, each PC will keep track of the set of relevant requests for the other PCs. On that account, each PC_k fills a request-coordinator correspondence matrix $Corr_k$ with L columns (referring to the total number of PCs and r rows (referring to the total number of requests being processed at the current period) as shown in (5).

$$Corr_x(i,j) = \begin{cases} 1, & \text{if } PC_j \in ZI_i \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

From the obtained matrix $Corr_k$, PC_k can extract two indicators:

- A coverage indicator Cov_i for each request req_i representing the number of coordinators within the zone of interest of vehicle v_i and eligible to answer its request, it is computed from (6).

$$\forall i \in \{1,..,r\},\ Cov_i = \sum_{j=1}^{L} Corr_k(i,j) \quad (6)$$

- A solicitation ratio Sol_j of each parking place coordinator PC_j indicating the number of relevant requests for the parking coordinator PC_j as shown in (7).

$$\forall j \in \{1,..,L\},\ Sol_j = \sum_{i=1}^{r} Corr_k(i,j) \quad (7)$$

Each PC_k will then compute, for each request req_i, the cost of positively answering that request based on (8).

$$U_k(req_i) = \alpha \frac{dist(v_i, PC_k)}{\sum_{j/Corr_k(i,j)=1} dist(v_i, PC_j)} + \beta \frac{dist(PC_k, PI_i)}{\sum_{j/Corr_k(i,j)=1} dist(PC_j, PI_i)} + \gamma RFC_k \quad (8)$$

where $RFC_k = Empty_k/Sol_k$ is a ratio indicating the capacity of the parking coordinator PC_k to fulfill the received requests. $Empty_k$ refers to the number of vacant parking places for PC_k. A ratio $RFC_k > 1$ indicates that there are enough vacant places to answer positively all the received requests.

For each request req_i, if the cost satisfies the condition stated in (9), then PC_k offers a parking slot to the vehicle v_i. Otherwise, it discards the request and processes the next one. This process goes on until all the relevant requests are satisfied or there are no more vacant parking slots.

$$U_k(req_i) = \underset{x \in L, PC_x \in ZI_i}{\operatorname{argmin}} \{U_x(req_i)\} \quad (9)$$

Once the PCs decide on the vehicles to which the spots are going to be assigned, a reservation mechanism will start between them and the vehicles in question. At the end of it, each spot will be reserved for a single vehicle. If by chance, the latter decline the offer, the spot will still be considered vacant and reassigned to the next prior vehicle. The reservation mechanism is beyond the scope of this paper and will be detailed in future work.

Also, we stress here that we assume all vehicles in the zone are using the designed scheme. Problems might occur if other vehicles in the area decide to act on their own and might take a spot already reserved to another vehicle. To reduce the magnitude of such problem, a penalty mechanism like the one presented in [2] can be adopted. If the infringing vehicle still ignores the penalty fee and occupies the spot, the PC will have to emergently assign another spot to the vehicle which originally reserved it. It has to forward it to the closest free spot among its own, if there are any available. In the negative case, it will check among its neighbors to see if any of them can offer it a free spot. Then, it forwards the new location of the parking spot to the vehicle in question.

V. PERFORMANCE EVALUATION

In this section, we present the simulation environment and describe the used scenario. Then, we analyze the obtained results of our variants. We note here, that the two variants of our solution were compared with the centralized model of reference. In this model, a single centralized authority assigns slots to vehicles. It aims to minimize the total network cost while satisfying the maximum number of requests with consideration of the parking places capacity. It abides to the same constraints listed in the previous section.

A. Simulation Environment

Our parking assignment problem can be seen as a variant of the reference task assignment problem. Thus, we ran our simulations by using the linear programming tool CPLEX [10]. The data files with the constraints were exported from MATLAB [11]. We chose our network topology as a grid of 500*500m on which respectively 100,150 and 200 vehicles are looking for parking places. The total number of available slots was fixed to 200 slots. We note here that we were unable to increase the number of vehicles and slots because of MATLAB and CPLEX computation limitations. However, our choices are realistic and match vehicles' density in some countries with a density varying from 20 vehicles per km of road (in countries were vehicles are affluent transport means) to 10 vehicles per km of road (in less crowded countries). The reader can refer to [9] for more details about the classification of countries by their vehicles' density per km of road. The positions of vehicles were randomly generated over the grid by using SUMO [10] to have realistic dispersion of vehicles over the network. The vehicles randomly aim for one of the 4 interest points. These interest points were dispersed randomly each in one of the network quarters which we refer to as regions. Hence, the number of considered regions is $nr = 4$.

Each region is managed by a single parking coordinator, which manages 50 of the total 200 slots.

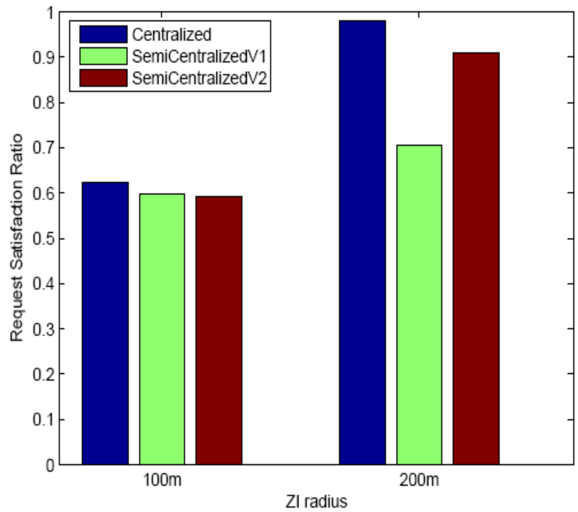


Figure 2. Impact of the variation of ZI radius on the Request Satisfaction Ratio

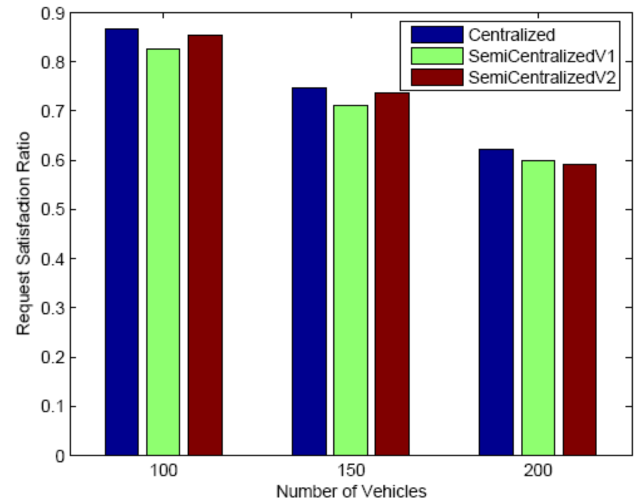


Figure 3. Impact of the vehicles' density on the request satisfaction ratio

B. *Evaluation parameters*

In order to evaluate the performance of our two proposals, we used the following parameters as performance indicators:

- The request satisfaction ratio: it is the quotient of the total answered requests divided by the total number of received requests.
- The fairness index: it indicates whether the slots are fairly shared inside the network regions among the vehicles. It is equivalent to Jain's Fairness Index introduced in [12] and is obtained from (10).

$$\text{FairnessIndex} = \left(\sum_{i=1}^{nr} x_i\right)^2 / \left(nr * \sum_{i=1}^{nr} x_i^2\right) \quad (10)$$

where nr is the network regions, and $x_i = nbReqAnsReg_i / nbReqRecReg_i$ is the fraction of answered requests with slots in region i over the total number of requests initially aiming for a point of interest inside the region i.

- The variance: it measures how far a set of numbers is spread out. It is computed by the formula in (11), where the expression of the variable x_i is the same as defined in the fairness index.

$$\text{Variance} = \left[\sum_{i=1}^{nr} x_i^2 - \left(\sum_{i=1}^{nr} x_i\right)^2\right] / nr \quad (11)$$

C. *Simulation Results*

The values given in the following section are the average values over multiple simulations ran for each point with a randomly generated position of both the slots and the point of interests. The vehicles' choice of the interest point to aim for is also randomly varied in each of the simulations.

1) *Impact of varying the radius of the zone of interest:*

First, we investigated the impact of varying the radius of the zone of interest of vehicles. We varied the radius for the same scenario consisting of 200 vehicles requesting slots in a network initially containing 200 vacant slots. We can see from Fig. 2 that increasing the radius from 100 meters to 200 meters increases the request satisfaction ratio. In fact, it increases the number of slots inside the zone of interest and hence more requests can be satisfied. We remark also that when the radius is of 100 meters, results given by both solutions are pretty similar. However, when increasing the radius to 200 meters, the second variant becomes very efficient and its satisfaction ratio is very closer to the centralized solution with almost 90% of the requests satisfied. In opposition, the first variant has a lower percentage of only 70 %. This difference between the two approaches is due to the fact that PCs act independently in the first variant. This results in having the same vehicle receiving several slots' offers while other vehicles' requests are left unanswered. Each vehicle getting more than one proposal will eventually choose the closest slot to it among them and reject the other offers. Thus, the network will end up with unassigned slots, which could have satisfied other vehicles' requests. In the second variant, the problem is solved since a parking coordinator would not offer a slot to a vehicle if it knows that another parking coordinator has a better offer. Henceforth, the request satisfaction ratio is higher for the second variant.

 ISBN: 978-1-61208-289-9

TABLE I. IMPACT OF VARYING THE RADIUS OF THE ZONE OF INTEREST

ZI Radius	*100m*			*200m*		
Solution	*Centralized*	*SemiCentralizedV1*	*SemiCentralizedV2*	*Centralized*	*SemiCentralizedV1*	*SemiCentralizedV2*
Fairness Index	0.977	0.973	0.838	0.984	0.977	0.974
Variance	0.014	0.0154	0.088	0.023	0.017	0.029

TABLE II. IMPACT OF VARYING THE VALUES OF THE FACTORS Α,Β AND Γ

α	1	0	0	0.5	0.25	0
β	0	1	0	0.5	0.25	0.5
γ	0	0	1	0	0.5	0.5
Request Satisfaction Ratio	0.7362	0.7364	0.7358	0.7369	0.7369	0.7364
Fairness Index	0.8628	0.8626	0.8632	0.8625	0.8625	0.8626
Variance	0.1045	0.1048	0.1036	0.1050	0.1050	0.1048

Increasing the radius of the zone of interest, increases the conflict between the parking coordinators and thus the strength of the second variant results over the first one becomes clearer.

Also, as shown in Table I, we remark that all three solutions give a fairness index values close to 1. This indicates a fair distribution of the resources (here the parking slots) among the vehicles. The low values of the variance close to zero, support the same conclusion. In fact, in the three solutions, the authorities (whether central or local) try to optimize the social welfare while ensuring a good distribution of the requests among the available slots with respect to the constraints.

2) Impact of varying the number of vehicles:

Next, we studied the impact of varying the number of vehicles on their requests' satisfaction ratio. For the same number of slots, and a radius of 100 meters, we varied the number of vehicles between 100, 150 and 200 vehicles to simulate respectively low, medium and high density. The obtained results are shown in Fig. 3. Obviously, a better request satisfaction ratio is obtained when the density is lower (the case for all solutions). In fact, when the number of vehicles is lower than the available vacant spots, the authorities are capable of satisfying most of the received requests that meet the constraints stated above. We also found that for the three densities, the request satisfaction ratio for the second variant of the semi centralized approach are slightly better than those of the first variant and hence tend toward the reference solution, i.e., the centralized one. This highlights the positive effect of exchanging information between the parking coordinators to make their decisions more efficient and avoid assigning several slots to the same vehicle while leaving others without any.

3) Impact of varying the values of the factors α, β and γ:

Finally, as described in Table II, we measured our performance parameters for different values of α, β and γ. The obtained results are almost similar to each other for the different values. This is probably due to two main factors:

- First, the value of the function as described in (1) is obtained by normalizing all the values, which highly reduces the differences between each fraction.
- Moreover, the size of our grid is not very big to show big difference between the distances between vehicles and the slots from one hand and the distance between the vehicles' point of interest and the slots from the other hand.

However, as we stated above, our choices were limited by capacities of MATLAB and CPLEX and chosen in order to have a realistic scenario as much as possible. A bigger scenario will probably show more clearly the impact of varying these factors.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a solution for parking slot assignment. Two variants of the solution were studied. In the first one, the parking coordinators decided independently of their surroundings. In the second variation, these PCs exchanged information with their neighboring ones and tuned their decisions accordingly. Simulation results showed that the second variant outperformed the first one especially when the number of conflicting requests between PCs increased. The results obtained for the second proposal were very close to the reference centralized solution while being more scalable than it. Preliminary results proved that cooperation between parking coordinators is highly recommended and its impact will be further studied in future work. Moreover, we still need to carry out further experiments to fine tune the values of α, β and γ parameters in order to improve the performance of our scheme. Furthermore, in the presented results only one single period (i.e., the time window during which PCs process received requests) was considered. However, the length of such period can highly affect the amount of requests processed. A short period will incur unnecessary load on the PCs whereas a relatively long period can affect the efficiency of the proposal since some requests might become obsolete by the time they are answered. Hence, the time window duration needs extensive simulation as to determine its best value. We note also that, in the case of the first proposed variant, the filling of the matrices would probably incur a high load on the network. Thus, we investigate to study the overhead induced by the proposed mechanisms. Finally, we plan to study the impact of learning mechanisms on the enhancement of our solution.

REFERENCES

[1] http://www.sfpark.org, [retrieved : July, 2013].

[2] G. Yan, W. Yang, D. B. Rawat, and S. Olariu, "Smartparking: a secure and intelligent parking system using notice", In proceedings of the international IEEE Conference on intelligent transportation systems. Beijing, China, October 2008, pp. 569-574.

[3] J. Campbell, P. B. Gibbons, and S. Nathand, "Irisnet: an internet-scale architecture for multimedia sensors", 13th annual ACM international conference on Multimedia(Multimedia'05), Singapore, 2005, pp. 81-88.

M. Caliskan, D. Graupner, and M. Mauve, "Decentralized discovery of free parking places", 3rd ACM International Workshop on Vehicular Ad Hoc Networks (VANET'06), Los Angeles, CA, USA, September 2006, pp. 30-39.

[4] S. Mathur and al., "Parknet: drive-by sensing of road-side parking statistics",8th Annual International Conference on Mobile Systems, Applications and Services (MobiSys 2010), San Francisco, USA, June 2010, pp. 123-136, doi:10.1145/1814433.1814448.

[5] T. Delot, N. Cenerario, S. Ilarri, and S. Lecomte, "A cooperative reservation protocol for parking spaces in vehicular ad hoc networks", 6th International Conference on Mobile Technology, Applications and Systems, ACM Press Library, Nice, France, September 2009, pp. 1-8, doi:10.1145/1710035.1710065.

[6] R. Lu, X. Lin, H. Zhu, and X. Shen, "Spark: a new vanet-based smart parking scheme for large parking lots", 28th Conference on Computer Communications (Infocom 2009), Rio de Janeiro, Brazil April 2009, pp. 1413-1421.

[7] D. Ayala, O. Wolfson, B. Xu, and B. Dasgupta, "Parking slot assignment games", 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACMGIS 2011), Chicago, USA, November 2011, pp. 299-308, doi:10.1145/2093973.2094014.

[8] http://data.worldbank.org/indicator/IS.VEH.ROAD.K1, [retrieved : July, 2013].

[9] http://sumo.sourceforge.net/ , [retrieved : July, 2013].

[10] www.ampl.com/CPLEX, [retrieved : July, 2013].

[11] www.mathworks.com/products/matlab/, [retrieved : July, 2013].

[12] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared computer systems", DEC Research Report TR-301, September 1984.

 ISBN: 978-1-61208-289-9

Innovations in User Interfaces for Pervasive Computing

Javier Oliver
Dept. of Computer Engineering
University of Deusto
Bilbao, Spain
javier.oliver@deusto.es

Begoña García
Life Research Group
DeustoTech
Bilbao, Spain
mbgarciazapi@deusto.es

Abstract— Pervasive computing needs efficient interaction techniques to reduce the burden of the user, and to provide more transparent computing environments. This can be achieved by taking into account the ergonomics, physical location and cognitive load involved in using the system. To improve some of these issues, one can try to innovate existing interaction techniques and in this paper, we mention two improvements on the traditional manipulation of small touch sensitive screens. One of these improvements is back of device interaction, in which the screen is not occluded by the finger of the user. Instead, the screen cursor is moved by touching on the back of the device. The other improvement that is explained is the use of magnetic field detectors to increase the active area of manipulation of small screens. Another approach to obtain more fluid and transparent interfaces is to use gesture (i.e., human body movements) as input device. In this paper, we also give an overview of some of the characteristics of the use of body movements as input device, including eye movement, facial expression and other movements.

Keywords-pervasive computing interfaces; gesture input

I. INTRODUCTION

Mature technologies reach a level of design efficiency that allows for complex devices being used in a very natural and transparent way. After more than a century of evolution, one might argue that automobiles have reached that state. Due to their flexibility and multipurpose nature, computers can be regarded as more complex to operate than cars, but Mark Weiser, in his frequently cited article [1] envisioned that computers of the XXIst century would be used with the same ease and naturalness as cars today. Although some aspects of this ubiquitous computing vision have been fulfilled, computers, in general, are still hard to use, obtrusive devices that tend to turn the attention of the user away from the real task.

If computers are really going to step back into the background and become unobtrusive aids for users to carry out their tasks, user interfaces have to be rethought, and not only existing technologies have to be combined in more creative ways, but also new technologies have to be developed [1]. Ubiquitous computing devices have to be physically and cognitively available [2]. Physical availability includes issues such as location of the device, and ergonomic design, whereas cognitive availability refers to the mental effort involved in using the equipment. Devices that are available in both senses disappear from the user consciousness and cease to be in the way of the user's aims and objectives, which is precisely the goal of ubiquitous computing [2].

In this broad sense, availability can be provided by making interaction more intuitive, natural and effortless. In the following sections, two approaches that can be taken will be discussed: the improvement of the interaction with small screens and the use of gesture as a way to achieve a more natural and less obtrusive interaction.

In Section II, the problem of manipulating small touch screens is described, and two possible solutions are presented. The first solution consists in moving the touch sensitive surface to the back of the device, and the second involves the use of a magnetic field detector. In Section III, the use of gesture as input device for pervasive computing systems is explored, and techniques such as eye movement and facial expression are described. Finally, in Section IV, conclusions are drawn.

II. THE *FAT FINGER* PROBLEM

Computer artifacts have been reducing their size in the last few years to the point where the limiting factor in their dimensions is not the battery size or the electronic circuit miniaturization, but the size of the input and output devices needed to control the computer [3]. A screen should be big enough to be comfortably read at normal viewing distance, and buttons –physical or virtual- should be big enough for a regular sized finger to press them. Even though the size of input devices keeps shrinking over time, the size of the average finger remains constant, hence the title of this section: the fat finger problem [4].

In order to increase the availability [2] of user interfaces, the fat finger problem has to be addressed. According to Fitts' Law [5], the smaller the target, the longer the interaction time. So, whenever a button is below its optimal size, we are making interaction unnecessarily slow. But, in the case of very small screens, the finger can cover most of it, making touch interaction almost unfeasible. Several solutions have been proposed for this situation. In the case of screens over one inch in diagonal, typical interfaces try to solve screen occlusion by offering the user an offset cursor, and showing a reproduction of the occluded area in a free part of the screen. But this approach does not work for screens below one inch in diagonal in which it has been proposed to move the touch sensitive surface to the back of the device, thus leaving the screen completely visible, and

 ISBN: 978-1-61208-289-9

showing the position of the finger by means of a small cursor [6].

Another solution for the interaction with small devices is the use of a magnetic field detector, which can be located behind the screen, thus leaving the screen visible. This type of device is capable of quite accurate positioning at a very moderate price. In a particular study, a screen of 1.5 inches in diagonal was used and the magnetometer offered an angular accuracy of about two degrees for the cost of less than five US dollars [7]. Users had to wear a small magnet in the index finger, which gave a practical range of about 10 cm. from the screen, offering a total active area of nearly 300 cm^2. With this technology, the original area of the screen was increased by a factor of more than 50, and the occlusion problem was solved.

III. GESTURE AS INPUT DEVICE

Pervasive computing needs user interfaces that do not make assumptions about what input and output devices are available for the user. The trend is to move from explicit use of traditional input devices, to more implicit forms of interaction, such as speech or gestures [8]. This implies that users will increasingly address computers in the same fashion as they address other people, taking advantage of their experience in human to human communication, in what has been considered an anthropomorphic approach. Currently, there are technologies that can input user movements into the computer, such as gloves and suits, but they are too cumbersome, expensive, difficult to calibrate and disruptive of the user's flow of action. New interfaces based on movement should be much more flexible in order to fully support the ubiquitous computing model.

Human movement is complex, involving the skeletal, muscular and neurological systems. The nature of the different types of movements depends on the class of muscle fibers involved, the shape and orientation of bones and joints, etc. Depending on the distance covered in these movements they can be classified as micro-movements (short distance) and macro-movements (longer distances) [8], and both types can be used in explicit and implicit interaction with ubiquitous systems.

The technologies used to capture body movements can be classified in two main groups:

- Wearable: the user carries different devices to monitor the body movements.
- Indirect: a variety of sensors can detect body movements at a certain distance.

Wearable devices can make use of accelerometers that detect movements in two or three dimensions. More innovative options include monitoring the electrical activity associated with the movement of the hand and video cameras that can track finger movements from a wrist mount [9].

A. Eye Movement

One of the most interesting micro-movements for interaction with computers is eye movement. Each eye is controlled by a group of six muscles that produce very fine and precise movements. Monitoring these movements is a promising area of research because they can be used to find out about the point of interest of the user in an unobtrusive way. In order to understand how eye movements reflect underlying cognitive processes, protocol analysis can be used. During the execution of a task, eye movements can be recorded. These recordings are the protocols, that can then be analyzed and can be used to understand user behavior in basic interaction tasks, to study the processing of information in users, or to predict the goals of users in real time interfaces [10].

The human retina has a very small area of high visual acuity, called the *macula*. This region, of about 1 mm^2, is near the back pole of the retina and the eyes are in constant movement, so that the point of interest always forms its image in the macula. These constant movements are called saccadic, and they are a good source of information to analyze cognitive actions such as reading or searching [8]. The monitoring of eye movements is a good technique to carry out certain input tasks in a natural and hands free manner. For example, users can look at an object on the screen and blink to select it. An interesting application of eye monitoring is the implementation of automatic scrolling. Normally, it is necessary to displace a rectangle on the screen to perform a scroll, but it is possible to do it by means of an eye gaze [11].

B. Facial Expression

The human face has a high number of muscles that are used to show a wide range of emotions. Traditionally, it has been thought that humans can show six basic types or emotion: fear, anger, disgust, sadness, surprise and happiness, and that these emotions were, in fact, an universal language, being understood worldwide. However, more recent research is showing that classification of emotions is not that straightforward, and that the expressions of emotions is probably culturally dependent [12]. If these findings are confirmed, it would mean that user interfaces based on facial expression would have to take into account cultural differences, and it would also mean that this type of interfaces would me more complex than initially thought. Computer systems are able to recognize a range of emotions, but recognition rates can be as low as 64% in some studies, so the technology is not mature yet to be used as the only type of input, and it should be avoided in critical applications [12]. It is possible, however, to design a multimedia application or videogame that modulates its content depending on the emotion that the user is showing in his face [8]. Apart from inferring the emotion felt by the user, facial configuration and movements have also been used to move a cursor on the screen (a technique mainly aimed at disabled people) or to control different types of programs such as audio or graphics applications [13].

C. Other Movements

Although until now facial expression has taken most of the research community's attention, arm and hand movements provide abundant information that enrich verbal communication, and different authors agree that emotion detection should be multimodal, using input from different body movements [14]. The different movements described so far can widely vary depending on the type of user, and they won't be the same, for example in the case of very young children, adult users or elderly people. The movements in the first group will typically be limited by the lack of neurological development, and in the last one, by the effects of age in the neuromuscular system. What seems clear is that taking into account the limitations of extreme age groups when designing user interfaces based on body movements, can result in better user interfaces for all age groups. Also, it seems probable that single design user interfaces will not adapt well to all situations, and customizable user interfaces will be needed [8].

D. *Examples of Applications*

The number of user applications that can take advantage of gesture for input is potentially endless. For example, a gesture-based interaction photo album (GIA) has been developed that uses gestures as the main source of input [15]. Y. K. Jin et al. started by benchmarking the existing album software for PCs and developing a set of user requirements. The system follows a physical photo album metaphor, and each album in GIA is similar to a folder in a file system. The opening screen shows a shelf with a number of albums on it. The thickness of each album is proportional to the number of pictures that it contains. GIA uses a gesture based interface based on a touch screen. For example, to turn pages, one only has to make a horizontal stroke on the screen. Depending on the length and speed of the gesture, the user can turn one or more pages with the same stroke. To avoid unwanted modifications of data, there is an edit mode, in which data can be modified, and a safe view mode, in which no data can be modified. The aim of the design of GIA was to appeal both to beginners and experts, and it uses two sets of gestures: symbols and characters. Users can choose an album by pretending to pick it up from the shelf. They can also modify the orientation and size of a picture by means of gestures. To define the sets of gestures, the authors analyzed different user groups, including children and elderly people, and they studied their interaction styles. The gesture sets were tested using expert evaluations, usability tests and focus group interviews. The expert evaluations were designed to check four dimensions: intuitive, efficient, fun and innovative. The first two try to measure the practical side of the gesture interface, and the last two deal more with the subjective and emotional side of it. After having a working prototype and before the final usability test, the system was shown at an exhibition where visitors were able to use it. The interaction techniques proved to be easy to remember, and after a first use, visitors had no problems manipulating the interface. In the final usability test, two different designs were compared: a traditional point and click interface based on menus and buttons and the gesture based interface. According to the authors, the gesture interface was, in general, preferred.

Looking over the shoulders to get a password or some other sensitive information is a problem difficult to solve. If a traditional input device such as mouse, keyboard or touch screen is used, a casual observer can obtain the secret information. In general, methods that make shoulder surfing more difficult are also an inconvenience for the user. A system called EyePassword, based on gaze tracking, can virtually eliminate all shoulder surfing attacks while maintaining a reasonable ease of input [16]. In gaze based password input, instead of typing on a keyboard or touch screen, the user looks at each of the desired characters one after the other. This technique can be used not only for alphanumeric passwords, but also for graphical passwords (i.e., different trajectories among the points in a matrix). The area associated with each character has to be large enough to avoid unwanted selection of characters. The most important factor to determine the ideal size of a character on the screen is the resolution of the eye tracker. For example, if a tracking device has a resolution of one degree of visual angle, this means that, for a normal viewing distance of 50 cm., the input of the eye tracker will have an uncertainty radius of about 33 pixels and each character should be at least 66 pixels wide. In EyePassword, a target size of 84 pixels was selected, with a spacing between characters of 12 pixels to reduce the possibility of erroneously selecting one character for another. In order to input the character, the user can fix the gaze of the target for a short period of time, or use a special trigger key like *return* or *space bar*. The fixation method is preferred because it is less error prone, and it is more secure, given that less information is available for the casual observer. Four different technique were implemented and evaluated:

1. Traditional keyboard input.
2. Gaze input combined with a trigger key.
3. Gaze input combined with a certain fixation time, and a qwerty layout of the characters.
4. Gaze input combined with a certain fixation time, and an alphabetic layout of the characters.

Typical passwords were used in the test, with 8 to 9 characters, including uppercase, lowercase, symbols and numbers. The error rate was significantly higher in the gaze plus trigger key combination, and although the input times were higher for gaze based input than for the keyboard, more than 80% of the subjects confirmed that they would prefer to use a gaze based method instead of a keyboard when entering a password in a public place.

IV. CONCLUSION

Pervasive computing needs interaction techniques that do not disrupt the workflow of the user. Computer devices should be available [2] both in the sense of physically available (i.e., ergonomics, location) and cognitively available (i.e., the mental effort required to use the device is minimized).

This broad sense availability can be reached by improving existing interaction techniques or by exploring new ones. In this paper, we have presented improvements for the classical interaction technique of manipulating small touch sensitive screens, such as back of device interaction and the use of magnetic field detectors. We have also

 ISBN: 978-1-61208-289-9

presented some of the recent techniques that use body movements and gestures as input for the computer system, including eye movements, facial expressions, and arm or hand movements.

REFERENCES

[1] M. Weiser, “The computer for the 21st Century. Scientific American,” Vol. 265, Nº 3, 1991, pp. 94-104.

[2] V. Waller and R. B. Johnston, “Making ubiquitous computing available,” Communications of the ACM, Vol. 52, Nº 10, 2009, pp. 127-130.

[3] T. Ni and P. Baudisch, “Dissapearing mobile devices,” 22nd Annual Symposium on User Interface Software and Technology, October 4-7, Victoria, British Columbia, Canada, 2009, pp 101-110.

[4] K. A. Siek, Y. Rogers, and K. H. Connelly, “Fat finger worries: how older and younger users physically interact with PDAs,” In Proc. INTERACT’05, 2005, pp. 267-280.

[5] T. Anthony, “Finger-friendly design: ideal mobile touch screen target sizes. Smashing Magazine,” Available at: http://uxdesign .smashingmagazine.com/2012/02/21/finger-friendly-design-ideal-mobile-touchscreen-target-sizes/, 2012 [retrieved: july, 2013].

[6] P. Baudisch and G. Chu, “Back-of-device interaction allows creating very small touch devices,” In Proc CHI 2009, pp. 1923-1932.

[7] C. Harrison and S. E. Hudson, “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.

[8] J. H. Abawajy, “Human-computer interaction in ubiquitous computing environments,” International Journal of Pervasive Computing and Communications, vol. 5, Nº 1, Special issue: Advances in pervasive computing, 2009, pp. 31-77.

[9] A. Vardy, J. A. Robinson, and L. T. Cheng, “The WristCam as input device,” San Francisco, CA, Proceedings of the 3rd International Symposium on Wearable Computers, 1999, pp. 199-202.

[10] D. D. Salvucci and J. R. Anderson, “Automated eye movement protocol analysis,” Human-Computer Interaction, Vol. 16 Nº 1, 2001, pp. 38-49.

[11] M. Kumar, T. Garfinkel, D. Boneh, and T. Winograd, “Reducing Shoulder-surfing by Using Gaze-based Password Entry,” Symposium On Usable Privacy and Security (SOUPS), July 18-20, Pittsburgh, PA, USA, 2007, pp. 13-19.

[12] M. Pantic and L. J. M. Rothkrantz, “Toward an affect sensitive multimodal human computer interaction,” Proceedings of the IEEE, Vol. 91 Nº 9, 2003, pp. 1370-90.

[13] G. C. DeSilva, M. J. Lyons, and N. Tetsutani, “Vision based acquisition of mouth actions for human-computer interaction,” Auckland, Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence, 2004, pp. 959-60.

[14] H. Gunes and M. Piccardi, “Automatic visual recognition of face and body action units,” Proceedings of the 3rd International Conference on Information Technology and Applications, 2005, pp. 668-73.

[15] Y. K. Jin et al. “GIA: design of a gesture based interaction photo album,” Personal and Ubiquitous Computing, Vol. 8 Nº 3/4, 2004, pp.227-233 [retrieved: july, 2013].

[16] M. Kumar, A. Paepcke, and T. Winograd, “EyePoint: practical pointing and selection using gaze and keyboard,” San Jose, CA., Proceedings of the CHI: Conference on Human Factors in Computing Systems, 2007, pp.421-430.

Mobile Ubiquitous System Adapted to Locate Disabled and Old People Living in Rural Areas

Raul Igual, Inmaculada Plaza, Carlos Medrano
EduQTech R&D&i group
EUPT– University of Zaragoza
Teruel, Spain
{rigual, inmap, ctmedra}@unizar.es

***Abstract*—People who depend on others to carry out daily activities face additional issues in rural areas. The design of assistive mobile technology adapted to the specific characteristics of this environment is a promising line of work. As an example, this paper presents a ubiquitous smartphone-based system to locate dependent people living in rural areas. A preliminary prototype has already been developed, which is aimed at both dependent users and their caregivers. Additionally, a 6-month pilot test has been conducted under real-life conditions, showing positive results.**

Keywords- ubiquitous systems; location; dependent people; rural areas; smartphones

I. INTRODUCTION

The number of people who depend on others to carry out daily activities is increasing in developed societies. In this sense, their quality of life can be affected by a range of factors including type of disability, time of onset, social support structures, life history and experiences, socio-economic factors, and each individual's level of psychological resilience [1].

The proportion of dependent people is higher in rural areas, which may be explained by the prevalence of older residents in remote environments [2]. In this context, the *Rural Assistance Centre* [3] argues that dependent people who live in rural areas face additional challenges. These include, among many others, poor and badly maintained roads, difficulties with medical appointments, extreme climatic conditions for parts of the year, difficulties in communication, lack of access to medical care or lack of educational opportunities.

In view of the significant number of rural dependent residents and the difficulties to which they are subjected, the use of appropriate assistive technologies is a major priority in this kind of region. In this sense, systems based on ubiquitous computing can help alleviate these problems.

Specifically, this work is focused on location systems. Lots of location applications to assist dependent people are available [4], although most of them were designed with the assumption that their operation is confined to urban areas where there are no communication barriers. However, when used in rural environments, they may experience from a mild to severe loss in performance, resulting in an abnormal operation of the system. If this situation persists, it could probably lead to device rejection by both the users and their caregivers. Specifically, there are three main factors that can compromise the operation of the ubiquitous locators in rural environments:

a) Few technical skills of the rural residents: Calvert et al. [5] identified a significant rural-urban digital divide in the day-to-day use of common technological devices, which was also confirmed by the study of Gandiya et al. [6]. They argued that there may be several potential barriers to the widespread implementation of communication technology to a general clinic population, particularly in remote and rural areas.

b) Internet access: Large differences exist in Internet access between rural and urban areas, which results in lower quality services at higher costs among the rural population. In 2007, broadband (DSL) coverage reached an average of 98% of the population in urban areas, while coverage of rural areas was limited to just 70% of the European Union rural population [7].

c) Mobile coverage: 3G mobile networks and, to a lesser extent, the brand new 4G networks, have a high level of penetration in urban areas, medium and small villages and the major transport routes of the European Union. However, coverage is limited in rural environments: blind areas (without mobile coverage) are common, many regions have only 2G coverage and 3G is exceptionally available.

To deal with all these additional issues, this paper presents a ubiquitous smartphone-based system to locate dependent people, even in remote areas. The system is adapted to be used by low skilled users and to operate in regions with frequent blind areas. Some strategies to infer the position of the user and to detect potentially dangerous situations are currently being developed.

The rest of this paper is organized as follows: Section 2 presents some related work, Section 3 shows the general structure of the system, Section 4 describes it in detail, Section 5 presents some preliminary results, and finally, Section 6 draws some initial conclusions.

II. RELATED WORK

Wearable location systems are a well-known technological support as demonstrated by the high number of existing devices [4]. As an example, we can cite either some research studies [8-10] or commercial products [11, 12]. However, among all these systems, we could not find any location assistive support specially adapted to operate in rural areas with communication restrictions.

If we change the search criteria to include not only location systems but also any other assistive technologies, we find that the work proposed by Havasi et al. [13] is one of the few studies that incorporate an adapted version of the prototype for rural areas. They present a PDA-centred medical sensor system that is suitable for the ambient assisting living homecare scenario. The PDA has been equipped with a network layer that can provide Internet access even in rural areas. Similarly, Sasaki et al. [14] propose a Life Support Network for elderly people living in a rural area. It consists of a type of Intranet, which incorporates a safety confirmation system, a remote healthcare system, an emergency information system and other useful daily network services in the closed life area. This work is just a proposal and only two of the sub-systems have been implemented.

Apart from these few works, we could not find many more assistive devices specially adapted to rural environments, which is a clear sign that this is an emerging line of work. Another proof of this is that there are some doctoral theses being conducted on this topic, such as the ones included within the TOPs project [15].

In this sense, the lack of rural-adapted assistive devices may be explained by the traditional reluctance of rural residents towards the use of technological tools. But as the digital generations of today reach old age, this trend is very likely to reverse. Thus, it is necessary to prepare for what is to come.

III. General Structure of the System

The proposed location system comprises three different parts:

a) A ubiquitous smartphone application to be used by the dependent people: It is in charge of recording the location information, adopting strategies to ensure the operation in remote areas. All this is accomplished transparently to the user.

b) A smartphone application and a Web application handled by the caregivers: Their main objectives are to manage, in a remote way, the users' smartphone applications and obtain feedback about their locations and the potentially dangerous situations that might occur.

c) A Web server with which the other two modules of the system communicate: It performs data processing, and manages the flow of information.

A general scheme of the system is shown in Figure 1.

IV. Description of the System

The following subsections describe each one of the three parts of the system.

A. Ubiquitous application for the dependent people

The smartphone application is in intermittent contact with the Web server, which allows it to obtain updated values of the following parameters:

- The time interval at which the location function should be active.
- The location frequency.

When, at a certain date and time, the location function is active, the device takes the corresponding coordinates and sends them periodically to the server.

This is a typical operation of the location systems, but the novelty lies in the way those data are sent to the server. The application considers the case where there is no coverage and the locations cannot be sent, thus they are stored in a 'stack'. When mobile networks are again available, the whole 'stack' is sent to the Web server, which has been programmed to deal with different input data formats. Together with the locations (latitude, longitude), the exact time at which they were recorded and a field indicating whether they were taken in a blind area or not are also sent to the server.

As the application is aimed at disabled and old people, it has been designed to run in the background, minimizing interaction of the users with smartphones. Thus, their participation is limited to carrying the devices with themselves and remembering to charge them.

Additionally, notifications to the caregivers reporting on low battery levels of the devices will be incorporated in future versions of the prototype.

B. The Caregivers' Management Tools

The caregivers' management module is composed of two complementary platforms: a Web server application, which has been almost completely implemented, and a smartphone application, which is still to be developed.

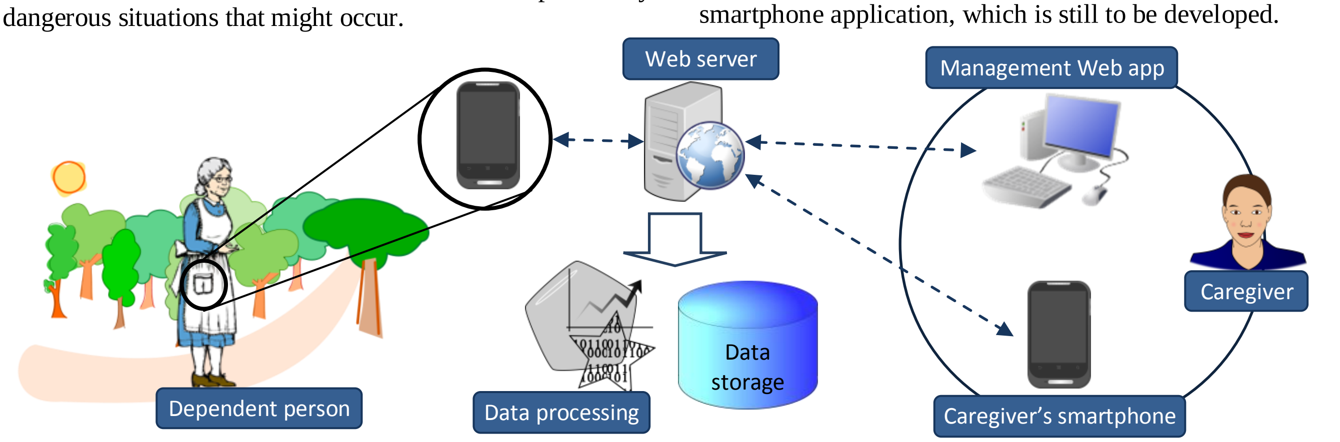


Figure 1. General structure of the system

On the one hand, the Web server application has three basic screens:

- A configuration screen to define the settings of the location function (date, time and frequency).
- A screen that displays the last location of the users on a map. It should also contain a notification area to report the caregivers on potentially dangerous situations (still to be done).
- A screen to visualize the routes performed by the users (see Figure 2).

On the other hand, the smartphone application should have a basic screen showing the current location of the dependent users. Additionally, the caregivers should receive notifications reporting on potentially dangerous situations.

C. *Web Server to Control the System*

The Web server is the central part of the system whose main purpose is to interact with the distributed applications. The Web server is composed of two different modules of increasing complexity: the storage and representation module and the processing module. The first module has already been developed, while the second part is still to be done.

On the one hand, the storage and representation module is in charge of receiving the location data from the ubiquitous applications of the dependent users and storing them in a database. It can also receive a request from the caregivers' management tools (either the Web application or the smartphone application) in order to display a specific route or location on a map. In this sense, it retrieves the information from the permanent storage and sends it back to the agent that made the request. Additionally, the transmission of the location settings, previously defined by the caregivers, to the smartphone applications of the dependent users is carried out by this module.

On the other hand, the processing module, which is still to be done, is more challenging and its main purpose is to add a further level of intelligence to the system in order to compensate for the limitations of the location function in remote areas. Its first phase will consist of training the system to obtain a pattern of the users' displacements. This stage may last for several weeks. The training dataset would be composed of the location data (latitude, longitude, time, a field indicating whether the point was taken in a blind area or not) provided by the smartphone applications. Machine learning strategies will be used for this purpose [16]. This approach is based on the assumption that dependent people usually follow similar paths during their walks. This does not mean that people have to do the same displacements at the same time every day, but it is true that their movements tend to follow a certain pattern to some extent.

Once the processing module was trained, it could provide two different outputs:

a) Estimate the position of the user: If a particular caregiver makes a request for information about the current location of a person and the smartphone application of the user cannot provide a proper response (it is in a blind area),

Figure 2. Screen of the Web application to visualize the routes on a map. The caregivers should select the target user and the date and time intervals.

the Web server may estimate the position of the user according to the model obtained after training. Of course, the provision of a response would be conditioned upon the previous locations fit with the expected pattern. It could happen that the Web server was not able to estimate the position (smartphone switched off, etc.), in which case it would return an alert message reporting this situation.

b) Notify potentially dangerous situations: The Web server should send a notification to both the Web application and the smartphone application of the caregivers in case something unusual was detected (the application does not communicate with the server for a long time, the user deviates from his/her usual path, the user remains for a long time in an unusual location, etc). Figure 3 shows an example of a use case that should trigger an alert.

V. Preliminary Results

A preliminary evaluation of this first prototype has been conducted under real-life conditions. Specifically, three users were selected to wear the location application continuously during six months. Additionally, three caregivers were in charge of managing the whole system. At the end of this period, we collected the views of the two groups using the ISO/IEC 2500X (SQUARE) family of standards (under development). The SQUARE model is based on the previous ISO/IEC 9126 *Software engineering – Product Quality* standards [17], which provide a quality model for the evaluation of software products.

Specifically, as we aimed to conduct a real-world evaluation, we focused on part four ISO/IEC TR 9126-4 *Quality in use metrics*. The quality in use metrics measure whether a product meets the needs of specified users to achieve specified goals with effectiveness, productivity, safety and satisfaction in a specified context of use. This can be only achieved in a realistic system environment.

The results of the evaluation are shown in Figure 4. A value in the range 0 to 5 is provided for each quality-in-use characteristic. Both safety and satisfaction were highly valued by the two groups, while there is still room to improve the effectiveness and productivity of the system, which may be because it is still a preliminary prototype rather than a final version.

 ISBN: 978-1-61208-289-9

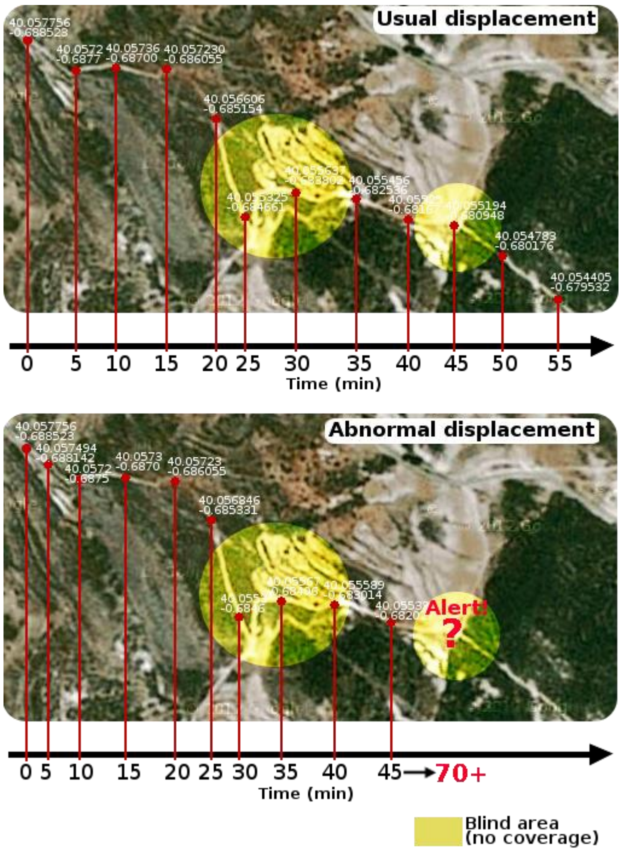


Figure 3. Use case that should trigger an alert. On top, a normal displacement pattern; On bottom, a potentially dangerous situation.

VI. CONCLUSION AND FUTURE WORK

This paper deals with a novel topic: the design of mobile ubiquitous assistive technology for rural environments. As an example, we have presented a smartphone-based location system to assist dependent people living in rural areas. This work, which is still in progress, has been evaluated under real-life conditions. We expected that the older residents were reluctant to use the prototype, but instead it was perceived as a useful support to help them stay in their own villages for longer. But although the preliminary results are promising, the most challenging stages of the project are still to be carried out.

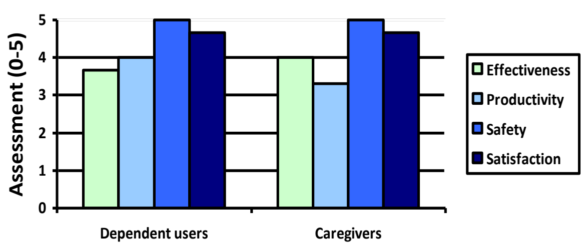


Figure 4. Preliminary evaluation of the system. The assessment of the quality in use for both the dependent users and their caregivers are shown.

In this sense, the demand for this kind of aids is likely to increase in the next decades since the digital generations of today will be the elders of tomorrow. Therefore, it is time to start preparing for the future, and the design of assistive ubiquitous technology adapted to rural environments is a promising line of work.

REFERENCES

[1] K. Murphy, E. O'Shea, A. Cooney, and D. Casey, The quality of life of older people with a disability in Ireland. Dublin: National Council on Ageing and Older People, 2007.

[2] L. Camarero et al., La población rural de España: De los desequilibrios a la sostenibilidad social. Barcelona: Fundacion la Caixa, 2009.

[3] Rural Assistance Center: Health and human Services Information for Rural America, Aging. Available at http://www.raconline.org/topics/aging/ [retreived: July, 2013]

[4] L. Martin, I. Plaza, M. Rubio, and R. Igual, "Localization systems for older people in rural areas: A global vision," in Adv. Intell. Soft-Comput., Eds Springer, March 2012, pp. 597-600.

[5] J.F. Calvert, J. Kaye, M. Leahy, K. Hexem, and N. Carlson, "Technology use by rural and urban oldest old," Technol. Health Care, vol.17, no. 1, May 2009, pp. 1-11.

[6] T. Gandiya et al., "Self-Reported Communication Technology usage in patients attending a cardiology outpatient clinic in a remote regional hospital," Telemedicine e-Health, vol. 18, no. 3, April 2012, pp. 219-224.

[7] Comission of the European Communities, Better access for rural areas to modern ICT. Available at http://europa.eu/ [retreived: July, 2013]

[8] G. Lachapelle, "Pedestrian navigation with high sensitivity GPS receivers and MEMS environments," Pers. Ubiquit. Comput., vol. 11, no. 6, Aug. 2007, pp. 481-488.

[9] Y. Chon and H. Cha, "LifeMap: Smartphone-based Context Provider for Location-based Services," IEEE Pervasive Comput., vol. 58, April 2011, pp. 58-67.

[10] J. Hightower and G. Borriello, "Location systems for ubiquitous computing," Computer, vol. 34, no. 8, Aug. 2001, pp. 57-66.

[11] Aerotel Medical Systems, GeoSkeeper. Available at http://www.aerotel.com/en/ [retreived: July, 2013]

[12] Laipac Tech, S-911 Bracelet Locator. Available at http://www.laipac.com/bracelet_application.htm [retreived: July, 2013]

[13] F. Havasi and A. Kiss, "Ambient Assisted Living in Rural Areas: Vision and Pilot Application," in Constr. Ambient Intell., Eds. Springer Berlin Heidelberg, 2008, pp. 246-252.

[14] J. Sasaki, K. Yamada, M. Tanaka, and Y. Funyu, "An experiment of Life-support network for elderly people living in a rural area," Int. J. Syst. Appl. Eng. Dev., vol.1, 2007, pp.120-124.

[15] Technology to support Older adults, Personal and Social interaction. Available at http://www.dotrural.ac.uk/tops/ [retreived: July, 2013]

[16] C. Piciarelli, C. Micheloni, and G.L. Foresti, "Trajectory-Based Anomalous Event Detection," IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 11, Nov. 2008, pp. 1544-1554.

[17] International Organization for Standardization, ISO/IEC 9126 Software engineering. Product quality. 2001.

Capture and Access Tools for Event Annotation and Visualisation

R. Hunter, M. P. Donnelly, D. D. Finlay, G. Moore
Computer Science Research Institute
University of Ulster
Newtownabbey, Northern Ireland
hunter-r9@email.ulster.ac.uk,
{mp.donnelly, d.finlay, g.moore}@ulster.ac.uk

N. Booth
Parents Education as Autism Therapists (PEAT),
Belfast, Northern Ireland
nichola@peatni.org

***Abstract* – With the increasing rate in diagnosis of developmental disorders in children, the availability of therapists does not satisfy demand. Parents are themselves taking on the role of the therapist as they turn to home-based therapy to manage their childrens' behaviour. Typically, this approach is supported by a Behaviour Analyst (BA) who periodically visits, assesses progress, and set new targets for development. Having access to accurate data is critical for determining suitable targets. This model presents several challenges such as maintaining a record of behavioural events over time, and providing suitable methods of representing such records. The current paper presents a suite of conceptual tools. Mobile Annotation of eVents *In Situ* (MAVIS) is a novel mobile-based tool for parents to use during home-based therapy sessions, allowing them to annotate significant behaviours *in situ*. Using a service called VISualisation of Annotated eVEnts (VISAVE), a BA could remotely access the annotations alongside synchronised environmental sensor data from the home to aid their understanding of specific behavioural events. This would enable them to provide succinct feedback to parents remotely, and to make recommendations for the future course of therapy.**

Keywords-event annotation; sensor data capture and storage; data visualisation; mobile; interface design.

I. INTRODUCTION

Behaviour monitoring is the process of observing people's interactions within an environment. The goal is to be able to model particular behaviours, or to observe changes in behaviour over time [1]. Behaviour monitoring is particularly useful for observing the behaviour of children with developmental disorders such as autism. One application of behaviour monitoring takes place within the context of home-based autism intervention. This involves BAs providing parents with the resources to conduct a tailored course of therapy at home. Children undergo intensive one-to-one sessions with a parent in order to increase or reduce particular behaviours. Traditionally, behaviour monitoring in this instance has been conducted using pen-and-paper, and stored in a paper-based file. This traditional method is prone to error particularly when an instance of behaviour is missed while the parent's attention is diverted as they write their notes from the previous behavioural event. Paper-based records also make retrieving past data a time-consuming challenge. One approach to overcome this challenge could be to introduce technology for adding annotations *in situ* [2]. This would be well suited to situations where behaviour monitoring already involves the presence of both an observer (parent) and a person being observed (child) such as in autism intervention. Due to the nature of observing and recording events *in situ*, it is postulated that mobile-technologies could offer a specific advantage that would be applicable in behavioural event annotation. Leveraging upon developments in touch screen technology, and the portability of these devices offers opportunities to augment traditional pen-and-paper annotation approaches. However, marking the occurrence of behavioural events is not enough on its own. For a BA to identify ways in which to address behaviours, they must understand why they occur in the first place. For this, they need to recognise what precedes each instance of behaviour, and the subsequent consequence. These details are seldom included in handwritten notes, and as a result there exists a real challenge to find improved ways to help augment traditional approaches for monitoring behaviour. One method for detecting the additional events surrounding specific behaviours is to embed sensors in the home to collect complementary data. In particular, video sensors have been used in recent research to support behaviour monitoring [3-5]. Video offers unprecedented insight into a person's movement within an environment, making it well suited to the application of behaviour monitoring. Using event annotation data from an intervention session, video data can be given context by adding visual markers to a session timeline to highlight significance segments of data for review.

This paper presents a conceptual suite of tools designed to support *in situ* mobile-based event annotation and the visualisation of synchronised sensor data. The MAVIS tool supports *in situ* annotation of behavioural events on a mobile device, and stores that data alongside sensor data gathered from heterogeneous sources within the home-therapy setting. VISAVE provides a visualisation service for BAs to remotely access therapy session data so they can observe changes, and make adjustments to a course of therapy. The remainder of the paper is structured as follows: Section II reviews the current state-of-the-art that is reported in the literature. Section III presents the opportunities and challenges that have been identified. Section IV describes a conceptual capture and access system before a discussion on the direction of the future work in Section V.

 ISBN: 978-1-61208-289-9

II. Related Work

There has been an increase in recent research into the design of capture and access tools, specifically for behavioural data. Here, we introduce some of the related work, from early studies in digital annotation to recently developed tools.

An early annotation tool was the Experimental Video Annotator (EVA) [2]. EVA led the way for digital, text-based, annotations; improving upon the previous system for writing on paper, which required later transcription. Marquee [6] was released five years later, and took the digital annotation process one step further by providing a platform for real-time pen-based annotation on a touchscreen display. Perhaps the closest to traditional paper-based annotation methods, pen-based annotation requires a user to draw or write on a touch sensitive surface with an augmented pen. This presented a more familiar interface to help those used to traditional annotation methods to transition to a digital system. It allowed users to not only write notes, but also draw symbols and sketches, allowing for more diverse styles of annotation.

In more recent years, the Notelook [7] client application runs on wireless pen-based notebook computers. Six microphone feeds across a room are amalgamated into a single audio stream, and video is captured at 15 or 30 fps with a resolution of 640 by 480 pixels. Both the audio and video streams are stored on the server and sent across a wireless network to be accessed on a notebook computer. With the video displayed on the touch-sensitive screen, the user writes freeform notes with a stylus. In an evaluation study, four subjects (each having undergone 20 minutes of training with the application) took notes in 13 sessions over a six-week period. Each of the subjects utilised the range features provided by the application and they reported that they would use Notelook again. Future considerations for the application include improving image quality and the generation of an electronic copy of the notes made.

Wearable sensors have also been explored in capture and access tools. Walden Monitor (WM) [8] is a system composed of a head-mounted camera and a tablet PC. It is designed to allow a user to observe behaviours in 10-second intervals, each time noting the child's behaviour through buttons on the interface. This is repeated 20 times, and all of the data (including video captured during each interval) is displayed in a timeline with the behaviour annotations given at the start of each interval. While WM effectively supports the gathering of behavioural data, it does not support customisation based on previously gathered data. A main strength of WM is that the first-person view of the video footage is able to provide rich data. However, an initial trial with the system indicated users' apprehension towards wearing a head-mounted camera. As a result of the trial, feedback expressed that the tablet PC was more cumbersome to carry around when compared to a clipboard for paper-based notes.

To overcome the apprehension towards a head-mounted camera, a non-invasive wearable device is used by the Multiple Perspective Behaviour Analyser (MuPerBeAn) [3]. This platform allows the review of footage captured from miniaturised cameras embedded within glasses. Footage is captured at multiple perspectives – capturing the personal view of the child, as well as the situation's context. The videos are then synchronised by aligning them at a frame which shows an object, for example – a watch displaying the time – that has been captured at the same time at multiple views. The videos can then be reviewed on a frame-by-frame basis, and annotated accordingly.

A more recent tool, Continuous Recording and Flagging Technology (CRAFT) [4], supports the *in situ* and *post hoc* 'flagging' of behavioural events within the home environment. Parents flag incidents of problem behaviour using a custom wireless device throughout the recording session. A BA then reviews and annotates the footage *post hoc* based on their professional criteria of problem behaviour. One study with CRAFT produced an average of 12 hours of video footage for each of the eight participants. It required 10 BAs to review the footage, and to annotate it accordingly. The main strength of this system is that the two sets of annotations can be compared to note agreements and disagreements. This allows the BA to train the parent in identifying what specifically constitutes problem behaviour. It also presents several weaknesses; the vast amount of video data requires extensive storage, and the system requires many hours to review the entirety of footage.

Another recent annotation tool is the Dynamic ANnotation Tool for smart Environments (DANTE) [5], which uses a pair of stereo cameras to monitor the movements of objects within a scene by the placement of markers and sensors. Using video technology, it was reported that activities of daily living (ADLs) could be verified when the video footage was reviewed alongside the sensor data. The DANTE system provides an activity recognition platform, reducing the need for individuals to keep manual records. Three subjects, with experience in assistive technology, were involved in a trial with DANTE. They each performed two ADLs three times, resulting in 18 video clips. The objects they interacted with were fitted with sensors and markers. DANTE detected the movement of the objects through the movement of the markers within the scene. At the end of the recording session, the video files were annotated, first based on the raw sensor data, and then supported by DANTE. The results showed that using DANTE for annotation took 6.5 minutes, compared to 12 minutes without using the tool.

Based on a review of the existing literature, there does not appear to have been a mobile solution that provides a suite of tools to support both the annotation and visualisation of behavioural events in synchrony with corresponding sensor data. MAVIS has been conceptualised to draw on the strengths of previously developed event annotation tools, while addressing their limitations. VISAVE is designed to

 ISBN: 978-1-61208-289-9

complement the mobile-based MAVIS tool, by adding a presentation layer to the sensor data for efficient review of data over time.

III. OPPORTUNITIES AND CHALLENGES

While the opportunity for the capture and retrieval of sensor data is clear, there is a challenge of storage. With many hours, days, and even several weeks' worth of sensor data, there is a clear challenge in deciding on a suitable storage medium for all data generated by a capture and access system. Traditional SQL databases are commonly used for sensor data storage and present a stable platform on which to store and query large volumes of data. However, with on-going challenges of scalability and performance, NoSQL databases are gaining more attention in the area of data storage [9]. NoSQL databases are more tolerant to network partitioning, making it easier to add servers as necessary, instead of increasing the capacity of a single server as would be the case for an SQL database, which is less tolerant to network partitions.

With a broad range of sensors being used to efficiently monitor behaviour within the home, it is inevitable that the resulting output is a vast quantity of heterogeneous data, which often lack interoperability [10]. The processes for the aggregation and synchronisation of sensor data are both common challenges in data handling [11]. The installation of a Network Time Protocol (NTP) time server provides a reliable and secure time synchronisation service for a network. With a NTP installed in a capture and access system, each segment of data can have an accurate timestamp, allowing efficient review of sequential events.

A further challenge exists surrounding the understanding of sensor data. Even when it is aggregated and synchronised, the ability to extract useful information from it can be a laborious manual task requiring many hours of review. Indeed, with video sensors alone there is a continuing challenge is to find ways to annotate and analyse a large volume of video data without having to review many hours of footage [12]. Szewcyzk *et al.* describe the challenge of annotating data by sensor data analysis as time consuming and subject to error, and outline that it is an on-going challenge. [13].

Traditionally, event annotation is completed *post hoc*. It requires a complete review of all the data, which is time-consuming and laborious. While *in situ* annotation offers significant advantages, it still presents a series of challenges. During *in situ* annotation, the observer who is making annotations may have a restricted view of events, or may be have their attention averted from observing live events as they focus on entering a previous annotation [14]. This means that the annotations may be incomplete, and would require a *post hoc* review. With additional data to support annotations, it is easy to confirm or reject events based on a review of the complementary data. In addition, real-time annotation causes a tendency to annotate as quickly as possible so as to not miss any events, which may results in annotations lacking sufficient detail.

There are several methods for adding annotations to data. The most popular is text entry, where a user has to type the details of the events. Audio annotation has also been utilised in some studies. While audio presents clear advantages of speed and attention required, it requires transcription to be turned in to useful information. The transcription process is one which remains inaccurate due to the variability of voice tones, accents and languages.

There is a clear opportunity for a novel capture and access suite of tools. By providing a mobile platform for the annotation of event data, and a visualization tool for illustrating annotations and sensor data, the process of behaviour analysis can be improved both from time and cost perspectives. This paper presents a conceptual system that aims to address the challenges and opportunities that have been discussed.

IV. SYSTEM DESCRIPTION

The conceptual use of MAVIS and VISAVE are shown in Figure 1. What follows is a detailed discussion of the system components.

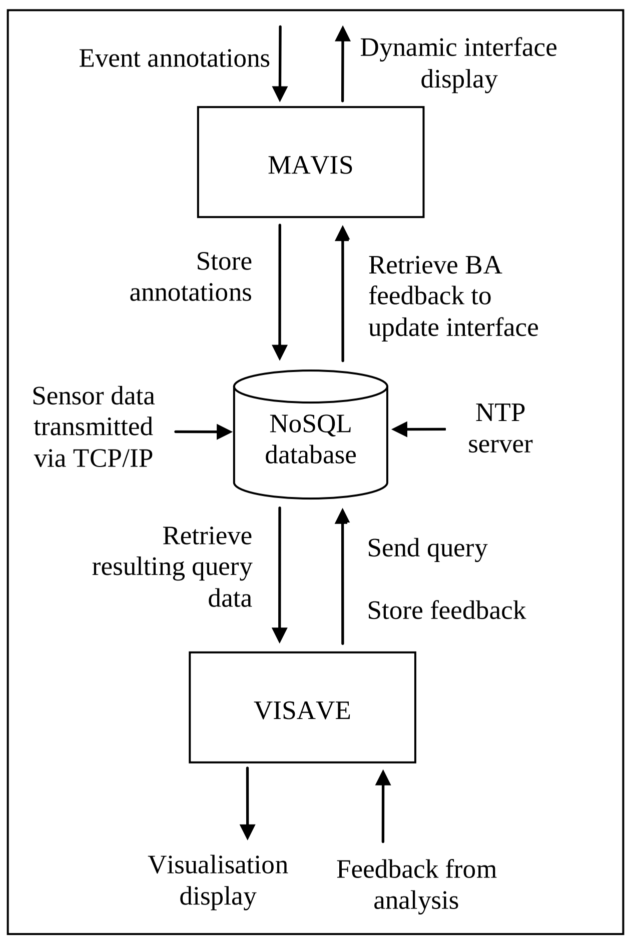


Figure 1 - A sequence diagram showing how data flows through the system.

 ISBN: 978-1-61208-289-9

A. Data Capture

Annotations are captured through interaction with the MAVIS tool in real-time. A parent loads the MAVIS tool through a web browser on their mobile device, and logs in with their username and password credentials. They are then presented with a dynamically generated interface with pre-defined buttons based on previous BA feedback. The buttons are labelled with specific events or behaviours being monitored at any one point throughout the course of therapy. When a parent touches a button on screen, the click event is detected and an annotation is initiated. The timestamp for the click event is retrieved from the NTP server, and the type of behaviour is determined based on the corresponding button label. This annotation data is added in real-time to an XML file stored on the server. At the end of the session, the XML file and the details of the session (including start and end time, and the child's client number) are sent to be stored in a database for later access.

In addition to annotations, the capture of sensor data is required to provide additional information about events. As sensor data is captured it is transmitted to the server via TCP/IP. The use of video sensors to record events as they occur is the most convenient way to provide a method for reviewing and analyzing behaviour events accurately. With the ability to capture a scene from several viewpoints simultaneously and store that data for later retrieval, video offers valuable data, which no other sensor type can provide [15]. Depending on the specific requirements for each child, an array of sensors can be used with the system such as physiological sensors to allow the automatic annotation of physiological events which may be unseen, such as a change in heart rate. Sensor-augmented toys can provide insight into how a child interacts with certain objects. For example, a parent may note that a child has thrown a toy, and with additional accelerometer data, it would be possible for the BA to determine the force with which it was thrown. When the session is initiated through the MAVIS tool, a message is sent to the server to signal the sensors to begin to record sensor data. Sensor data is captured and stored directly to the database through the network.

B. Data Storage

For a novel approach to the storage of sensor data and annotations, the NoSQL database CouchDB [16] will be implemented for the system. This will give the advantages of scalability and performance over traditional relational databases (RDBMS). CouchDB will allow the storage of heterogeneous sensor data within the database without having performance issues that would commonly be caused by video data, for example, when stored in an RDBMS.

The data stored in the database consists of parental data including login credentials and contact information; session data including start and end time, and data about the child; annotation data including event type and a timestamp; sensor data including sensor type, sensor data and a timestamp; and feedback data indicating the progress of each child.

C. Data Access

With the vast amount of sensor data gathered from heterogeneous sources, there is a requirement to have a service in place for retrieving and visualising relevant data to produce useful information. VISAVE is a web-based application that has access to the database containing session data archives for each user. The service is accessed by the BA remotely on a desktop computer. They access the data by logging in with their credentials. A list of MAVIS users with which they are associated is given, and they select the name of the child requiring review. Depending on the specified course of therapy for that individual child, the BA identifies the events they wish to review. They can customise their selection by behavioural event type, by session date, and by timescale. Depending on their selection, they will be presented with a single-screen interactive visualisation illustrating the annotations and synchronised sensor data. They are able to navigate quickly to areas of importance as highlighted by the annotation markers along a timeline, and observe the surrounding sensor data. This can help the BA to understand why certain behaviours may occur; including observation of environmental factors or other events which may be triggering repeat behaviours. After reviewing the data, they decide if a change in the course of therapy is to be recommended, or if they require the parents to annotate additional events during future sessions. They can input this information as feedback for the parents to review when they access the tool.

D. Interface Considerations

The prototype interface for MAVIS (Figure 2) has been designed to allow even novice users to efficiently utilise the tool *in situ*. As mentioned in the description, parents do not need to set up the tool for each session; they simply have to follow instructions from the BA, and note events as they occur through on screen buttons. The configuration is carried out by the BA, who specifies the parameters for the interface through the VISAVE tool, ahead of a therapy session. The simple button interface allows a parent to maintain their attention on conducting the therapy, as a touch of a button will not avert their attention nor prove to be a distraction for the child.

The key design consideration for the VISAVE interface (Figure 3) is to consider ways in which to display a large volume of heterogeneous data in a single visualisation. The data that a BA requests is to be presented full screen so in a single glance they can tell if the current course of therapy is altering a child's behaviour, and the rate at which that change is occurring. It is envisaged that one or more timelines will display markers for annotations and significant sensor data changes, allowing a BA to quickly browse to a segment of interest. The surrounding event data is played back in synchrony in windows on the interface. The interface layout will adjust accordingly for multiple sensor streams. VISAVE will also have a feedback facility where a BA can type notes to the parents, to guide them in a suitable course of therapy. It will support a range of personalisation features to allow the BA to update the MAVIS interface with suitable data for a specific therapy session. This feedback would be sent to the server for MAVIS to retrieve when next accessed.

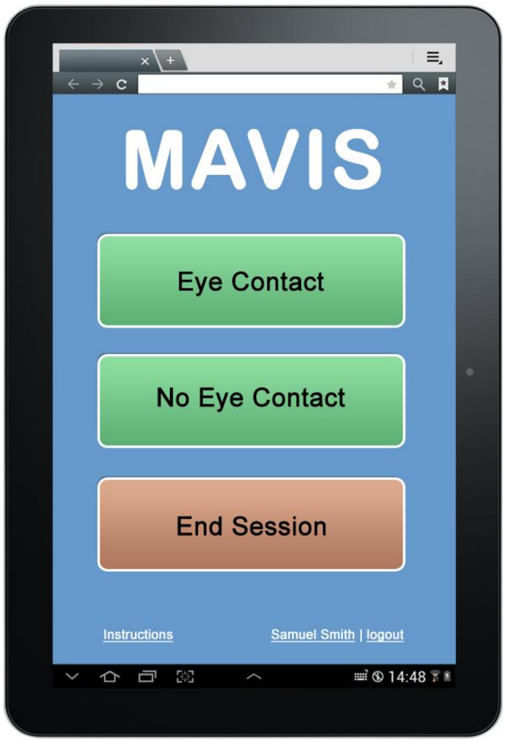


Figure 2 – The MAVIS prototype interface displayed on a touchscreen tablet computer.

Figure 3 – The conceptual VISAVE interface displayed on a desktop computer.

E. Case Study Presentation

To help contextualise the potential utilisation of the proposed system, an autism intervention scenario is presented in the following case study.

John is an 8-year old boy who has been diagnosed with autism. He has impairments in social interaction and communication; notably struggling with maintaining attention, eye contact, and speech development. His father, Samuel, has been researching his condition, and has decided that he wants to take on the role of a therapist at home, and guide John through a program of interventions designed for him by Laura, a qualified BA.

Laura attends the home to meet John, and prepares a series of interventions to address his impairments. She designs an intervention to improve John’s eye contact, and gives Samuel a program to follow once a day. Samuel is to sit across from John at a table, and say “John, look at me”. If he follows the command within 10 seconds, he receives verbal praise and a reward. If he does not, Samuel must repeat the phrase again. This process should continue for 5 minutes during each intervention.

Laura specifies the text for event buttons via VISAVE on her laptop. For the eye contact intervention, she wants to specifically see positive and negative responses, so adds two buttons. She shows Samuel how to use MAVIS on his personal tablet, and points out the button interface that has been set up to indicate the behaviours he should annotate.

The dining room in their house is set up with two cameras pointing at the table, and Laura has provided a set of sensor-augmented objects to use during the therapy. At the start of the session, Samuel presses a ‘Start Session’ button, which triggers the start of the capture of sensor data. During the intervention, he uses the MAVIS tool on his tablet to annotate John’s responses.

The following week, Laura logs on to the VISAVE service and reviews the data. At a glance, she is able to observe how often John followed Samuel’s command during each intervention session, and whether or not his attention remained on the task or if he was interacting with the sensor-augmented objects. The visualisation shows the video and other sensor data in synchronisation with Samuel’s annotations. By clicking through the annotations on the video, Laura can observe the eye contact taking place, and the surrounding conditions. She is able to identify that John is showing an improvement in eye-contact, and can recommend a slight change to the intervention to make further improvement. The following week, Samuel is to repeat the intervention, but now present John with a toy at the table before repeating the “look at me” command.

The process of regularly adjusting the course of intervention helps John’s impairments to improve. Samuel continues his role as a therapist at home, and Laura’s remote input ensures the interventions are effective.

V. Conclusion

MAVIS is currently in prototype phase. As development continues, the next step will be to utilise MAVIS to annotate pre-recorded events post hoc. By comparing the processes for online and offline annotation, the effectiveness of the MAVIS tool will be realised. A usability study will follow where users will experiment with the tool for testing and analysis purposes. The interface will be updated based on the feedback provided. A further study will explore the methods for storing heterogeneous sensor data alongside annotations, as well as effective query and retrieval methods. The VISAVE service will need to be developed with user needs in mind. It will be adapted to suit any range of queries that a BA may require. With testing complete, the entire MAVIS system will be trialled across a suitable cohort.

References

[1] T. Ploetz et al., “Automatic Assessment of Problem Behaviour in Individuals with Developmental Disabilities,” Proceedings of the 14th ACM International Conference on Ubiquitous Computing, Pittsburgh, USA, Sep. 2012, pp. 391-400.

[2] W. E. Mackay and G. Davenport, “Virtual Video Editing in Interactive Multimedia Applications,” Communications of the ACM, vol. 32, no. 7, Jul. 1989, pp. 802-810.

[3] P. Alt, B. Pfleging, A. Bungert, A. Schmidt, and M. Havemann, “Supporting Children with Special Needs through Multi-Perspective Behaviour Analysis,” Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia, Beijing, China, Dec. 2011, pp.81-84.

[4] N. Nazneen et al., “Supporting parents for in-home capture of problem behaviours of children with developmental disabilities,” Personal and Ubiquitous Computing, vol. 16, no. 2, Feb. 2012, pp. 193-207.

[5] M. Donnelly, T. Magherini, C. Nugent, F. Cruciani, and C. Paggetti, “Annotating Sensor Data to Identify Activities of Daily Living,” Toward Useful Services for Elderly and People with Disabilities - Proceedings of the 9th International Conference on Smart Homes and Health Telematics, Montreal, Canada, Jun. 2011, pp. 41-48.

[6] K. Waber and A. Poon, “Marquee: A Tool For Real-Time Video Logging,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 94), Boston, USA, Apr. 1994, pp.58-94.

[7] P. Chiu, A. Kapuskar, S. Reitmeier, and L. Wilcox, “NoteLook: Taking Notes in Meetings with Digital Video and Ink,” Proceedings of the 7th ACM International Conference on Multimedia, Orlando, USA, Nov. 1999, pp. 149-158.

[8] G. R. Hayes, J. A. Kientz, K. N. Truong, D. R. White, and G D. Abowd, T. Pering, “Designing Capture Applications to Support the Education of Children with Autism,” Proceedings of the 6th international conference on Ubiquitous Computing, Nottingham, UK, Sep. 2004, pp. 161-178.

[9] J. S. van der Veen, B van der Waaij, and R. J. Meijer, “Sensor Data Storage Performance: SQL or NoSQL, Physical or Virtual,” Proceeding of the 5th IEEE International Conference on Cloud Computing (CLOUD), Honolulu, USA, Jun. 2012, pp. 431-438.

[10] H. A. McDonald, C. D. Nugent, G. Moore, and D. D. Finlay, “An XML Based Format for the Storage of Data Generated both inside and outside of a smart home environments,” in Proceedings of the 10th IEEE International Conference on Information Technology and Applications in Biomedicine, Corfu, Greece, Nov. 2010, pp.1-4.

[11] H. Lee, K. Park, B. Lee, J. Choi, and R. Elmasri, “Issues in data fusion for healthcare monitoring,” in Proceedings of the 1st International Conference on Pervasive Technologies Related to Assistive Environments (PETRA), Athens, Greece, Jul. 2008, Article no.3.

[12] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce, “Automatic annotation of human actions in video,” in Proceedings of IEEE 12th International Conference on Computer Vision, Kyoto, Japan, Oct. 2009, pp. 1491 - 1498.

[13] S. Szewcyzk, K. Dwan, B. Minor, B. Swedlove, and D. Cook, “Annotating Smart Environment Sensor Data for Activity Learning,” Technology and Healthcare, vol. 17, no. 3, Aug. 2009, pp. 161-169.

[14] P. Alt, B. Pfleging, A. Bungert, A. Schmidt, and M. Havemann, “Supporting Children with Special Needs through Multi-Perspective Behaviour Analysis,” in Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (MUM 11), Beijing, China, Dec 2011, pp. 81-84.

[15] D. Ding, R. A. Cooper, P. F. Pasquina, and L. Fici-Pasquina, “Sensor technology for smart homes,” Maturitas, vol. 69, no. 2, Mar. 2011, pp. 131-136.

[16] J. Lennon, “Exploring CouchDB:A document-oriented database for Web applications”, 31 Mar 2009 [Online]. Available: http://www.ibm.com/developerworks/opensource/library/os-couchdb/ [Accessed on 2013-03-01]

Unconventional TV Detection using Mobile Devices

Mohamed Ibrahim, Ahmed Saeed
and Moustafa Youssef
Department of Computer Science and Engineering
Egypt-Japan University of Science and Technology(E-JUST)
{mibrahim,ahmed.saeed,moustafa.youssef}@ejust.edu.eg

Khaled A. Harras
Computer Science Department
School of Computer Science
Carnegie Mellon University Qatar
Email: kharras@cs.cmu.edu

***Abstract*—Recent studies show that the TV viewing experience is changing giving the rise of trends like *"multi-screen viewing"* and *"connected viewers"*. These trends describe TV viewers that use mobile devices (e.g., tablets and smart phones) while watching TV. In this paper, we exploit the context information available from the ubiquitous mobile devices to detect the presence of TVs and track the media being viewed. Our approach leverages the array of sensors available in modern mobile devices, e.g., cameras and microphones, to detect the location of TV sets, their state (ON or OFF), and the channels they are currently tuned to. We present the feasibility of the proposed sensing technique using our implementation on Android phones with different realistic scenarios. Our results show that in a controlled environment a detection accuracy of** 0.978 **F-measure could be achieved.**

***Index Terms*—TV detection; ubiquitous sensing; mobile computing; audio fingerprinting; computer vision**

I. Introduction

TV viewers' profiling is an important functionality for both advertisers and service providers. Traditionally, the detection techniques of TV viewers' habits are concerned more about the collective preferences of the viewers and rely mainly on focus groups [16] or special hardware connected to the TV (e.g., set top devices) [19]. Recent studies show that 52% of cell phone owners use their phones while watching TV [17] and 63% of tablets owners use their tablets while watching TV [1] in what was called *"Connected Viewers"*. The rise of these *"Connected Viewers"* opens the door for a new unconventional approach for TV viewers' profiling based on the ubiquitous mobile devices and their equipped sensors. Such approach can provide ubiquitous fine-grained information about the user's TV viewing preferences leading to new possibilities for advertisers and service providers on both the TV and mobile sides.

In this paper, we present the design, implementation and evaluation of a system that can leverage the large array of sensors currently available in smart phones and other mobile devices to accurately detect the presence of TV sets. In particular, our implementation currently depends on the acoustic context analysis and visual surroundings detection using microphones and cameras embedded in mobile devices to identify (1) the presence and locations of the TV sets, (2) whether they are ON or OFF, and (3) the channels they are currently tuned to. We test the system's ability to differentiate between a TV set's acoustic and visual fingerprints on one side and other sources of similar fingerprints such as people having a conversation and laptops playing audio/video files on another side. The results showed that a typical mobile device can reach an F-measure of 0.978 in a *controlled environment*.

Our goal is to develop a novel system that could be deployed on mobile devices to passively detect the TV viewing preferences of the mobile devices owners. This TV viewing history information, analogous to web browsing history information, can assist with enhancing the user mobile and web browsing experience. This information can also give way to new social applications that connect users with the same TV viewing preferences. On the other hand, this information will be invaluable to advertisers by providing fine-grained audience measurement, tracking mobile users' preferences through their TV viewing habits which can enable a new generation of targeted mobile ads and more informed planning of TV ads.

This approach sets itself apart from earlier work that detects TV shows or identifies playing music based on acoustic fingerprints [2], [7], [8], [14], [20] by allowing for passive detection of TVs and the shows they are playing. Conventional popular applications (e.g. IntoNow [2]) are interactive applications that require the user to operate it to detect the TV show playing which assumes the presence of an operating TV set. This approach is not appropriate for a passive application that aims at tracking the user's TV view preferences. Moreover, all these audio detection approaches focus on identifying the content regardless of its source and hence cannot determine the audio source type (whether it is a laptop, people talking, or TV). On the other hand, our proposed system addresses the challenges of detecting the presence of a TV set and determining whether it's turned on or off before determining the show playing.

The rest of the paper is organized as follows: In the next section, we provide our vision and architecture for the novel sensing approach. Section III presents and evaluates our mobile TV detection service. Then, we briefly discuss related work in Section IV. Finally, Section V discusses the challenges and gives directions for future work.

II. System Architecture

Figure 1 shows the proposed architecture. Standard mobile devices with embedded sensors (such as mic, camera, accelerometer, Global Positioning System (GPS), etc.) submit their location tagged sensory information to the system's server. The server has three main components: Mobile TV Detection Service, TV Detection Supporting Services and TV Detection Applications.

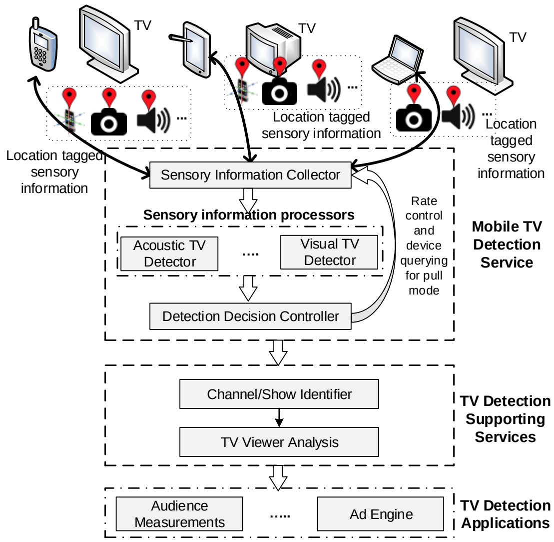
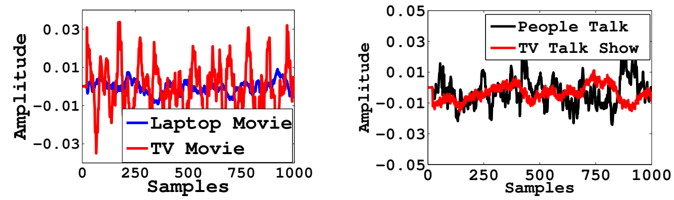


Fig. 1. System architecture.



(a) Movie on a TV and laptop

(b) TV show and normal people talking

Fig. 2. Acoustic time-domain raw data amplitude.

Mobile TV Detection Service is responsible for the collection and processing of the sensory information. This service contains different modules responsible for the detection of TV sets based on information collected from different sensors. It is also responsible for the fusion of the detection decision made by the different sensors. Moreover, this service is responsible for controlling the rate at which the sensors collect their information.

TV Detection Supporting Service is responsible for further processing of the information collected about the detected TV sets. It connects to TV streaming servers, online schedules, and media databases to detect the current channel. It depends on the comprehensive previous techniques for detecting TV shows, music, commercials, channels [3], [5], [6], [14], [20]. Other possibilities include interaction with social network sites to access information about the user preferences.

TV Detection Applications use the TV sets information collected by other services to provide different services either to the mobile user (e.g., personalization) or to third party applications (e.g., audience measurement systems and targeted ads systems).

For the rest of the paper, we focus on the detection of the presence of TV sets using mobile phones. We present the design, implementation and evaluation of the *Mobile TV Detection Service*.

III. Mobile TV Detection Service

We implemented the service on different Android phones and used it while watching different TV sets made by different manufacturers. We tested our implementation in a *controlled environment* using two sensors: microphone and camera. We address the challenge of differentiating the visual and acoustic signature of TV sets and other sources that could be confused with the TV. For example, the sounds coming from a TV set could be confused with a laptop or normal people talking. Moreover, the visual signature of a TV set (i.e., rectangular-shaped object with varying content) could be confused with picture frames and windows.

A. Acoustic TV Detector

The main challenge for acoustic TV detection is extracting unique features for the acoustic fingerprint that would enable the differentiation between TV sets and other sources that could be confused with it. We collected an acoustic dataset composed of 91 audio recordings for training and 60 independent audio recordings for testing. Each audio recording is 30 seconds long. We had different configurations including the TV volume, phone relative distance to the TV, position of the phone (in pocket, on couch, etc.), show type (movie, talk show, sports, etc.), gender and talking level of the actor/anchor. Also, we collected a data set under the same different configurations for the laptop and normal people talk classes. Our goal in the rest of this section is to identify time and frequency domain features that can differentiate between the TV case on one hand and the laptop and people talking case on the other hand. Figures 2 and 3 show sample raw data obtained from our acoustic dataset.

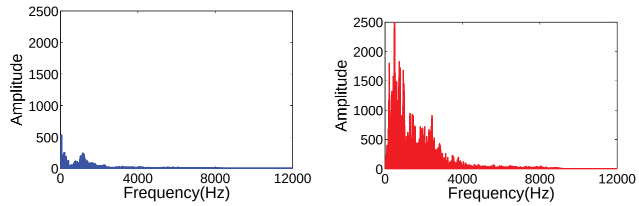
1) Time domain features: Figure 2(a) shows the raw time domain acoustic amplitude for listening to a movie on a TV and on a laptop, whereas Figure 2(b) shows the same signal while listening to a TV show and listening to a group of people talk. The figure shows that there is a noticeable difference between the two cases in each figure. This is intuitive, as a person listening to a movie or show on a laptop will usually have a lower volume than the case of listening to the same show on the TV. On the other hand, people talking will tend to lower the volume of the TV.

Based on Figure 2, we extract features that capture the amplitude variations of the acoustic signal in the time domain. One of these key features is the Zero Crossing Rate (ZCR) [13] that represents the rate at which the signal crosses the x-axis (zero amplitude).

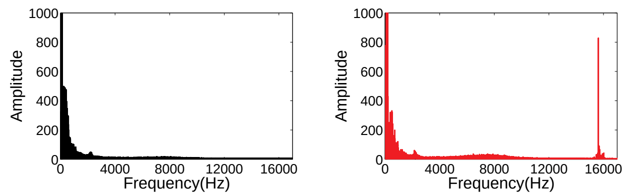
$$ZCR = \frac{1}{N}\sum_{i=1}^{n}|sign(data(i)) - sign(data(i-1))|$$

$$sign(x) = 1, x > 0 \quad (1)$$
$$sign(x) = 0, x = 0$$
$$sign(x) = -1, x < 0$$

 ISBN: 978-1-61208-289-9



(a) Watching a movie on laptop vs. TV



(b) Normal people talk vs. Watching a talk show on TV

Fig. 3. Acoustic frequency-domain raw data.

ZCR is used to estimate the fundamental frequency of the signal. Therefore, it is used as an indication of the noisiness of the signal. Another time domain feature is the Short Time Energy (STE) [13] that captures the loudness of the signal and is computed as the average of the square amplitude of the signal:

$$STE = \frac{1}{N}\sum_{i=1}^{n} data^2(i) \quad (2)$$

2) Frequency domain features: Figure 3 shows the frequency domain signal for the same example as in Figure 2. The figure shows that the frequency domain response of the signal differs from the TV and other classes. From the figure, it could be observed that media streamed to laptops are lower quality in terms of bit rate compared to media displayed on the TV. This observation leads to the conclusion that the acoustic fingerprint of laptops will have a lower bandwidth as compared to TV sets. Similarly, comparing the acoustic fingerprint of a TV set and normal people talk, it could be observed that the TV set's fingerprint is a combination of people talk (4 KHz) and other sources (e.g, music (16 KHz)). This observation also leads to the conclusion that people conversations will have a lower bandwidth as compared to TV sets in the frequency domain. Based on these observation, we use the following frequency domain features: Spectral Centroid (SC) and Spectrum Spread (BW) [13].

$$SC = \frac{\sum_{i=1}^{n} f(i) * A(i)}{\sum_{i=1}^{n} A(i)} \quad (3)$$

$$BW = \frac{\sum_{i=1}^{n} (f(i) - SC)^2 * A(i)}{\sum_{i=1}^{n} A(i)} \quad (4)$$

where $f(i)$ is frequency and $A(i)$ is the amplitude values at index i. These features represent the spectrum by its center of mass and its spread around that center. We also use the Mel-frequency Cepstral Coefficients (MFCC) [13] which are

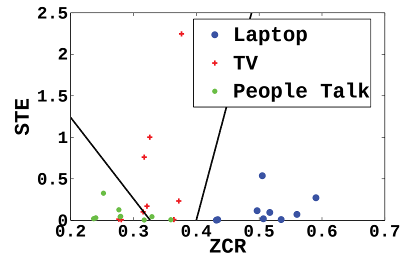


Fig. 4. SVM discriminant function using two features.

a set of features, where each feature represents a portion of the spectrum in the Mel scale.

3) Acoustic fingerprint classification: After extracting the features, we use a Support Vector Machine (SVM) classifier to distinguish TVs from the other two classes. Figure 4 shows a sample result using the classifier for two features (ZCR and STE). As the figure shows, the three classes are linearly separable, except for some TV talk shows and people talk samples, then the SVM classifier, which is a discriminate based classifier, can easily classify them.

B. Visual TV Detector

Acoustic detection may confuse the presence of TV sets with other sources of similar acoustic fingerprints, e.g., stereo players. To reduce this ambiguity, we consider the usage of cameras as a source of detection information. Our approach is based on recent statistics that show that a smart phone user holds the smart phone at 12.5 inches from her eyes while surfing and at 14.1 inches away from her eyes while texting [4]. At these distances, if the smart phone user is watching the TV, the TV will either partially or fully appear within the camera's frame. We collected 26 shots by normal users using their phones, e.g., to text or browse the Internet, with each shot composed of 8 consecutive frames. Fourteen out of the twenty six shots were taken in different light conditions in different locations with a mix of shots showing the TV as a whole or partially. The remaining 12 shots had no TV sets but rather objects that could be confused with TV sets using our proposed algorithm (e.g., windows, doors, picture frames, and shots with moving people and objects).

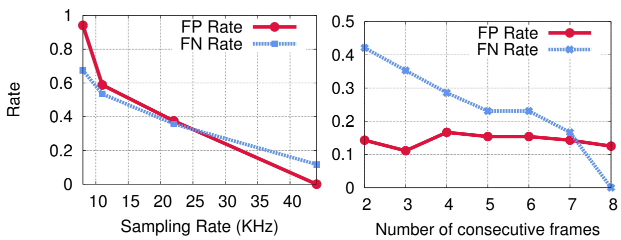
We use a simple algorithm that detects the characteristics of a TV in a sequence of frames captured either through a recorded video or sequence of captured images. The algorithm works in three steps, summarized in Figure 5. The first step, Figure 5(a), detects changing parts in the image sequence, which represent the dynamics of the scene on a TV set. This is performed using a simple background extraction algorithm [22]. In the second step, Figure 5(b), it determines rectangle-shaped contours within each image filtering out small contours (smaller than 5% of the total area of the image) and large contours (larger than 70% of the total area of the image). The rectangle shapes detection works in two steps: the first finds all contours in the image using the algorithm proposed in [18]. In the second step, all contours are simplified by reducing

(a) The changing areas between different frames are detected as foreground.

(b) All rectangles in picture are detected then small rectangles and large rectangles are filtered out.

(c) The intersection of the previous two steps is performed to detect the existence of the TV.

Fig. 5. TV detection steps using the camera.



(a) Audio sampling rate effect on detection accuracy.

(b) Camera frames number effect on false positive and false negative rates.

Fig. 6. Summary of Individual Sensor Results

TABLE I
COMPARISON BETWEEN DIFFERENT TV DETECTION APPROACHES.

Approaches	Acoustic	Visual	**Fused**
False Negative Rate	0.13	0	**0**
False Positive Rate	0	0.125	**0.042**
F-measure	0.928	0.933	**0.978**

the number of points forming them using the Ramer-Douglas-Peucker algorithm [9]. Reduced convex contours composed of only four points are selected as rectangle-shaped areas. Finally, an intersection step between the previous two stages is performed to produce the final output (Figure 5(c)). In particular, the rectangle with the smallest area that encloses all the recorded foreground contour centers is declared to be a TV.

C. Detection Decision Controller

This module is responsible for fusing the decisions made based on different sensors. Furthermore, it is also responsible for setting the frequency by which the sensory information are collected (e.g., acoustic sampling rate and number of captured frames per second) to control both the accuracy and energy efficiency of the system. The sensors fusion is based on the assumption that avoiding false negatives is more important than false positives, as not detecting a TV set wastes opportunities of detecting context information. Therefore, we fuse the results of the audio and video modules using the a simple OR rule: If the two techniques result in two opposite results, then the fused results will always be positive, i.e., a TV is detected.

Figure 6(a) shows the effect of the acoustic sampling rate on the false positive and false negative rates. The figure shows as the sampling rate increases, more information is extracted from the audio source and lower false positive and false negative rates are achieved. Figure 6(b) shows the effect of increasing the number of consecutive frames on the visual detection algorithm. Table I summarizes the results. The acoustic approach achieves a zero false positive rate and a 0.13 false negative rate (0.928 F-measure) with most of the errors in mixing a quiet talk show on the TV with normal people talking. On the other hand, the visual detection approach achieves a detection accuracy of zero false negative rate and a 0.125 false positive rate (0.933 F-measure). The results of the fusion approach is summarized in Table I. This approach results in a zero false negative rate and 0.042 false positive rate (0.978 F-measure). Note that this can also be further enhanced by combining the detection results from different nearby devices and other sensors.

IV. RELATED WORK

Extensive work has been done in detecting real time TV shows, commercials, and channels [3], [6]. This involves scene boundary detection [15] and TV shows recognition [10]. Another line of work depends on audio as their data source for TV shows, music and channel identification [5]–[8], [14], [20]. E. Bisio et al. [6] showed how to detect in real time what people are watching on TV using audio signals recorded by their smart phones. The IRCAM audio Fingerprint framework [14] enhances the accuracy of two commercial systems: Philips [11] and Shazam [20]. In [8], a Filtering approach was proposed to extract music from background noise and identify it. However, all these audio detection approaches focus on identifying the contents regardless of its source and hence cannot determine the audio source type (whether it is a laptop, people talking, or TV).

On another perspective, earlier work investigating the detection of TV sets, e.g., [21], relied on special devices for sensing. This work detects the power leakage of a receiver's local oscillator in order to detect the presence of a TV set. This technique required the usage of special hardware that needed to be setup in the vicinity of the TV set. Such systems do not scale and are harder to deploy.

V. Conclusion and Future Work

A. User Privacy

Protecting the user privacy can be achieved by local processing of the raw data on the user mobile device and forwarding only the anonymized detection results. This can be extended by forwarding the data from the mobile phone to a more powerful device, such as the user laptop for processing before forwarding to the back end server. Another possibility is to provide an option for users to avoid visual recordings and use only acoustic recordings for better privacy settings. Moreover, secure computations can be used on encrypted data [12], which is still an area with a space for huge improvement. Attacks on the system, e.g., by submitting incorrect data, should also be accounted for.

B. Incentives

To encourage the users to deploy the proposed system, different incentive techniques can be used including providing coupons, recommendation services, among other traditional incentive systems.

C. Using Other Sensors

The proposed approach can be extended to use other sensors. For example, the inertial sensors (e.g., accelerometer, gyroscope and compass) can be used to better trigger the acoustic and visual detection sensors based on the detected user activity. Other sensors, such as WiFi, can be used to obtain the device location indoors and hence provide better context information about the user actions and the TV location.

D. Energy Efficiency

Continuous sensing on a mobile device can quickly drain the scarce battery resource. Automatically setting the sensing rate and which devices to sense based on their remaining battery, the device context and location, and required information are different steps to address this issue. This is one of the main functionalities of the *Detection Decision Controller*. In addition, offloading the computations to a more powerful user device can also help alleviate this concern.

E. Large Scale Deployments

The results presented in the paper were conducted as a proof-of-concept in a controlled environment. A large scale deployment of the application should be conducted to measure the performance of the proposed system in a real environment. A major challenge of such a large scale evaluation include determining the ground truth of whether the user is actually viewing the TV or not.

References

[1] "Tablets and Multi-Tasking," *The GfK MRI iPANEL Reporter*, 2012.

[2] "IntoNow," http://www.intonow.com/ci, Last Accessed July 2013.

[3] A. Albiol, M. Ch, F. Albiol, and L. Torres, "Detection of TV commercials," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 3. IEEE, May 2004, pp. iii–541.

[4] Y. Bababekova, M. Rosenfield, J. Hue, and R. Huang, "Font size and viewing distance of handheld smart phones," *Optometry & Vision Science*, vol. 88, no. 7, pp. 795–797, 2011.

[5] I. Bisio, A. Delfino, G. Luzzati, F. Lavagetto, M. Marchese, C. Fra, and M. Valla, "Opportunistic estimation of television audience through smartphones," in *Performance Evaluation of Computer and Telecommunication Systems (SPECTS), 2012 International Symposium on*, 2012, pp. 1–5.

[6] I. Bisio, A. Delfino, F. Lavagetto, and M. Marchese, "A Television Channel Real-Time Detector using Smartphones," *IEEE Transactions on Mobile Computing*, vol. 99, no. PrePrints, p. 1, 2013.

[7] M. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668 –696, April 2008.

[8] C.-Y. Chiu, D. Bountouridis, J.-C. Wang, and H.-M. Wang, "Background music identification through content filtering and min-hash matching," in *ICASSP 2010*, 2010, pp. 2414 –2417.

[9] D. Douglas and T. Peucker, "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature," *Cartographica*, vol. 10, no. 2, pp. 112–122, 1973.

[10] M. Fink, M. Covell, and S. Baluja, "Social-and interactive-television applications based on real-time ambient-audio identification," in *Proceedings of EuroITV*. Citeseer, 2006, pp. 138–146.

[11] J. Haitsma, "A highly robust audio fingerprinting system," in *ISMIR 2002*, 2002, pp. 107–115.

[12] M. Naehrig, K. Lauter, and V. Vaikuntanathan, "Can homomorphic encryption be practical?" in *Proceedings of the 3rd ACM workshop on Cloud computing security workshop*, ser. CCSW '11, 2011, pp. 113–124.

[13] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," Tech. Rep., 2004.

[14] M. Ramona and G. Peeters, "Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection," in *ICASSP 2011*, May 2011, pp. 477 –480.

[15] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2, 2003, pp. II – 343–8 vol.2.

[16] C. A. Russell and C. P. Puto, "Rethinking television audience measures: an exploration into the construct of audience connectedness," *Marketing Letters*, vol. 10, no. 4, pp. 393–407, 1999.

[17] A. Smith and J. L. Boyles, "The rise of the Connected Viewer," *Pew Internet & American Life Project*, 2012.

[18] S. Suzuki *et al.*, "Topological structural analysis of digitized binary images by border following," *Computer Vision, Graphics, and Image Processing*, vol. 30, no. 1, pp. 32–46, 1985.

[19] W. L. Thomas, "Television audience research technology, today's systems and tomorrow's challenges," *Consumer Electronics, IEEE Transactions on*, vol. 38, no. 3, pp. XXXIX–XXLII, 1992.

[20] A. L. Wang, "An industrial-strength audio search algorithm," in *ISMIR 2003*, 2003.

[21] B. Wild and K. Ramchandran, "Detecting primary receivers for cognitive radio applications," in *DySPAN 2005*, 2005, pp. 124 –130.

[22] Z. Zivkovic, "Improved adaptive Gaussian mixture model for background subtraction," in *ICPR 2004*, vol. 2. IEEE, 2004, pp. 28–31.

Mobile Architecture for Identifying users in Ubiquitous Environments Focused on Percontrol

Valderi R. Q. Leithardt[1,2,3], Guilherme A. Borges[1], Iván M. Carrera[1],
Anubis G. M. Rossetto[1], Carlos O. Rolim[1], Cláudio F. R. Geyer[1]
[1] Institute of Informatics, GPPD, UFRGS
[2] Institute of Technology Senai - POA, FATEC
[3] University of the Vale do Rio do Sinos, UNISINOS
Porto Alegre, RS, Brazil
{vrqleithardt, gaborges, ivan.carrera, anubis.rossetto, carlos.oberdan, geyer}
@inf.ufrgs.br

David Nunes, Jorge Sá Silva
University of Coimbra
Coimbra, Portugal
dsnunes@student.dei.uc.pt; sasilva@dei.uc.pt

***Abstract*—In pervasive and ubiquitous environments, the process of identifying users can be done using various devices and techniques like radio frequency identifiers (RFID), bluetooth devices, smartphones and others. In this context, it is necessary to implement management techniques and control of these environments, as well as customize the environment characteristics to meet user preferences.**

In this work, we present a solution for control and manage various users, devices and environments. We base our research on Percontrol (a system for pervasive user management), which was only intended to identify users using Wi-Fi, and now it is capable of managing temperature, luminosity and other preferences, measured by a Wireless Sensor Network (WSN) embedded to Percontrol, and the data treatment is done by a Artificial Neural Network (ANN).

***Keywords*—Ubiquitous Communication; Intelligent applications; Artificial Neural Networks**

I. INTRODUCTION

In 1991, Mark Weiser [1] claimed that the most useful technologies would become so integrated onto our daily lives that it would be as if they just vanished, since using them would become second-nature. With this idea in mind, Weiser proposed a new form of computation, known as pervasive computation. By vanishing, it is meant that the device becomes imperceptible, i.e., the user no longer needs to worry about the technological resource in use. From this perspective, pervasive computing frees users from worrying about secondary tasks and allows them to focus on their primary task.

The development of microelectronics and wireless communications has been for great benefit for Wireless Sensor Networks (WSNs), in a pervasive/ubiquitous environment. WSNs are usually composed by a large number of wireless nodes spread through an area where the events or phenomenon to be monitored is located [2]. A sensor node is a device with reduced size meant to be cheap and easily deployed in risky locations or with difficult access. The small size of the nodes severely limits their hardware capabilities; sensors usually present many limitations with their battery, processing power, storage and communication interfaces [3]. These restrictions with sensor nodes affect the amount of nodes present in a given network, and said network's performance, leading to a need for developing communication protocols that could handle these limitations while being adequate for the application and its associated routing mechanism.

In these pervasive/ubiquitous environments, routing is a task handled by the sensors themselves and not managed by a specific routing device. Sensors can carry data throughout the network in an power efficient way through wireless radios, or light pulses cooperating with each other. Sensorial data is transmitted throughout the network in hops, i.e., a sensor transmits its data to the node that is closer to final recipient of the data, trusting that the transmission will have no interruptions. This transmission mechanism is useful to distribute the transmission's energy costs throughout the whole network, avoiding the need to use high power signals [2]. This makes the use of WSNs in pervasive/ubiquitous environments a more complex challenge, since these environments are populated with many other types of devices and communication mechanisms [1]. All this diversity in devices and communication types is managed by the pervasive/ubiquitous application; an example this kind of application is Percontrol [4], a pervasive system for controlling user attendance through mobile devices and wireless networks.

The main contribution of this work is an analysis of the viability to use Artificial Neural Networs (ANN), as an improvement of a previous work [5], by using a base technique to handle user profiles and manage devices on a pervasive environment also using WSNs to capture environmental data. The results from this analysis are innovative and serve as initiative to further research. The present work is organized as follows: section II presents the Percontrol system and introduces several important concepts regarding ANNs and WSNs. Section III describes some concepts of routing in WSNs. Section IV gives an overview of the state-of-the-art in the area while Section V presents the application scenario. Section VI contains information about an example of use and shows some preliminary results. Finally, Section VII gives a conclusion and alludes to future work.

II. PERCONTROL

Percontrol, as described in [4] and [5], is a system that automatically manages and keeps track of user attendance.

 ISBN: 978-1-61208-289-9

The system detects the entrance and exit of users within an academic or business environment. Percontrol also improves user discovery and localization service, within the local environment based on bluetooth, Wi-Fi and RFID identifiers.

In an academic environment, Percontrol monitors student attendance without teacher's intervention, and in a business environment, it controls the entrance and exit times of employees without the need of their intervention. By resorting to certain aspects of pervasive computing, the system developed in [5] can guarantee reliability and transparency in the detection and localization of users, resulting in minimal impact on both environments. The system identifies users through their mobile devices and controls the environment based on their previously-defined profiles. The user/device presence on a certain location generates an event, which is sent to a WebService listener application. The listener application first identifies the user within the environment and then configures the venue in accordance to the user's profile stored in a database. Environmental data such as temperature and luminosity are acquired through SunSpot wireless sensors [6] in conjunction with Arduino development kits [7].

In the work described in [5], an earlier version of Percontrol was already being developed, but the management of different user profiles and devices on a single environment was yet to be supported. To support this feature, an artificial intelligence module was necessary, which is now presented by using Artificial Neural Networks. Thanks to its new-found intelligence, the system is now capable of adapting the environment to fit the demands of several users, simultaneously. The ANNs receive data from each user, perform pattern recognition and, by resorting to previous event data, make a decision to configure the environment in a way that satisfies most users.

Solving problems with ANNs is an attractive approach. According to [8], the way these problems are internally represented in the network and the natural parallelism associated with ANNs architecture create the possibility of superior performance when compared with traditional machine-learning methods.

In ANNs, the usual procedure for problem-solving begins with a learning phase, where a set of examples is presented to the network from which are extracted the necessary characteristics to correctly represent the information provided [9]. These characteristics are then used to generate answers to the problem in question. An ANN is composed by several simple processing units connected through communication channels that have a certain associated cost associated. Each processing unit performs computational operations on the fed data that through its connections. The intelligent behavior resulting from an ANN comes from the interactions between the network's processing units.

The most important property of ANNs is the ability to learn from their environment and improve their own performance. The entire knowledge contained within the network is stored in the links between the networks artificial neurons; in other words, the networks behavior is mostly conditioned by the weights associated to the connections between individual neurons. The development of a computational solution that not only infers a system's current state but also predicts its future condition seems to be the solution for the environment management problem previously mentioned.

Considering the characteristics metioned above, a neural network is a prediction mechanism that seems to fit Percontrol needs since it can use raw contextual data, learn from it and predict configuration values that will better fit a future situation. The solution proposed by Silva [10] was used as basis to adapt a neural network to the current application scenario. This effort demanded an analysis of the characteristics of WSNs in order to find the best configuration and techniques to associate ANNs parameters to Percontrol's WSNs.

The main components of sensor networks are the phenomenon, the sensor and the observer. The phenomenon is the event of interest to an observer, and it is monitored and processed by a WSN. Several different phenomena can be observed concurrently by a single WSN. Monitoring is performed under certain criteria determining the performance requirements for the nodes to be used. The sensor device is used for collecting the data generated by the phenomenon and transmit it to the next node through wireless communication. Nodes respond to changes in the environment where they are currently in operation, e.g., climatic and eviromental changes. The characteristics and capabilities of the sensor depend on the application and each node's sensibility may differ, depending on the distance to or the exposure time to the event being monitored. The observer is the element that receives the information collected by the WSN; depending on the application's needs, there may be more than one observer at any given time [11]. WSNs can be classified according to their configuration, sensing mechanisms, communication media and processing type [2]. The type of routing is one of the most important research topics for reducing energy consumption and increase performance in wireless sensor applications.

III. Routing

Due to the specific features of their utilization, (e.g., their limited battery life), communication between nodes deployed in an external environment is subjected to many obstacles that may stop its effectiveness. Most applications that use WSNs communication are directed towards configurations with stationary characteristics, which differ from traditional ad-hoc networks; routing can have its focus on addresses or data, the latter being a more recent approach, which main advantage is the fact that data aggregation manages to reduce the amount of packets exchanged in the network, and thus, reduce the overall energy consumption of the system [12].

Another approach for routing in WSNs is related to the TinyOS operating system, developed by the University of Berkeley. TinyOS does not use a complete addressing scheme like traditional IP-based ad-hoc networks [12], [13]. Traditional ad-hoc networks implement a routing protocol that constructs a Minimum Spanning Tree from an initial sink or access point [13]. Routing in WSNs can be categorized as: based on Medium Access Control (MAC), plane routing, hierarchical routing and geographical routing. There is also Geographical and Energy Aware Routing (GEAR) and Greedy Perimeter Stateless Routing (GPSR).

For this project, SunSpot wireless sensors were used for the tests. The SunSpot sensors use the ZigBee specification for wireless protocols, which use Ad hoc On-Demand Distance Vector Routing, that can be categorized as based on MAC.

In this section, we only mention the categories of routing protocols used in WSNs. An exhaustive and comparative description of routing protocols for WSNs is presented in [14], which was used as theoretical basis to formulate the tests and obtain the preliminary results that serve as validation for the present work.

IV. State-of-the-art

As previously mentioned, this article's proposal refers to an implementation of ANNs in a pervasive control system for pervasive/ubiquitous environments in conjunction with WSNs (composed by SunSpot wireless sensors for testing purposes). The system deals with concurrent data and adapts the environment's configuration to match its users' profile, according to the devices used to perform login onto the environment.

In [15], we can find several academic works that focus on the detection of devices in many application areas within the domain of pervasive systems and wireless sensor networks, as introduced by Mark Weiser. The work in [15] shows a system that implements several integrated systems that allow the monitoring of patients in their own home; however, it does not address the integration of different forms of communication within the same environment nor does it present mechanisms for handling user profiles. In [8], an extensive proposal that focus on elder users in ubiquitous environments is presented; the Internet is shown as a form of interaction but this work does not specify which languages, devices and forms of communication should be used. In [16], it was proposed a project that consisted on the creation of an artificial intelligence system capable of leading a robot through certain routes. The system was tested using a robot vehicle with freedom of horizontal movements. The environment was scanned using sensors where data were fed to the artificial intelligence system, and where an algorithm evaluated the situation and acted on the locomotion engines, making the robot avoid possible obstacles and follow the desired route.

Despite being an innovative proposal, a target audience and usage scenario are not defined, therefore not adapting to the proposed model. Fonseca [17] defines a solution that uses neural networks and RFID, partially achieving the desired results.

Percontrol presents all necessary features for pervasive/ubiquitous environment usage, considering different types of devices, and forms of communication. This work presents a significant improvement over previous versions in [4], and [5], while addressing all the primary objectives previously set. Table 1 presents the main characteristics and functionalities of works researched in the available literature.

V. Application Scenario

The application scenario shows the potential pervasive/ubiquitous computation for improving efficiency in workplaces. It also attempts to illustrate different possible perspectives one can have on a single pervasive scenario. Initial versions of Percontrol did not anticipate the use of WSNs or ANNs [6]; such versions only intended to automatize student attendance tasks in classrooms. Percontrol focuses on performing automatic detection of people that enters an environment.

This work proposes an extension of the work developed in [6], and [7], increasing the pervasive functionalities available in this user-tracking system, with the objective of increasing control over environmental conditions through user's mobile devices. Using SunSpot wireless sensors [14] and Arduino kits [7], Percontrol can sense and manage the temperature and luminosity of an environment; and by using ANNs, the system can also attempt to adjust the values of these environmental properties to fit user preferences and the number of people in the environment, turning it into an intelligent location.

The sequence diagram in Fig. 1 shows the primary interactions between all parts of the system, as well as the messages exchanged since a user is detected until the environment adapts to its preferences. When the application detects the entrance of a device in the environment, a Web service that manages the associations between users and devices is accessed. The device is identified through its BDA (Bluetooth Device Address), Wi-Fi or RFID. The application maintains a module called BlueID which holds a list of all devices that were ever detected. Each time the application verifies the devices currently present in the environment, it performs a comparison with the previously stored list; newly detected devices generate an "entry" event while missing devices are associated with an "exit" event.

When accessed, the webservice returns the username to the application, and also associated device resources and personal preferences through the HTTP protocol and an XML format message. The application also communicates with the SunSpot sensors to fetch the room temperature, luminosity, humidity or other environmental data that may be used at a later time. The following format was used to communicate with the sensor: messageID : sensorTypenn. Both messageID and sensorType are numerical values. The messageID field is used to associate sensor response with the respective BlueID request, an important step since communication is asynchronous.

The sensorType represents the type of data being sent (luminosity, humidity, etc); the "\n" character is used to mark the end of a message, while the ":" character is used as a data separator. The current environmental state is compared with user's environmental preferences in order to decide the needed changes to be done. After a decision is reached, the environment sends commands to the actuator controllers, connected through USB to an operating computer, to change the environmental characteristics (e.g., turning the A/C unit on and change the room temperature). The extension of Percontrol's functionalities translated into a more complex architecture shown in Fig. 2.

Initially, the prototype application and its respective transmitters were tested with a Windows operating system, an environment that benefited from the use of SunSpot sensors [6] and Arduino hardware [7]. There were many other advantages that led us to choose the Arduino boards, namely the embedded input/output ports, low cost and strong modularity. The main idea behind the use of Arduinos was to test their viability for middleware development in pervasive environments, not excluding the possibility of having these boards completely replace several individual sensors for an integrated, single-

 ISBN: 978-1-61208-289-9

TABLE I. COMPARISON BETWEEN RELATED RESEARCH

Support and Control Works	Scope	Devices Transmission
ARAUJO	ubiquitous computing	Not defined
CARVALHO	ubiquitous computing	Environmental sensors, measuring devices, electronic devices, Internet
RIBEIRO	Neural Networks	Sensor end-of-course and sonar
FONSECA	Neural Networks, pervasive systems	RFID
PERCONTROL	Neural Networks, systems, pervasive/ubiquitous	Wi-Fi/ RFID / Bluetooth / Wireless Sensor Networks

Support and Control Works	Public	Routing Protocol	Language
ARAUJO	Not defined	Not defined	Not defined
CARVALHO	Elderly people who require constant monitoring of their health	Not defined	Not defined
RIBEIRO	Not defined	Not defined	C/C++/C#, Java
FONSECA	Laboratory of the University	Not defined	Delphi
PERCONTROL	Educational and business environments	ZigBee	.Net

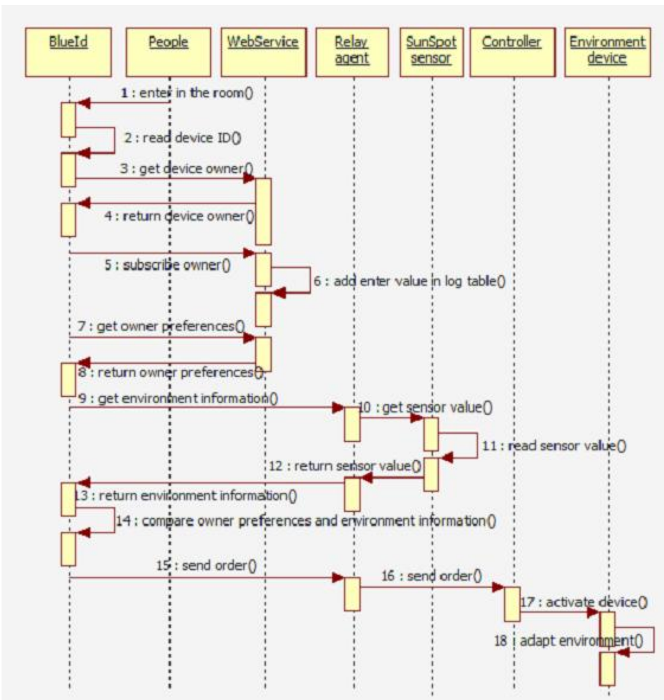


Fig. 1. Sequence Diagram

board solution connected to a computer. Fig. 3 illustrates the operation of Percontrol.

From Fig. 3, it can be seen that a central controller is missing for allowing the exchange of profiles in the environment, according to user preferences. The main challenge here is controlling the number of parameters generated by the application while using WSNs and ANNs; the number of parameters increase with the amount of nodes and this means larger energetic and resource demands, as well as an more complex neural network processing, which may compromise the ANN's response time.

In the previous version of Percontrol [5], there was no control of the different ways for communicate and indentify users. As it is shown in Fig. 3, it is possible to visualize that the used can be identified through wireless communication protocols, and later this information is treated individually, based on parameters defined by the ANN. Said parameters can be definitions of user preferences as luminosity, temperature and the actual user indentification in the environment. In Fig. 3, there also can be seen the utilization of temperature

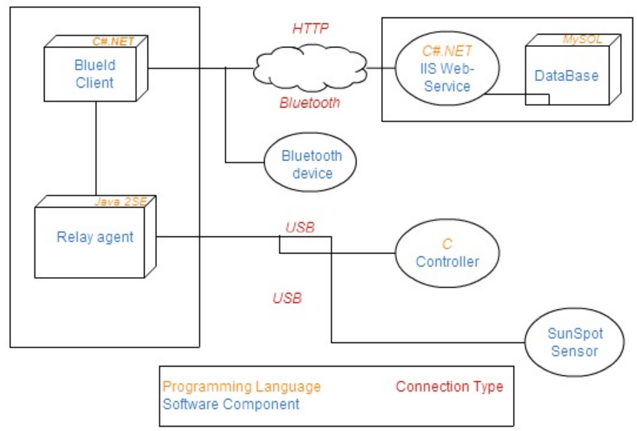


Fig. 2. System's Architecture

controller, which has access to the air conditioning system. It is possible also to update the location of several users in different environments, by sending messages to a REST web server that stores users and environment data and preferences.

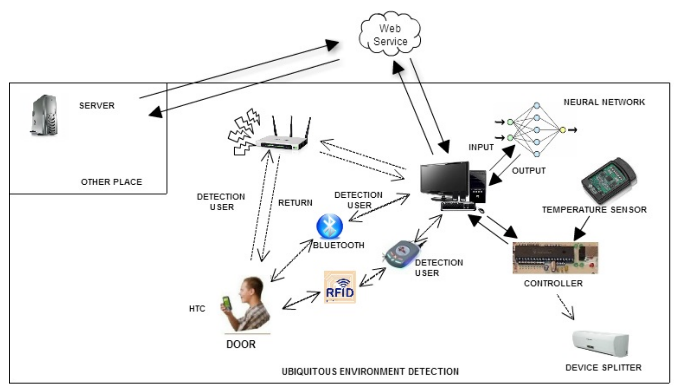


Fig. 3. Functioning of Percontrol

An efficient monitoring of the network performance is necessary to guarantee a good quality of service. An example of this can be observed in the amount of time necessary to obtain information regarding a monitored environment; if it takes too long to obtain environmental information, this information may lose its value from an application perspective. Management of performance may provide means for the application to define proper quality metrics. These may be influenced by node density, exposition time, amount of dissipated energy and other factors, as specified in the description of the study presented in section III.

A mechanism that evaluates the level of importance of information is necessary for the management of quality of service. For example, a sensor detecting a temperature of 20° C during spring is a normal occurrence, but a the measurement of 50° C under the same circumstances is an abnormal event, which would turn it into a relevant situation that would require extra attention; it would implicate an artificial intelligence mechanism that could compare the abnormal value with other measurted values by other sensors to see if the information is reliable and determine the proper course of action. Information that is of great importance to the normal function of the system should imply a greater effort for proper delivery. That is, energy consumed in communications should vary depending on the importance of the data.

Another relevant management aspect concerns the installation of ad-hoc networks in unknown areas, where the behavior of wireless communications can be highly unpredictable, with high error-rates and considerable delays, which may compromise the value of the information provided to the application. Performance management usually includes quality assurance, performance monitoring, control and analysis [2]. The QoS management process begins with the detection of performance degradation and ends when the source of the problem is ceased or removed. In between, the process has many intermediary stages of situation analysis [18]. Initially there were used only 2 Sunspot tests kits containing two wireless sensor nodes communicating with each base station connected to the computer via USB. With that comes the need to conduct a comparative study of routing protocols for use in different environments and with numbers over wireless sensors. To this end, several techniques exist to treat this problem and also allied service discovery, one of the most important, by the Service Location Protocol (SLP) [14], which basically consists of maintaining a directory that contains the services available to whom it is offering them to. However, it is necessary to study thoroughly the operation of routing protocols in order to verify the protocol that best fits the pervasive control system, it is not in the scope of this work the study of routing protocols. Therefore, for this work there were conducted only some tests to validate the survey and obtained results that demonstrate the feasibility of work and their implementation and use with Percontrol, contributing to the improvement of the system and data so that other researchers can use it.

VI. Example of Use and Preliminary Results

One issue when having multiple users on the same system is the problem of concurrent data; e.g., the configuration of an air-conditioning unit may be influenced by every user that registers in this environment, since each user might have its own preferred temperature, and the temperature itself is general throughout the whole environment. In order to bypass this problem, the decision-making process for selecting the best "average" temperature must take into account the individual preferences from all users within the environment. A widely used solution [10][19][20] that has shown great results is the use of AI, in particular ANNs [9]. A neural network bases itself on real data that has occurred in the past and has been stored within the system for posterior access and use.

The main objective of this work is not the choice of proper protocols or AI tools, but the creation of novel help mechanisms for Percontrol. Our choice for an AI mechanism dwelt on neural networks, while communication mechanisms were TCP/IP and ZigBee. These choices are supported by published works in routing protocols [14], artificial intelligence [20], and comparison and use of neural networks [21] and [22].

The neural network loads the entire history of a device being handled within the environment, and uses its historical data as training, in order to identify decision patterns that were assumed in a recent past. Considering our air-conditioning example, these patterns include the temperature that each user wants for a certain environment and what temperature was actually used when all users were taken into consideration. This type of analysis is crucial for the network's decision-making process. Fig. 4 presents part of the source code used to define the desired and assumed temperatures. These values are fixed for testing, but in a real scenario they are fetched from a database or an archive.

```
// Initializes with
//   1 randon
//   3 perception neurons (entries)
//   6 hidden layer neurons
//   1 output neuron (output)
net.Initialize(1,3,6,1);
// Learning rate
net.LearningRate = 3;
// iterations
iterations = 10000;
// Train the network
net.Train(input, output, TrainingType.BackPropagation,
iterations);
```

Fig. 4. Part of the neural network's source code

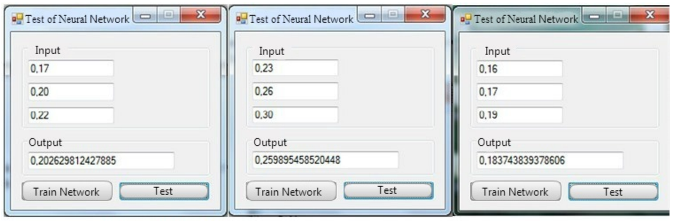
The code shown in Fig. 4 is used to train the neural network. After training phase the next step is to test the network to determine if it is well-suited to solve the problem of finding the ideal temperature using past event data; in order to perform the testing, a graphical interface was developed. The interface receives the values for current data and returns the ideal temperature estimation, as shown in Fig. 5.

```
private void Treina()
{
        int iterations;
        double[][] input, output;
        #region Initialization

        // Instantiates the network
        net = new NeuralNet();

        // Assemble the historic Network for Train-there
        input = new double[5][];
        input[0] = new double[] { .30, .25, .27 }; // A very cold day
        input[1] = new double[] { .17, .18, .19 }; // A hot day
        input[2] = new double[] { .16, .20, .22 }; // A cold day
        input[3] = new double[] { .18, .19, .24 }; // A cold day
        input[4] = new double[] { .26, .28, .30 }; // A very cold day

        output = new double[5][];
        output[0] = new double[] { .26 }; // expected temperature
        output[1] = new double[] { .18 }; // expected temperature
        output[2] = new double[] { .20 }; // expected temperature
        output[3] = new double[] { .22 }; // expected temperature
        output[4] = new double[] { .27 }; // expected temperature

        #enderegion
```

Fig. 5. Training the ANN

After the neural network's training, we could identify the network's response time after a user enters the environment, as shown in Fig. 6.

Fig. 6 presents response times of the ANN for the cases with 1, 3 and 10 distinct users identified by the pervasive




```
private void test()
{
//Defines the parameters to test the network
net.PerceptionLayer[0].Output = Convert.ToDouble(txtE1.Text)
net.PerceptionLayer[1].Output = Convert.ToDouble(txtE2.Text)
net.PerceptionLayer[2].Output = Convert.ToDouble(txtE3.Text)
//Pulses network for feedback
net.Pulse();
txtS.Text = net.OutputLayer[0].Output.ToString();
```

Fig. 6. Neural network's performance and response time

environment, where X axis represents the number of users and Y axis represents the elapsed time. With this prototype we obtained not very satisfactory results, for example, for a single user, the ANN took 3 seconds to process the information contained in the user's profile, returning an average temperature with a value equal to the one defined by the user (since it is just a single person). For three users, the neural network took 5 seconds to respond, and for ten users it took 13 seconds. In Fig. 7, a screenshot shows information on the users identified by the system, as well as on the devices associated with them.

To perform the identification of different environments we used an Arduino Duemilanove [7]. It is a microcontroller board based on ATmega328, with 14 digital input/output pins, 6 analog inputs, a 16MHz crystal oscillator, a USB connection, a power jack, an ICSP header, and a reset button. The board contains the necessary assets to support the microcontroller and its use is as simple as connecting it to a computer with a USB cable or powering it with an AC-to-DC adapter or battery.

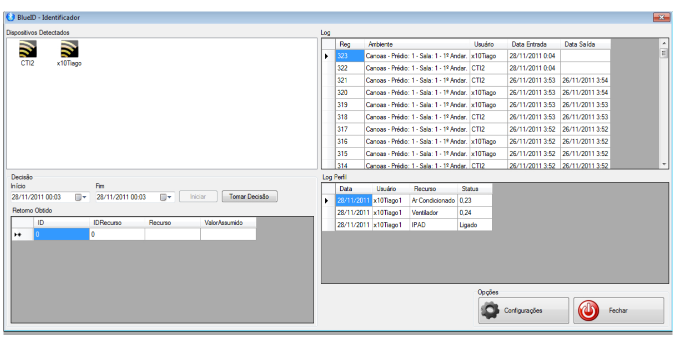


Fig. 7. User identification screen

With this board it was possible to detect devices via Bluetooth, Wi-Fi and, after being integrated with an appropriate card reader, RFID. The reader fetched the RFID a card's serial number that can be cross matched with the user's registration on the database. For this purpose, a RFID card reader model YHY502CTG was used in conjunction with the Arduino board. After obtaining the necessary application data and performing the necessary adjustments the system aws validated using a didactic MultiPIC development Kit, which possesses its own internal programmer. The didactic development Kit was connected to a stepper motor that simulated a ventilator. The stepper motor can be put in action with different speeds; initially we defined 3 different speeds that corresponded to 3 different profiles. Fig. 8 shows a picture of the assembled device.

Fig. 8. Validation equipment

The software used for simulating a ventilator with 3 speeds and controlling the board and stepper motor was developed in C language using the development software from Microchip MTLab, and transferred to the microcontroller with the IC-Prog 1.06C, the software used to compile the source code onto the MultiPIC kit's processor.

On the performed tests the ANNs computed the average temperature from the user profiles and used current environmental information from the SunSpot sensors to correctly manage the ventilation system. From these tests we conclude that Percontrol managed the pervasive environment in a satisfactory manner and that the primary objectives of this research were met, although there is still much room for improvements.

VII. Conclusion and Future Work

The choice of a routing protocol should be based on the restrictions inherent to the observed phenomena, which defines the monitoring environment and minimum requirements for the sensoring hardware. The main purpose of the work and results herein presented are to be used as basis for future research work that improves the area of pervasive/ubiquitous computation, namely in the use of ANNs with low response times in environments with thousands of users.

The performed tests confirmed the viability of device detection with Wi-Fi, Bluetooth and RFID, an improvement over previous Percontrol versions. Nevertheless, there is still some latency in registering new devices on the system, which may be reduced by further adjustments of the parameters sent to the ANNs.

The main contribution of this work is an improvement of [5], using an artificial intelligence technique, as well as treating several users in environments with heterogeneous devices. Using ANNs made possible to automatically treat user profiles, which is a new feature of Percontrol. This work also represents a significant contribution since it covers different areas and technologies within pervasive computation. On top of several parameters and definitions still waiting implementation, future

 ISBN: 978-1-61208-289-9

work needs to define a model and implement a mechanism for privacy control, for both users, their devices as well as for the pervasive environment itself.

VIII. ACKNOWLEDGMENTS

This work is associated with the UbiArch project - Ubiquitous Architecture for Context Management and Application Development of the Federal University of Rio Grande do Sul. We would like to thank the Conselho Nacional de Desenvolvimento Cientifico e Técnologico CNPq for its support. We would also like to thank Sun Microsystems for donating the SunSpot development kits, which allowed us to perform the necessary tests for our study. We kindly thank the students Sebastien Skorupski, Thiago Tomasel, Marceli Stein and Jeison Nunes da Silva, who contributed towards the completion of this work.

REFERENCES

[1] M. Weiser, "The computer for the twenty-first century," *Scientific American*, pp. 94–10, Sep 1991.

[2] L. Ruiz, J. Nogueira, and A. A. F. Loureiro, "Manna: a management architecture for wireless sensor networks," *Communications Magazine, IEEE*, vol. 41, no. 2, pp. 116–125, 2003.

[3] K. Kwon, M. Ha, T. Kim, S. H. Kim, and D. Kim, "The stateless point to point routing protocol based on shortcut tree routing algorithm for ip-wsn," in *Internet of Things (IOT), 2012 3rd International Conference on the*, 2012, pp. 167–174.

[4] A. G. M. Rossetto, C. O. Rolim, V. R. Q. Leithardt, M. A. R. Dantas, and C. F. R. Geyer, "An adaptive fault tolerance approach to enhance the execution of applications on multi-cluster grid configurations from mobile grid interfaces in wireless networks." *IJHPSA*, vol. 3, no. 4, pp. 202–215, 2011.

[5] V. Leithardt, C. Rolim, A. Rossetto, C. Geyer, M. Dantas, J. Silva, and D. Nunes, "Percontrol: A pervasive system for educational environments," in *Computing, Networking and Communications (ICNC), 2012 International Conference on*, 30 2012, Feb. 2 2012, pp. 131–136.

[6] O. Labs, "Sun spot world," May 2013, last Accessed May 2013. [Online]. Available: http://www.sunspotworld.com

[7] A. BRASIL, "Arduino brasil," Apr. 2013, last Accessed April 2013. [Online]. Available: http://www.arduinobrasil.com/

[8] S. T. Carvalho, A. Copetti, and O. G. Loques, "Sistema de computação ubíqua na assistência domiciliar á saúde," *Journal of Health Informatics (JHI)*, vol. 3, no. 2, pp. 51–57, 2011.

[9] C. O. Rolim, A. G. Rossetto, V. R. Leithardt, and C. F. Geyer, "Analysis of a hybrid neural network as underlying mechanism for a situation prediction engine," *Journal of Applied Computing Research*, vol. 2, no. 1, pp. 22–31, 2012.

[10] R. Silva, V. R. Leithardt, J. S. Silva, C. Geyer, J. Rodrigues, and F. Boavida, "A comparison of approaches to node and service discovery in 6lowpan wireless sensor networks," in *Proceedings of the 5th ACM symposium on QoS and security for wireless and mobile networks*. ACM, Oct 2009, pp. 44–49.

[11] C. H. P. Augusto, C. B. Carvalho, M. W. R. da Silva, and J. F. de Rezende, "Reuse: A combined routing and link scheduling mechanism for wireless mesh networks," *Computer Communications*, vol. 34, no. 18, pp. 2207–2216, 2011.

[12] S.-H. Lee, D.-W. Kum, K.-W. Lee, and Y.-Z. Cho, "Energy-efficient data-aware routing protocol for wireless sensor networks," in *Mobile Ubiquitous Computing, Systems, Services and Technologies, 2007. UBICOMM '07. International Conference on*, 2007, pp. 101–106.

[13] A. T. A. Gomes, A. Ziviani, L. Lima, and M. Endler, "Dichotomy: A resource discovery and scheduling protocol for multihop ad hoc mobile grids," in *Cluster Computing and the Grid, 2007. CCGRID 2007. Seventh IEEE International Symposium on*, 2007, pp. 719–724.

[14] W. Khan, N. Saad, and M. Aalsalem, "An overview of evaluation metrics for routing protocols in wireless sensor networks," in *Intelligent and Advanced Systems (ICIAS), 2012 4th International Conference on*, vol. 2, 2012, pp. 588–593.

[15] R. B. ARAUJO, "Computação ubíqua: Princípios, tecnologias e desafios. departamento de computação – universidade federal de s. carlos (ufscar)," Sep. 2003, last Accessed September 2013. [Online]. Available: http://im.ufba.br/pub/MAT570FG/LivroseArtigos/045_AraujoRB.pdf

[16] D. R. d. S. Ribeiro, K. L. d. S. Machado, and L. C. A. Perereira, "Rede neural perceptron embarcada em robótica," *Engenharia de Computacão em Revista*, p. 1, 2008.

[17] L. O. Fonseca, "Sistema de localização rfid de usuários visando à racionalização de energia em ambientes inteligentes," *Universidade de Brasília*, p. 1, Feb 2011.

[18] V. Pereira, J. Silva, J. Granjal, R. Silva, E. Monteiro, and Q. Pan, "A taxonomy of wireless sensor networks with qos," in *New Technologies, Mobility and Security (NTMS), 2011 4th IFIP International Conference on*, Feb 2011, pp. 1 –4.

[19] A. de Pádua Braga, *Redes neurais artificiais: teoria e aplicações*. LTC, 2007. [Online]. Available: http://books.google.com.br/books?id=R-p1GwAACAAJ

[20] H. Qiao, J. Peng, Z.-B. Xu, and B. Zhang, "A reference model approach to stability analysis of neural networks," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 33, no. 6, pp. 925–936, 2003.

[21] R. Silva, J. S. Silva, C. Geyer, V. Leithard, and F. Boavida, "Use of gps and 6lowpan in mobile multi-sink wireless sensor networks - issues and feasibility," in *8th International Information and telecommunication Technologies Symposium (I2TS)*. IEEE R9 Latim America, 2009, pp. 154–160.

[22] C. O. Rolim, F. Schubert, A. G. Rossetto, V. R. Leithardt, C. F. Geyer, and C. B. Westphall, "Comparison of a multi output adaptative neuro-fuzzy inference system (manfis) and multi layer perceptron (mlp) in cloud computing provisioning," *X Workshop de Computação em Grade e Aplicações - WCGA*, 2012.

Security Adaptation Based on Autonomic and Trust Systems for Ubiquitous mobile network and Green IT

Tewfiq El-Maliki
HEPIA Geneva
University of Applied Sciences Western Switzerland
tewfiq.elmaliki@hesge.ch

Jean-Marc Seigneur
University of Geneva
Advanced Systems Group
Jean-Marc.Seigneur@trustcomp.org

Abstract— Security in Wireless Sensor Network (WSN) or Internet of Things has become a hot research topic due to their wide deployment and the increasing new runtime attacks they are facing. Thereby, applications are unaware of what security mechanisms are as well as dynamic changing attacks. Accordingly, the concept that must cope with this new security challenge has to satisfy an overall performance such as power consumption, being actually a key issue for internet of things. This objective is completely compatible with green computing (Green 1.0). This research investigates methods mainly based on autonomic computing and adaption security to build a framework capable of determining appropriate security means for a highly dynamic wireless network with respect to context-aware, self-management, self-optimization and self-protection paradigms of autonomic system. Trust computing was the means used to mitigate the influence of attackers. Extensive simulations using agent based approach have been also conducted for a case study of pollution detection in Geneva city in concordance with Green 2.0. We have proved the performance of the framework in the case of mobile sensor network in the presence of different mobile attacks. The results clearly show that SARM is efficient in terms of survivability, overall network utilization and power consumption.

Keywords-Security adaptation; Autonomic; Trust; Green IT

I. INTRODUCTION

Ubiquitous computing is becoming more popular than ever, mainly by Internet of Things highlighted by large-scale embedded sensor devices. Indeed, trends are best observed by sensors and Radio Frequency Identification (RFID) in everyday life such as cars, refrigerators or even animals to track some useful information about them. This overflow of information coming from sensing and communication is fitting squarely within complex system and autonomic computing [1]. Furthermore, mobility of these devices exposes them to different security vulnerabilities.

However, there will be no acceptance of these new paradigms without security methods, which are main concerns of industry and consumers.

The increasing complexity of communication system and also of attacks makes the conventional static security almost obsolete. Whereby, it is resources consuming to maintain required security level. The overhead cost reaches high rate. Thus, new mechanisms need to be set up to achieve principle of adaptation security based on autonomic system in the field of security of Internet of Thing.

On the other hand, sustainable development has propelled the efficiency in using resources, which is one of the fundamental principles of green computing (Green 1.0). Indeed, it studies practices of efficient use of computing resources, motivated by reducing the use of materials harmful to nature, maximizing energy efficiency and life product [2]. In addition, Green IT (2.0) [3] initiative is also a good mean to contribute indirectly to catalyze economy of energy by using smart protocols and communications to reduce emission of other technologies and business sectors. In short, we would like to minimize the overall energy consumption when using security mechanisms (Green 1.0) as well as using ICT to contribute indirectly in reducing gas emission (Green 2.0).

In this paper, we introduce our security autonomic framework based on the concept of adaptation security, Green IT and explain its components and functionalities. In addition, we have evaluated the framework in the case study of pollution detection in Geneva based on mobile WSN to manage transport traffic and thus energy. In Section 2, the related work is reviewed. Section 3 gives the problem statement, highlighting the motivation of our work. Section 4 shows our proposed framework and Section 5 explains our Green-SARM, experiments and simulation implementation to validate our Framework for the case study. Our simulation results and performance analysis are presented in Section 6 and our conclusion is to be found in Section 7.

II. RELATED WORK

Security in sensor networks is complicated by the constrained capabilities of sensor node hardware and the properties of the deployment [4]. Individual sensor nodes in a WSN have the inherent limitations in resources, which make the design of security procedures more complicated. Each of these limitations is due in part to the two greatest constraints: limited energy and physical size [5].

Other security issues include security-energy assessment, data assurance, survivability, Trust, end to end security, Security Support for data centric sensor networks and node compromise distribution [5]. It is very important to study these areas due to a sensor network's special character, such as battery limitation, high failure probability nodes, easier compromised nodes, unreliable transmission media, etc.

 ISBN: 978-1-61208-289-9

Mobility greatly exacerbates the problem. Until now, there have been only a few approaches available, and more studies are needed in these areas. Furthermore, Trust [6] is a good path to explore because it gives in some cases better results. That's why we have carried out extreme simulation based on Trust to assess security requirements of our case study.

The best way to overcome this constraint is to implement the Framework capable of adapting security to the context based on the ideas similar to those described in [7] and consequently having an overall security control. We have been inspired by the concept of autonomic computing that was develop by IBM [1] to propose a new Security Adaptation security Frameworks based on autonomic computing and "Green IT" [2,3].

III. Motivation

A. General Motivation

IT systems are resource consuming especially battery power [8]. This is because a lot of smart and useful mobile applications need significant power consumption, such ones using geo-localization. However, all actors avoid using or arming security means in attempt to reduce power and resource consumption. Accordingly, they put users' security at risk and thus facilitating also distributed attacks to be successful. In addition, in the smart 2020 report [3], it is illustrated the scale of the opportunity for IT to drive efficiency across the economy and deliver emission savings of 15% of global emissions in 2020. One of the biggest challenges is overcoming the lack of information about the emissions impacts of products and services, especially in the context of complex configurations and integration.

Furthermore, conventional security mechanisms such as cryptography are unable to protect against new attacks such as jamming mainly in WSN. Ref. [9] talk about hard security for conventional security mechanisms such as authentication versus soft security measures for trust and reputation systems.

Investigation of new techniques to deal with the trade-off between the use of security mechanisms and performance are highlighted as essential to computing [10, 11].

We also argue in this article that the spare processing and transmission resources are wasted in mobile environments if security is over-provisioned. Hence, the trade-off between security and performance is essential in the choice of security services. Adaptive security mechanisms are also found in flexible protocol stacks for wireless networks [12], context-aware access control systems [13] and security architectures [14]. This prompted us for the implementation of a completely reconfigurable architecture [15], which is fundamental to adapt the architecture to the terminal and network variability of the context and particularly in the security field [16]. J-M Seigneur [6] has introduced autonomic security pattern in his security design but only at the authentication level.

In [17], the author listed the main and typical problems for the security in complex system.

a) Inefficient and inadequate usage of available security methods and tools

b) Scattering of resources when trying to solve a lot of special security problems at the same time

These problems need efficient solutions, which lead to high demand for adaptive security methods.

We propose a generic framework called Security Adaptation Reference Monitor (SARM) as a compelling solution for this problem, because it uses autonomic paradigm and is developed especially for highly dynamic wireless network.

IV. Framework Description

We would like with SARM to fine-tune security means as best as possible taking into account the risk of the current environment and the performance of the system especially regarding the optimization of its energy consumption. All these are under policies and user real time intervention constraints. Thereby, our system differs from others by its [10]:

a) autonomic computing security feedback control system,

b) dynamic and evolving security mechanisms related to context-monitoring,

c) explicit energy consumption management,

d) dealing with mobility of attackers

The concept of isolating various functions and restricting their access to specific system can also be applied to security in wireless environment integrated in the operating system itself. The best way to overcome the nonrealistic constraint of implementing the framework in each communication program is to integrate it in the kernel and consequently having an overall security control. Thus, all communication programs go through SARM at some stage in order to gain access to communication resources.

Our framework could work as a cross-layers program and thus it is not related to any layer. However, the best place to implement is in the kernel to avoid any compromising overall security.

To reduce the system complexity and to make the system incremental, we propose a feedback loop framework as introduced in [18] at the authentication level, that is, the system automatically tunes to its best configuration based on the current monitored context, thus avoiding any static decision making. Hence, we split SARM into two units with feedback loop. One unit called management or monitor unit is for monitoring the context by evaluating and analyzing risks, performances, and energy consumption, which are significant for detecting attacks and tuning the adequate security means using the second module called functional unit.

Security means are defined as any algorithm or mechanism that could ensure security. We have carried out a good synthesis survey on identity management security initiative used as means in [19].

We have depicted in Fig. 1 the different components of SARM and their interconnections.

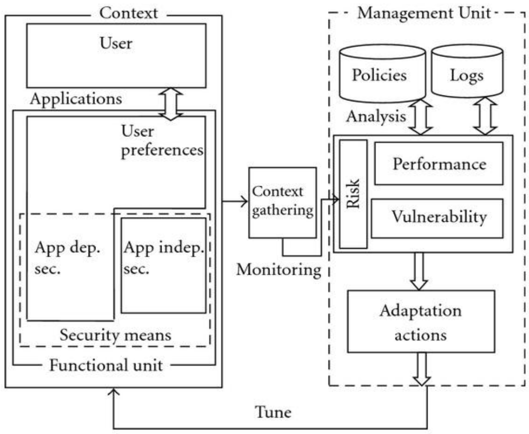


Figure 1. *SARM*.High level view

Security means can be application dependent such as a localized trust [20] or a distributed trust [21]. In [10], we have explicit all the details about SARM.

V. Green-SARM

We have used an adapted version of SARM called Green-SARM to the application domain of wireless sensor network applied to a case study of pollution detection in the city of Geneva, which is an application for both Green 1.0 and 2.0 initiatives.

A. *Sensors Energy Consumption*

A typical sensor node processor is of 4 MHz at 916 MHz frequency [5]. Table I shows that receiving costs almost half the energy of sending, which are almost 5 to 25 time average computing energy. That is why, our computing energy for trust and detection is really insignificant compared to transmission.

TABLE I. CHARACTERISTIC DATA FOR THE MICA2DOT SENSOR(

Field	Value
Effective data rate	12.4 kbps
Energy to transmit (5dBm)	59.2 µJ/byte
Energy to receive	28.6 µJ/byte

A. *The Main Problem : use case*

Geneva is an international financial city, and worldwide centre for diplomacy and for UN agency. It is the most popular and the second most populous city in Switzerland.

During summer, traffic jam is making the city unfortunately polluted. To apply Green IT 1.0 and 2.0 objectives, we would like to minimize this pollution by implementing:

a) A sensor network will be deployed based on a cheap Zigbee transmitter to monitor the level of pollution (CO2).

b) Every transmitter is placed in a car or a pedestrian

c) There are many fixed base stations that collect information about CO2 from sensors

d) Some information is exchanged within a fixed range and an evaluation of data consistency is rated as a trust

e) Obtained data will be aggregated by weighted average according to the trust on collected information

f) Therefore, recommendations will be overspread to conductor and walker to avoid the places where the pollution overpasses a certain level and simultaneously a trust about every transmitter will be send to the participant.

g) Finally, the traffic is managed in real time to initiate action plans crescendo capable of reducing the emission and thus avoiding in advance pollution peaks.

B. *Implementatin of Green-SARM*

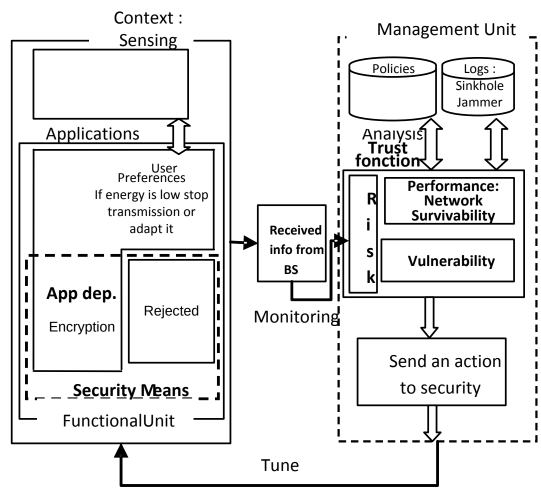


Figure 2. Adapted SARM.: Green-SARM

The control system of SARM Framework is ideal for a collaborative environment where the decision making trust function for the security must interact with other users to find the adequate decision. In Fig. 2, we describe module by module, how SARM is applied to the application domain of our validation, becoming the Green-SARM version. First of all, the security means are tuned efficiently by SARM. However, the application could use or not authentication and encryption with all rejected sensors. The application preference is to maximize the usage time whilst keeping enough security. The gathering context module is used to collect and distribute trust values between the Base Station (BS) and Nodes (sensors). These values represent the trust of a sensor about its neighbors (summarized in Table II)

TABLE II. BEHAVIOR AND RECOMMENDED VALUE SENT BY BASE STATION TO SENSOR UNDER SINKHOLE AND JAMMER ATTACKS

Sensor Behavior over neighbors According to a scenario	Recommended value to Sensor
Normal	Good (1) = Pos
Sinkhole or Jamming to neighbors “Who are you”	Bad (-1) = Neg
Unkwon	Neutral (0.5)= Neu

Attackers are detected when they are within a given range to the BS. The network is configured with messages following a given communication protocol to establish a connection as describe in Fig 3. In fact according to it, only the BSs and the attackers send messages "Who Are You?" As the BSs are too far from each others, if a message of this type is received it means that it is send necessarily by an attacker. In this case, the base sends a "Bad" message to everyone to indicate an intruder is present and BS updates its trust in a centralized database.

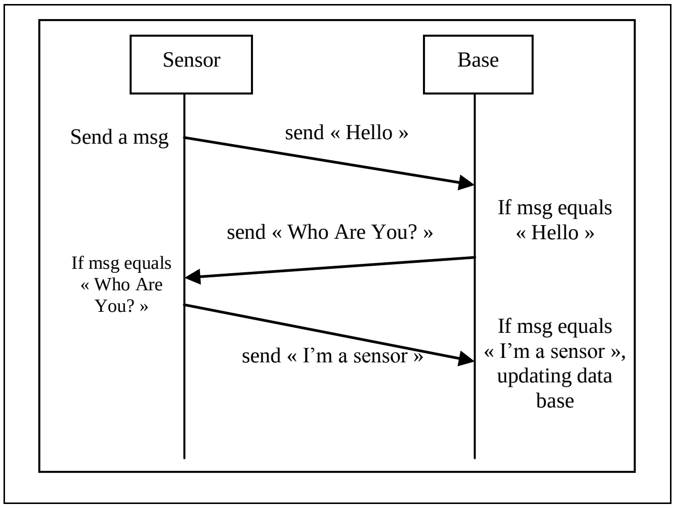


Figure 3. Communication protocol based on soft security

The values are sent to the management unit for analysis using a Trust Function (TF) that will assert the fact which algorithm has to be used or not. In addition, the performance is fixed as energy saving in accordance with Application Preference, which is lifespan maximizing.

One of our used TF is explained in "(1)":

$$TF_i = (2*Pos-2*Neg)/ (2*Pos+2*Neg+ Neu) \quad (1)$$

For all j sensors

if (TF_i > **threshold**)

Accept connection

else{ TF_i =< **threshold**}

then {rejected and use encryption and authentication if required}

End for

The TF will be used to calculate the weighted average of pollution values gathered from sensors and it will be also used to minimize overall energy consumption.

The implementation and the system analysis are difficult and complex. This comes from the fact that every sensor acts independently from others. Therefore, our model will be studied using simulation tools in order to compare it with reference cases. Indeed, each Sensor sends data packets to a number of Sensors within a defined range according to threshold used as policy. Thanks to its context gathering module the TF has all information to evaluate the trust.

B. Attackers

The behavior of a node is fixed in the starting of the simulation based on a uniform distribution, which has an average equal to rate of attackers .There are many attacks but we will consider only two attacks:

- Jamming attack: given the sensitivity of the wireless medium to noise, a node can cause a denial of service by transmitting signals at a certain frequency. It is implemented just by sending packet repeatedly.
- Sinkhole attack: the node tries to attract to it the most possible path like a concentrator to have control over most of the data through the network. To do this, the attacker must appear to others as being very attractive, presenting optimal routes. It is implemented by imitating any BS or any good node.

Note that we do not treat Sybil attacks.

C. Metrics

Due to the characteristics of WSN, their major objective is to fulfill their mission even though some nodes are out of use due to attackers. Indeed, this means to ensure survivability of WSN, which can be defined as: "the capability to fulfill its mission, in a timely manner, in the presence of intrusions, attacks, accidents and failures" [22] Gain of survivability is the ratio between: Duration time when 75% of nodes are out of battery using our framework and same case but without using our framework

VI. Implementation and Validation Methodology

We have implemented Green-SARM and validated it in a mobile Sensor wireless network simulation developed with AnyLogic, which is a simulation tool that supports all different simulation methodologies. It is based on Real-time UML, Java object-oriented language and agent based model.

A. Model Set-up

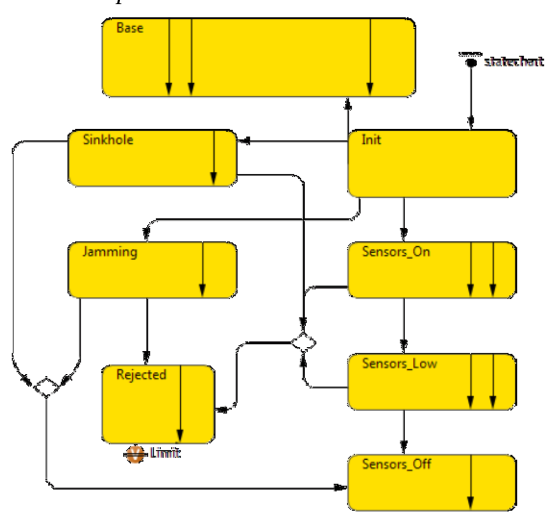


Figure 4. Statechart of sensors

Each Sensor is associated to a given agent matching with its location and behavior.

Setting up our security model using Table II, we can take advantage of state chart by monitoring the behavior of agents. The state of our agents is controlled by state-charts, which represents the exact behavior of sensors, as shown in

Fig. 4. Using AnyLogic as implementation platform agents and state-charts can be programmed very conveniently. Particular modifications and/or extensions of the final model can be handled in a simple way.

In Fig. 4, each Agent (Sensor) starts at state Init. The Agents are switched to their relative state (Sinkhole, Jamming, Base Station, Sensors) according to percentage of attackers, BS and Sensors. They are then added to a list of the sensor whenever they are within his range.

We used Agents having one of the following behaviors:

a) Normal state and

b) Sinkhole and Jamming as attakers

Each Agent is then processed depending on the decision of the monitor unit to choose a security means or not. When arriving at a minimum defined level of Energy according to the real consumption of Sensor (see table [1]), the Agent transits to Sensor_low; and then transit to Sensor_off when the Agent has not enough energy to transmit data. The Agent transits to Rejected state when its TF is lower than a Threshold. The state-chart Trust updates the trust each time the Base Station received information about a sensor. The BS is not limited in energy and has also access to trust database to spread it over its range.

In order to have a deep study of the model, we have introduced factors :

a) ImmunityRate: when TF of a Sensor reaches this value it has double influence on its values "Good". It is a catalyst to accelerate convergence.

b) Limit: it is the Threshold of TF to reject sensors.

c) JudgementError: it defines an error rate among the large number of messages received and helps to study the robustness of the model. It is useful in our case where errors are very frequent due to the media.

B. Validation Methodology

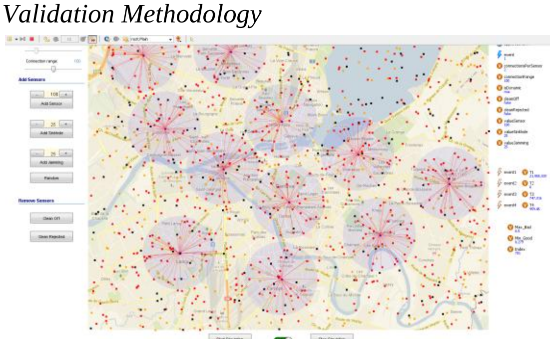


Figure 5. Simualtion interface for Geneva pollution use case

Fig. 5 depicts the context of the simulation interface. In this respect, we have carried out simulations under 0%, 20%, 40%, 60% and 80% half-half sinkhole and jammer attackers. Furthermore, the network topology was set to random spreading of sensors. We have taken a uniform packet distribution over the neighbors.

The BSs are randomly spread over the network. In our experiments, we have validated our proposed solution and analyzed the extended performance under a range of various scenarii, where we have fixed to ten BSs. All sensors are spread over a square topology and operating over 35 days of simulation time. The Base Station coverage a large circle but it has no contact over the coverage of other BSs. We deployed the Sensors in an incremental mode, from S1 to Sn. As the device is not static, we have modeled his mobility using random variables model. The movement pattern of mobile clients was totally randomized, in order to comply with a real application. To achieve this, we used the Random WayPoint (RWP) mobility model [23]. All nodes are mobile and their pause time is a randomly uniform variable. The time is in minutes and is in a range [0; 50'000] adapted to the battery of sensors.

C. Scenarii

We have used three scenarii to validate our model. In our scenarii, sensors (agents) were divided in four categories.

A normal behavior, they are composed of N sensors set in the range of [100; 1000]. The trust threshold is optimized after many series of simulation to 0.3.

Sinkhole and jammer attackers are composed of the same amount of sensors. In the first scenario, we fixed the percentage as: N of normal behavior and 10 % of sinkhole and 10% jammer attackers. In the second scenario, we fixed the percentages as: N of normal behavior and 20 % of sinkhole and 20% jammer attackers. In the third scenario, we fixed the percentages as: N of normal behavior and 40 % of sinkhole and 40% jammer attackers

VII. RESULTS ANALYSIS

During our analysis, we firstly studied the performance of Green-SARM in the three defined scenarii where sensors were arranged at random. Secondly, we studied the scalability of the model. Thirdly, we studied robustness of the model based on the factor Error.

For comparison purpose, we plotted the Green-SARM for 1000 sensors in Fig. 6 and without trust in Fig. 7 for the same first scenario. We can easily conclude that SARM is largely better than normal case without security trust; a ratio of 6.5 is reached and we can see that in short time 1000min compared to the maximum time 50'000min. Indeed, we have obtained the desired effect of the feedback mechanism of the SARM. An example of 500 sensors without Green-SARM security is plotted in Fig. 8. We see clearly that the attackers diminish quickly the survivability of the network.

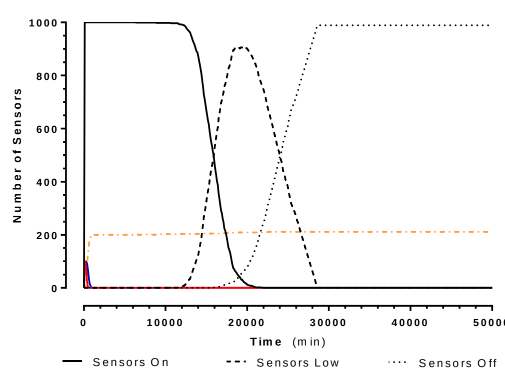


Figure 6. 1000 sensors for the first scenario (Green-SARM)

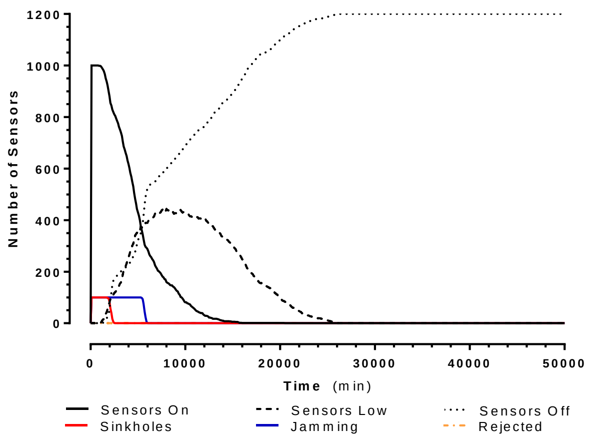


Figure 7. Network of 1000 Sensors for the first scenario (without SARM)

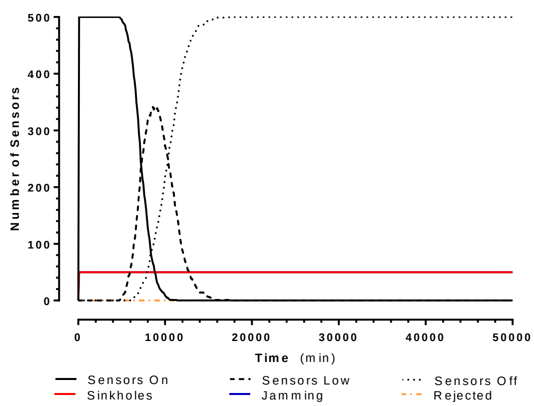


Figure 8. 500 sensors for the first scenario (without SARM)

In Fig. 9, we plotted the scalability of the model in function of number of sensors using Green-SARM for the first scenario. This graph shows clearly that our model is scalable. The number of sensors goes higher as the network survivability goes higher.

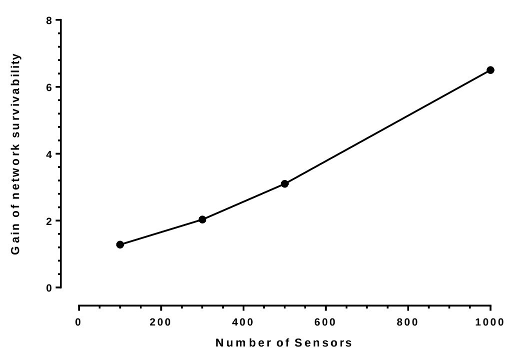


Figure 9. Gain of survivability for 20 % of attackers for number of sensors from 100 to 1000

For comparison purpose, we plotted the Green-SARM under 20% of sinkhole attackers (the second scenario) using our TF in Fig. 10 and also without trust in Fig. 11. The results clearly demonstrate that the survivability is boosted.

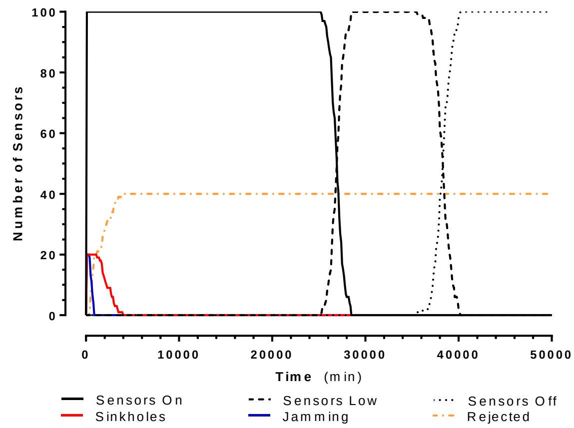


Figure 10. Number of 100 sensors for the 2nd scenario using Green-SARM

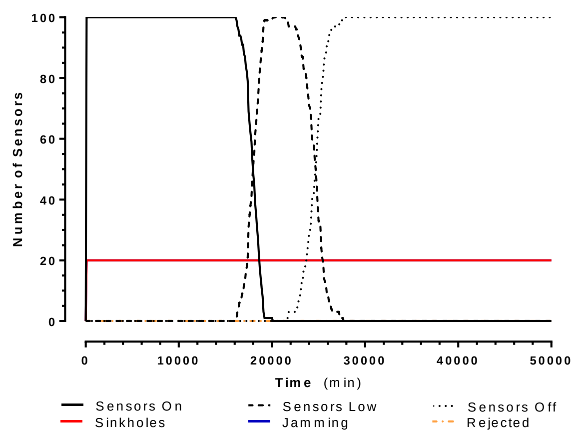


Figure 11. 100 sensors for the 2nd scenario (without Green-SARM)

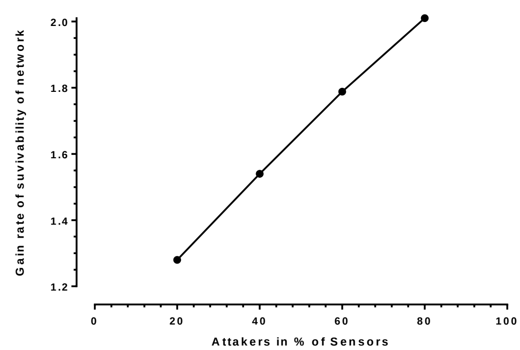


Figure 12. Survivability of 100 sensors for different scenarii

In Fig. 12, we plotted the gain of survivability of the model in function of the percentage of attackers using Green-SARM. Therefore, we have the proof that Green-SARM is

 ISBN: 978-1-61208-289-9

ameliorating the survivability to up a rate of about 2 when the network is under more attackers compared to a passive security network.

The *JudgmentError* factor, which designs errors on TF, was set to 10% but the variation of results was less than 1%. This shows clearly the robustness of the Green-SARM.

All the results show clear advantages of Green-SARM even at 80% of sinkhole and jammer attackers. We can conclude that our security monitor helps the WSN to operate even under 80 % of attackers thanks to the looping system connected to the context gathering monitor and the TF.

VIII. CONCLUSION & FUTURE WORK

We have proposed a Security Adaptation Reference Monitor based on security adaptation and the Autonomic Computing Security pattern to support both context monitor and behavior control. This paper also presents the validation of SARM in WSN applied to support Green IT concepts. The results clearly show that Green-SARM copes with survivability and network Energy loss under sinkhole and jamming attacks even at 80% of attackers. Indeed, Green-SARM constitutes a good Platform within the Base Station to detect any sinkhole and eliminate it from its connections and put it into log file, thanks to the context gathering monitor and the feedback control and regulation system. Therefore, we show that our Framework is efficient in this context and is tuning to achieve the best trade-off between security and performance according to application preferences. In addition, the network is well energy balanced. These results encourage us to implement the model in a tamper-resistant security module based on a Secure Digital card.

ACKNOWLEDGMENT

I would like to thank my students O. Oudom and M. De Azevedo who have worked in computer model implementation.

REFERENCES

[1] M. Parashar and S. Hariri, “Autonomic Computing : An overview”, Spring-2005

[2] L. Wilbanks, “Green, My Favorite Color”, IT Pro v.10, n.6., 2008, pp.63-64. IEEE Comp Soc. Press

[3] SMART 2020, “Enabling the low carbon economy in the information age”, Global e-Sustainability Initiative, 2008

[4] H. K. D. Sarma and A. Kar, “Security Threats in Wireless Sensor Networks”, In proceeding of: Carnahan Conferences Security Technology, Oct. 2006, pp. 243-251, Proceedings 40th Annual IEEE International, 2006

[5] X. Chen, K. Makki, K. Yen, and N. Pissinou, “Sensor Network Security: A Survey”, IEEE Communications Surveys & Tutorials, Vol. 11, No. 2, Second Quarter 2009, pp. 52–73, 2009.

[6] J.-M. Seigneur, “Trust, Security and Privacy in Global Computing”, PhD Thesis, 2005.

[7] B. Badrinath,. A. Fox, L Kleonrock, G. Popek, and M. Satyanarayanan, “A conceptual Framework for Network and Client Adaptation”, Mobile Networks and applications, v.5 n.4, Dec. 2000, pp. 221-231, 2000.

[8] A. S. Wander, N. Gura, H. Eberle, V. Gupta, and S. C. Shantz, “Energy Analysis of Public-Key Cryptography for Wireless Sensor Networks”, Proceedings of the Third IEEE International Conference on Pervasive Computing and Communications, March 2005, pp. 324-328, 2005

[9] A. Josang, R. Ismail, and C. Boyd, “A Survey of Trust and Reputation Systems for Online Service Provision”, Decision Support Systems, 43(2) 2007, pp. 618-644, 2007

[10] T. El-Maliki and J.-M. Seigneur, “A Security Adaptation Reference Monitor (SARM) for Highly Dynamic Wireless Environments”, The International Conference on Emerging Security Information, Systems, and Technologies SECURWARE, Jul. 2010, pp. 63-68

[11] A.V. Taddeo, “Gradual Adaptation of Security for Sensor Networks”, IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks, June. 2010, pp. 1-9

[12] C. Hager, “Context Aware and Adaptive Security for Wireless Networks”, PhD thesis, Virginia Polytechnic Institute and State University, 2004

[13] M. Lacoste, G. Privat, and F. Ramparany, “Evaluating Confidence in Context for Context-Aware Security”, European Conference on Ambient Intelligence (AmI'07), Nov. 2007, pp. 211-229.

[14] J. Al-Muhtadi, D. Mickunas, and R. Campbell, “A Lightweight Reconfigurable Security Mechanism for 3G/4G Mobile Devices”, IEEE Wireless Communications, 9(2), April. 2002, pp. 60–65

[15] E2R Deliverable D2.2, “Equipment Management Framework for Reconfiguration: Architecture, Interfaces, and Functions”, Technical Report, Dec. 2005.

[16] M. Lacoste, T. Jarboui, and R. He, “A Component-Based Policy-Neutral Architecture for Kernel-Level Access Control,” Annals of Telecommunications, vol. 64, no. 1-2, Feb. 2009, pp. 121–146

[17] A. Shnitko, “Practical and Theoretical Issues on Adaptive Security”, WOLFASI, Turku, Finland, 2004.

[18] D. M. Chess, “Security in Autonomic Computing”, IBM Thomas J. Watson Research Center, 2005.

[19] T. El-Maliki, J.-M. Seigneur, “A survey of user-centric identity management technologies”, Emerging Security Information, Jul. 2007, pp. 12-17, ieeexplore.ieee.org, 2007.

[20] L. Eschenauer, V. Gligor, and J. Baras, “On Trust Establishment in Mobile Ad-Hoc Networks”, Proc. 10^{th} International Workshop of Security Protocols, Springer Lecture Notes in Computer Science (LNCS), Apr. 2002, pp. 44-66.

[21] A. Rahman and A. Hailes, “A Distributed Trust Model”, New Security Paradigms Workshop, 1997, pp. 48-60, ACM, 1997.

[22] R. Ellison, D. Ficher, R. Linger, H. Lipson, T. Longstaff, and N. Mead, “Survivable Network System: An Emerging Discipline”, CMU/SEI-97-TR-013, 1997

[23] C. Bettstetter, “Stochastic Properties of the Random Waypoint Mobility Model”, 2004.

A Mathematical Model for Improving Lightweight Security with Network Coding

Alireza Esfahani
Instituto de Telecomunicações
Aveiro, Portugal
Email: *alireza@av.it.pt*

Alberto Nascimento
Mathematics and Engineering Department
University of Madeira
Funchal, Portugal
Email: *ajn@uma.pt*

Du Yang
Instituto de Telecomunicações
Aveiro, Portugal
Email: *duyang@av.it.pt*

Jonathan Rodriguez
Instituto de Telecomunicações
Aveiro, Portugal
Email: *jonathan@av.it.pt*

***Abstract*—As a promising technology of efficiently utilizing resources, network coding is capable of improving many performance such as throughput, security, etc. The main purpose of this paper is joining network coding and cryptography methods. In this paper, not only we propose a new mathematical method to use in network coding but also a security framework with slightly increased complexity and overhead. We use Greatest Common Divisor and Lowest Common Multiplication instead of sending original data packets, so as to provide additional guards against eavesdropping attack.**

Keywords-Security; Network Coding; Random Linear Network Coding; Eavesdropping attack.

I. INTRODUCTION

As wireless communication has become a part of people's daily life in modern society, providing efficiency, convenience and security in communications network is becoming a very crucial aspect of modern digital infrastructure. There are many aspects to security, addressing a range of natural adversaries and malicious threats. Nevertheless, one ultimate goal of security is to remain data confidential impervious to malicious or accidental eavesdropping.

Network coding was introduced for the first time by Ahlswede et al. [1] and they showed that data throughput and network robustness can be considerably improved by allowing the intermediate nodes in a network to mix different data flows through algebraic combinations of multiple datagrams. In [2], Random linear Network Coding (RLNC) was studied as a fully distributed method for performing network coding. They mentioned there is a possibility that each node in the network independently and randomly selects a set of coefficients and uses them to make linear combinations of the data symbols. In other words, RLNC allows each node in the network to make local decisions [3]. The original data packet represented as X is divided into n blocks $X_1, X_2, ..., X_n$ and each node computes and forwards linear combined packet $P = \sum_{i=1}^{n} C_i X_i$ with randomly chosen coefficients $C = (C_1, C_2, ..., C_n)$. When sufficient number of linearly are received, a node would be able to decode the original X.

Security for network coding has also been an active research area recently. The problem of secure network coding was first studied by Cai and Yeung [4]. They proposed a secure network coding scheme based on a given decodable linear network code over a sufficiently large field.

Although there are a lot of works, which have been done in security via network coding recently [5-9], network coding is still vulnerable to eavesdropping attacks. Lima et al. [10] have made cryptographic mechanism with maintaining the NC properties. They have used locked and unlocked coefficients through multiple paths from source to sinks. They have showed a noticeable reduction in computational overhead which needs for performing encryption. Motivated by gaining supplementary security approach, in this paper, we present a mathematical lightweight cryptography scheme. We mainly deal with eavesdroppers which have full access to the information about decoding and encoding. We will use properties of number theory and finite field against eavesdropping attacks and maintaining RLNC's properties.

The remainder of this paper is organized as follows. Section II presents the history of Network Coding, Preliminaries and Definition of Greatest Common Divisor (GCD) and Lowest Common Multiplication (LCM). Section III introduces the new security scheme. Section IV discusses the achievable performance. Section V concludes the paper.

II. PRELIMINARIES AND DEFINITION

In this section, we summarize the history of Network Coding; then, we discuss GCD and LCM. Also, we explain how to decompose a number to prime factors. Moreover, we examine the complexity of finding GCD and LCM in the end of this section.

A. Network Coding

After the first Max-Flow Min-Cut theorem [13], [14] presentation, Ahlswede et al. [1] elaborated a version of this theorem for information flow. Against to traditional and classical commodity flow, in which information is only routed or replicated, information flow can also employ coding operations at the nodes and we have known this method as Network Coding. After, that Linear Network Code (LNC) was appeared. In LNC, for all nodes (except source nodes) the outgoing packets are always linear combinations of the incoming ones. Yeung et al. [15] explained the relation between a linear dispersion and a generic Network Coding; also they defined a relation on

 ISBN: 978-1-61208-289-9

the sizes of the base fields of the code. A different approach to Network Coding was presented in [16]. It proposed a completely algebraic framework, with the consequent possibility to apply the mathematical theorems of algebra on Network Coding. The objective of this framework is the definition of the transfer matrix M, which includes all the characteristics of the network itself. The translation of the Max-Flow Min-Cut theorem into the new framework modified the Network Coding problem into the problem of finding a point on an algebraic variety. Ho et al. [5] provided two results for solving the LNC multicast problem. After that and considering the results of the previous work, they proposed Random Linear Network Coding (RLNC) [2]. RLNC was studied as a fully distributed method for performing network coding.

B. LCM and GCD

In concept of arithmetic and number theory, we can define the least common multiple of two integers a and b that usually denoted by $LCM(a, b)$, as the smallest positive integer that is divisible by both a and b. Also greatest common divisor (GCD) of two integers is the largest positive integer that divides the numbers without a remainder. According to the fundamental theorem of arithmetic we can have a positive integer number which is the product of prime numbers [11], and it could be defined as:

$$n = 2^{n_2} * 3^{n_3} * 5^{n_5} * ... = \prod p^{n_p} \quad (1)$$

$2, 3, 5, ...$ are prime numbers and $n_2, n_3, n_5, ...$ are the exponents of those prime number, which are non-negative integers $(n_i \geq 0)$. By considering two integer numbers, e.g., $a = \prod p^{a_p}, b = \prod p^{b_p}$, we can define GCD and LCM as the following relations:

$$GCD(a, b) = \prod p^{min(ap,bp)} \quad (2)$$

$$LCM(a, b) = \prod p^{max(ap,bp)} \quad (3)$$

$$LCM(a, b) = (a * b)/(GCD(a, b)) \quad (4)$$

Crandall *et al.* [11] explained above equations in details. Finally we have to mention that prime factorization is the decomposition of a composite number into smaller non-trivial divisors, whose multiplied result equals the original integer. We need to decompose the LCM and GCD at the end of the process which we will explain in the next section.

C. Time Complexity of LCM and GCD

In the following of this section, we present two pseudo codes for calculating GCD and also for finding prime factors of two integer numbers. LCM can be obtained from Equation (4) using a, b and $GCM(a, b)$. Regarding these two algorithms which we want to use in source and sink nodes, we can calculate the time complexity of our method.

Homer and Selman [17] have shown a lot of methods for calculating time complexity in different algorithms. According to their approach, we have one loop in both algorithms (Algorithm 1 and Algorithm 2), and all the steps of each loop can be calculated according to logarithmic-scale. Considering to the time complexity of Gaussian Elimination which is exponential-scale; it is clear that the time complexity of both algorithms which have been used in source and sink nodes, are less than exponential-scale. So our noticeable result is that the time complexity is close to the logarithmic-scale behaviour, and we can consider $O(logn)$ as the time complexity of our approach.

Algorithm 1 Algorithm for calculating Greatest Common Divisor

```
GCD(x,y)
if y == 0 then
  return x;
else
  return GCD( y, x MOD y )
end if
```

Algorithm 2 Algorithm for finding prime factors

```
Input(A)
for all i = 2 to i = A do
  if A MOD I == 0 then
    write A
    A = A / i
  else
    Continue
  end if
end for
```

Algorithm 3 Algorithm for recovering original numbers

```
Input(G,L,S)
if (G==L) then
  return (G,L)
else
  X= The difference prime numbers
  for all i = 1 to i = X do
    Choosing P1, P2
    if S == P1+P2 then
      return P1,P2
      exit
    else
      Continue
    end if
  end for
end if
```

D. Finding unique result

As recovering the original data, the two integer a and b, through the GCD , LCM does not have a unique result, so we need to use addition information for having the unique result. Thus we want to send the summation of the original numbers before calculating GCD , LCM. So, by having three parameters, i.e., GCD, LCM, and summation of a and b, we could recover original numbers.

 ISBN: 978-1-61208-289-9

III. Our Scheme: M*athematical Secure* N*etwork* C*oding* (M*e*N*o*)

It is about twelve years after the emergence of the first example of network coding and specifically the butterfly network coding; a lot is already known about network in particular for the case of network multicast. Network multicast refers to simultaneously transmitting the same information to multiple receivers in the network [1][12]. The Secure Practical Network Coding (SPOC) which has been proposed in [10] is a secure algorithm with keeping the RLNC properties. The authors have tried to achieve confidentiality by protecting the locked coefficients without impairing any of the operations of practical network coding protocols. They could reduce the complexity of their algorithm in comparing to the traditional end to end encryption approaches. They have used two kinds of coefficients (locked and unlocked), but as we want to maintain the RLNC properties. It's clear whenever one node acts as an eavesdropper and hears all or part of the information which send to the sink so there is possibility that he discovers the original packet. We show the packet format scheme in Fig. 1. For instance, we suppose the basic scheme which has been used in Fig. 2. If node 4 appears as an eavesdropper so he has this chance that finds original data.

Fig. 1. Proposed packet format

In this work, not only we show that our scheme does not bring more complexity but also we create more security framework to ensure that data are sent completely in secure condition. We propose Mathematical Secure Network Coding (MeNo), a more secure framework for improving security mechanism. As we considered that eavesdrop attacker is a person who can hear information and also he has full access to all the information about coding and decoding, in this case eavesdrop attacker knows that source node has used RLNC properties and has made a linear combination of original packets, so there is possibility that this node decodes original packets. In our work, we want to use GCD and LCM instead of sending original packets. For instance if we want to send two data, e.g., $a = 4$ and $b = 10$, the first step is calculating the GCD and LCM and for this, we need using the algorithm was explained in Section II-B. In fact, we can calculate $GCD = 2$ and $LCM = 20$.

These values are all sent to the receiver to guarantee correctly recovering the original data. Having the knowledge of $GCD(a, b)$ and $LCM(a, b)$, we could have two possible recovering results $(a = 4, b = 10)$, and $(a = 2, b = 20)$. Aided with the additional knowledge of sum of a and b, it is evident that the original value of (a, b) can only be 4 and 10. Now

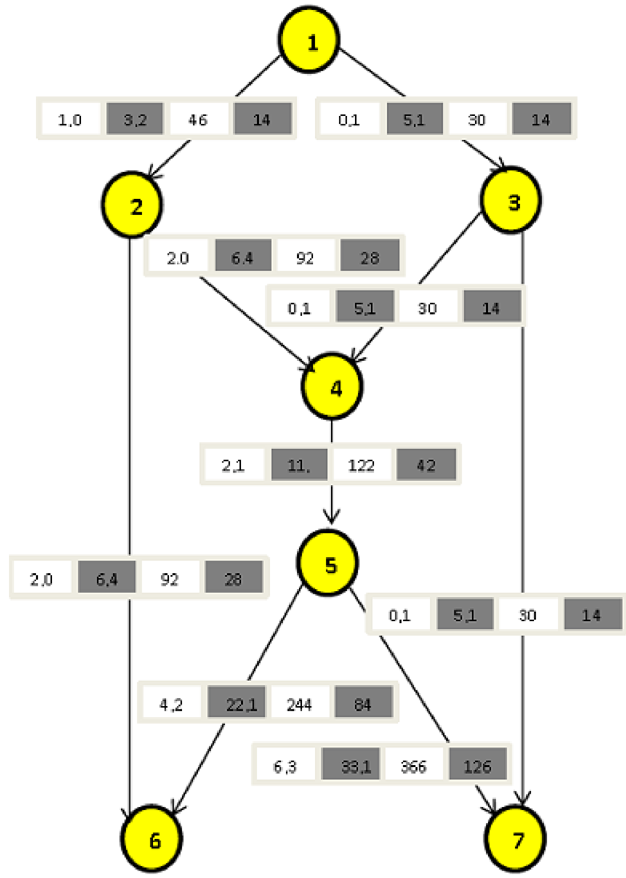


Fig. 2. Basic scheme (Encrypted information are in Gray color)

TABLE I
Summary of MeNo

Node	Operation
Source	Calculation GCD , LCM and generating RLNC
Intermediate	Combination by maintaining RLNC properties
Sink	Gaussian Elimination ,Recovering and Finding original data

we can use RLNC to create random linear equations and also use locked and unlocked coefficients. For recovering unique results over GCD and LCM we have to send the summation of original numbers to the sink. But, if attacker wants to discover data, by knowing (2,20), in fact he knows nothing or in the other word he realizes meaningless data, and it is the advantage of our algorithm. So, we can protect the native information from eavesdrop attackers and we can guarantee we will have completely security method without noticeable overhead. We summarize our algorithm in table I.

According to Vilela et al. [10], the stream of packets divides into some generations of size h, and just those packets which belong to the same generation number could be combined with each other. Before moving to next section and presenting performance evaluation, we have to mention again that the kind of attack which we have considered in our approach is an eavesdropping attack which has the ability of hearing to all the

TABLE II
TIME COMPLEXITY OF MENO

Node	Operation	Detailed cost	Total Cost
Source	Generation	$O(h^2)$	$O(h^2)$
	GCD and LCM	$O(Log(h))$	
Intermediate	Combination	$O(nh)$	$O(nh)$
Sink	Gaussian Elimination	$O(n^3)$	$O(n^3)$
	Recovering	$O(Log(h))$	
	Original data	Negligible	

information and it has full access to information of coding and decoding. On the other hand, considering to RLNC properties, they know source node has generated random coefficients, but as if they can recover these coefficients, solve the equations and recover packets, but they do not know about the process which we need to find original packets through GCD, LCM and summation of original data. We show the process for recovering original data for instance for node 6 as follows:

$$\left|\begin{array}{cc|cc} 2 & 0 & 6 & 4 \\ 4 & 2 & 22 & 10 \end{array}\right|$$

Then by implementing Gaussian elimination:

$$\left|\begin{array}{cc|cc} 1 & 0 & 3 & 2 \\ 0 & 1 & 5 & 1 \end{array}\right|$$

$\Longrightarrow (3,2),(5,1)$

$$2(3G+2L)=92 \tag{5}$$

$$4(3G+2L)+2(5G+L)=244 \tag{6}$$

Equation (5) and (6), we have $G=2, L=20$. G represents the GCD value, and L represents the LCM value. In addition with the knowledge of sum value S, we could recover the original data $a=4, b=10$.

IV. PERFORMANCE EVALUATION

Although the proposed algorithm needs to calculate GCD and LCN, the additional complexity is in log-scale, and is much smaller compared to Guassian Elimination in exponential-scale, Hence, the complexity of the proposed scheme is slightly higher than the one proposed in [10]. The detail information is summarized in Table II.

V. CONCLUSION AND FUTURE WORK

We proposed a new mathematical secure framework in additional to the lightweight secure network coding by utilizing GCD and LCM. The proposed scheme is capable of providing additional security against eavesdrop attack, with slightly increased computational complexity only at the source and sink nodes.

For future work, as homomorphic hash function brings more computational overhead, we are going to join our model with this approach for providing authentication and lower overhead.

ACKNOWLEDGMENT

The first author acknowledges financial support of FCT through a scholarship given by Universidade da Madeira. The research leading to these results has received funding from the Fundao para a Cincia e Tecnologia (PTDC/EEA-TEL/119228/2010 - SMARTVISION).

REFERENCES

[1] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, ”Network information flow”. IEEE Transactions on Information Theory, 2000, pp. 1204−1216.
[2] T. Ho, M. Medard, R. Koetter, D.R. Karger, M. Effros, J. Shi, and B. Leong, ”A random linear network coding approach to multicast”, IEEE Transactions on Information Theory, vol. 52, no. 10, 2006, pp. 4413−4430.
[3] T. HO, M. Medard, R. Koetter, D. Karger, M. Effros, S. Jun, and B. Leong, ”A random linear network coding approach to multicast,” IEEE Transactions on Information Theory 52, 2004, pp. 4413-4430.
[4] N. Cai and R. W. Yeung, ”Secure network coding,” IEEE International Symposium on Information Theory, Lausanne, Switzerland, , July. 2002, p. 323.
[5] L. Lima, S. Gheorghiu, J. Barros, M. Mdard, and A. L. Toledo, ”Secure Network Coding for Multi-Resolution Wireless Video Streaming,” Journal of Selection Areas in Communication, vol. 28, 2010, pp. 377-388.
[6] Y. Hongyi, D. Silva, S. Jaggi, and M. Langberg, ”Network Codes Resilient to Jamming and Eavesdropping,” in Network Coding (NetCod), IEEE International Symposium on, 2010, pp. 1-6.
[7] G. Zhenzhen, Y. Yu-Han, and K. J. R. Liu, ”Anti-Eavesdropping Space-Time Network Coding for Cooperative Communications,” Wireless Communications, IEEE Transactions on, vol. 10, 2011, pp. 3898-3908.
[8] S. Jaggi and M. Langberg, ”Resilient network codes in the presence of eavesdropping Byzantine adversaries,” IEEE International Symposium on information theory, 2007, pp. 541-545.
[9] C. Gkantsidis and P. Rodriguez, ”Cooperative security for network coding file distribution,” In Proceeding of IEEE INFOCOM, Barcelona, April. 2006, pp. 1-13.
[10] J. P. Vilela, L. Lima, J. Barros, ”Lightweight Security for Network Coding,” Proceeding of the IEEE International Conference on Communications (ICC 2008), Beijing, China, May. 2008, pp. 1750-1754.
[11] R. Crandall, C. Pomerance, ”Prime Numbers: A Computational Perspective”, New York: Springer, ISBN 0-387-94777-9, 2001.
[12] Q. Li, J. C.S. Lui, and D. Chiu, ”On the Security and Efficiency of Content Distribution via Network Coding,” IEEE Transactions on Dependable and Secure Computing, vol. 9, no. 2, March/April. 2012, pp. 211-221.
[13] P. Elias, A. Feinstein, and C. E. Shannon, ”A Note on Maximum flow through a Network,” IRE Transactions on Information Theory, vol. IT, no. 2, 1956, pp. 117-179.
[14] L. R. Ford and D. R. Fulkerson, ”Maximal flow through a network”, Canadian Journal Mathematics,vol. 8, 1956, pp. 399-404.
[15] S.Y. R. Li and R. W. Yeung, ”Linear Network Coding,” IEEE Transactions on Information Theory, vol. 49, no. 2, 2003, pp. 371-381.
[16] R. Koetter and M. Mdard, ”An algebraic approach to network coding,” IEEE/ACM Transactions on Networking, vol. 11, no. 5, 2003, pp. 782-795.
[17] S. Homer and A. Selman, ”Computability and Complexity Theory”, Springer, 2000.

MAC Spoofing Attack Detection based on EVM in 802.11 WLAN

Gaeil An and Shin-hyo Kim
Cyber Security Research Laboratory
Electronics and Telecommunications Research Institute (ETRI)
Dajeon, Korea
{fogone, shykim}@etri.re.kr

***Abstract*— Wireless LAN is very vulnerable to MAC spoofing attack. This paper proposes an Error Vector Magnitude (EVM)-based MAC spoofing attack detection scheme. The proposed scheme identifies wireless devices by using EVM as feature for distinguishing them. The proposed scheme can detect MAC spoofing attacks in real-time by using a prototype hardware system which has been developed in this paper to capture radio signal of wireless devices and extract EVM. Through experiment, we have confirmed that our scheme is excellent in detecting MAC spoofing attacks that employ wireless devices with different Wi-Fi chipset from legitimate one.**

Keywords-MAC spoofing attack; EVM; Error Vector Magnitude; Wireless device; WLAN

I. INTRODUCTION

As mobile devices have increased explosively with appearance of smartphone and performance of Wireless Local Area Network (WLAN) has been highly enhanced with realization of new technology such as channel bonding and Multiple-Input Multiple-Output (MIMO) in 802.11n, the importance of WLAN in wireless communication network is growing bigger and bigger. Currently, one of the greatest barriers that obstruct development of WLAN is problems with security. In WLAN, there exist a lot of security vulnerabilities such as eavesdropping, wireless spoofing, DoS, session hijacking, man-in-the-middle attacks, data modification, and so on. In this paper, we focus on security attack that spoofs MAC address of wireless device among many security threats that occur in WLAN. In this paper, a wireless device indicates a computer device with 802.11 WLAN card such as smart phone and Access Point (AP).

IEEE 802.11 implicitly trusts the MAC address of L2 frame and does not basically provide any mechanism for verifying it, just like IEEE 802.3 Ethernet. So, WLAN is very vulnerable to MAC spoofing attack, which can raise de-authentication, dis-association, power-saving, rogue AP, and rogue client attacks[1][2].

To react against MAC spoofing attack, various kinds of schemes have been proposed such as authentication, sequence number analysis, Received Signal Strength (RSS) fingerprinting, Radio Frequency (RF) fingerprinting, and so on. Authentication scheme is a way to defeat the attack by authenticating wireless devices or frames through cryptography [3]. Even if it is excellent in defeating MAC spoofing attack, it is not perfect solution because it is not a scheme for device authentication, but for user and message authentication.

Sequence number analysis scheme [4] detects the attack by checking consistency between the sequence numbers of L2 frames with the same MAC address. Even if it works well, it is known to be vulnerable to impersonation attack because it is a kind of software-based scheme. RSS fingerprinting scheme [5][6] distinguishes wireless devices by using Received Signal Strength Indication (RSSI). RSS fingerprinting scheme is based on a fact that the RSS pattern of a legitimate node is different from that of the spoofed one because their transmission power or location is different from each other. RSS fingerprinting scheme can effectively detect rogue AP, but not good at detecting rogue mobile client. Finally, RF fingerprinting scheme [7-10] is a way to identify a wireless device by using distinctive physical layer characteristics exhibited by the device. Even if RF fingerprinting scheme need hardware equipment to extract physical layer characteristics, its performance is known to be better than any other existing schemes.

This paper proposes an Error Vector Magnitude (EVM)-based MAC spoofing attack detection scheme, which is a kind of RF fingerprint scheme. The proposed scheme identifies a wireless device by using EVM as feature for distinguishing wireless devices. EVM is related to digital modulation error and defined as vector magnitude difference between an ideal reference signal and measured signal. The existing RF fingerprinting schemes are not real-time in that they collect feature information with the help of wireless measurement instrument such as Agilent. Our scheme has a capability that can detect MAC spoofing attack in real-time. For this, we have developed a prototype hardware system which can capture radio signal of wireless devices and extract EVM.

The rest of this paper is organized as follows. Section 2 introduces MAC spoofing attack in WLAN and the existing detection schemes. Section 3 describes our scheme, EVM-based MAC spoofing detection scheme. The performance of the proposed scheme is measured and evaluated in Section 4. Finally, conclusion is given in Section 5.

II. RELATED STUDY

Generally, attacker spoofs MAC address of wireless device to hide his or her presence or to use network resource illegally. MAC spoofing attack can be used to raise other attacks such as de-authentication attack, dis-association attack, power-saving attack, Rogue AP, and rogue mobile device.

First of all, in IEEE 802.11 a client uses authentication message to authenticate itself to AP to join. De-authentication message is used to release authentication relationship between the connected nodes. Similarly, a client uses association message to setup a L2 link with a wireless node. Disassociation message is used to release the link between the associated nodes. De-authentication attack and disassociation attack can illegally terminate connections between legitimate nodes by using a spoofed de-authentication and disassociation message with MAC of victim node.

Power-saving attack makes use of device power conservation function. The device power conservation function allows clients to enter sleep mode for the purpose of battery conservation. The data sent to the client by other nodes is temporally stored into its AP. To check whether the data for the client is buffered by its AP, the client periodically wakes up from sleep mode and receives a traffic indication map (TIM) from its AP. If through the TIM the client gets known that its AP buffers its data, it sends its AP a poll message to receive its data. Power-saving attack is used to prevent clients from communicating normally with their APs. For example, power-saving attacker can send a spoofed poll message to remove the buffered data or a spoofed TIM message to prevent a client from receiving its data from its AP.

Rogue AP attack is one that lures clients into connecting to a fake AP instead of a legitimate one by spoofing MAC and SSID of legitimate one. Finally, rogue client attack is one that spoofs the MAC of a legitimate mobile device to bypass an MAC address-based access control system. Through this attack, rogue client can gain illegally access to a WLAN.

To protect WLAN from such MAC spoofing-based security threat, various kinds of schemes have been proposed such as authentication scheme, sequence number analysis, RSS fingerprinting, and RF fingerprinting. First of all, authentication scheme provides IEEE 802.11i and 802.11w as a way that authenticates wireless devices or L2 frames. The standard IEEE 802.11i is designed to provide secure communication based on cryptography in WLAN. The standard 802.11w is designed to provide cryptographic protection to IEEE 802.11 management frames such as de-authentication and disassociation. Even if authentication scheme is very excellent in defeating MAC spoofing attack, it is not perfect solution because it is not a scheme for device authentication, but for user and message authentication.

Secondly, sequence number analysis scheme is based on a fact that there is no consistency between the sequence numbers of L2 frames from the legitimate device and those from the spoofed device with the MAC of the legitimate one. If sequence number analysis system finds that there is a jump in the sequence number of the received frame, it regards the transmitter of the frame as a MAC spoofing device. Even if sequence number analysis scheme works well, it is vulnerable to impersonation attack because it is a kind of software-based scheme.

RSS fingerprinting scheme distinguishes wireless device by using RSSI. RSSI is correlated with transmission power, distance between transmitter and receiver, and radio environment that raises multipath and absorption effects. RSS fingerprinting scheme detects MAC spoofing devices by making use of a fact that the RSS pattern of a legitimate node is different from that of a spoofed one because their transmission power or location is different from each other. RSS fingerprinting scheme has a merit that it can exactly detect MAC spoofing attack performed by fixed wireless devices such as AP. But, its performance is not good at detecting MAC spoofing attack performed by mobile devices.

Finally, RF fingerprint scheme is a way to identify a wireless device by using distinctive physical layer characteristics exhibited by the device, such as modulation error or radio signal pattern. Our scheme belongs to RF fingerprint scheme. RF fingerprint scheme is expensive because it requires hardware equipment that can capture radio signal and extract RF feature. However, RF fingerprint scheme has been known as one that can detect MAC spoofing attacks most accurately among the existing ones. According to [7] and [8], the accuracy of RF fingerprinting in wireless device identification is more than 99%.

III. EVM-BASED MAC SPOOFING ATTACK DETECTION

A. Background

To detect MAC spoofing attack, we should be able to identify a wireless device. The scheme proposed in this paper identifies a wireless device by using digital modulation error which is distinctive physical layer characteristics exhibited by the device.

To transfer messages over wireless network, a transmitter should perform digital modulation. Digital modulation is to transfer digital bit streams over analog channel. The number of data bits to encode is determined by each modulation scheme. For example, Quadrature Phase Shift Keying (QPSK) encodes two bits using a symbol, 8-state quadrature amplitude modulation (8-QAM) does three bits, and 16-QAM does four bits. When digital modulation is performed, there exists a slight error because of manufacturing imperfections.

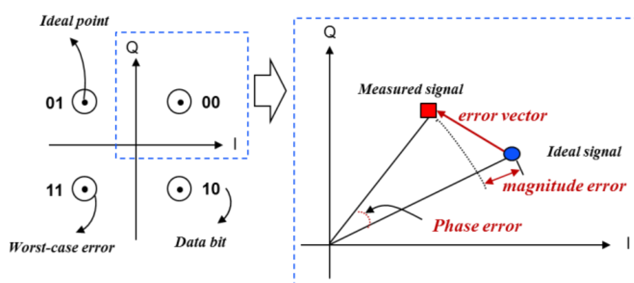


Figure 1. Modulation Error in QPSK.

Fig. 1 shows digital modulation error in QPSK. QPSK can send two bits of digital information at a time. In QPSK, data is encoded using two independent carrier components, in-phase (I) and quadrature phase (Q). The carrier components are separated in phase by π /2. If data is sent by transmitter using digital modulation, the signal measured in the receiver should be mapped to the ideal point as shown in Fig. 1. However, the measured signal does not match with ideal point because of various reasons such as hardware impairments, channel characteristics, and noise at the receiver. Modulation error becomes the most important element for identifying wireless devices because wireless transceivers made not only by different manufacturer, but also even by same manufacturer have different values. To quantify the performance of a digital radio transmitter or a receiver, it is can be used several metrics such as EVM, magnitude error, phase error, as shown in Fig. 1. In this paper, we use EVM as a feature for distinguishing wireless devices. EVM is vector magnitude difference between an ideal reference signal and measured signal.

B. Architecture and Algorithm

Fig. 2 shows architecture for MAC spoofing attack detection proposed in this paper. The architecture includes four components: signal acquisition, EVM extraction, training, and MAC spoofing detection.

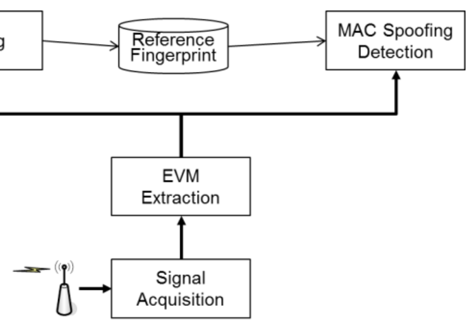


Figure 2. Architecture for MAC spoofing attack detection.

The signal acquisition is a hardware component. It monitors WLAN and captures the radio signal of wireless devices. The EVM extraction component has a responsibility for extracting EVM and MAC address from the radio signal. The training component is performed before the MAC spoofing detection module starts. It measures EVM values for authorized wireless devices and stores them into the reference fingerprint DB. Finally, MAC spoofing detection component accepts or rejects newly detected devices by comparing the measured EVM and the reference EVM.

$$d_1(x) = \sum_{i=1}^{k_1} \frac{ed\left(x - NN_i(x)\right)}{k_1} \quad (1)$$

$$d_2(x) = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2} \frac{ed\left(NN_i(x) - NN_j(NN_i(x))\right)}{k_1 \cdot k_2} \quad (2)$$

In this paper, we employ K-NNDD (K-Nearest Neighbor Data Description) [11] as an algorithm for training EVM for authorized wireless devices and for comparing the measured EVM and the reference EVM. K-NNDD has been proposed to solve one-class classification. In this paper, K-NNDD is used to decide whether a new device is authorized one or attacking one with spoofed MAC address because it can calculate distance between the measured fingerprint and the reference fingerprint.

K-NNDD is explained through (1) and (2). The distance between the measured fingerprint and the reference fingerprint is calculated by (1). The distance between the reference fingerprints is calculated by (2). If the distance calculated by (1) is less than the distance calculated by (2), the measured fingerprint is regarded as the same one as the reference fingerprint. In (1) and (2), *ed* means Euclidean distance. $NN_i(x)$ is the ith nearest reference EVM value to the measured one, x. K_1 indicates the number of reference EVM values to be compared with the measured EVM value. K_2 indicates the number of reference EVM values to be compared with each of K_1-number of reference EVM values. In (1), $d_1(x)$ means average distance between the measured EVM and its nearest reference EVM fingerprints. In (2), $d_2(x)$ means average distance between the reference EVM fingerprints selected in (1) and their nearest reference EVM fingerprints. If $(d_1(x) / d_2(x))$ is greater than a threshold, then our scheme regards the wireless device with the EVM value x as one with a spoofed MAC address because the distance between the measured EVM and the reference EVM is farther than the distance between the reference EVM values. Typically, the value of the threshold in K-NNDD is set to 1.

IV. PERFORMANCE EVALUATION

A. Experimental environment

We have implemented a prototype system called a MAC spoofing detection sensor which provides EVM-based MAC spoofing attack detection.

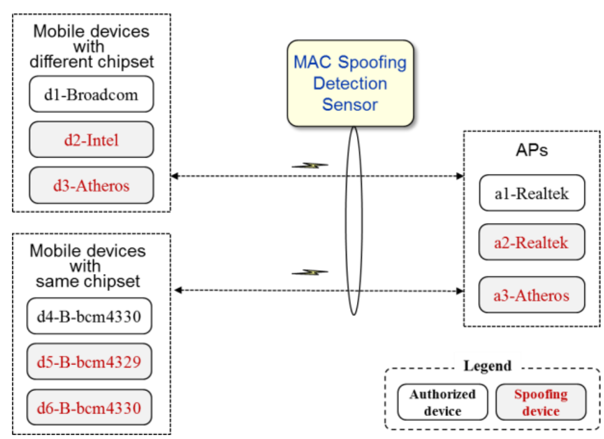


Figure 3. Experimental network.

The sensor consists of Athoros 9380 wireless LAN chipset, Intel atom CPU, 4GB memory, PoE, and so on. The Athoros chipset in the proposed sensor monitors WLAN and captures the radio signal of wireless devices. To extract EVM and MAC address from the radio signal, we have modified a WLAN device driver called compat-driver. In our scheme, the training and the MAC spoofing detection components have been implemented as application software. The extracted EVM values include abnormal ones. So, the training and the MAC spoofing detection components choose the best 15 EVM values among the received 20 ones.

Fig. 3 shows an experimental network for measuring and evaluating the performance of the proposed scheme. The experimental network consists of a MAC spoofing detection sensor, mobile devices with different Wi-Fi chipset, mobile devices with same Wi-Fi chipset, and APs. As mobile devices with different chipset, d1-Broadcom, d2-Intel, d3-Atheros each indicates a smart-phone with Broadcom chipset, a laptop computer with Intel chipset, and a laptop computer with Atheros chipset, respectively. As mobile devices with same chipset, d4-B-bcm4330 and d6-B-bcm4330 are smart-phone (iphone-4s) with Broadcom's BCM 4330 chipset. The d5-B-bcm4329 means smart-phone (iphone-4) with Broadcom's BCM 4329 chipset. As AP, a1-Realtec and a2-Realtec both are ipTIME AP with Realtek RTL8198. The a3-Atherros means Netgear AP with Atheros AR7161. The APs stay away from each other about 1m. In Fig. 3, white box and gray box mean authorized wireless device and MAC spoofing device, respectively.

B. *Experiment Results*

The experiment results of the proposed scheme are described in Fig. 4, 5, and 6. In Fig. 4 and 5, x axis indicates a device identification execution number and y axis indicate distance/dissimilarity between the measured EVM for a device and its reference EVM.

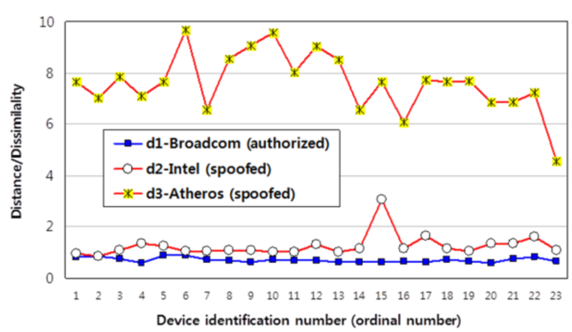


Figure 4. Performance of MAC spoofing attack detection against mobile devices with different Wi-Fi chipset.

Fig. 4 shows the performance of MAC spoofing attack detection against mobile devices with different Wi-Fi chipset. In Fig. 4, d1-Broadcom (device with Broadcom chipset) is an authorized device and its EVM fingerprint is trained and stored into DB before the MAC spoofing detection operation is started. The d2-Intel and the d3-Atheros both are attacking devices that spoof the MAC address of d1-Broadcom.

In the average distance ($d_1(x)$ / $d_2(x)$) between the measured EVM and the reference EVM, d1-Broadcom, d2-Intel, and d3-Atheros are about 0.7, 1.3 and 7.6, respectively, as shown in Fig. 4. This indicates that the wireless devices with different Wi-Fi chipset have different values in EVM. When we set the threshold for deciding between the authorized device and the spoofed device to 1, our scheme decides d1-Broadcom as authorized one, and d2-Intel and d3-Atheros as attacking one with spoofed MAC. Therefore, Fig. 4 means that the proposed scheme can almost perfectly detect MAC spoofing attack in case that attacker employs wireless devices with different Wi-Fi chipset from that of the authorized device. In Fig. 4, False Accept Rate (FAR) is 4.35% and False Reject Rate (FRR) is 0.0%. EER (Equal Error Rate) is 1.09 % (when threshold = 0.93). The 'false accept' means that the system mistakes device with spoofed MAC as authorized one. On the contrary, the 'false reject' means that mistakes authorized device as one with spoofed MAC. EER indicates the value where FAR and FRR are equal.

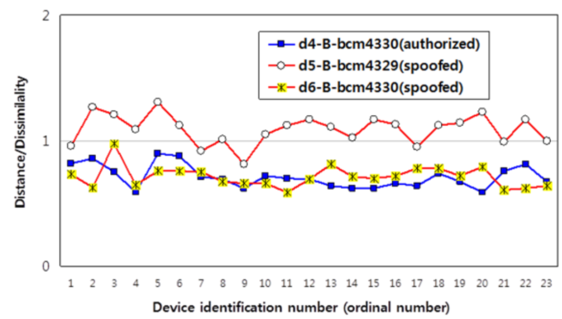


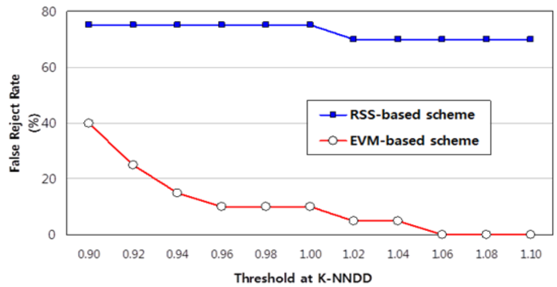
Figure 5. Performance of MAC spoofing attack detection against mobile devices with same Wi-Fi chipset.

Fig. 5 shows the performance of our MAC spoofing attack detection against mobile devices with same Wi-Fi chipset. In Figure 5, d4-B-bcm4330 is authorized device. D5-B-bcm4329 and d6-B-bcm4330 both are ones with spoofed MAC address. The average distances in d4-B-bcm4330, d5-B-bcm4329, and d6-B-bcm4330 are about 0.7, 1.1 and 0.7, respectively, as shown Fig. 5. Fig. 5 indicates that even if our scheme is not bad in case that attacker employs device with same chipset but different model (i.e., d5-B-bcm4329), it has great difficulty in detecting attack devices with same chipset and same model (i.e., d6-B-bcm4330). In Fig. 5, FAR is 63.04% and FRR is 0.0%. EER is 31.52% (when threshold = 0.74).

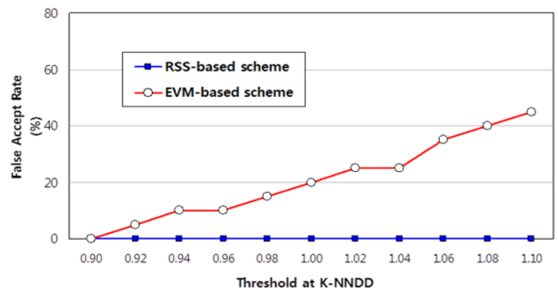
Fig. 6 shows the performance of our detection scheme against rogue AP that spoofs MAC address. In this experiment, our scheme (i.e., EVM-based scheme) is compared with RSS-based scheme. RSS-based scheme is commonly used to detect rogue APs because it is comparatively easy to measure RSSI on WLAN and APs that are located in different place have different RSSI values. In Fig. 6, x axis indicates threshold value at K-NNDD algorithm and y axis indicates the error rate of MAC spoofing detection schemes.

Fig. 6-(a), (b), and (c) shows FRR of two detection schemes against the authorized AP (a1-Realtec), FAR of them against the spoofed AP with same Wi-Fi chipset (a2-Realtec), and FAR of them against the spoofed AP with different Wi-Fi chipset (a3-atheros), respectively. In Fig. 6-(a), our scheme is much better than RSS-based scheme in terms of FRR. When the threshold is set to 1.0, FAR of our

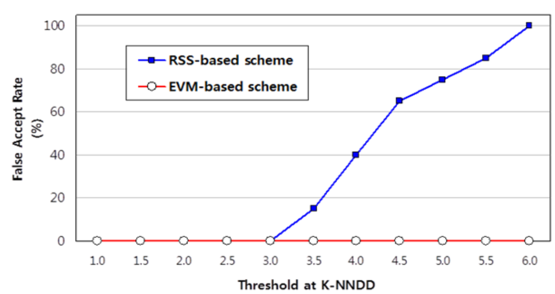
scheme is 10% while that of RSS-based scheme is 75%. We think that this happens because RSSI is more sensitive to radio frequency interference than EVM.



(a) Against the authorized AP (a1-Realtec)



(b) Against the spoofing AP with same Wi-Fi chipset (a2-Realtec)



(c) Against the spoofing AP with different Wi-Fi chipset (a3-atheros)

Figure 6. Performance of the spoofing AP detection by RSS-based and EVM-based scheme.

In case of FAR of the detection schemes against the spoofing AP with same Wi-Fi chipset, RSS-based scheme is better than our scheme, as shown in Fig. 6-(b). As mentioned above, our scheme is not good in detecting device that spoofs MAC address using same Wi-Fi chipset. However, the performance between our scheme and RSS-based scheme does not make big difference. When the threshold is set to 1.0, FAR of our scheme is 20% while that of RSS-based scheme is 0%. In Fig. 6-(c), our scheme is better than RSS-based scheme in terms of FAR against spoofing AP with different Wi-Fi chipset. As mentioned above, this is because our scheme is very good in detecting device that spoofs MAC address using different Wi-Fi chipset.

V. CONCULSION AND FUTURE WORK

In this paper, we have proposed an EVM-based MAC spoofing attack detection scheme. The proposed scheme belongs to RF fingerprinting one. To measure and evaluate the performance of the proposed scheme, we have developed a prototype hardware system. The experiment results prove that our scheme is excellent in detecting MAC spoofing attack that employs a wireless device with different Wi-Fi chipset from that of legitimate one. However our scheme is not good in case that attacker employs a wireless device with same Wi-Fi chipset as that of legitimate one.

Our future work is to add new other features besides EVM to enhance the performance of the proposed scheme as highly as it can detect the sophisticated attacks using a wireless device with same Wi-Fi chipset from that of legitimate one. Another future work is to construct a large-scale of experiment environment that consists of lots of and various kinds of wireless devices to verify and refine the proposed scheme.

ACKNOWLEDGMENT

This research was funded by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013.

REFERENCES

[1] G. Lackner, U. Payer, and P. Teufl, "Combating Wireless LAN MAC-layer Address Spoofing with Fingerprinting Methods," International Journal of Network Security,Vol.9, No.2,pp.164-172, Sept. 2009.

[2] B. Danev, D. Zanetti, and S. Capkun, "On Physical-Layer Identification of Wireless Devices," ACM Computing Surveys, Vol. 45, No. 1, Nov. 2012.

[3] Z. Yang, A. C. Champion, B. Gu, X. Bai, and D. Xuan, "Link-Layer Protection in 802.11i WLANs with Dummy Authentication," ACM WiSec, pp. 131-138, March 2009.

[4] F. Guo and T. Chiueh, "Sequence number-based MAC address spoof detection, " Lecture Notes in Computer Science Volume 3858, pp 309-329, 2006.

[5] D. Madory, "New Methods of Spoof Detection in 802.11b Wireless Networking," A Thesis Submitted to the Faculty in partial fulfillment of the requirements for the degree of Master of Science, Darmouth College, June 2006.

[6] Y. Sheng, "Detecting 802.11 MAC Layer Spoofing Using Received Signal Strength," INFOCOM 2008, pp. 1768-1776, April 2008.

[7] B. Danev and S. Capkun, "Transient-based identification of wireless sensor nodes," International Conference on Information Processing in Sensor Networks (IPSN), pp. 25-36, April 2009.

[8] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, "Wireless Device Identification with Radiometric Signatures," ACM MobiCom, pp. 116-127, Sept. 2008.

[9] Y. Shi and M. A. Jensen, "Improved Radiometric Identification of Wireless Devices Using MIMO Transmission," IEEE Transactions on Information Forensics and Security, Vol. 6, Issue 4, pp. 1346-1354, Dec. 2011.

[10] B. Danev, H. Luecken, S. Capkun, and K. El Defrawy, "Attacks on Physical-layer Identification," 10th ACM Conference on wireless network security (WiSec), pp. 89-98, March 2010.

[11] J. Son and S. Kim, "kNNDD-based One-Class Classification by Nonparametric Density Estimation," Journal of the Korean Institute of Industrial Engineers, Vol. 38, No. 3, pp. 191-197, Sep. 2012.