MOBILITY 2014

The Fourth International Conference on Mobile Services, Resources, and Users

ISBN: 978-1-61208-366-7

July 20 - 24, 2014

Paris, France

MOBILITY 2014 Editors

Josef Noll, University of Oslo & Movation, Norway
Yan Cimon, Université Laval, Canada

MOBILITY 2014

Foreword

The Fourth International Conference on Mobile Services, Resources, and Users (MOBILITY 2014), held between July 20-24, 2014, in Paris, France, continued a series of events dedicated to mobility-at-large, dealing with challenges raised by mobile services and applications considering user, device and service mobility.

Users increasingly rely on devices in different mobile scenarios and situations. "Everything is mobile", and mobility is now ubiquitous. Services are supported in mobile environments, through smart devices and enabling software. While there are well known mobile services, the extension to mobile communities and on-demand mobility requires appropriate mobile radios, middleware and interfacing. Mobility management becomes more complex, but is essential for every business. Mobile wireless communications, including vehicular technologies bring new requirements for ad hoc networking, topology control and interface standardization.

We take here the opportunity to warmly thank all the members of the MOBILITY 2014 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to MOBILITY 2014. 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 MOBILITY 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that MOBILITY 2014 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the areas of mobile services, resources and users.

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

MOBILITY 2014 Chairs:

MOBILITY Advisory Committee

Josef Noll, University of Oslo & Movation, Norway
Pekka Jäppinen. Lappeenranta University of Technology, Finland
Abdulrahman Yarali, Murray State University, USA
Alexandre Caminada, Laboratoire Systèmes et Transport (UTBM) - Belfort, France
Yan Cimon, Université Laval, Canada
Masashi Sugano, Osaka Prefecture University, Japan
Einar Snekkenes, Gjovik University College (HiG), Norway
Claudia Linnhoff-Popien, Ludwig-Maximilians-University, Germany

MOBILITY Industry/Research Chairs
Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Kamill Panitzek, Telecooperation Lab, Technische Universität Darmstadt, Germany
Matthias Trojahn, Volkswagen AG, Germany
Ulrich Meissen, Fraunhofer FOKUS, Germany
Xun Luo, Qualcomm Inc. - San Diego, USA
Jingli Li, TopWorx, Emerson, USA
Jiankun Hu, Australian Defence Force Academy - Canberra, Australia
Knut Øvsthus, Høgskolen i Bergen (HiB), Norway
Massimo Paolucci, DOCOMO Euro-Labs, Germany
Andrey Somov, CREATE-NET, Italy
Danny Soroker, IBM T.J. Watson Research Center, USA
Sonja Schmer-Galunder, SIFT, USA
Tim Strayer, BBN Technologies, USA
Lars Svensson, German National Library, Germany

MOBILITY Special Area Chairs
Video
Mikko Uitto, VTT Technical Research Centre of Finland, Finland
Mobile Wireless Networks
Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Masashi Sugano, Osaka Prefecture University, Japan
Mobile Web / Application
In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea
Mobile Internet of Things and Mobile Collaborations
Nils-Olav Skeie, Telemark University College, Norway
Vehicular Mobility
Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy
Mobile Cloud Computing
Chunming Rong, University of Stavanger, Norway

MOBILITY 2014 Publicity Chairs
Waqas Aman, Gjøvik University College, Norway
Swagato Barman Roy, School of Computer Engineering, Singapore
Esua Kinyuy Jaff, University of Bradford, United Kingdom
Hiroyuki Hatano, Utsunomiya University, Japan

MOBILITY 2014

COMMITTEE

MOBILITY Advisory Committee

Josef Noll, University of Oslo & Movation, Norway
Pekka Jäppinen. Lappeenranta University of Technology, Finland
Abdulrahman Yarali, Murray State University, USA
Alexandre Caminada, Laboratoire Systèmes et Transport (UTBM) - Belfort, France
Yan Cimon, Université Laval, Canada
Masashi Sugano, Osaka Prefecture University, Japan
Einar Snekkenes, Gjovik University College (HiG), Norway
Claudia Linnhoff-Popien, Ludwig-Maximilians-University, Germany

MOBILITY Industry/Research Chairs

Filipe Cabral Pinto, Telecom Inovação S.A., Portugal
Kamill Panitzek, Telecooperation Lab, Technische Universität Darmstadt, Germany
Matthias Trojahn, Volkswagen AG, Germany
Ulrich Meissen, Fraunhofer FOKUS, Germany
Xun Luo, Qualcomm Inc. - San Diego, USA
Jingli Li, TopWorx, Emerson, USA
Jiankun Hu, Australian Defence Force Academy - Canberra, Australia
Knut Øvsthus, Høgskolen i Bergen (HiB), Norway
Massimo Paolucci, DOCOMO Euro-Labs, Germany
Andrey Somov, CREATE-NET, Italy
Danny Soroker, IBM T.J. Watson Research Center, USA
Sonja Schmer-Galunder, SIFT, USA
Tim Strayer, BBN Technologies, USA
Lars Svensson, German National Library, Germany

MOBILITY Special Area Chairs

Video
Mikko Uitto, VTT Technical Research Centre of Finland, Finland

Mobile Wireless Networks
Mohammad Mushfiqur Chowdhury, University of Oslo, Norway
Masashi Sugano, Osaka Prefecture University, Japan

Mobile Web / Application
In-Young Ko, Korea Advanced Institute of Science and Technology (KAIST), Korea

Mobile Internet of Things and Mobile Collaborations
Nils-Olav Skeie, Telemark University College, Norway

Vehicular Mobility
Gianluca Franchino, CEIICP - Scuola Superiore Sant'Anna - Pisa, Italy

Mobile Cloud Computing
Chunming Rong, University of Stavanger, Norway

MOBILITY 2014 Publicity Chairs

Waqas Aman, Gjøvik University College, Norway
Swagato Barman Roy, School of Computer Engineering, Singapore
Esua Kinyuy Jaff, University of Bradford, United Kingdom
Hiroyuki Hatano, Utsunomiya University, Japan

MOBILITY 2014 Technical Program Committee

Jemal Abawajy, Deakin University - Geelong, Australia
Ioannis Anagnostopoulos, University of Central Greece, Greece
Payam Barnaghi, University of Surrey, UK
Mostafa Bassiouni, University of Central Florida - Orlando, USA
Alessandro Bazzi, IEIIT-CNR, Italy
Yuanguo Bi, Northeastern University, China
Evangelos Bekiaris, CERTH/HIT, Greece
Paolo Bellavista, University of Bologna, Italy
Rajendra V Boppana, University of Texas - San Antonio, USA
Paolo Bouquet, University of Trento, Italy
Carlos Carrascosa Casamayor, Universidad Politécnica de Valencia, Spain
Ciro Cattuto, Data Science Lab - ISI Foundation, Italy
Ioannis Christou, Athens Information Technology, Greece
Yan Cimon, Université Laval, Canada
Cherita Corbett, Johns Hopkins University, USA
Klaus David, University of Kassel, Germany
Claudia de Andrade Tambascia, CPqD Foudation, Brazil
Amnon Dekel, Hebrew University of Jerusalem. Israel
Emanuele Della Valle, Politecnico di Milano, Italy
Raimund Ege, Northern Illinois University, USA
Gianluigi Ferrari, University of Parma, Italy
Gianluca Franchino, TeCIP - Scuola Superiore Sant'Anna - Pisa, Italy
Xiaoying Gan, Shanghai Jiao Tong University, China
Thierry Gayraud, Université de Toulouse, France
Chris Gniady, University of Arizona, USA
Javier Manuel Gozalvez Sempere, Miguel Hernandez University of Elche, Spain
Mesut Günes, Freie Universität Berlin, Germany
Richard Gunstone, Bournemouth University, UK
Jiankun Hu, Australian Defence Force Academy - Canberra, Australia
Peizhao Hu, NICTA, Australia
Muhammad Ali Imran, University of Surrey, UK
Jin-Hwan Jeong, ETRI (Electronics and Telecommunications Research Institute), Korea
Vana Kalogeraki, Athens University of Economics and Business, Greece

Georgios Kambourakis, University of the Aegean, Greece
Vasileios Karyotis, National Technical University of Athens (NTUA), Greece
Kinda Khawam, University of Versailles / PRISM, France
Moritz Kessel, Ludwig-Maximilians-Universität München, Germany
Nikos Komninos, Athens Information Technology - Peania, Greece
Ioannis Krikidis, University of Cyprus, Greece
Abderrahmane Lakas, United Arab Emirates University, United Arab Emirates
Qinxue Li, Guangdong University of Petrochemical Technology, China
Jingli Li, TopWorx, Emerson, USA
Xun Luo, Qualcomm Research Center, USA
Yuanjia Ma, Guangdong University of Petrochemical Technology, China
Stephane Maag, Telecom SudParis, France
Dario Maggiorini, University of Milano, Italy
Charif Mahmoudi, LACL - Paris 12 University, France
Philipp Marcus, Institute for Informatics - Ludwig Maximilian University of Munich, Germany
Kirk Martinez, University of Southampton, UK
Barbara M. Masini, CNR - IEIIT, University of Bologna, Italy
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Stefan Michaelis, TU Dortmund University, Germany
Masayuki Murata, Osaka University, Japan
Simin Nadjm-Tehrani, Linköping University, Sweden
Priyadarsi Nanda, University of Technology Sydney, Australia
Fatemeh Nikayin, Delft University of Technology, The Netherlands
Ryo Nishide, Ritsumeikan University, Japan
Antonino Orsino, University "Mediterranea" of Reggio Calabria, Italy
Shumao Ou, Oxford Brookes University, UK
Knut Øvsthus, Høgskolen i Bergen (HiB), Norway
Massimo Paolucci, DOCOMO Euro-Labs, Germany
Evangelos Papapetrou, University of Ioannina, Greece
Symeon Papavassiliou, National Technical University of Athens, Greece
Marco Picone, University of Parma, Italy
Laurence Pilard, Université de Versailles, France
Przemyslaw Pochec, University of New Brunswick, Canada
Stefan Poslad, Queen Marry University of London, UK
Daniele Puccinelli, University of Applied Sciences of Southern Switzerland (SUPSI), Switzerland
Joel Rodriques, University of Beira Interior - Covilhã / Instituto de Telecomunicações, Portugal
Anna Lina Ruscelli, TeCIP Institute - Scuola Superiore Sant'Anna, Italy
Djamel Sadok, Federal University of Pernambuco, Brazil
Farzad Salim, Queensland University of Technology - Brisbane, Australia
Abdolhossein Sarrafzadeh, Unitec Institute of Technology, New Zealand
Stefan Schmid, TU-Berlin, Germany
Christelle Scharff, Pace University - New York City, USA
Minho Shin, Myongji Unversity, South Korea
Behrooz Shirazi, Washington State University, USA
Lei Shu, Guangdong University of Petrochemical Technology, China
Sabrina Sicari, Università degli studi dell'Insubria, Italy
Andrey Somov, CREATE-NET, Italy
Danny Soroker, IBM T.J. Watson Research Center, USA

Tim Strayer, BBN Technologies, USA
Masashi Sugano, Osaka Prefecture University, Japan
Lars Svensson, German National Library, Germany
Javid Taheri, The University of Sydney, Australia
Vahid Taslimi, Wright State University, USA
Matthias Trojahn, Otto-von-Guericke University Magdeburg, Germany
Wantanee Viriyasitavat, Mahidol University, Thailand
Miao Wang, Free University Berlin, Germany
Wei Wang, University of Surrey, UK
Rainer Wasinger, The University of Tasmania, Australia
Stephen White, University of Huddersfield, UK
Hui Wu, University of New South Wales, Australia
Zheng Yan, Xidian University, China / Aalto University, Finland
Lei Ye, No SQL Database Service - Amazon AWS, USA
Chansu Yu, Cleveland State University, USA
Ting Zhu, State University of New York, USA

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

A Design of Mobile Trusted Module for Application Dedicated Cryptographic Keys

Daewon Kim, Yongsung Jeon, and Jeongnyeo Kim
Cyber Security Research Department
Electronics and Telecommunications Research Institute
Daejeon, Korea
emails: {dwkim77, ysjeon, jnkim}@etri.re.kr

***Abstract*—Normally, users encrypt data with cryptographic keys to protect original contents from various hackings. The use of cryptographic keys means that the protection of cryptographic keys is also an important problem as much as that of the encrypted data. A common way for protecting the keys is to authenticate user's key access authorities through some key passwords. However, nowadays the passwords can be easily exposed to a variety of password hacking techniques. The facts that the encrypted data is stored in unsafe storage, such as hard disk drivers or secure digital memory cards and that the cryptographic keys are accessed with any passwords mean that the encrypted original contents are no longer safe from the hackings. It is because hackers can decrypt user's encrypted data with the acquired passwords after they modify user's original applications or create new malicious applications. To solve this issue, we have developed a new mobile trusted module chip and management middleware based on the architecture with a key access mechanism dedicated to an application. In this paper, we present the design and operation of mobile trusted module chip and middleware together with some experimental results.**

Keywords-trusted platform module; mobile trusted module; hardware security module; integrity verification; cryptography.

I. INTRODUCTION

Data encryption is a common way to protect original contents from data hackings. For encrypting, some cryptographic keys and the authorities, which can be passwords for using the keys, are required. Normally, the encrypted data and keys are stored in some unsafe storage devices, which may be Hard Disk Drivers (HDD) or Secure Digital (SD) memory cards. Moreover, the passwords with authorities for accessing the keys can be easily exposed to various password hacking techniques. It means that the encrypted data can be decrypted by the malicious applications modified or created by hackers who already know the passwords of cryptographic keys. Finally, the traditional systems for managing cryptographic keys and passwords cannot guarantee the confidentiality of original contents included in the encrypted data from hackers.

From a few years ago, some hardware modules [1]-[5] have been used for encrypting data and for managing cryptographic keys. They are representatively Hardware Security Module (HSM) [1]-[4], Trusted Platform Module (TPM) [5], Mobile Trusted Module (MTM) [5], and so on. The hardware modules independently attached to user's devices include access control functions and require any authority information such as a password to access the critical data, which may be cryptographic keys, in the modules. Therefore, although hackers acquire the privileged authority of a target device itself, they cannot directly get important data in the modules.

The hardware modules can partially provide the safe storage for important data such as cryptographic keys and the access control functions commonly based on passwords. However, the reliability of password safety is gradually decreasing by a variety of password hacking techniques such as key hooking, screen capturing and social engineering methods. Additionally, if the password inputs are frequently requested, it can make normal users uncomfortable. In our previous work [9], we discussed some problems of password-based key access and proposed a mechanism verifying the integrity of application executed for accessing cryptographic keys in our MTM.

Another consideration of key management based on previous hardware modules is that the cryptographic keys can be accessed by each other applications. It means that malicious applications created by hackers can use the keys as well. Moreover, if several regular applications share a cryptographic key, due to a password exposed to any security vulnerability of an application the encrypted data of other applications sharing the key can also be at risk of information leakage. As the mobile work environment of Bring-Your-Own-Device (BYOD) is spreading more, the issue needs to be treated carefully.

As our considerations of the above problems, in this paper, we describe the design and operation of our prototype MTM and management software. The paper also includes a brief of our previous work related to the authority for accessing keys in the MTM. In the MTM, there are cryptographic keys dedicated to the authorized application and the application only can use the keys in the MTM. It means that hackers cannot use the keys to decrypt target encrypted data through the applications maliciously made by them although they know the passwords of keys. Additionally, our prototype basically supports the default cryptographic keys per application for users. Therefore, users do not need to create and manage the cryptographic keys with complex options for the simple and quick encryptions. It is sure that the users can

 ISBN: 978-1-61208-366-7

also create and manage the keys with various options suitable to their purposes.

The rest of paper is organized as follows; Section II represents the related works about hardware modules to be attached to user devices for securities. Section III describes the operations and features of our MTM architecture together with our previous work [9]. Section IV shows the detailed operations of our mechanism with some experiment results. Finally, Section V concludes this paper with a summary and future works.

II. RELARED WORKS

HSMs [1]-[4] are hardware modules including cryptographic hardware engines for specific services. They are commonly composed of cryptographic key generator, public key cryptography engine, symmetric key cryptography engine, the composition engine of two cryptographies, random number generator, and so on. Internal critical keys of HSMs do not be exposed to any outside and for accessing or using the keys, application users have to input any passwords to the HSMs. Normally, additional safe storage is not supported in them.

TPM and MTM [5] are hardware modules for user device security introduced by Trusted Computing Group (TCG), and TCG is documenting specifications for hardware modules and software stacks. HSM is a hardware module specified to a service, and otherwise TPM and MTM have the feature of common platform with the standardized Application Programming Interfaces (APIs) for data protection. They have an Endorsement Key (EK), which is a key embedded from the factory, and a Storage Root Key (SRK) is generated from the EK if a user gets the ownership of them. Other cryptographic keys are generated by the random number generator and cryptographic key generator in TPM and MTM, and after the keys are encrypted by key chains started from the SRK, the encrypted keys are stored into the file system of user device. Data encrypted by the cryptographic keys are confidential because the EK and SRK are not exposed to the outside of TPM and MTM. For using the keys that the platform of TPM and MTM manage, application users have to input passwords through TCG key management APIs.

There are a few researches [6]-[8] for simplifying the complex processes for accessing the keys in TPM and MTM. They provide wrapper APIs that integrate TCG APIs with commonly needed functions and minimize the effort that developers write applications with the standardized TCG APIs. However, user-friendly minimization of TCG APIs has a limit because its integration level is bounded under the basic hardware operations of TPM and MTM.

Our previous work [9] described a mechanism to authenticate an application for accessing the important data and keys in the new MTM designed by us. When an application with a certificate is installed into a user mobile device with the MTM, the integrity information of application is stored into the MTM. The application executed by a user is verified with the integrity information in the MTM and acquires the authority for accessing the application dedicated data in the MTM. The dedicated data such as cryptographic keys and private information can be accessed only by the verified application.

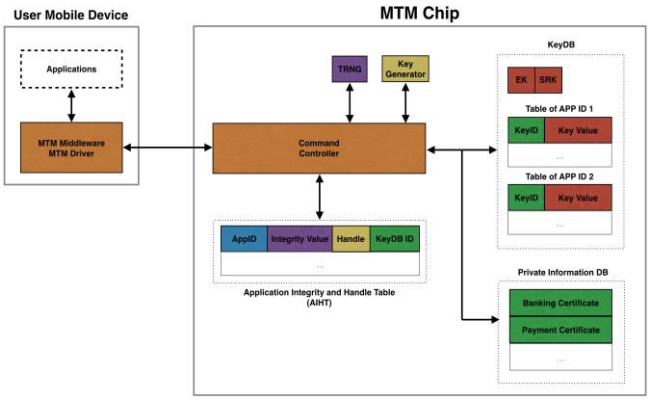


Figure 1. The overview of MTM architecture.

III. THE MTM ARCHITECTURE FOR APPLICATION DEDICATED CRYPTOGRAPHIC KEYS

A. Overview

Fig. 1 shows the architecture for presenting the operation mechanism of our MTM. The detailed mechanism for verifying application integrity has been described in our previous work [9]. By the previous work, if an application is regally installed in a user mobile device, the AppID and the measured integrity value of the application are stored into Application Integrity and Handle Table (AIHT) of our MTM and a default Storage Key (SK) of KeyID 0 is created in an APP ID Table of KeyDB. The SK is for the reserved key slot and is not used currently. The Table ID of KeyDB is also recorded to the KeyDB ID related to the integrity-verified application in AIHT.

When the installed application is executed by the user, the MTM middleware with an Integrity Measurement Agent [9] verifies the integrity of application and inserts an application-specific handle value randomly generated by the middleware into the MTM. The handle value is related to a communication channel between the middleware and the application. The middleware adds the handle value to the commands requested by the application and transfers them to the MTM. Through the handle value, MTM can access a table with APP ID in KeyDB and Private Information DB. Malicious applications cannot get a handle value from the middleware and cannot directly access to the MTM because they cannot know current handle values in the MTM.

B. Features of Our Prototype MTM

1) Integrity-based MTM Access Authority (IAA): If the integrity of application installed or executed by an user is successfully verified, Command Controller in our MTM inserts a handle value received from the MTM middleware into AIHT. It means that regal user applications can access some information, such as keys in the MTM without any passwords. Therefore, malicious or tampered applications

 ISBN: 978-1-61208-366-7

Figure 2. Experiment environment.

cannot access to the MTM due to none of handle values in MTM.

2) Application-dedicated cryptographic Key access Authority (AKA): The integrity-verified application has an authority for accessing a dedicated cryptographic key table in KeyDB. The middleware adds a pre-allocated handle value to every command messages received from the application. The application can access own keys by cryptography-related commands with the handle value. Therefore, applications cannot access keys of each other applications.

3) Default cryptographic Keys Ready (DKR): MTM prepares three default keys in the application-dedicated key table in KeyDB for supporting cryptography using MTM. If an application is verified and installed in the user mobile device, MTM automatically generates a SK in the key table. The KeyID of SK is zero, and it is first default key for reserving the zero of KeyID and is not used yet. KeyID is a index that the application accesses a key. Applications, normally, can use symmetric and asymmetric cryptographies, which are representatively Advanced Encryption Standard (AES) [10] and Rivest Shamir Adleman (RSA) [11]. If for the first time the application uses a cryptography command, our middleware and MTM processes the command after they automatically generates an symmetric or asymmetric key set. We defined the generated default keys with a 256-bit symmetric key of KeyID 1 and 2048-bit asymmetric keys of KeyID 2. Therefore, the application can simply use 256-bit AES and 2048-bit RSA, and it can also create new keys with user-defined KeyIDs.

IV. EXPERIMENTS

For verifying the cryptography feasibility of our MTM and middleware, we have experimented with a few examples in Android environment. Fig. 2 shows the experiment environment. In Fig. 2, a left upper part is the board that MTM chip has been mounted and a left lower is the interface board connecting user mobile device to the MTM board.

A. Specifications

For MTM, we use a smart card chip which includes CPU, OS, memory, and so on. The size of MTM chip is width 5 mm and height 5 mm. The MTM can support each 10 key sets about 300 dedicated applications except for normal applications. In this paper, we do not explain other detailed

```
public static class BindInfo
{
  int in_KeyID;
  int in_DataSize;
  byte[] in_Data;
  int out_DataSize;
  byte[] out_Data;

  public BindInfo()
  {
    // in_KeyID = 2;
    in_KeyID = MTM_KEYID_DEFAULT_ASYMMETRIC;
  }
}
```

(a)

```
MTM mtm = new MTM();
MTM_Crypto.BindInfo bind_info = new MTM_Crypto.BindInfo();

mtm.Open();

// Bind
bind_info.in_DataSize = 4;
bind_info.in_Data = new byte[] {
  (byte)0xaa, (byte)0xbb, (byte)0xcc, (byte)0xdd};
mtm.Bind (bind_info);

// User processing

// UnBind
bind_info.in_DataSize = bind_info.out_DataSize;
bind_info.in_Data = bind_info.out_Data;
mtm.UnBind (bind_info);

...

mtm.Close ();
```

(b)

Figure 3. The sample codes for estimating performance: (a) java class for Bind/UnBind and (b) the example code for Bind/UnBind.

- Kernel version: 3.0.15 SMP PREEMPT
- Android version: 4.0.4
- Core: ARM Cortex-A9 based Dual CPU
- Clock Speed: 1.2GHz
- Internal RAM: 128MB
- Memory: mobile DDR2 1GB, embedded Multi Media Card (eMMC) 16GB

Figure 4. Android mobile device specifications for experiments.

specifications of our MTM chip because the information is confidential yet. Fig. 4 shows user mobile device specification that applications are executed.

B. Experiment Scenario

For cryptography testing, we have experimented RSA encryption and decryption commands. We defined the commands as Bind and UnBind, which are same API names with TCG. The (a) and (b) of Fig. 3 shows real sample codes for Bind and UnBind. Like (b), users and applications do not need to manage keys for basic cryptography because the application dedicated default keys are supported through our middleware and MTM. Fig. 5 shows the message flows in (b) of Fig. 3 between an application and our MTM. The middleware adds a handle value on every command and sends a command for creating a default key to the MTM if there is not an appropriate key in the MTM.

C. Performance Estimation

1) Key Creation: We measured the elapsed time for creating a key set of RSA 2048 bits. As the experiments of 100 times, we presents the average time in Table I. Through the experiments, we also measured the overheads of our

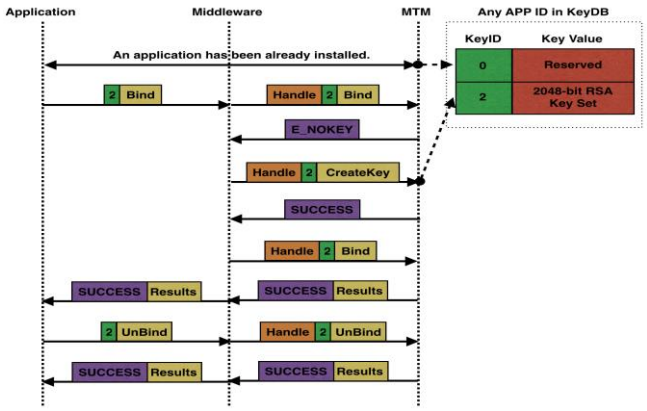


Figure 5. The message flows in (b) of Fig. 3. The middleware can support multiple applications communicating with MTM simultaneously.

softwares. In Table I, MTM Service Provider (MSP) is a client library linked to the application and MTM Core Service (MCS) is a service daemon with middleware functions. The time for generating RSA keys is variable because of the algorithm features of RSA key generation. We are trying to enhance the performance and the issue for waiting users will be treated as future work.

TABLE I. THE PERFORMANCE ESTIMATION OF CREATEKEY

Command Name	Round Trip Time (ms)		
	Application to MTM	*Driver to MTM*	*MSP+MCS Overhead*
CreateKey	5486	5485	1

2) Bind: We measured the elapsed time for encrypting the data of 1 and 214 bytes. In current, our MTM for RSA 2048 bits encryption with Optimal Asymmetric Encryption Padding (OAEP) mode can process maximum 214-byte data per one command message. As the experiments of 1000 times in Table II, we present the average Round Trip Time just from the application to MTM because the software overhead (MSP+MCS) 1 ms is a small part of full elapsed time.

3) UnBind: We measured the elapsed time for decrypting 256-byte data encrypted as RSA 2048 bits. As the experiments of 1000 times, we present the average time in Table II.

TABLE II. THE PERFORMANCE ESTIMATION OF BIND AND UNBIND

Command Name	Application to MTM RTT (ms)		
	1 Byte	*214 Bytes*	*256 Bytes*
Bind	88	116	-
UnBind	-	-	289

Currently, our prototype has about 30 commands including the above three commands.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced the design and operation examples of our MTM and middleware. They have the features of Integrity-based MTM Access Authority (IAA), Application-dedicated cryptographic Key access Authority (AKA), and Default cryptographic Keys Ready (DKR). The features have the advantages that users can access the cryptographic keys, private information, and the secure operation of financial services user-friendly to the MTM. The comprehensive functions are not supported in other hardware security modules yet.

In current prototype, the time for generating RSA keys is too long to wait users when the default RSA key is used for the first time. In future work, we will create the default RSA keys at the time for installing an application instead of the time for using the keys. Additionally, the hardware logics for creating the RSA keys will be optimized.

ACKNOWLEDGMENT

This work was supported by the IT R&D program of MSIP/KCA. [12-912-06-001, Development of the Security Technology for MTM-based Mobile Devices and next generation wireless LAN].

REFERENCES

[1] T. Souza, J. Martina, and R. Custodio, "Audit and backup procedures for Hardware Security Modules," Proc. of the 7th symposium on Identity and trust on the Internet, 2008, pp. 89-97.

[2] B. Rosenberg, "Handbook of Financial Cryptography and Security," Chapman and Hall/CRC, 2010.

[3] J. Kang, D. Choi, Y. Choi, and D. Han, "Secure Hardware Implementation of ARIA Based on Adaptive Random Masking Technique," ETRI Journal, vol. 34, no. 1, Feb. 2012, pp. 76-86.

[4] M. Wolf and T. Gendrullis, "Design, implementation, and evaluation of a vehicular hardware security module," Proc. of the International Conference on Information Security and Cryptology (ICISC 2011), Springer Berlin Heidelberg, 2012, pp. 302-318.

[5] Trusted Computing Group. TPM main specification. Main Specification version1.2 rev116, Trusted Computing Group, March 2011.

[6] G. Cabiddu, E. Cesena, R. Sassu, D. Vernizzi, G. Ramunno, and A. Lioy, "The Trusted Platform Agent," IEEE Software, vol. 28, 2011, pp. 35-41.

[7] C. Stuble and A. Zaerin, "uTSS – A Simplified Trusted Software Stack," Proc. of the 3rd International Conference on Trust and Trustworthy Computing, 2010, pp. 124-140.

[8] R. Toegl, T. Winkler, M. Nauman, and H. Theodore, "Specification and standardisation of a java trusted computing api," Software: Practice and Experience, vol. 42, no. 8, 2012, pp. 945-965.

[9] D. Kim, J. Kim, and H. Cho, "An Integrity-Based Mechanism for Accessing Keys in A Mobile Trusted Module," Proc. of the International Conference on ICT Convergence, 2013, pp. 780-782.

[10] J. Daemen and V. Rijmen, "The Design of Rijndael: AES - The Advanced Encryption Standard," Springer, 2002.

 ISBN: 978-1-61208-366-7

[11] R. Rivest, A. Shamir, and L. Adleman, “A method for obtaining digital signatures and public-key cryptosystems,” Commun. ACM, vol. 21, no. 2, 1978, pp. 120-126.

 ISBN: 978-1-61208-366-7

Predicting Destinations with Smartphone Log using Trajectory-based HMMs

Sun-You Kim, Sung-Bae Cho
Dept. of Computer Science
Yonsei University
Seoul, Korea
sykim@sclab.yonsei.ac.kr, sbcho@yonsei.ac.kr

***Abstract*—With the spread of smartphones, it is easy to obtain sensor data from users, and location-based service (LBS) becomes the most common service in mobile industry. Predicting the user's destination can lead to a variety of services in mobile devices. In addition to the user's final destination, the locations during movement are also important in LBS. In this paper, we propose a destination prediction method based on hidden Markov models for representing the paths using sensor data from smartphone and identifies the destination and intermediate locations in future moving with visiting probabilities. In order to demonstrate the usefulness of the proposed method, we compare it with Dynamic Time Warping (DTW), a method of template matching. Experiments with the data collected by 10 college students for five months confirm that the proposed method results in 12.67 times faster and 2.88 times more accurate than the DTW.**

Keywords-destination prediction; forecasting; hidden Markov model; location-based service.

I. INTRODUCTION

The proliferation of smartphones facilitates to obtain various sensor data from the users, and a wide range of services using a variety of sensor data are introduced. Location-based service (LBS) is the most common service for utilizing the sensor information. It can extract key locations and identify the exact coordinates using user location provided by smartphone. Also, it predicts where the user moves next and provides the information required in the future in advance [1]. As a result, the technology to predict the destinations and movements of the user is required in mobile environment.

Users make a trajectory by moving locations along the flow of time. A trajectory based on the information of locations visited can be obtained. Prediction of destination is to find out the next location in future based on the information of movements until now. In other words, when the information of location movement is $Trajectory_t = \{L_1, L_2, \cdots, L_t\}$ until the current time t, the prediction of location is to find out the location L_{t+n} at time $t+n$.

In order to search L_{t+n}, we should look for a movement which has the same pattern with $Trajectory_t$ in the past [2]. As a path is a subsequence of trajectory, the path becomes a moving pattern. Path is a set of locations, each of which is constructed by temporal and spatial information.

The problem of predicting destination can be defined using the path. The past path $P_{optimal}$ with the highest similarity using the current moving path $P_{present}$ is determined, and the endpoint of path $P_{optimal}$ is said to be the destination $L_{destination}$.

In this paper, we propose a destination prediction method which utilizes hidden Markov models for representing the paths using sensor data from smartphone, estimates the destination L_{t+n} based on $P_{present}$, and finds out the intermediate locations $L \in \{L_{t+1}, L_{t+2}, \cdots, L_{t+n-1}\}$. The problem of destination prediction is shown in Fig. 1.

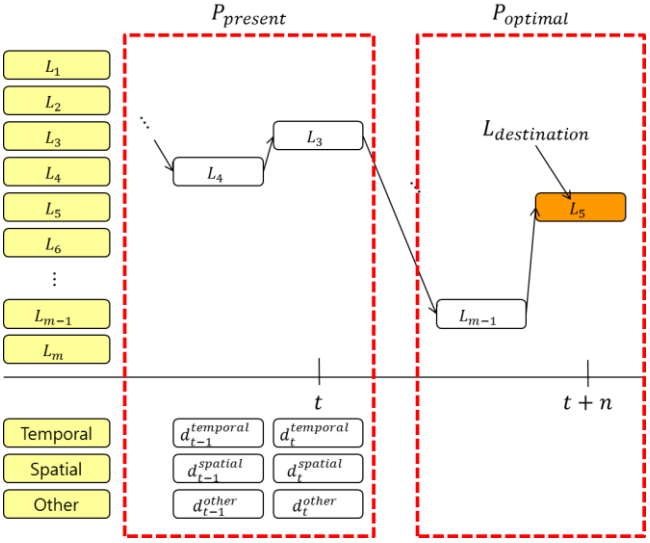


Figure 1. The problem of destination prediction

The rest of this paper is organized as follows. Section II describes the related works in destination prediction. Section III describes the proposed method. The proposed method consists of constructing model and predicting destination. Section IV addresses the result of experiments. Section V summarizes the paper and draws a conclusion.

II. RELATED WORKS

The study of destination prediction is related with comparing the information of previous visits with the current location of visit along the flow of time, in order to identify the destination. The trajectory of locations along the flow of time can be expressed as time series data. There are three types of classification methods for time series data. First, feature-based classification finds the decision boundary. Second, the sequence distance-based classification method classifies using the class of closest distance of time-series data. Third, the

 ISBN: 978-1-61208-366-7

model-based classification method identifies the most probable class after creating the probability model [3].

TABLE I. PREVIOUS STUDIES ON DESTINATION PREDICTION

Year	Author	Data	Method	Method category
2013	Do, et al.	Apps, Call history, Bluetooth	Linear regression	Feature-based
2012	Lu, et al.	GPS, Bluetooth	SVM	Feature-based
2012	Mathew, et al.	GPS	HMM	Model-based
2012	Gambs, et al.	GPS	Mobility Markov chain	Model-based
2011	Kim, et al.	GPS	Bayesian network	Feature-based
2009	Monreale, et al.	GPS	Trajectory pattern tree	Feature-based
2009	Lee, et al.	GPS	DTW	Sequence distance-based
2008	Burbey, et al.	GPS	PPM	Model-based
2007	Akoush, et al.	Cell ID, Cell history, time	Bayesian neural network	Feature-based

A. *Feature-based classification*

The studies which incorporated feature-based classification classified the next destination by using the context information of the current state and the current location. Do, et al. proposed a destination prediction method based on linear regression using location, Apps, call history, and Bluetooth [4]. Lu predicted destinations using SVM with place, Bluetooth, WLAN, and call history as inputs [5]. Monreale, et al. predicted a destination using the method of Trajectory Pattern Tree that uses GPS trajectory [6]. Also, Akoush, et al. performed a destination prediction by applying Bayesian neural networks with cell ID, cell history, and time as variables [7]. In addition, Kim, et al. predicted the user's destination using a Bayesian network created from history information of visited locations in the past [8]. However, the studies using feature-based classification method predicted the next location with fragmented information only. Moreover, because it does not consider the path, it is impossible to know the location information of the intermediate paths.

B. *Sequence distance-based classification*

Sequence distance-based classification, with the use of time series data, is a classification method which works by determining the class which has the highest similarity among the stored templates. Lee, et al. classified the user's destination by using dynamic time warping, which is a method of determining the similarity of the pattern of the two GPS paths [9]. However, because the processing speed increases as the number of templates in the pattern becomes larger, the problem of template management is usually encountered when using this method.

C. *Model-based classification*

The studies using a classification model-based approach is a way to model the sequence data and find the model most similar through matching to the new input. Mathew, et al. made a model using HMM to a sequence of visited locations to derive the destination [10]. Gambs, et al. determined the destination by using the mobility Markov chains from the sequence of the POI (Point of Interest) to make the model [11]. Burbey, et al. predicted the destination by applying the PPM using the residence time and visit time and location [12]. The studies using the model-based classification predict the destination by using only spatial information, such as GPS. However, when using the spatial information only, there is a problem that the prediction is biased to the lower information that lacks of initial movement.

III. THE PROPOSED METHOD

When it comes to represent the path using only the spatial information, prediction methods cannot resolve the problem of partially overlapped paths as shown in Fig. 2. This means that there are main trajectories for users to move. When the user passes through the main trajectory, using only spatial information will result in predicting an incorrect destination which overlaps the main trajectory. In order to work out this problem, we extend the path information by adding time and other context data. Location information, including temporal and spatial among others, is collected by sensors in smartphone environment. Time and date of the smartphone are used as temporal information, and latitude and longitude from the GPS sensor are utilized as spatial information. Also, other context data from accelerometer, magnetic, and orientation sensors determine the mode of transportation of the user.

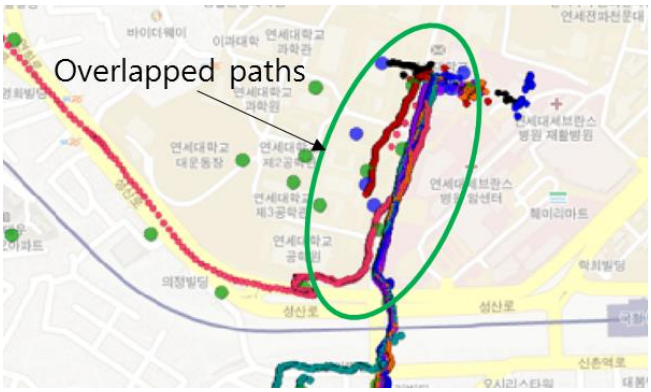


Figure 2. An example of overlapped paths

In this paper, in order to predict the destination, the path information is stored using hidden Markov model (HMM). HMM is a statistical model characterized by Markov process with unknown parameters, modeling observations to determine these hidden parameters. It is a widely used technique that stochastically models sequence data of the time series. It is mainly composed of the state transition probabilities, and the probabilities that select the observation value at each state.

 ISBN: 978-1-61208-366-7

About the path $P = \{L_1, L_2, \cdots, L_n\}$ of length n, the information of location L_t at time t is affected by the information of L_{t-1}. Therefore, it can be assumed to be a Markov process. In this method, because it uses a variety of information of context, we make a model using HMM, which probabilistically represents many features. A HMM is defined by state transition probability A, probability distribution of observed symbols B, and probability distribution of initial state Π. One HMM, λ, is expressed as (1).

$$\lambda = \{A, B, \Pi\} \quad (1)$$

HMM consists of only one probability distribution which is made from various sequence data. Therefore, it eliminates the storage of unnecessary sequence data because new input sequences can be compared with only one model, as opposed to comparing with many sequence data. Thus, HMM is suitable in mobile environment which has limitations in processing time and storage.

In this method, when m_i is the ith HMM that makes a model with the same paths as the departure and destination about $P_{input} = \{L_1, L_2, \cdots, L_t\}$ which is moving path to the time t, we find a HMM model $\hat{m}$ which is the most similar to P_{input}. Destination of $\hat{m}$ is predicted destination. Also after finding the $P_{optimal}$ which is the most similar to P_{input}, we pick up the future visiting locations based on the $P_{optimal}$. The structure of the proposed method is shown in Fig. 3.

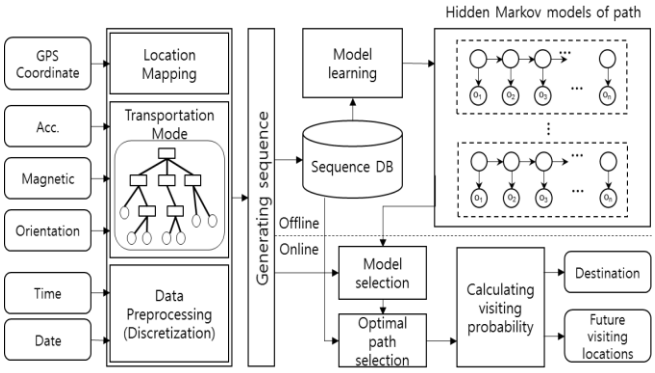


Figure 3. The structure of the proposed method

A. Construction of path sequences

The path sequence is a set of locations along the temporal flow. Each element of the sequence represents the place. The ith location of the path sequence P is represented by (2).

$$L_i = \{time_i, weekday_i, place_i, transportation_i\} \quad (2)$$

The L_i used for the observed symbol of HMM is quantized as follows.

The temporal information of the place, to generate the observed values, separated by time and day of the week. Because the user works in a different pattern on the basis of the time of day and day of the week, we extract the two features in the time information. The time, in the representation of the time zone, has a total of 6 values through the vector quantization, and the day of the week has 7 values.

Spatial information is formed by the latitude and longitude coordinates of a large number. However, as it is not possible to use up all coordinates, it is necessary to extract the key locations. The key locations may be the departure and destination locations or a location to visit during moving such as crossroad, bus stop, subway station, etc. In other words, a key location means user's meaningful location or intermediate location to it. Previous studies to extract key locations use K-means clustering which finds the centers of crowded area in GPS data [12].

However, this method should determine the number of locations for extracting key locations. Also, it cannot guarantee the performance of extraction, because the criteria of density are uncertain. Therefore, we used G-means clustering [13] for extracting key locations which follow a Gaussian distribution in GPS data [14]. The G-means clustering is a clustering method to test each cluster whether Gaussian distribution through statistical verification and repeat the K-means clustering until all clusters follow Gaussian distribution. Extracted locations were labeled by discrete value as the observation symbols for HMM.

Other context information is used by transportation in location. In order to determine the transportation mode, this method classifies the transportation based on accelerometer, magnetic, orientation sensors in smartphone. For converting high-level data to low-level data, this method uses decision tree algorithm which is suitable in mobile device because the recognition speed is faster than other methods. Transportation mode is classified into 4 states, such as Staying, Walking, Running and Vehicle [1].

TABLE II. QUANTIZATION OF INPUT DATA

Type	Low-level data	Quantized data
Temporal information	Time (0-23)	6 separate units for 4 hours
	Day of week	Day of week
Spatial information	GPS coordination	Labeled location
Other context information	Accelerometer sensor value	4 state (Staying, Walking, Running, Vehicle)
	Magnetic sensor value	
	Orientation sensor value	

B. Building path model based on HMM

A path which is a subsequence of trajectory can be generated from many different paths. Because it is not possible to create a model for all paths, we construct a HMM based on the paths of start and end points. The HMM has information about the start and end points of a path. In this way, the number of HMM is reduced, but is not enough because of the large number of locations. When the number of locations is n, $n \times (n-1)$ HMMs are required to create models. So, we build the models only with start and end points of a path for departure and destination.

The HMM with path information is learned by Baum-Welch algorithm [15], which is a learning method typically to

represent the probabilistic information of multiple sequences. Fig. 4 shows the generation process of HMM-based path model in a real data. In this figure, a HMM is created by a pair of departure and destination. This is an example to use the different sequences to model as the same path.

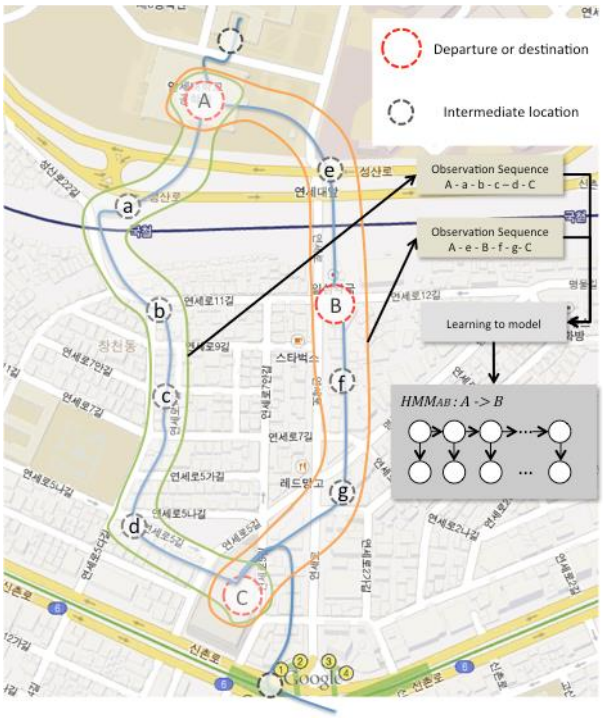


Figure 4. An example of building path model

C. Predicting destination

If a user reaches at the present location with path $P_{present}$ at time t, which is the same with the moved path of length T, P_{past} until time t in past, we assume that moving location of $P_{present}$, in the future, is the same with moving path of P_{past} on $t+1$ to T. Based on this, we define that destination of $P_{present}$ is a destination of HMM model which includes the most similar path with $P_{present}$ until time t.

The similarity of path can be represented by the probability that $P_{present}$ is observed until time t in HMM_{AB} whose departure is A and destination is B. When the sequence of states until time t is $Q = \{q_1, q_2, \cdots, q_t\}$ on λ_{AB} which is the model parameter of HMM_{AB}, the probability of observing $P_{present}$ can be expressed by (3).

$$\mathrm{P}(P_{present}|\lambda_{AB}) = \sum_Q \mathrm{P}(P_{present}, Q|\lambda_{AB})$$

$$= \sum_Q \mathrm{P}(P_{present}|Q, \lambda_{AB})\mathrm{P}(\mathrm{Q}|\lambda_{AB}) \quad (3)$$

$\mathrm{P}(P_{present}|Q, \lambda_{AB})$ is the probability to appear $P_{present}$ given sequence of state and model parameter of HMM_{AB}. Also, $\mathrm{P}(\mathrm{Q}|\lambda_{AB})$ is the probability to select Q given model parameter of HMM_{AB}. Therefore, the probability of observing $P_{present}$ in HMM_{AB} is the sum of $\mathrm{P}(P_{present}|Q, \lambda_{AB})\mathrm{P}(\mathrm{Q}|\lambda_{AB})$ for all the state sequences. Equation (4) is expressed by using Markov process on (3).

$$\sum_Q \mathrm{P}(P_{present}|Q, \lambda_{AB})\mathrm{P}(\mathrm{Q}|\lambda_{AB})$$

$$= \sum_Q (b_{q_1 L_1} \cdots b_{q_t L_t})(\pi_{q_1} a_{q_1 q_2} \cdots a_{q_{t-1} q_t}) \quad (4)$$

$\mathrm{P}(P_{present}|\lambda_{AB})$ is calculated using (2). After calculating the probability of observing $P_{present}$ all about HMM, $HMM_{optimal}$ which has the highest probability is selected. Destination of $P_{present}$ is a destination of the selected $HMM_{optimal}$.

D. Calculating visit probabilities

Based on the destination which is determined by $HMM_{optimal}$, the probabilities of visiting desination and intermediate locations are calculated. First, we find out a path sequence $P_{optimal}$, which is the same with a departure and a destination of $HMM_{optimal}$ and includes the current path $P_{present}$. By the assumption of section C, the location movements of $P_{optimal}$ from time $t+1$ decide the future location movements of $P_{present}$.

Determining a sequence of future movements of the locations allows to find out optimal state sequence $\hat{Q}$ from $HMM_{optimal}$ about path $P_{optimal}$ and calculate the probabilities of visiting locations based on $\hat{Q}$. The method of calculating the optimal state sequence $\hat{Q}$ is as (5).

$$\hat{Q} = max_{Q=q_1 q_2 \cdots q_T} P(Q|P_{optimal}, \lambda_{optimal}) \quad (5)$$

This can be calculated using the Viterbi algorithm [16]. When the observation sequence O is given in the HMM, Viterbi algorithm searches for a state sequence Q as shown the best. That is, it is possible to find the state sequence $\hat{Q}$ that maximizes the probability of discovery of $P_{optimal}$ from the $HMM_{optimal}$.

If it finds the state sequence $\hat{Q}$ based on the Viterbi algorithm, when $HMM_{optimal}$ is given, the probability that path $P_{optimal}$=$\{L_1, L_2, \cdots, L_T\}$ and sequence of states $\hat{Q}$ are found together (joint-probability) is the same as (6).

$$\mathrm{P}(P_{optimal}, \hat{Q}|\lambda_{optimal})$$

$$= \prod_{t=1}^{T} P(\hat{Q}_t|\hat{Q}_{t-1}, \lambda_{optimal})P(L_t|\hat{Q}_t, \lambda_{optimal}) \quad (6)$$

Based on (6), when the user has actually moved to the location of the ith path $P_{optimal}$ whose length is T, the probability of the location L_j of the jth $(j > i)$ may be calculated by the following equation (7).

$$\mathrm{P}(L_j) = \prod_{k=i}^{j-1} P(\hat{Q}_{k+1}|\hat{Q}_k)P(L_{k+1}|\hat{Q}_{k+1}) \quad (7)$$

It is possible to calculate the probabilities of the intermediate locations and a destination previously visited using (7).

 ISBN: 978-1-61208-366-7

IV. EXPERIMENTS

The experiments were performed with the sensor data that 10 university students in their 20s collected during the five months with the smartphone (Samsung Electronics, smartphone SHV-E300K). The description of each user's data is shown in Table III.

TABLE III. DESCRIPTION OF DATA

	#Location	#Path	Size of storage
User 1	16	193	2.44GB
User 2	20	268	2.62GB
User 3	32	149	1.46GB
User 4	50	288	3.41GB
User 5	42	309	3.66GB
User 6	32	233	1.34GB
User 7	28	236	1.48GB
User 8	24	294	3.21GB
User 9	36	237	2.37GB
User 10	14	189	2.08GB

TABLE IV. ACCURACY ACCORDING TO ADVANCEMENT OF THE PATH

	0% (Departure)	20%	40%	60%	80%	100%
User 1	48.05	62.34	75.32	87.01	88.31	90.91
User 2	51.11	60.00	73.33	91.11	93.33	97.78
User 3	50.00	58.06	74.19	87.1	93.55	95.16
User 4	38.46	58.46	67.69	78.46	87.69	89.23
User 5	84.62	87.18	89.74	94.87	94.97	94.87
User 6	74.21	77.24	81.75	85.32	87.86	93.38
User 7	68.90	74.71	78.89	82.36	86.48	88.11
User 8	62.20	67.89	74.71	79.54	85.61	90.71
User 9	53.90	56.74	59.24	65.96	70.71	75.67
User 10	48.12	52.87	68.12	78.22	83.47	93.11
Average	57.96	65.55	74.30	83.00	87.20	90.89

In order to evaluate the accuracy of the proposed detination prediction method, we measured the accuracy according to the progress of path. Prediction result of the advancement of the path is illustrated in Table IV. Looking at the prediction accuracy in accordance with the progress of the path, as the path is largely moves, it can be seen that the prediction accuracy becomes higher because the information of the location movement is increased. 0% progression of the path, that is, is capable of predicting only location information from the starting location. HMM showed accuracy of 57.96% on average only with the information of departure.

Fig. 5 shows the accuracies for the data of 10, indicating the average of the predicted test results with and without the use of context information and progress of the path. When using all context information, the accuracy is the the highest. When using only the spatial information, the accuracy is the lowest. The difference in the case of not using the transportation information and time information is not large, and by using the information of the day, it can be seen that the accuracy is significantly increased.

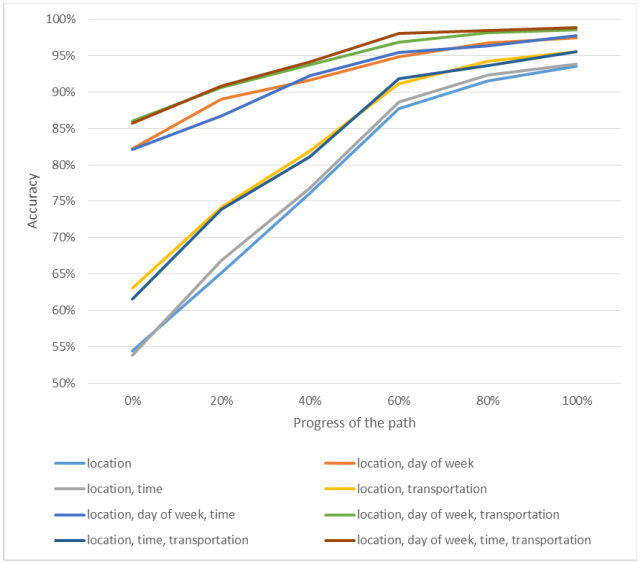


Figure 5. Accuracy according to context contribution and advancement of the path

To demonstrate the usefulness of HMM for the prediction of the destination, we compared it with dynamic time warping (DTW) method [8] which is a template pattern matching method. Fig. 6 shows the average accuracy of prediction based on the data of the 10 users. When the path of the progress is 0% in DTW, because of the shorter length of the input, the prediction is impossible. In the case of HMM, $P_{present}$ matches up the part of the path of the past. However, in DTW, because it matches the full path of the past, it is shown in very low prediction probabilities.

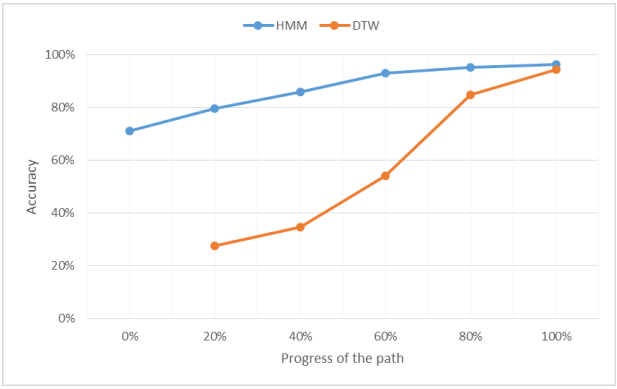


Figure 6. Comparison of accuracy of HMM and DTW

 ISBN: 978-1-61208-366-7

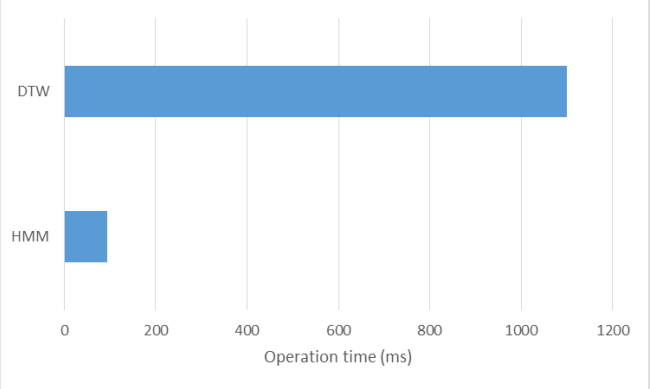


Figure 7. Comparison of processing time of HMM and DTW

Fig. 7 illustrates the prediction time of DTW and HMM. It shows the overall average of time that has been used for predicting each input path. It is possible to see the HMM 12.67 times faster than DTW. In the DTW, the amount of calculation per one template is bigger than HMM because of matching the sequence of the entire path in many cases. Also, it consumes a lot of time, since input data are matched to all the patterns which are the same as departure and destination. However, in the HMM, because it calculates the similarity only until current moving time and it makes one model for all the paths which have the same departure and destination, its running time is shorter.

V. CONCLUSIONS

In this paper, we have proposed a destination and intermediate location prediction method using user's smartphone sensor data. A path is the changes in the location due to human judgment. Based on this, we represent the path model using the HMM where the user moves, and predict a destination. The pre-processing for destination prediction includes extracting key locations, and classifying transportation mode using smartphone sensor data. After making a HMM of paths using pre-processing data, HMM is to learn the parameters. When new input comes, this method finds out the optimal HMM and decides a destination. Also, it calculates the probabilities of visiting destination and intermediate locations. When evaluated with the data of 10 users' destinations, by using not only the spatial information, but a variety of context information improves the accuracy significantly. Also, when compared to the other methods, this method yielded higher accuracy and showed fast running time.

ACKNOWLEDGENMENTS

This work was supported by Samsung Electronics, Inc.

REFERENCES

[1] Y. J. Kim, and S.-B. Cho, "A HMM-based location prediction framework with location recognizer combining k-nearest neighbor and multiple decision trees," 8th Int. Conf. on Hybrid Artificial Intelligent Systems, pp. 618-628, 2013.

[2] I. Burbey and T. L. Martin, "Predicting future locations using prediction by partial match," First ACM Int. Workshop on Mobile Entity Localization and Tracking in GPS-less Environments, pp. 1-6 , 2008.

[3] Z. Xing, J. Pei, and E. Keogh, "A brief survey on sequence classification," ACM SIGKDD Explorations Newsletter, vol. 12, no. 1, pp. 40-48, 2010 .

[4]] T. M. T. Do, and D. Gatica-Perez, "Where and what: Using smartphones to predict next locations and applications in daily life," Pervasive and Mobile Computing, pp. 79-91, 2013.

[5] Z. Lu, Y. Zhu, V.W. Zheng, and Q. Yang, "Next place prediction by learning with multiple models," Mobile Data Challenge Workshop, 2012.

[6] A. Monreale, F. Pinelli, and R. Trasarti, "WhereNext: a location predictor on trajectory pattern mining," 15th Int. Conf. on Knowledge Discovery and Data Mining, pp. 637-646, 2009.

[7] S. Akoush, and A. Sameh, "Mobile user movement prediction using Bayesian learning for neural networks," 2nd Int. Conf. on Systems and Networks Communications, pp. 191-196, 2007.

[8] B. Kim, J. Y. Ha, S. Lee, S. Kang, and Y. Lee, "AdNext: A Visit-Pattern-Aware mobile advertising system for urban commercial complexes," 12th Workshop on Mobile Computing Systems and Applications, pp. 7-12, 2011.

[9] S.-H. Lee, and B.-K. Kim, "A path prediction method using previous moving path and context data," Int. Symposium on Advanced Intelligent Systems, pp. 199-202, 2009.

[10] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden Markov models," ACM Conf. on Ubiquitous Computing, pp. 911-918, 2012.

[11] S. Gambs, M. O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility Markov chains," First Workshop on Measurement, Privacy, and Mobility, p. 3, 2012.

[12] A. J. Dou, V. Kalogeraki, D. Gunopulos, T. Mielikinen, V. Tuulos, S. Foley, and C. Yu, "Data clustering on a network of mobile smartphones," IEEE/IPSJ Symposium on Applications and the Internet , pp. 118-127, 2011.

[13] G. Hamerly, and C. Elkan, "Learning the k in k means," Advances in Neural Information Processing Systems, vol. 16, pp. 281, 2004.

[14] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces," Int. Conf. on Computer Communications, vol. 6, pp. 1-13, 2006.

[15] L. E. Baum, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," Ann. Math. Statist, vol. 41, pp.164 – 171, 1970.

[16] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proc. IEEE, vol. 77, pp. 257–286, 1989.

 ISBN: 978-1-61208-366-7

TEDS: A Trusted Entropy and Dempster Shafer Mechanism for Routing in Wireless Mesh Networks

Heng Chuan Tan
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
htan005@e.ntu.edu.sg

Maode Ma
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
emdma@ntu.edu.sg

Houda Labiod
INFRES
Telecom ParisTech
Paris, France
Labiod@telecom-paristech.fr

Peter Han Joo Chong
School of Electrical and Electronic Engineering
Nanyang Technological University
Singapore
ehjchong@ntu.edu.sg

***Abstract*— Wireless Mesh Networks (WMNs) have emerged as a key technology for the next generation of wireless networking due to its self-forming, self-organizing and self-healing properties. However, due to the multi-hop nature of communications in WMN, we cannot assume that all nodes in the network are cooperative. Nodes may drop all of the data packets they received to mount a Denial of Service (DoS) attack. In this paper, we proposed a lightweight trust detection mechanism called Trusted Entropy and Dempster Shafer (TEDS) to mitigate the effects of blackhole attacks. This novel idea combines entropy function and Dempster Shafer belief theory to derive a trust rating for a node. If the trust rating of a node is less than a threshold, it will be blacklisted and isolated from the network. In this way, the network can be assured of a secure end to end path free of malicious nodes for data forwarding. Our proposed idea has been extensively tested in simulation using network simulator NS-3 and simulation results show that we are able to improve the packet delivery ratio with slight increase in normalized routing overhead.**

Keywords- wireless mesh networks; information fusion;trust system; blackhole attacks.

I. INTRODUCTION

Wireless Mesh Network (WMNs) are fast gaining popularity as the next generation of wireless networking due to their low setup cost, ease of implementation, good network coverage and self-management capabilities [1]. A WMN is made up of two types of nodes: the mesh routers and the mesh clients. The mesh routers are statically deployed and they form the wireless mesh backbone to provide network access for the mesh clients such as your laptops, smart phones or tablets, etc. The mesh clients on the other hand, can be static or mobile with simpler hardware and software requirements. Together, the mesh routers and the mesh clients cooperate to carry out packet forwarding via multi-hop communications to ensure proper data delivery.

However, due to the openness of the wireless medium and the multi-hop nature of communications in WMNs [2], we cannot assume all the nodes in the network are cooperative and well-behaved. Nodes may act selfishly by not forwarding the data packets in order to conserve their scarce resources, such as power and bandwidth. Second, the use of cryptographic techniques, although can deny unauthorized users access to the network, it may not be a viable solution as the mesh clients are resource limited. Also, if the nodes are compromised, they can retrieve the public and private keys used for communications and break the cryptography systems. Subsequently, they may conduct internal attacks by dropping packets to mount a Denial of Service (DoS) attack. If the compromised nodes drop 100% of the data packets, it is called a blackhole attack.

Several works have been proposed in literature [3]-[9] to deal with packet droppers or blackhole attacks. Zhang et al. [9] use a reputation driven mechanism called EigenTrust [10] to evaluate the trust of a node and integrates the trust information into an anomaly detection system to identify packet droppers in the network. This method, however, assumes the presence of prior trustworthy nodes which is not practical in WMNs because pre-trusted nodes may misbehave to protect their own interests. Proto et al. [7] use EigenTrust to compute the trust of a node via a path-wide approach. The trust values are transformed into a weighting metric in Optimized Link State Routing (OLSR) protocol [19] to determine the best path trustworthiness. One issue with this approach is that well-behaved nodes are treated unfairly. One misbehaved node in the forwarding path will result in the decrease of the reputation values of all other nodes along the path.

Marti et al. [6] proposed two mechanisms to detect misbehavior in Mobile Ad hoc NETworks (MANETs): Watchdog and pathrater. Watchdog uses overhearing

 ISBN: 978-1-61208-366-7

technique to identify misbehaved neighboring nodes whereas pathrater is to keep state about the goodness of other nodes in order to decide the most reliable routes among the nodes. This method however, suffers from badmouthing attack as an attacker can malign a good node and cause other nodes to avoid the good node. Shila et al. [8] further extend the Watchdog mechanism by enabling both upstream and downstream traffic monitoring to enhance the detection capabilities in the presence of wireless losses. The disadvantage is that misbehaved nodes can only be detected by the source node based on the receipt of the PROBE ACK message. If PROBE ACK is not received by the source, then the source needs to initiate a hop by hop query for the PROBE and PROBE ACK packets which is going to increase the computational load.

In [3]-[5], the authors used subjective logic developed by Josang [11] to qualify trust where trust is represented by an opinion having belief, disbelief and uncertainty. The motivation behind this idea is that no one can determine with absolute certainty that a proposition is true or false. Hence, we can only form subjective opinion which contains certain degree of uncertainty regarding the truth of the proposition. It allows for better expressiveness and clarity than traditional probabilistic logic thereby allowing users to specify situations like "I don't know" or "I'm not sure". While subjective logic is effective in providing accurate opinions as it takes into account uncertainty, aggregation of trust opinions is complex and it requires high memory storage as each node needs to store the belief, disbelief and uncertainty parameters.

Motivated by the limitations of existing approaches, we propose a trust based mechanism called Trusted Entropy and Dempster Shafer (TEDS), through incorporating with Ad hoc On demand Distance Vector (AODV) routing protocol [19] to find a secure end to end path free of non-trusted nodes to safeguard against blackhole nodes.

The rest of the paper is organized as follows. Section II describes the threat model and assumptions. Section III presents the proposed TEDS design. Section IV presents the simulation results to demonstrate the performance of TEDS. Section V gives the conclusion and future works.

II. THREAT MODEL AND ASSUMPTIONS

WMNs are exposed to security threats at any layer of the internet protocol stack which can cause the network to degrade or malfunction [13][14]. In this paper, we only focus on blackhole attacks at the network layer and state our assumptions for our TEDS design.

A. Blackhole Attacks

In a blackhole attack, the malicious node will advertise itself as having the best route to the destination even though it does not have a route to it. It does this by sending a route reply (RREP) packet immediately to the source node [14]. The source node upon receiving this malicious RREP assumes the route discovery is complete and ignores all other RREP from other nodes and selects the path that now includes the malicious node as a relayer to forward the data packets. Subsequently, the malicious node will drop all the traffic received. The malicious node thus creates a blackhole in the network. An illustration is given in Fig. 1.

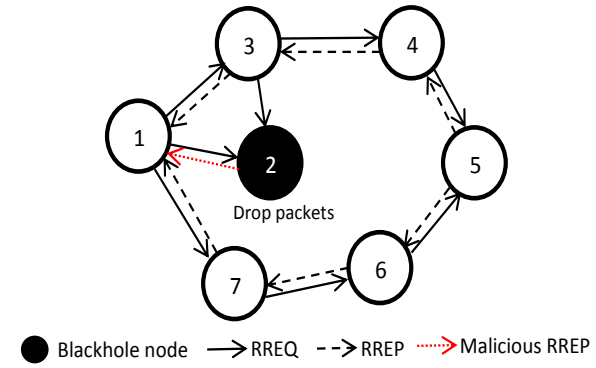


Figure 1. Blackhole Attack

B. Assumptions

We assume the mesh routers and mesh clients in the network are statically deployed to model communications in infrastructure based WMNs and the majority of the mesh routers are cooperative and well-behaved. We assume that the sources and destinations are fully trusted. We further assume that the network is strongly connected where there are many other alternate paths from the source to the destination free of malicious nodes. Next, we assume the malicious nodes act independently and they do not collude. Therefore, our model is free of attacks, such as cooperative blackholes or wormhole attacks or any other forms of collusion attacks. The problem of colluding nodes is left as a future work. Lastly, trust value in our model is defined in the range $[0,1]$, where trust value of 0 means the node is untrusted and trust value of 1 means the nodes is fully trusted. Lastly, all nodes at the start of the network assume an initial trust value of 0.5.

III. TEDS DESIGN

In this section, we present the details of TEDS. Lastly, we present our simulation settings with results and discussions.

A. TEDS Overview

Our proposed model makes use of trust metric to assess the trustworthiness of a node in the network and it is based on three mechanisms. The three mechanisms are Watchdog [6], Shannon entropy function [12] and Dempster Shafer Theory [15]. In our scheme, we assume each node starts off with a trust value of 0.5 and each node is installed with Watchdog functionality to monitor the next hop forwarding behavior. The observations gathered from overhearing will be used to compute the forwarding probability of each neighbor node in the network. We then applied Shannon entropy function to compute the uncertainty of this

 ISBN: 978-1-61208-366-7

forwarding probability thereafter, compute the direct trust value of each neighbor node. This direct trust value represents a node's direct experiences with its one hop neighbor. Next, we apply Dempster Shafer theory to combine conflicting trust values coming from a node's direct interactions as well as indirect interactions to determine the overall trust value. The indirect interactions in our context are defined as indirect trust derived from recommendation trust values from other nodes. This overall trust value will be calculated periodically *(every trust interval)* which will be feedback to the routing protocol for routing decisions.

B. Calculation of Forwarding Probability

Each node is installed with Watchdog functionality to monitor the number of packets sent and the number of the packets overheard. To prevent the trustor *(node responsible for the evaluation of trust of its downstream node)* from falsely accusing its neighbor of misbehaving due to the inevitable collisions at the sender [6], we proposed that the trustor continues to monitor its downstream node for an extended period of time which we set to 2 seconds. If the trustor still fails to overhear the packet sent out by its neighbor after 2 seconds, the trustor concludes that the downstream node has dropped the packets maliciously. An example of how the watchdog works is given in Fig. 2. Each node keeps track of all sent packets by maintaining a table containing the ID of the node that the packet is directed to, the packet ID and the expiration time which is 2 seconds. When the trustor overhears a packet and finds a match in its corresponding table, the table entry for the overheard packet ID will be deleted. At every trust interval, we compute the forwarding probability, f_p using (1).

$$f_p = \frac{\text{\# of overheard packets sent by n}_i}{\text{\# of packets send to n}_i \text{ for forwarding}} \quad (1)$$

C. Calculation of Direct Trust

The next step is to formulate the direct trust value and we proposed using the Shannon binary entropy function defined in (2) together with a set of mapping equations defined in (3) to compute the direct trust values. This will ensure that the trust value is bounded by and confined in the range [0,1] and such that low forwarding probabilities correspond to lower trust values while high forwarding probabilities corresponds to high trust values.

$$H_b(p) = -p\log_2 p - (1-p)\log_2(1-p) \quad (2)$$

$$DT = \begin{cases} 1-0.5H(p), & \text{for } 0.5 \le p \le 1 \\ 0.5H(p), & \text{for } 0 \le p < 0.5 \end{cases} \quad (3)$$

where DT denotes the trust value of a node, $H_b(p)$ denotes the binary entropy function in (2) and p denotes the forwarding probability derived in Section B.

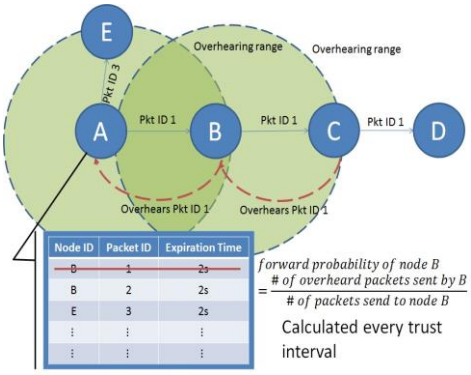


Figure 2. Watchdog Operation

D. Update of Direct Trust

To better reflect the currency of direct trust values, we use Exponential Moving Weighted Average (EMWA) [16] to combine the past historical trust values of a node with the current measured trust values. EMWA applies exponentially decreasing weighting factors to each data point so as to smooth out the direct trust value and hence it provides a better representation of trust values over a period of time. EMWA for our trust model is given by (4).

$$DT_t = \alpha \cdot DT_t + (1-\alpha) \cdot DT_{t-1} \quad (4)$$

where α is a constant smoothing factor between 0 and 1, DT_t represents the current trust value to be evaluated and DT_{t-1} represents the previous trust value recorded by TEDS. If the smoothing factor α is large, it discounts the older observations faster. For our scheme, the smoothing factor α is selected to be 0.667. This means that a higher weightage is placed on a node's current trust value compared to the past. The choice of α is chosen to match the average age of data in simple moving average (SMA) and the formula is given by (5). The proof can be found in [17].

$$\alpha = \frac{2}{N+1} \quad (5)$$

where N is the number of samples or records considered in EMWA.

E. Calculation of Indirect Trust

Beside direct trust values, a node could also form an indirect trust value based on the recommendation trusts from other neighboring nodes, as shown in Fig. 3. Let us assume node A's trust of node B denoted by DT_{AB} is 0.667 and node B's trust of C denoted by DT_{BC} is 0.3. To determine the indirect trust value of node A about C taking into consideration node B's recommendation about node C,

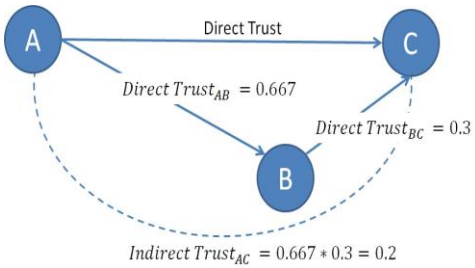


Figure 3. Formulation of Indirect Trust

node B's trust of node C has to be weighted by node A's trust of node B according to (7). This is similar to transitive trust described in Eigentrust [10] and in subjective logic [11]. The purpose is to discount node B's direct trust of node C based on node A's observation of node B so as to mitigate the effects of badmouthing. Using (7), the weighted trust value *(indirect trust value denoted as IDT)* for the example given in Fig. 3 will be equal to 0.2.

$$IDT_{AC} = DT_{AB}.DT_{BC} \quad (7)$$

F. Combining Direct Trust and Indirect Trust

Dempster Shafer theory (DST) [15] is used to combine a trustor's direct trust of a node and indirect trust values received from other recommending nodes to arrive at the overall trust value of a node. DST is used because it allows us to quantify uncertainty in our trust computation instead of being forced to use prior probabilities to add up to 1 based on traditional probability logic. For instance, if the trust value of a node is 0.6, its complement probability which is distrust will be 0.4 according to traditional probability theory. Sometimes, it is unrealistic to make that claim because the lack of knowledge about an event is not regarded as evidence supporting the distrust of a node. Instead, DST classifies 0.4 as uncertainty which means a node can either be trusted or untrusted. Hence, DST can better reflect the behavior of the node and can improve on the trust evaluation.

First, nodes are classified into two states: Trusted (T) and Untrusted (UT). So the frame of discernment Θ in DST consists of $\{T, UT\}$. The power set denoted by 2^{Θ} contains these four sets:

$$2^{\Theta} = \left[\{T\}, \{UT\}, \{T, UT\}, \{\phi\}\right] \quad (6)$$

The set represented by $\{T, UT\}$ denotes uncertainty in DST, which means that a node can be trusted or untrusted. We apply the direct trust values we obtained from (4) and indirect trust values obtained from (7) as Basic Probability Assignment (BPA) to denote the strength of evidence pertaining to a particular subset of 2^{Θ}. In our scheme, trust value of 0.5 and above will be classified as trusted whereas trust value less than 0.5 will be classified as untrusted. For example, if the trust value of a node is 0.6, BPA for the set $\{T\}$ will be 0.6. The remaining belief mass of 0.4 will be allocated to the set $\{T, UT\}$. Following this rule and using Dempster's rule of combination in (8), direct trust and indirect trust can be combined to compute the overall trust value of a node.

$$m_{1,2}(S) = \frac{1}{1-K} \sum_{A \cap B = S \neq 0} m_1(A) m_2(B) \quad (8)$$

where $K = \sum_{A \cap B = 0} m_1(A) m_2(B)$ is the normalization factor to ensure the total sum of combined masses, $m_{1,2} = 1$; $m_1(A)$ and $m_2(B)$ each represent the BPA assigned to direct trust and indirect trust, respectively.

G. Decision Making

The overall trust value of a node is feedback to the routing protocol for routing decisions. If the overall trust value is ≥ 0.5, we conclude that node is trusted, else the node is misbehaving. If a node is detected as misbehaved, the trustor will add the misbehaved node to the blacklist. At the same time, it will broadcast a message throughout the network to inform other nodes of the blacklist nodes. Other nodes upon receiving the broadcast message will also put the misbehaved node in blacklist to avoid using it for all future communications. The trustor will also send a route error (RERR) message to notify the source node where another route discovery will be initiated to find a path free of malicious nodes.

IV. SIMULATION

All our simulations are performed using Network Simulator NS3 (v3.17) [18]. TEDS is integrated into AODV [20] of NS-3 [18]. Although AODV is used, our proposed trust system is independent of the underlying routing protocols. It is compatible with other routing protocols as it runs on top of any routing protocol. TEDS only triggers the underlying routing protocol to re-initiate a new route discovery upon detection of malicious nodes. To evaluate the performance of TEDS, we compared it with basic AODV that is without any trust mechanisms.

A. Simulation Environment

Our simulation environment consists of 100 nodes placed in a square grid manner. The distance between each adjacent node is 150m and the radio range of each node is 250m. The source and destination nodes are located on the left and right side of the square grid. All nodes in our simulation environment are assumed static to model the WMN backbone infrastructure. We simulate a total of 10 CBR flows between randomly selected source nodes on the left and randomly selected destination nodes on the right.

 ISBN: 978-1-61208-366-7

The maximum packet per flow is configured as 300 with a packet generation rate of 4 packets/s. The start time of each flow is uniformly distributed between 30 seconds and 200 seconds. A random number generator is used to randomly select the source and destination nodes as well as to locate the malicious nodes in the forwarding path between each source and destination pair. Simulations were performed for a period of 300s and each data point is an average of 10 runs unless otherwise stated. More simulation parameters are given in Table I.

TABLE I. SIMULATION PARAMETERS

Simulation tool	NS-3
Grid spacing	150m
Data rate	16kbps
Transmission range	250 m
Network area	1500 m x 1500 m
No: of nodes in the network	100
Traffic	10 source-destination pairs
Packet size	512 B
Packet generation rate	4 packets/s
Simulation time	300 s
Traffic type	CBR
Transport protocol	UDP
Routing protocol	AODV (disable HELLO)
Mac protocol	IEEE 802.11b
Propagation loss model	RangePropagationLossModel
Physical layer	YansWifiPhy channel
Mobility	Static

The performance of TEDS is evaluated for the following cases. First, we study the performance when the network is under blackhole attack. Next, we examine the sensitivity of the trust interval on the performance of TEDS in terms of packet delivery ratio and routing overhead. The selection of trust interval determines how often TEDS computes and propagates the trust value in the network. For all these experiments, we made two assumptions. We assume that all source and destinations are trustworthy and the second assumption is that all nodes start off with a default trust rating of 0.5 at the initial state.

B. Results and Discussions

1) Performance Analysis under Blackhole Attacks

We first evaluate the performance of TEDS under the effects of blackhole attacks and we compare the result with basic AODV without any trust mechanisms. We are interested in the Packet Delivery Ratio (PDR) and normalized routing overhead of TEDS compared to basic AODV. To simulate the blackhole attacks in NS-3, we configure the malicious nodes to send a RREP packet with a high sequence number such that it will be selected by the source node during route discovery. Subsequently, the malicious nodes will drop 100% of the data packets. The comparison was done by varying the number of blackhole nodes in our network and the blackhole nodes are randomly selected from the network area of 80 nodes discounting the

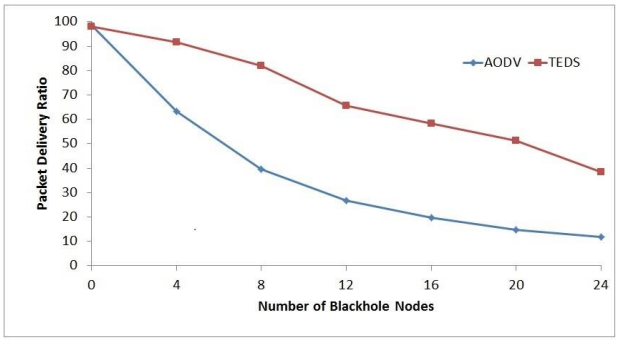


Figure 4. PDR performance in the presence of blackhole nodes

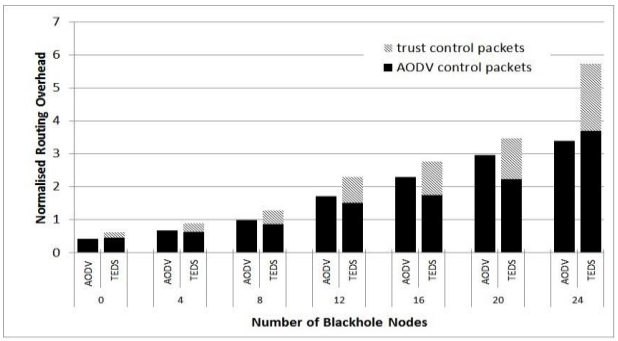


Figure 5. Normalized routing overhead in the presence of blackhole nodes

source and destinations which are on opposite sides of the square grid.

Fig. 4 shows the average PDR of TEDS compared with basic AODV. As shown in Fig. 4, the PDR decreases when the number of malicious nodes increases. When there are no malicious nodes in the network, we are able to achieve similar PDR performance for TEDS and AODV. As the number of malicious nodes increases, the PDR of TEDS decreases at a slower rate but still able to achieve about 30% improvement on the PDR compared to AODV. This shows that TEDS is capable of detecting and isolating malicious nodes using the trust mechanisms introduced. As more and more blackhole nodes are introduced into the network, TEDS's performance also starts to decrease gradually; this is because as more blackhole nodes are being identified and isolated by TEDS, there remain fewer alternatives available for choosing the forwarding paths considering our network is a static environment.

Fig. 5 compares the normalized routing overhead for TEDS and basic AODV when the network is under blackhole attacks. Based on Fig. 5, two observations can be made. First, the normalized routing overhead incurred by TEDS is higher than AODV. Second, the normalized routing overhead increases with increasing blackhole nodes in the network. The increase in normalized routing overhead is due to the following reasons: (1) periodic exchange of trust information with neighboring nodes, (2) broadcast of trust control message to notify other nodes of blackhole nodes, so that they can isolate them and not use them for packet forwarding and (3) re-initiation of a new route

 ISBN: 978-1-61208-366-7

discovery upon detection of blackhole nodes. We further show the breakdown of the normalized routing overhead for TEDS into control packets introduced by our trust mechanism and control packets introduced by AODV. We conclude that the increase in normalized routing overhead for TEDS is mainly due to the trust related control packets which also increases proportionally with the increase of blackhole nodes in the network. This increase in overhead is in the acceptable range in exchange for higher security and higher PDR.

2) Selection of Trust Interval

In this experiment, we assess the sensitivity of the trust interval on the performance of TEDS. The trust interval determines how frequent we perform the trust computations and propagation of trust values in the network. We compare the PDR vs. trust interval with 10% blackhole nodes in the network. The trust interval is varied from 10s to 60s and each data point on the plot represents an average of 20 runs.

Fig. 6 shows the PDR for TEDS by varying the trust interval period. The curve shows that the PDR is the highest when trust interval is 10s. Beyond 10s, the PDR starts to drop. The PDR is around 85% at trust interval of 10s compared to 60% when the trust interval is configured as 60s. One reason is that, when trust query period is small, a malicious node will be detected earlier and hence less number of packets will be lost. As trust interval increases, more packets are lost due to the dropping behavior of the malicious nodes and that attributed to the drop in PDR as trust interval increases.

Next, we examine the effects of trust interval on the routing overhead. The routing overhead in our case consists

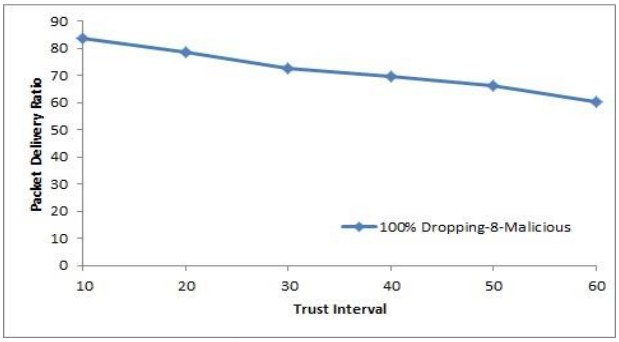


Figure 6. PDR performance under varying trust interval

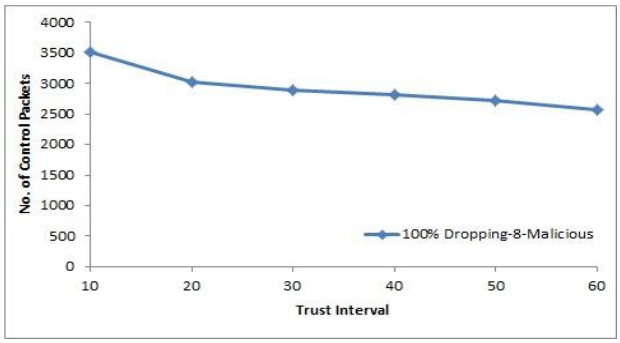


Figure 7. Routing overhead under varying trust interval

of control packets generated by TEDS (trust-related control packets) and AODV related control packets for route discovery and route maintenance (RREQ, RREP, RERR). From Fig. 7, we see that the number of routing control packets is the highest when the trust interval is at 10s and that it tapers off as the trust interval increases. The reason is that at smaller trust interval, nodes need to exchange and disseminate trust related control packets more frequently which resulted in the increase in routing overhead. Based on Fig. 6 and Fig. 7, we can have the following observation. There is a trade-off between PDR and routing overhead vs. the trust interval. High PDR is achieved when the trust interval is small but at smaller trust interval, the routing overhead is high which makes TEDS less efficient and more costly in resources. Here, we conclude that the optimum trust interval based on simulation is 20s.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel lightweight trust mechanism called TEDS to detect blackhole nodes in the forwarding path of a node. TEDS exploits the broadcast nature of wireless medium to listen promiscuously to the next node's transmission to detect packet droppers. It uses Shannon entropy to derive the direct trust value of a node and we demonstrated how to combine direct trust and indirect trust to form a shared belief using Dempster's rule of combination. Benefits of TEDS are that it is lightweight; uncertainty is quantified in the trust computations and it is portable. TEDS can be integrated into any routing protocols. Our simulation results show that TEDS can detect packet droppers and improved the packet delivery ratio of the network as the number of malicious nodes increases with minimum increase in normalized routing overhead. We further show that the optimum trust interval is 20s through simulation where TEDS is able to ensure high PDR with reasonable routing overhead. For future work, we plan to study collusion attacks such as cooperative blackhole or wormhole initiated by multiple attackers and to study the impact of node mobility on the performance of TEDS.

REFERENCE

[1] I. F. Akyildiz and W. Xudong, "A survey on wireless mesh networks," Communications Magazine, IEEE, vol. 43, 2005, pp. S23-S30.

[2] N. Ben Salem and J. P. Hubaux, "Securing wireless mesh networks," Wireless Communications, IEEE, vol. 13, 2006, pp. 50-55.

[3] K. Kane and J. C. Browne, "Using uncertainty in reputation methods to enforce cooperation in ad-hoc networks," presented at the Proceedings of the 5th ACM workshop on Wireless security, Los Angeles, California, 2006, pp. 105-113.

[4] H. Lin, J. Ma, J. Hu, and K. Yang, "PA-SHWMP: a privacy-aware secure hybrid wireless mesh protocol for IEEE 802.11s wireless mesh networks," EURASIP Journal on wireless communications and networking, vol. 2012 , 2012, pp.1-16.

[5] Y. N. Liu, K. Q. Li, Y. W. Jin, Y. Zhang, and W. Y. Qu, "A novel reputation computation model based on subjective logic for mobile ad hoc networks," Future Generation Computer Systems-the

 ISBN: 978-1-61208-366-7

International Journal of Grid Computing and Escience, vol. 27, May 2011, pp. 547-554.

[6] S. Marti, T. J. Giuli, K. Lai, and M. Baker, "Mitigating routing misbehavior in mobile ad hoc networks," presented at the Proceedings of the 6th annual international conference on Mobile computing and networking, Boston, Massachusetts, USA, 2000, pp.255-265.

[7] F. S. Proto, A. Detti, C. Pisa, and G. Bianchi, "A Framework for Packet-Droppers Mitigation in OLSR Wireless Community Networks," in Communications (ICC), 2011 IEEE International Conference on, 2011, pp. 1-6.

[8] D. M. Shila, C. Yu, and T. Anjali, "Mitigating selective forwarding attacks with a channel-aware approach in WMNS," Wireless Communications, IEEE Transactions on, vol. 9, 2010, pp. 1661-1675.

[9] Z. Zhang, P.-H. Ho, and F. Naït-Abdesselam, "RADAR: A reputation-driven anomaly detection system for wireless mesh networks," Wireless Networks, vol. 16, 2010/11/01 2010, pp. 2221-2236.

[10] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina, "The Eigentrust algorithm for reputation management in P2P networks," presented at the Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary, 2003, pp. 640-651.

[11] A. Jøsang, "Artificial reasoning with subjective logic," in Proceedings of the Second Australian Workshop on Commonsense Reasoning, 1997.

[12] Y. L. Sun, Y. Wei, H. Zhu, and K. J. R. Liu, "Information theoretic framework of trust modeling and evaluation for ad hoc networks," Selected Areas in Communications, IEEE Journal on, vol. 24, 2006, pp. 305-317.

[13] S. Glass, M. Portmann, and V. Muthukkumarasamy, "Securing Wireless Mesh Networks," Internet Computing, IEEE, vol. 12, 2008, pp. 30-36.

[14] J. Sen, "Secure routing in wireless mesh networks," arXiv preprint arXiv:1102.1226, 2011.

[15] G. Shafer, A Mathematical Theory of Evidence: Princeton University Press, 1976.

[16] S. Buchegger and J.-Y. Le Boudec, "A robust reputation system for mobile ad-hoc networks," Technical Report IC/2003/50, EPFL-IC-LCA, 2003.

[17] S. Nahmias, *Production and Operations Analysis*, 6th ed.: McGraw-Hill Education, 2013.

[18] NS-3 network simulator homepage. Available: http://www.nsnam.org, retrieved: June, 2014.

[19] P. Jacquet, P. Muhlethaler, T. Clausen, A. Laouiti, A. Qayyum, and L. Viennot, "Optimized link state routing protocol for ad hoc networks," in Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International, 2001, pp. 62-68.

[20] C. Perkins, E. Royer, and S. Das, "RFC 3561 Ad hoc On-Demand Distance Vector (AODV) Routing," 2003.

 ISBN: 978-1-61208-366-7

A Non-GPS Low-Power Context-Aware System using Modular Bayesian Networks

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

***Abstract*—The proliferation of smartphones has lead to the development of a large variety of applications and the inverstigation on the use of various sensors through context-awareness, in order to provide better services. However, smartphone battery capacity is extremely limited, so that the applications cannot be effectively used. In this paper, we propose a low-power context-aware system using modular Bayesian networks. Bayesian networks are known to respond flexibly to uncertain situations. However, probabilistic models, such as Bayesian networks, have high time complexity, resulting in high power consumption. To reduce the time complexity, we modularize the network based on the Markov boundary, and eliminate the use of GPS because it consumes a lot of power. We compare the accuracy of the system using a combination of sensors and confirm the decrease in the time complexity. Experiments with the real data collected show that the proposed Bayesian networks yield an accuracy of 92.47%.**

Keywords-Low-power system; context-awareness; modular Bayesian network; Markov boundary.

I. INTRODUCTION

With the widespread production of smartphones along with its diverse array of applications and sensors, the trend of smartphone applications has shifted towards the personal and intelligent service route. Many researchers have investigated the possibility of intelligence techniques for context-awareness. The battery capacity of a smartphone, however, is well behind the development of applications. In a typical case, the user has to carry an extra battery or charge it frequently. There is the critical issue of how to reduce battery consumption for the context-awareness on the smartphone [1].

In this paper, we propose a low-power context-aware system using Bayesian networks. We focus on the configuration of input sensors and the inference time of the system. In the configuration of input sensors, the GPS sensor is a very important tool for situational context-awareness because the state of the user is closely related to the location. However, GPS usually consumes the highest amount of power and induces the need for a large location database in order to convert semantic location. For this reason, we choose to infer the user position not by GPS, but by determining whether the user is indoors or outdoors using the temperature and humidity sensors. In terms of reduction of time complexity, the Bayesian Network (BN) is modularized based on the Markov boundary [2].

We compare the accuracy using different combinations of sensors and evaluate the inference time of the proposed method against a monolithic network. In addition, we verify the low-power consumption feature of the proposed method in a real smartphone environment using the power tutor application.

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

II. RELATED WORKS

A. Context aware services in smartphone

A context can be defined as information that can be used to characterize the situation of an entity, such as the person, place, or device that is considered relevant to the interaction between the user and the application [3]. Context-aware applications recognize the situation and provide services. The services have been studied using various sensors, as shown in Table I.

TABLE I. CONTEXT AWARENESS IN SMARTPHONE

Authors	Sensors	Services
Otebolaku, et al. [4]	Accelerometer, Orientation, Rotation	Mobile media content recommendation
Wang, et al. [5]	Accelerometer, Gyroscope, Brightness, Bluetooth, GPS	Music recommendation for daily activities
Santos, et al. [6]	Accelerometer, Brightness, Temperature, Humidity, Microphone, Time	Social networking application
Phithakkitnukoon, et al. [7]	Accelerometer, Microphone, GPS	Alert mode control
Chon, et al. [8]	GPS, GSM, Wi-Fi, Accelerometer, Thermometer, Digital compass	Location-based service

 ISBN: 978-1-61208-366-7

Otebolaku and Andarade developed a context-aware mobile application [4]. The system used classifiers for recognizing high-level contexts from low-level sensor data. Wang, et al. presented a probabilistic model to integrate contextual information with music content analysis to offer music recommendation for daily activities [5]. Santos et al. described the architecture, operation and potential applications of a prototype system developed within the User-Programmable Context-Aware Services (UPCASE) project [6]. Phithakkitnukoon and Dantu proposed a three-step approach in designing the model based on the embedded sensor data for controlling alert mode [7]. Chone and Cha presented the smartphone-based context provider [8]. The system used the activity, connectivity, location and environment for inferring the current context. However, these previous works focus on the high accuracy of awareness using as much information as possible. These researches do not consider the power consumption of the context-aware system.

B. Power consumption problem

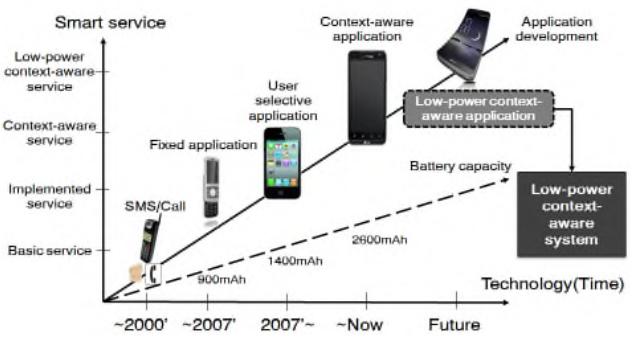


Figure 1. Trend of smartphone service

Various sensors have been implemented in smartphones, and the applications are utilizing these sensors. Although the battery capacity of smartphone has improved, it is insufficient to freely use a variety of applications. Fig. 1 shows the trend of smartphone services and the improvement of battery capacity. The energy consumption of these sensors is various, depending on the kind of smartphone. Abdesslem, et al. measured the energy consumption of different sensors [9]. Each sensor was run continuously on a Nokia N95 8GB smartphone until the battery was depleted. In this research, the power consumption of GPS was 623mW, which is 10 times more than the power consumption of the accelerometer senor. Wang, et al. measured the energy consumption using Nokia N95, as well [10]. In this research, the power consumption of GPS was 0.3308W, which is 7 times more than the power consumption of accelerometer. Although GPS sensor is more important for detecting user location, it has high energy consumption.

Many researchers proposed low-power applications using context-awareness to solve sensor problems. Herrmann et al. proposed to use context knowledge to dynamically adapt the behavior of sensing applications running on smartphones [11]. In the system, a context-aware application manager starts, suspends and changes the sampling rate of the sensors. Bareth and Kupper proposed a hierarchical positioning algorithm [12]. The algorithm dynamically deactivates different positioning technologies and only activates the positioning method with the least energy consumption. Seo et al. proposed a context-aware configuration manager for smartphones [13]. The system changes the configuration settings of a smartphone in response to changes in context, according to user-defined policy rules. These previous pieces of research focused on the management of the sensor, and they were not interested in reducing the power consumption of the inference module.

C. Context-aware service using Bayesian networks

Bayesian networks is a graph-theoretic concept for representing uncertain and incomplete knowledge using Bayesian statistics. BN has a structure of a directed acyclic graph which represents the link relations of the node, and has conditional probability tables (CPT). Assume that nodes are independent of each other.

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

The conditional probability distribution of variable A can be represented as $P(A \mid pa(A))$, where $pa(A)$ denotes the set of parent variables of variable A, where $U = A_1, A_2, \ldots, A_n$ is a set of nodes, and the joint probability distribution is computed by the chain rules as equation (1). For each child node, conditional probabilities are allocated for each combination of states in their parent nodes, so that the size of each CPT depends on the number of parent nodes and the number of their states, as follows:

$$size(CPT) = S\prod_{i=1}^{n} P_i \quad (2)$$

where S is the number of states, and P_i is the number of states in the i th parent node. Therefore, the size of the CPT can increase considerably with the number of parents, which can make the process of calculating the CPT intractable, especially if this is done through expert elicitation [14].

There are two approaches to identify the structure and parameter of a Bayesian network model. The first approach is the learning from the data on problem domains. The learning of structure is useful if we have a lack of understanding about the system. The method requires a sufficient amount of data, but it is not easy to obtain reliable data in many real-world problems.

The second one is to construct the model based on the domain knowledge. The experts identify the structure and set of parameters according to their knowledge, if we do not have enough data in the domain. In the context-aware service,

 ISBN: 978-1-61208-366-7

Bayesian networks have been used for sensor fusion [15]. Lee and Cho proposed two-layered Bayesian networks for inference on a mobile phone [16]. This network was designed with the domain knowledge. The performance of Bayesian network using the domain knowledge can be evaluated through the scenarios or collected data [17]. In the field of context-awareness on the smartphone, the number of the used sensors is proportional to the number of nodes in BN. Therefore, selecting the sensor used for context-awareness is very important to reduce the time complexity.

III. LOW-POWER CONTEX-AWARE SYSTEM

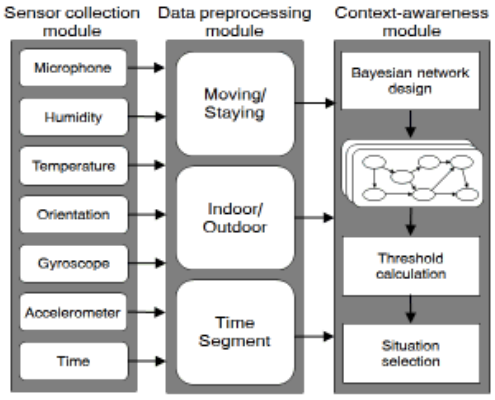


Figure 2. System archtecture

In this paper, we propose a low-power context-aware system. The system considers the power consumption and the sensors combination which are needed for inferring the user situation. Fig. 2 illustrates the system architecture for context-awareness. The system consists of three modules: sensor collection, data preprocessing, and context-awareness. In this proposed architecture, we do not use the GPS sensor because of the energy consumption. The sensor collection module obtains continuous sensor data in the smartphone. The data are sent to the data preprocessing module that discretizes them using a decision tree and a naïve Bayes classifier. The context-awareness module infers the user situation using a Bayesian network that is modularized based on Markov boundary. If the result of inference is higher than the threshold, it is the current situation. However, if all the results of a situation are less than the threshold, the module does not infer the current situation.

A. Data preprocessing

The purpose of data preprocessing is to reduce the time complexity of the Bayesian network modules. The number of states of input node is proporional to the size of CPT, because the parameter S in equation (2) is selected as the number of states. There are two discretization methods for input data. First, the input can be divided into a predefined number of intervals of equal width such as 0~1, 4~8, 8~12, 12~16, and so on. Second, it can be divided using statistical methods. For instance, a range between 22~27℃ represents 'normal', 10~22℃ is 'cold' and 'hot' indicates between 27℃ to 34℃. Table II shows the result of preprocessing. In this paper, we use the two preprocessing methods: Decision tree and naïve Bayes classifier.

TABLE II. RESULT OF PREPROCESSING

Method	Type	Input sensor	Result
Decision tree	Indoor/ Outdoor	Temperature	{Very high, High, Normal, Low, Very low}
		Humidity	{Very high, High, Normal, Low, Very low}
	Noise	Microphone	{Very high, High, Normal, Low, Very low}
	Time	Time	{Morning, Afternoon, Evening, Night}
Naïve Bayes classifier	User state	Accelerometer	{Stay, Moving}
		Gyroscope	
	User position	Accelerometer	{Sitting, Lying, Standing}
		Orientation	

A decision tree is a powerful and popular tool for classification. This method makes rules which can be understood by humans. The inputs such as temperature, humidity, microphone, and so on, make the rules as a range of the division using the decision tree, because the input data do not need to change into the semantic data. It just needs to divide each range.

A Naïve Bayes (NB) classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumption. We infer the user position and state using NB classifier, because some input data need to be converted into semantic data, because if each input data are independent, we do not need Bayesian network. We use the two methods for preprocessing because the methods have low complexity.

B. Tree structure modular Bayesian network

This section presents a modular method of the types of situation. We design the tree structure Bayesian network for reducing the time complexity. We identify the three types of input nodes: The set of situation nodes $S = \{s_1, s_2, ..., s_i\}$, the set of type nodes $T = \{t_1, t_2, ... t_j\}$, and the set of preprocessing result nodes $R = \{r_1, r_2, ..., r_k\}$, where the type node t consists of the set of associated result nodes; $\{r_1, r_2, ..., r_m\} \in t$. The situation node S consists of the associated set of type nodes.

Neil proposed a method on how to separate large-scale Bayesian networks [18]. We define the two types of structure according to this research.

- Definition 1 (Result-Type structure): The network is used to reason about the uncertainty we may have about our own judgments. This structure represents uncertainties. We have result of preprocessing.
- Definition 2 (Type-Situation structure): Inferring the situation needs to reconcile independent source of evidence about a single attribute of a single entity, where these sources of evidence have been produced by

different measurements or prediction methods. In this domain, the type nodes have different attributes and components. Therefore, the network is to reconcile independent response to the type node.

Locality of causal relation in a BN can facilitate decomposition of inference processes. A *d*-separation concept describes how different parts of BN can be rendered conditionally independent [2]. Pavlin et al. analyzed locality of causal relation with the *d*-separation concept [19].

- Definition 3 (Markov boundary of a set of variables): Markov boundary $B(V_i)$ of a set of variables $V_i \subset V$ in BN is the union of parents of set V_i and parent of children of V_i.

$$V(V_i) = Pa(V_i) \cup Ch(V_i) \cup (\bigcup_{\gamma \in Ch(V_i)} Pa(\gamma)) \qquad (3)$$

where $Ch(V_i)$ represents the children of V_i and $Pa(\gamma)$ is parent of γ.

We apply this locality of causal relation to the situation domain which has the response to type nodes as independent. For the independence, each module uses the sensor that is directly related to the situation instead of all sensory information. In the modules, the relations of nodes consist of two types: Type-Situation and Result-Type.

- Definition 4 (Type-Situation structure): The situation nodes represent the probability of the current situation. The type nodes denote the factors for inferring situation. For instance, the location, noise, time, and user is represented the type node. The situations such as viewing, moving, and studying are represented with the situation node.

- Definition 5 (Result-Type structure): The result nodes denote the obtained value of the sensor. For instance, the result node of the user state node denotes the values of user state: staying, and moving.

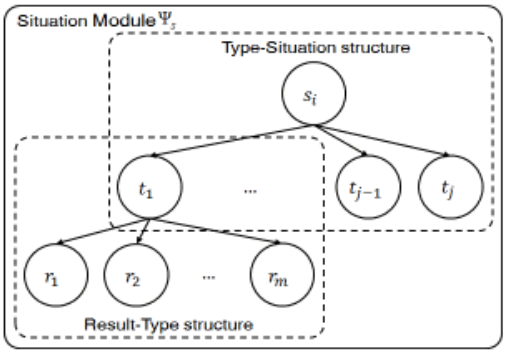


Figure 3. Modular structure for inferring user situation

Fig. 3 represents the situation module. Each situation module consists of one Type-Situation structure and one or more Result-Type structures.

C. *Situation inference process*

Input: Sensor data
Output: User situation

Preprocessing_sensor data

Repeat: all BN modules complete inference
Setting preprocessed sensor data
Executing propagation using (1)
Getting result of inference R
If ($R \geq$Threshold)
Saving R to current situation queue
Else
Next module

Selecting the highest probability of R in current situation queue

Figure 4. Situation inference process

The inference process of the proposed context-aware system consists of multiple steps, as shown in Fig. 4. First, the system preprocesses the sensor data. Next, the system selects a Bayesian network module randomly. Then, the preprocessed sensor data set the values as evidence. The module propagates the probability and gets the result. If the result is larger than the threshold, the result is pushed to a current situation queue. The current situation queue has the candidate of the current situation. If the result is smaller than the threshold, the result is discarded. This step is repeated up to infer all BN modules. Finally, the system selects the highest probability of the result.

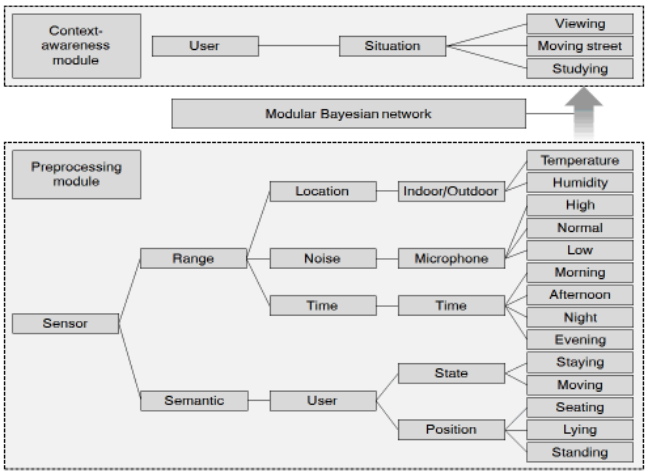


Figure 5. Formation of the context

Statistics Korea surveyed the time Korean people use to perform different activities [20]. In this survey, the students used a lot of time for sleeping, viewing, studying, and moving, in that order. We define the three network modules:

 ISBN: 978-1-61208-366-7

moving in the street, viewing, and studying to infer the situation of the student, because the user does not use the smartphone when he is sleeping. Fig. 5 shows the formation of context. In the preprocessing module, the input data include six types: temperature, humidity, microphone, time, state, and position. The sensor data are prepossessed and sent to the context-aware module. The module infers the three situations: viewing, moving in the street, and studying.

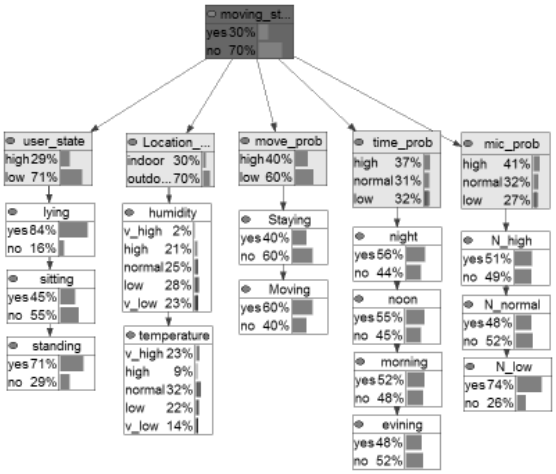


Figure 6. Example of moving street network

Fig. 6 shows an example of moving in the street network. In the network, the white color nodes are result nodes, the right gray nodes are type nodes, and the black node is situation node. The parameters are trained using Maximum Likelihood Estimation (MLE). This is the well-known method for learning the structure and parameters from data [2]. We collected the data for training parameters using MLE. The details about the data will be explained in the experimental section.

IV. EXPERIMENTS

This section describes the experiments conducted to evaluate the usefulness of the proposed method. For the experiments, the evaluation of the proposed method is conducted on Android phone datasets. The data was collected from five graduate students for one week. We used Samsung Galaxy S4. Android phone collected sensor data two times per second, and the amount of the collected data is 66,849 line. We collected data in three situations: studying, moving in the street, and viewing. The students selected the situation and conducted it. Android phone was put into their pocket.

A. Combination of sensors

Accuracy comparison for each combination of sensors verifies the accuracy of the method using sensor combinations except GPS. We conduct experiments of changing from one sensor to five sensors. Fig. 7 shows the accuracy comparison for each combination of sensors. As the result, the network using only GPS sensor has the highest accuracy among the networks using one sensor, because the GPS sensor can collect user location and relates closely to the user situation. However, the network uses about 0.3*W* per one time, which is not low. The accuracy of networks using two and three sensors combination, except GPS is less that of using GPS. The accuracy of network of four sensors; accelerometer, gyroscope, temperature, and humidity, is 94.25% whose power consumption is less.

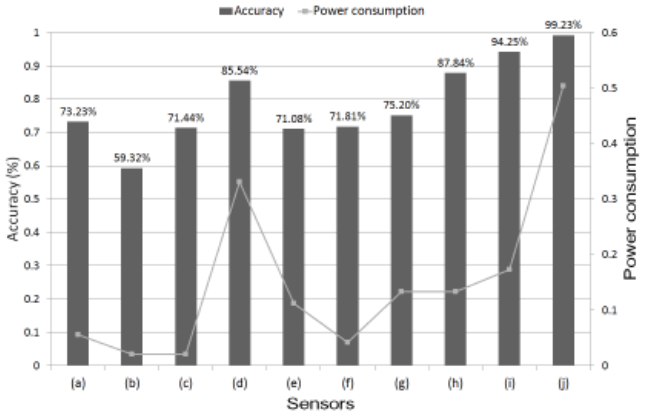


Figure 7. Accuracy comparision for each combination of sensors. (a): Accelerometer, (b): Temperature, (c): Humidity, (d): GPS, (e): Acelerometer+Orientation, (f):Temperature+Humidity, (g) Accelerometer +Gyroscope+Temperature,(h): Accelerometer+Gyroscope+Humidity, (i): Accelerometer+Gyroscope+Temperature+Humidity, (j): Accelerometer+ Gyroscope+Temperature+Humidity+GPS

The accuracy of the proposed method is higher than the accuracy of the network using GPS sensor only, because the proposed method collects user location using temperature and humidity sensors. Although the network using all sensors has the highest accuracy, its power consumption is very high too.

B. Energy consumption

To verify the relation of the inference time and power consumption, we calculate them. The PC configuration is Intel® Core™ i7-2600L CPU, 16.0 GB RAM, Window7. The inference time is calculated in the PC configuration. Table III shows the result of inference time, resulting in that the proposed method reduces more time of about 34% than monolithic BN. The structure of monolithic BN is trained using the Expectation Maximization (EM) algorithm [2].

TABLE III. RESULT OF INFERENCE TIME

BN type	Monolithic BN (BN)	Proposed BN (MBN)	Ratio (BN/MBN)
Moving street	1.178*ns*	0.095*ns*	12.3
Studying		0.084*ns*	14.0
Viewing		0.078*ns*	15.1
All module inferences		0.789*ns*	1.5

To verify the relation of power consumption, we calculate the power consumption using the power tutor application in Samsung Galaxy S4 [21]. The application infers 100 times per second. Fig. 8 confirms the difference of the time consumption of the BN and MBN. The BN

consumes 45,241mW for an hour, whereas the MBN consumes 31,732mW for an hour.

C. Accuracy of proposed method

We conduct 10-fold cross validation to calculate the accuracy of each network, as shown in Fig. 9.

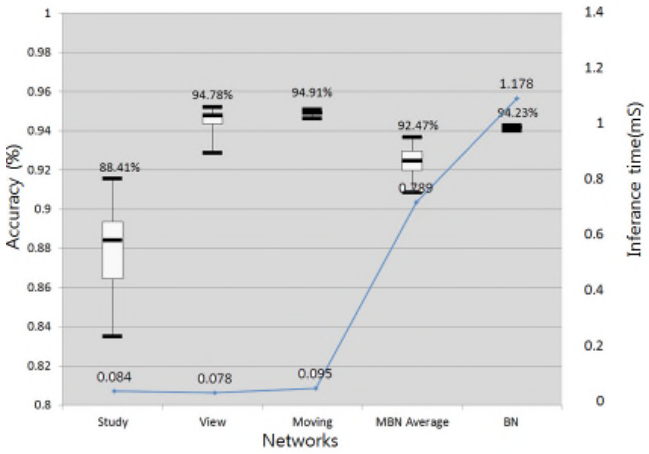


Figure 8. Accuracy and time comparion for each network

The accuracy of the average of MBN is 1.76% less than the BN while the inference time is 0.389mS more than the BN. The system for low-power context-awareness has to consider the inference time and the power consumption, the proposed method is better than the BN.

V. CONCLUDING REMARKS

In this paper, we have proposed a low-power context-aware system considering the power consumption and sensor combination. The system uses temperature and humidity to infer user location instead of using GPS sensor, because the GPS sensor has high power consumption. We use the Bayesian network considering the uncertainty of the situation and modularize to reduce the inference time. The system infers three situations: moving in the street, viewing, and studying. To verify the efficiency of the proposed method, we compare the accuracy of BN of the combination of sensors and calculate the inference time. In addition, we confirm the power consumption using power tutor application and verify the system has lower power consumption than the conventional method.

We will modify parameters using obtained data and will select the optimal time interval of inference. The system will apply to various context-aware service applications.

ACKNOWLEDGMENT

This work was supported by Samsung Electronics, Inc.

REFERENCES

[1] G. Chen D. Kotz, “A survey of context-aware mobile computing research,” Technical Report, Dept. of Computer Science, Dartmouth College, vol. 1, no. 2, pp. 1–16, 2000.

[2] F. V. Jensen, “Bayesian Networks and Decision Graphs,” Springer, 2007.

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

[4] A. M. Otebolaku, M. Abayomi, and M. T. Andrade, “Recongizing high-level contexts form smarphone built-in sensors for mobile media content recommendation,” IEEE Conf. on Mobile Data Management, vol.2, pp. 142–147, June 2013.

[5] X. Wang, D. Rosenblum, and Y. Wang, “Context-aware mobile music recommendation for daily activities,” 20th Int. Conf. on Multimedia, pp. 99–108, November 2012.

[6] A. C. Santos, G. M. P. Cardos, D. R. Ferreira, P. C. Diniz, and P. Chainho, “Providing user context for mobile and social networking applications,” Pervasive and Mobile Computing, vol. 6(3), pp. 324–341, 2010.

[7] S. Phithakkitnukoon and R. Dantu, “Context-aware alert mode for a mobile phone,” Pervasive Computing and Communications, vol. 6(3), pp. 1–23, 2010.

[8] J. Chon and H. Cha, “Lifemap: A smartphone-based context provider for location-based services,” IEEE Pervasive Computing, vol. 10(2), pp. 58–67, 2011.

[9] F. B. Abdesslem, A. Phillips, and T. Henderson, “Less is more: Energy-efficient mobile sensing with senseless,” Proc. of ACM Workshop on Networking, Systems, and Applications for Mobile Handhelds, pp. 61–62, Auguse 2009.

[10] Y. Wang, J. Lin, and M. Annavaram, “A framework of energy efficient mobile sensing for automatic user state recognition,” Proc. of Int. Conference on Mobile Systems, Applications, and Services, pp. 179–192, June 2009.

[11] R. Herrmann, P. Zappi, and T. Rosing. "Context aware power management of mobile systems for sensing applications,” Proc. of Int. Workshop on Mobile Sensing, April 2012.

[12] U. Bareth and A. Kupper, “Energy-efficient position tracking in proactive location-based services for smartphone environments,” Computer Software and Applications Conf. (COMPSAC), pp. 516–521, July 2011.

[13] S.-S. Seo, A. Kwon, J. M. Kang, J. Strassner, and J. W. Hong, "PYP: Design and implementation of a context-aware configuration manager for smartphones,” Proc. of Int. Workshop on Smart Mobile Applications, pp. 12–15, June 2011.

[14] B. G. Marcot, J. D. Steventon, G. D. Sutherland, and R. K. McCann, “Guidelines for developing and updating Bayesian belief networks applied to ecological modeling and conservation,” J. Forest Research, vol. 36, pp. 3063–3074, 2006.

[15] P. Korpipaa, M. Koskinen, J. Peltola, S.-M. Makela, and T. Seppanen, “Bayesian approach to sensor-based context awareness,” Personal and Ubiquitous Computing, pp. vol. 7, pp.113–124, 2003.

[16] Y.-S. Lee, S.-B. Cho, “Mobile context inference using two-layered Bayesian networks for smartphones,” Expert Systems with Applications, vol. 40, no. 11, pp.4333–4345, 2013.

[17] S. H. Chen and C. A. Pollino, “Good practice in Bayesian network modelling,” Environmental Modelling and Software, vol. 37, pp. 134 –145, 2012.

[18] M. Neil, N. Fenton, and L. Nielson, “Building large-scale Bayesian networks,” The Knowledge Engineering Review, vol. 15(3), pp. 257–284, 2000.

[19] G. Pavlin, P. de Oude, M. Maris, J. Nunnink, and T. Hood, “A distributed approach to information fusion systems based on causal probabilistic models,” Intelligent Autonomus Systems Technical Report IAS-UVA-07-03, pp. 1–36, 2007.

[20] “Time use survey”, Strtistics Korea, 2009. http://kostat.go.kr/portal/english/surveyOutlines/3/2/index.static

[21] “A power monitor for android-based mobile platforms,” http://powertutor.org/.

 ISBN: 978-1-61208-366-7

Smart TV – Smartphone Cooperation Model on Digital Signage Environments: An Implementation Approach

Francisco Martinez-Pabon
Telematics Engineering Group
University of Cauca
Popayán, Colombia
fomarti@unicauca.edu.co

Jaime Caicedo-Guerrero
Telematics Engineering Group
University of Cauca
Popayán, Colombia
jcaicedo@unicauca.edu.co

Jhon Jairo Ibarra-Samboni
Telematics Engineering Group
University of Cauca
Popayán, Colombia
jjibarra@unicauca.edu.co

Gustavo Ramirez-Gonzalez
Telematics Engineering Group
University of Cauca
Popayán, Colombia
gramirez@unicauca.edu.co

Mario Muñoz-Organero
GAST Group
Carlos III University of Madrid
Madrid, España
mario.munoz@uc3m.es

Angela Chantre-Astaiza
GITUR Group
University of Cauca
Popayán, Colombia
achantre@unicauca.edu.co

***Abstract*—Modern pervasive digital signage environments demand capabilities beyond the interaction schemes, frequently implemented throughout personal area network technologies or touchscreen features. Smart TV emerges as an interesting alternative model for public displays and Smartphone cooperation, in order to implement a multi-screen approach that complements the task of ads recommendation algorithms for a group of people watching the screen. This paper introduces an implementation approach for a Smart TV – Smartphone cooperation model in digital signage environments using a multi-screen paradigm.**

Keywords - Pervasive advertising; digital signage; Smart TV; cooperation model.

I. INTRODUCTION

Advertising has played an important role in the commerce from its origins; as part of the promotion, one of the marketing areas, the advertising is defined as "any paid form of non-personal presentation and promotion of ideas, goods or services by an identified sponsor" [1]. Recently, a new paradigm known as pervasive advertising, which refers to the use of pervasive computing technologies for advertising purposes [1], has arisen as a promising bet for modern advertisers and consumers. Although most of pervasive advertising approaches has been addressed to mobile devices (Smartphones or tablets), even the public spaces are very interesting for the industry, taking into account that the 75% of the purchase decisions are taken at the purchase places or near of them [2]; this field, known as Digital Signage, it is related to digital content display using public screens [3].

Traditionally, the public screens have been static and non-personalized devices, but modern approaches have enabled the public displays inclusion on pervasive environments. Specifically, the Smart TV model emerged in 2010 throughout the initiative of big vendors such as Samsung, LG, Sony and Intel to build televisions and set-top boxes with more processing power and a better Internet integration [4]. The results for this emergent model are not only limited to free internet access and customization capabilities throughout applications download, but also a valuable capability for connecting and sharing content via standards like Universal Plug and Play (UPnP) or Digital Living Network Alliance (DLNA) [5] with other devices like Smartphones or tablets; this feature is extremely interesting for advertising environments. Although several researches have developed interaction schemes between public displays and mobile devices using personal area network technologies like Bluetooth or Near Field Communication (NFC), even the research for Smart TV model incorporation is incipient. On the other hand, pervasive digital signage environments face other challenges related to customized ads for a group of people watching the screen; traditionally, this issue has been addressed from Recommender System (RS) perspective, which applies search and information filtering techniques to provide users with personalized suggestions about a set of items in a particular domain [6]. However, the perceived serendipity and accuracy about ads recommendations may be not only a matter of the RS algorithm itself, but also a better display strategy issue. Most of public displays interaction initiatives do not consider multi-screen approaches were the content is distributed between the screens in a complementary way; a screen content replication has been used instead.

This paper proposes an implementation approach for a Smart TV – Smartphone cooperation model in digital signage environments using a multi-screen paradigm supported on a flexible protocol. Section 2 summarizes some related work. Section 3 presents a reference architecture for the proposed cooperation model and summarizes some aspects related to the protocol design. Section 4 describes some experimentation results from the user perceived satisfaction perspective. Finally, Section 5 provides some conclusions and future work.

 ISBN: 978-1-61208-366-7

II. STATE OF ART

The research about pervasive advertising involves several interesting topics related to the most important challenges in its implementation. Although several research works have focused on mobile environments where mobile devices are the main tool for advertising purposes, the modern digital signage environments demand the study of new interaction schemes between Smartphones and public displays. Next, some relevant works related to the context of this article will be presented.

At first, some interesting experiences about interactive public displays have been developed around the world. For example, the University of Oulu, in Finland, installed thirteen interactive LCD-screens in several public places across the city, which have been updated and studied since 2009 [7]. The users could interact with the screens by using the touchscreen or their mobile devices. In a similar way, the University of Lancaster, England, installed interactive displays in the small village of Wray [8]. The displays allowed people to upload photos about the village's history and later some capabilities for advertisements uploading about village's services and products were added. Meanwhile, the University of Stuttgart in Germany developed an interesting study about the factors for a successful public digital display environment for advertising purposes [7]. However, these experiments neither consider the use of Smart TV as Interactive Public Display nor interaction schemes throughout Smartphones that enable multi-screen features.

On the other hand, several approaches have been studied for years about content adaptation and collaboration schemes. For example, [9] analyses the interaction between Smartphones and public displays throughout gestures that are used during a screen replication and [10] introduces a touch screen interaction supported on NFC capabilities, but a collaborative interaction between devices is not considered in both approaches. Otherwise, [11] outlines an overview of some functions for a Smart TV – Smartphone interaction, but a reference implementation is not provided. Other researches consider some collaboration models for including zoom functions for the main screen content on the mobile devices [12][13] and some of them consider some phone sensor functionalities [14][15]. Nonetheless, the pervasive advertising in public spaces faces other challenges related to personalized recommendations when a group of people is watching the main screen. Traditionally, this issue has been addressed from the Recommender Systems perspective. Regarding to group recommendations several approaches have been developed: Jameson [16] analyses the issues related to groups recommendations and Masthoff [17] discuss some strategies known as aggregation techniques, which try to aggregate (averaging) individual preferences models in order to create a group model to deliver the recommendations. Other systems related to this approach, like Polylens [18], a particular MovieLens system version, recommends movies based on an algorithm that combines recommendation lists for individual users and sort them in decreasing order. Other similar approaches may be found in [19][20]. Carolis [21] developed a proposal for a pervasive advertising environment using an aggregation approach for recommending ads to the people working at a gym. The ads were displayed on public screens an also on mobile devices but basically they were replied and there was no an interaction mechanism between them. The context of this paper analyzes the personalized ads recommendations provision in public spaces throughout a Smart TV – Smartphone cooperation approach that complements the RS perspective, in order to improve the users perceived satisfaction using interaction and multi-screen display approaches that favors the RS algorithms results assimilation.

III. COOPERATION MODEL REFERENCE ARCHITECTURE

According to the state of art, traditionally the digital signage spaces use public displays with static information following a broadcast approach or some interaction approaches using mobile devices, but they do not use a multi-screen paradigm, so the smartphone or tablet screen capabilities are somewhat wasted. The cooperation model architecture proposes a Smart TV – Smartphone cooperation scheme in which the Smart TV behaves as a public screen displaying recommendations adapted for the group of users that are in front of the TV as long as the Smartphone screen is used to display ads according to the preferences of each user individually. The purpose of this scheme is to take advantage of the full capabilities of each device, using a multi-screen cooperation paradigm, so the TV information is not replied to the Smartphone screen; instead, a complementary information about ads is always displayed on the mobile device, offering more details about a specific offer in the TV screen or giving a more personalized set of items according to the individual profile of the user. Figure 1 shows the reference architecture proposed for Smart TV-Smartphone cooperation model.

A Recommender System (RS) is used for ads suggestion on both Smart TV and mobile device screen. The RS applies search and information filtering techniques to provide users with personalized suggestions about a set of items in a particular domain, in this case advertising. Specifically, a User x User collaborative filtering approach [6] was used for customizing ads on Smartphones and aggregation techniques were used to deliver the recommendations on the Smart TV; in simple terms, the aggregation techniques try to average individual preferences models in order to create a group model [17] as it was previously defined. Although the RS techniques description is out of the scope of this paper, from the RS perspective, the expected results are extremely interesting because this display scheme favors the precision for the ads in the Smartphone. Also, a larger extent of serendipity may be perceived for the recommendations on the TV screen, because the RS uses aggregation techniques for trying to satisfy the preferences of a group of users. So, the recommendations may not be enough accurate for each user, but they may result on novel ads instead, something valuable for pervasive advertising environments with persuasion purposes.

 ISBN: 978-1-61208-366-7

A reference implementation for this model was developed using the Apache Mahout framework [22] for the RS and Samsung Smart TV SDK [23]; mobile application was implemented over the Android platform as long as a simple Representational State Transfer (REST) Application Programming Interface (API) was developed to handle the communication between RS and applications of Smart TV and Smartphone. The communication between the Smartphones and the Smart TV for interaction purposes was enabled throughout the UPnP protocol [24].

According to previous description, a loosely HTTP-based protocol was designed to support multiple user interactions. Standard Simple Service Discovery Protocol (SSDP) [25] messages were used to discover Smart TV Devices in order to be compliant with Samsung Smart TV SDK and Convergence Framework restrictions. A full protocol messages description may be complex for the scope of this paper, so a special emphasis will be done about the discovery, authentication and pairing mechanisms, which are essential processes to start a two-way interaction between the Smart TV and Smartphones under a multi-screen paradigm.

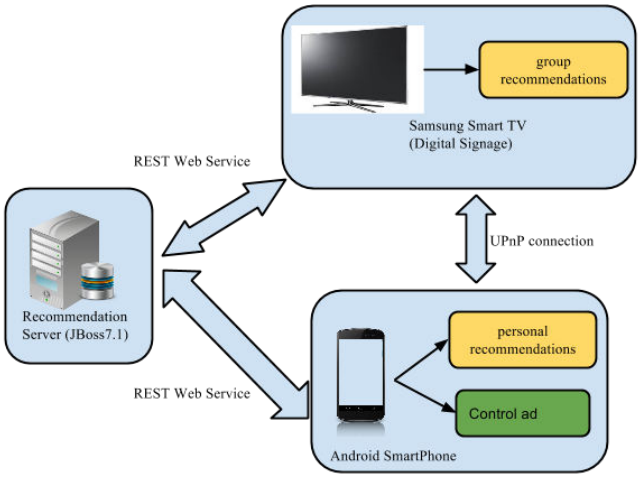


Figure 1. Smart TV – Smartphone cooperation model.

A. *Discovery*

Throughout this process, the Smartphone application looks for available Smart TV devices. SSDP messages defined by the UPnP standard must be used according to Samsung Smart TV Convergence Framework. A list of discovered devices is shown to user at the end of process. The complete messages flow is shown in Figure 2.

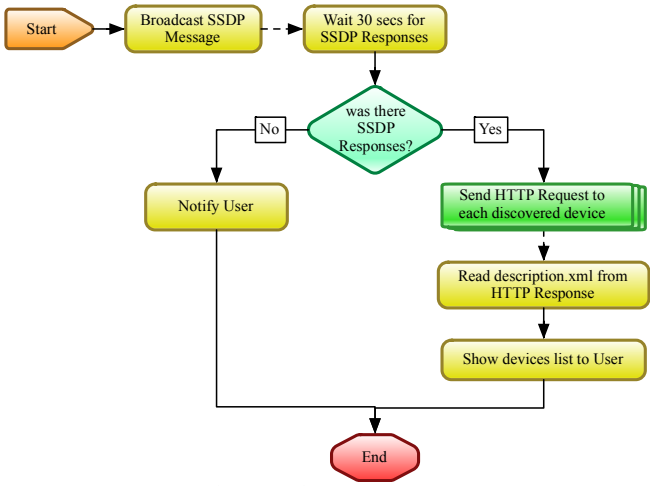
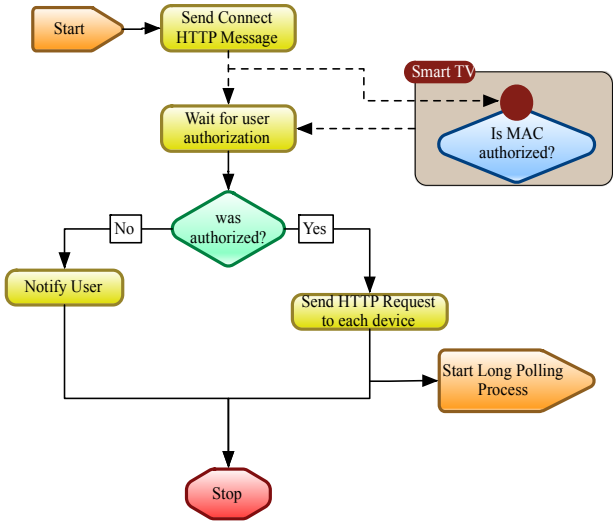


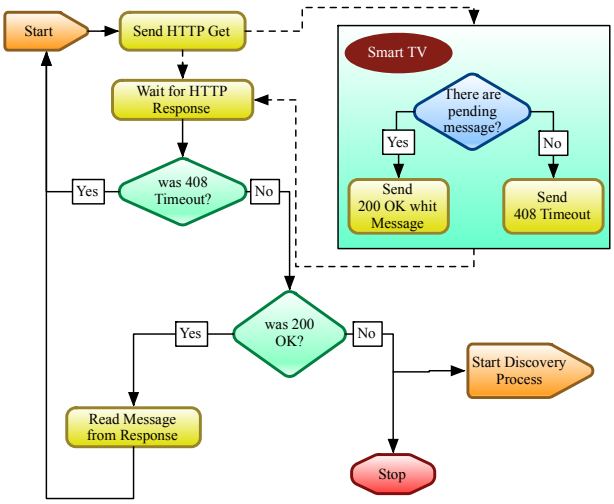
Figure 2. Discovery process.

B. *Authentication and pairing*

In this process, the Smartphone application requests a connection with the Smart TV. In order to be compliant with Samsung Smart TV Convergence Framework, the Smart TV device authenticates the mobile device using the MAC address (Figure 3.a). Once authorized, Smart TV and Smartphone start a long polling process to keep the connection alive (Figure 3.b).



(a)



(b)

Figure 3. Authentication and pairing.

C. *Login*

A login process is required to identify the user in front of the Smart TV screen, so the ads may be customized for the group and individual profiles by the RS accordingly. A "login with Facebook" approach was used to make the process easier and transparent for the users. Once the user login has been completed successfully, the system assigns a color id to each connected smartphone in order to identify all users interaction with the ads on the screen (Figure 4).

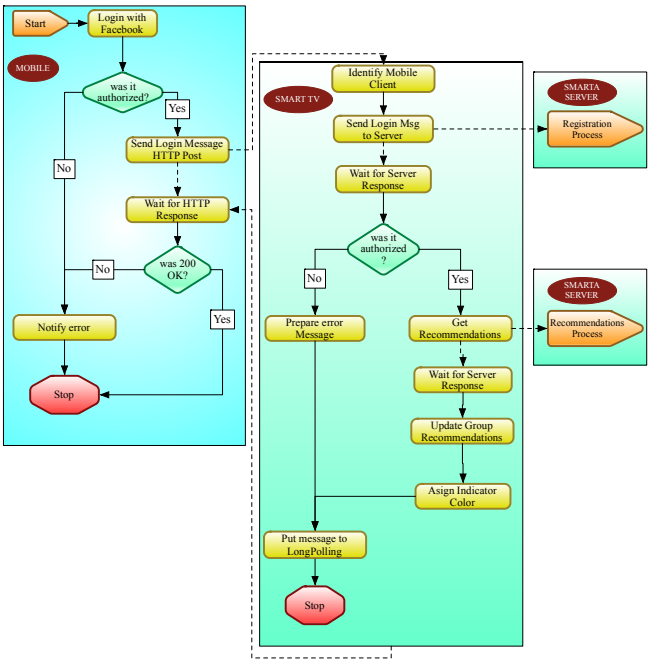


Figure 4. Login.

The other protocol messages were designed as a template for the business logic messages required for a particular advertising application. The messages transport user actions information such as ads retrieval, ads rating or ads publishing to screen, by encoding this content using Java Script Object Notation (JSON). Table I shows a detailed package example for posting ads from the Smartphone to the Smart TV.

IV. EXPERIMENTATION

A small prototype for experimentation purposes was developed and tested in an academic environment. The prototype basis was to implement an electronic alternative to a traditional static ads board, where users post ads using paper posters; these boards are frequently found in small shops or academic campus. The alternative implementation replaces the old board by a new cooperative Smart TV – Smartphone model, where both devices screens are offering ads to users under different but complementary approaches: ads recommendations for group profiles on TV screen and ads recommendations for individual profiles on the Smartphones screens. Moreover, interaction capabilities between both devices were provided to change the static behavior of the traditional board.

In summary, the prototype application included the following functionalities: ads recommendations for a group of users watching the TV screen; ads recommendations according to individual preferences on the Smartphone screen; basic interaction between Smartphone application and Smart TV to go over the ads on TV screen from the mobile application and detail the information for a particular ad in TV screen on the Smartphone screen; post ads to system from the mobile application, mark ads as favorites and rate ads (Figure 5). Alternatively, the interaction protocol detects the users activity, so a list of top ads is displayed on Smart TV when no interaction is detected.

TABLE I. AD POST MESSAGE.

Ad post message request		
Source – destination	Smartphone - Smart TV	
Type	HTTP POST	
Path	http://TV_IPADDR/ws/app/SMARTA/connect	
Path parameters	TV_IPADDR: Smart TV IP address	
HTTP Headers	SLDeviceID	Random sequence of 10 alphabetic characters
	ProductID	SMARTDev
	VendorID	VendorMe
	msgNumber	Auto incremental sequence number for the message
	Content-Type	application/json
	Connection	Keep-Alive
	Accept	*/*
	Accept-Encoding	gzip, deflate, sdch

 ISBN: 978-1-61208-366-7

	Accept-Language	es,en-US;q=0.8,en;q=0.6
	Transfer-Encoding	chunked
	Content-Length	Number of sent bytes in JSON format
Content	{ "type":"post", "title":"post title", "content":"post content", "img_url":"http://someurl.com" }	

Ad post message reply		
Reply codes	200	Accepted connection
	403	Not authorized to receive messages.
	404	Unknown APP_ID.
Reply content	{ "type":"post", "posted":"true/false" }	

At a first stage of the experiment, a group of 26 students from the Tourism program of University of Cauca, posted items and rated almost all of them using an alternative system, even without interacting with the Smart TV and Smartphone cooperation framework. The objective of this first phase was to build a low sparse and editable dataset to make offline tests about the performance of some RS algorithms, previously to online tests. In a second phase, a group of about 50 students posted and rated items in a 1 to 5 scale, using the Smart TV - Smartphone cooperation structure. During this phase, groups of four people were configured randomly to interact with the Smart TV using an Android Smartphone during about five minutes. During this time, users watched the ads recommendations on both screens and used the prototype functions described previously; the number of people per group was a restriction imposed by the UPnP protocol handling of Samsung SDK, but it was considered enough by experimentation purposes. Figure 6 shows some latency results for Smartphones – Smart TV connections and Figure 7 shows the estimated latency for getting ads detailed information in the Smartphone from server, once the user has chosen and ad from Smart TV screen. The tests were performed using a WiFi connection, a Samsung Smart TV 6 Series and Samsung Galaxy and LG Nexus phones with Android 4.2 or above.

At the end of the session, each student filled out a survey form where they were asked explicitly about the perceived satisfaction regarding to the ads delivered on both Smart TV and Smartphone screen. The following analysis will be focused on these results, taking the three aggregation techniques defined by Masthoff (Table II) and tested during the experiment as a starting point; the purpose is try to infer which technique supposes a better connection between the proposed Smart TV – Smartphone cooperation model and the Recommendation System.

Figure 5. Smart TV - Smartphone prototype.

The survey results showed good levels of perceived satisfaction about the ads accuracy using this cooperation scheme between TV and Smartphone screens. However, a deeper analysis was performed for the perceived satisfaction from a group profile perspective, a challenging issue for this kind of digital signage environments. Table III shows the results for the perceived satisfaction for the three techniques.

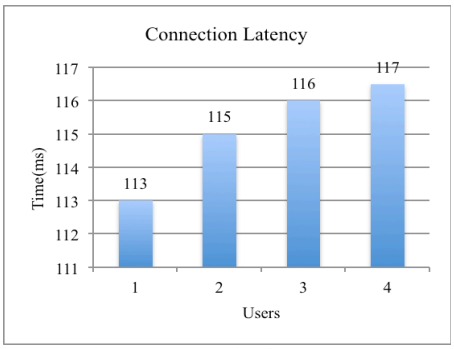


Figure 6. Connection latency.

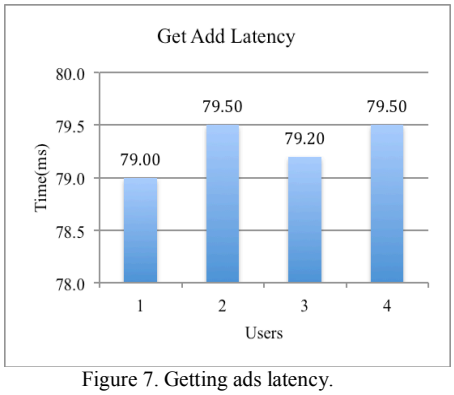


Figure 7. Getting ads latency.

TABLE II. SATISFACTION PERCEPTION.

Test No.	Aggregation technique	Description
1	Additive	Ratings are added; the larger the sum the earlier the alternative is recommended
2	Less Misery	Make a list of ratings with the minimum of the individual ratings; items are recommended based on the rating on that list, the higher the sooner. The idea is that a group is as happy as its least happy member
3	Most pleasure	Make a list of ratings with the maximum of the individual ratings; items are recommended based on the rating on that list, the higher the sooner.

The mean value suggests that less misery technique offers a best-perceived satisfaction value for users; however, it is important to find out if this difference is meaningful. Each technique was tested for a different group of users, so it is required to compare the mean difference in two independent samples. The goal is to define if the satisfaction expressed by users using a less misery technique is significantly higher than Additive or Most pleasure techniques respectively; this analysis was performed using the Two sample t technique, which is frequently used to compare whether the average difference between two groups is really significant [26]. According to this, a null hypothesis should be considered to carry out the test; in this case there are two null hypotheses: the first one proposes that the satisfaction perceived when additive technique is used, is the same when a less misery technique is used; in the same way, the second one involves less misery and most pleasure techniques. In both cases there is an alternative hypothesis, which proposes that less misery strategy offers the best-perceived satisfaction value.

TABLE III. SATISFACTION PERCEPTION.

Aggregation technique	Sample size	Mean	Standard deviation
Additive	16	uA = 3,375	0,957
Less misery	18	uLM = 4,056	0,725
Most pleasure	14	uMP = 3,5	0,941

Table IV shows the analysis results for both hypotheses using a significance level (p) of 5%; a p-value lower than 0.05 is a statistical evidence that approves the alternative hypothesis over the null hypothesis.

TABLE IV. 2 SAMPLE T RESULTS FOR PERCEIVED SATISFACTION.

Test	Null hypothesis	Alternative hypothesis	p-value
1	uLM = uA	uLM > uA	0,014
2	uLM = uMP	uLM > uMP	0,040

For both tests, p-values were less than 0.05; it means there is a meaningful difference about the perceived satisfaction by the users using less misery in contrast with additive or most pleasure techniques. These results suggest that perceived satisfaction about ads recommendations for the group improves when less misery is provided to the least satisfied of its members. In this sense, the contribution of the Smart TV – Smartphone cooperation model is very relevant, because even the least satisfied member of the group has the alternative to find more customized ads recommendations on his Smartphone screen, which seems to increase the perceived satisfaction.

V. CONCLUSION AND FUTURE WORK

This paper proposes a Smart TV – Smartphone cooperation model for digital signage environments using a multi-screen and interactive approach for ads recommendations display. The design of a loosely coupled and simple protocol based on HTTP using a RESTful style and UPnP packets, provides a scalable and compatible framework for the cooperation model implementation.

On the other hand, although the recommendations to groups have been a challenging issue on digital signage environments, it has been addressed frequently from RS algorithms perspective, but the Smart TV – Smartphone cooperation model offers an alternative to complement the RS algorithms task from an information display approach. The experimentation suggests that less misery aggregation technique offers the best results in this kind of digital signage environments regarding to perceived satisfaction by users, taking into account that even the least satisfied user has the chance to improve the perceived accuracy of group recommendations displayed on TV, using his Smartphone screen.

Future work is related to the cooperation model evolution towards a more distributed middleware that overcome some restrictions related to the number of supported users and some intermittent behavior during pairing and connection mechanisms observed during the experimentation with Samsung Smart TV Convergence Framework.

ACKNOWLEDGMENT

This work was supported by the University of Cauca throughout the projects VRI 3593 “SMARTA: Modelo para el despliegue de publicidad en entornos de computación ubicua soportado en un esquema de cooperación Smart TV - Smartphone” and VRI 4045 “MANTISS: Modelo para la adaptación de contenidos publicitarios en entornos n-screen interactivos soportados en un esquema de colaboración Smart TV – Smartphone”. Francisco Martinez is funded by Colciencias Doctoral scholarship N. 567.

REFERENCES

[1] J. Müller, F. Alt, and D. Michelis, “Pervasive Advertising,” Springer London, pp. 1–29, 2011.

[2] C. Bauer and S. Spiekermann, “Conceptualizing context for Pervasive Advertising,” in Pervasive Advertising, J. Müller, F. Alt, and D. Michelis, Eds. Springer London, pp. 159–183, 2011.

[3] U. Stalder, “Digital Out-of-Home Media: means and effects of digital media in public space,” in Pervasive Advertising, J.

Müller, F. Alt, and D. Michelis, Eds. Springer London, pp. 31–56, 2011.

[4] F. Jewet., "Why Smart TV is the next big thing". Available: http://www.uievolution.com/mobileconnect/Mobile_Connect_June_2011.pdf, [Accessed: 10-Dic-2013].

[5] DLNA.org, "Digital Living Network Alliance," July, 2014. Available: http://www.dlna.org, [Accessed: 03-Jul-2014].

[6] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in Recommender Systems Handbook, F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds. Springer US, pp. 1–35, 2011.

[7] F. Alt, A. Schmidt, and A. Schmidt, "Advertising on public display networks," Computer, vol. 45, no. 5, May. 2012, pp. 50–56, doi:10.1109/MC.2012.150

[8] T. Ojala, et al., "Multipurpose interactive public displays in the wild: three years later," Computer, vol. 45, no. 5, 2012, pp. 42–49, doi:10.1109/MC.2012.115

[9] The Next Web, "Google's Open Project lets you beam apps to an external display using only your smartphone's camera," Available: http://thenextweb.com/google/2013/09/26/google-researchs-open-project-lets-you-beam-apps-to-an-external-display-using-only-your-smartphones-camera/, [Accessed: 06-Nov-2013].

[10] G. Broll, E. Vodicka, and S. Boring, "Exploring multi-user interactions with dynamic NFC-displays," Pervasive and Mobile Computing, vol. 9, no. 2, Apr. 2013, pp. 242–257, doi:10.1016/j.pmcj.2012.09.007

[11] C. Yoon, T. Um, and H. Lee, "Classification of N-Screen services and its standardization," Proc. 14th International Conference on Advanced Communication Technology (ICACT), 2012, pp. 597–602.

[12] P. Baudisch, N. Good, and P. Stewart, "Focus plus context screens: combining display technology with visualization techniques," Proc. 14th Annual ACM Symposium on User Interface Software and Technology, ACM Press, 2001, pp. 31–40.doi:10.1145/502348.502354

[13] S. Boring, M. Jurmu, and A. Butz, "Scroll, tilt or move it: using mobile phones to continuously control pointers on large public displays," Proc. 21st Annual Conference of the Australian Computer-Human Interaction Special Interest Group: Design: Open 24/7, 2009, pp. 161–168, doi:10.1145/1738826.1738853

[14] R. Ballagas, M. Rohs, and J. G. Sheridan, "Sweep and point and shoot: phonecam-based interactions for large public displays," Proc. HCI '05 Extended Abstracts on Human Factors in Computing Systems, 2005, pp. 1200–1203, doi:10.1145/1056808.1056876

[15] R. Hardy and E. Rukzio, "Touch & Interact: touch-based interaction of mobile phones with displays," Proc. 10th International Conference on Human Computer Interaction with Mobile Devices and Services, 2008, pp. 245–254, doi:10.1145/1409240.1409267

[16] A. Jameson, "More than the sum of its members: challenges for group recommender systems," Proc. Working conference on Advanced visual interfaces, 2004, pp. 48–54, doi:10.1145/989863.989869

[17] J. Masthoff, "Group modeling: selecting a sequence of television items to suit a group of viewers," User Model User-Adapt Interact, vol. 14, Feb. 2004, pp. 37–85, doi:10.1023/B:USER.0000010138.79319.fd

[18] M. O'Connor, D. Cosley, J. A. Konstan, and J. Riedl, "PolyLens: a recommender system for groups of users," in ECSCW 2001, W. Prinz, M. Jarke, Y. Rogers, K. Schmidt, and V. Wulf, Eds. Springer Netherlands, pp. 199–218, 2002.

[19] J. K. Kim, H. K. Kim, H. Y. Oh, and Y. U. Ryu, "A group recommendation system for online communities," in International Journal of Information Management., vol. 30, no. 3, Jun. 2010, pp. 212–219, doi:10.1016/j.ijinfomgt.2009.09.006

[20] I. A. Christensen and S. Schiaffino, "Entertainment recommender systems for group of users," Expert Systems with Applications, vol. 38, no. 11, Oct. 2011, pp. 14127–14135, doi:10.1016/j.eswa.2011.04.221

[21] B. D. Carolis, "Adapting News and Advertisements to Groups,"in Pervasive Advertising, J. Müller, F. Alt, and D. Michelis, Eds. Springer London, pp. 227–246, 2011

[22] Apache, "Apache Mahout: Scalable machine learning and data mining," Available: http://mahout.apache.org/, [Accessed: 05-Jul-2013].

[23] Samsung, "Samsung Smart TV Apps Developer Forum," Available: http://www.samsungdforum.com/Devtools/Spec, [Accessed: 31-Jan-2014].

[24] UPnP Forum, "UPnP Forum," Available: http://www.upnp.org/, [Accessed: 11-Feb-2014].

[25] UPnP Forum, "Universal Plug and Play Device Architecture," Available: http://upnp.org/specs/arch/UPnP-arch-DeviceArchitecture-v1.1.pdf, [Accessed: 11-Feb-2014].

[26] S. Wellek, Testing Statistical Hypotheses of Equivalence, 1st ed., Chapman and Hall/CRC: Florida, 2002.

 ISBN: 978-1-61208-366-7

A Flow Aggregation Scheme for Seamless QoS Mobility Support in Wireless Mesh Networks

Dario Gallucci, Steven Mudda, Salvatore Vanini
Information Systems and Networking Institute
SUPSI
Manno, Switzerland
Email: [dario.gallucci,steven.mudda,salvatore.vanini]@supsi.ch

Radoslaw Szalski
Institute of Control and Information Engineering
Poznan University of Technology
Poznan, Poland
Email: radoslaw.szalski@put.poznan.pl

***Abstract*—Current solutions for network mobility support in wireless mesh networks lack Quality of Service (QoS) capabilities. Thus, they are not well suited for supporting services with QoS requirements (e.g., Voice over IP or Video on Demand). WiOptiMo is a solution, originally designed for seamless handoff management in the Internet, that was adapted for seamless internetworking in wireless mesh networks. In this paper, we show how its basic infrastructure was modified in order to meet the QoS expectations of mobile users running heterogeneous applications on a wireless mesh network. Specifically, QoS support is provided by aggregating application traffic flows with the same characteristics to limit overhead and by relaying compressed aggregated flows to the appropriate mobility provider. We experimentally evaluate the performance of our aggregation scheme and demonstrate that link utilization is optimized and QoS is improved.**

***Keywords*-Wireless Mesh Networks; Seamless Handover; QoS Mobility Support; Flow Aggregation; Flow Classification.**

I. INTRODUCTION

Recent years have witnessed a significant reduction in the costs of mobile computing platforms (e.g., laptops and smartphones), especially the hardware used in WiFi devices and has led to a widespread use of Wireless Mesh Networks (WMNs). WMNs provide multiple services to people using their mobile devices via a combination of fixed and mobile nodes, interconnected via wireless links to form a multi-hop ad-hoc network. WMNs are a cost-effective solution to extend the range of wired infrastructure networks with the help of easy to deploy wireless nodes. For example, the backbone of a telecom service provider can be easily expanded utilizing mechanisms to manage resources of wireless nodes [1] [2]. Existing mechanisms work only in scenarios where wireless connection stability can be ensured. For example, CARMNET [3] [4] utilizes the WMN paradigm to enable nearby wireless devices communicate with each other and proposes a distributed resource management method that can be easily integrated with a telecom IMS software infrastructure. This method (implicitly) assumes that the underlying network connectivity is not affected by topological changes (e.g., gateway changes) caused by the mobility of network's nodes. During those changes, packets for a given application flow might be rejected because of the change of the IP address, or they might be lost due to out-dated routing information. As a consequence, the quality and performance of correspondent applications can significantly decrease. Traditional mobility management schemes designed for IP-based networks are not suitable for WMN architectures. For example, Mobile IP [5] focuses on keeping the IP identity of a mobile node only. However, it introduces network overhead due to the protocol signaling and, consequently, causes a degradation of TCP throughput. On the other hand, since mobility support in pure ad-hoc networks focuses on rerouting (i.e., finding an alternative path in a timely manner, so that a flow can be handed off to the new path upon link disruption), these schemes perform poorly in WMNs. To overcome these limitations, several works have proposed different approaches to provide QoS and seamless mobility support in WMNs. However, many of them are not designed to manage multimedia services with QoS requirements—e.g., Voice over IP (VoIP) or Video on Demand (VoD). In this paper, we present an extension of our WiOptiMo [6] framework (described in section III) to provide generalized QoS mobility support in WMNs. In sections IV and V, we describe our enhanced framework and flow aggregation scheme to provide the required QoS to different types of applications in a WMN scenario. Finally, in section VI, we evaluate the performance improvement with respect to its standard configuration for WMNs.

II. RELATED WORK

The existing work on mobility management in WMNs focuses on providing network-layer mobility support. RFC 4886 [7] specifically addresses the issue of network mobility. The different solutions presented in literature focus on managing the address of a mobile node due to the handoff process. In general, we can distinguish between intra-domain and inter-domain mobility. The first refers to handoffs inside the same network domain, the second to handoffs between different network domains. MobileNAT [8] addresses both intra- and inter-domain mobility. MobileNAT requires a modification at the network layer stack of a mobile node and changes to the standard DHCP protocol, which introduces network latency. SyncScan [9] is a Layer-2 procedure for intra-domain handoff in 802.11 infrastructure mode networks. It achieves good performance at the expense of a required global synchronization of beacon timings between clients and access points (AP). iMesh [10] provides low handoff latency for Layer-3 intra-domain handoffs between APs of a WMN. However, the hand-

 ISBN: 978-1-61208-366-7

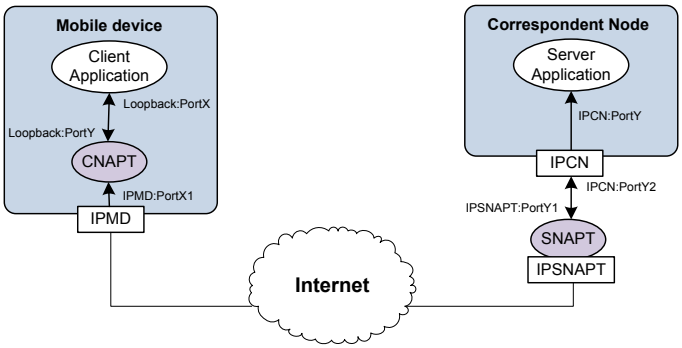


Figure 1: WiOptiMo's CNAPT and SNAPT IP decoupling.

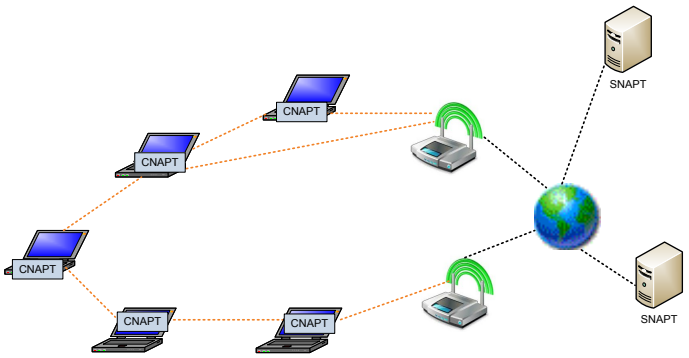


Figure 2: WiOptiMo configuration for a WMN.

off latency depends on the number of nodes between the new and old AP. BASH [11] focuses on the design of a intra-domain Layer-2 seamless handoff scheme for 802.11 WMNs, but the handoff protocol requires modifications at every mobile client. Authors of [12] use tunneling, as well as the standard Mobile IPv6 solution [13] and most of the existing network-layer mobility management schemes based on Mobile IP, such as Mobile Party [14] and AODV-PRD [15]. Tunneling introduces extra delay for the encapsulation/decapsulation of packets and has intrinsically low flexibility. Finally, SMesh [16] provides a 802.11 mesh network architecture for both intra-domain and inter-domain handoffs. For intra-domain handoffs, SMesh generates high network overhead, which grows linearly with the number of mobile clients. In case of inter-domain handoffs, network overhead generated by SMesh is proportional to the number of connections of a mobile client. The WiOptiMo framework provides mobility support by separately managing each application's flow, to meet the QoS expectations of all applications. In [6], we describe the architecture of WiOptiMo and present how it is adapted to handle a WMN context in [17]. In the next sections, we show how its architecture has been modified to handle efficiently multiple application's flow with different QoS requirements.

III. The WiOptiMo framework

WiOptiMo enables handoffs initiated by a mobile device. It manages the mobility of every device with the help of two software modules: Client Network Address & Port Translator (CNAPT) and Server Network Address & Port Translator (SNAPT). Together, these two components provide decoupling between the IP address assigned to a mobile device and the IP address used to access a service on the Internet. CNAPT and SNAPT hide any change of the IP address when a mobile host moves between different access networks, inside the same domain or between different domains. In Figure 1, we present a scenario where a mobile device with IP address IPMD has an active TCP session to a corresponding node with IP address IPCN. The TCP data packets are first relayed to the local CNAPT, which in turn relays them to the SNAPT. Upon receiving packets, the SNAPT (processes and) forwards them to the IPCN address. When the mobile device moves to a new network and gets a new IP address, the change in IP does not affect the application layer because the application packets are sent to the the local CNAPT, which relays them to the SNAPT with fixed IP address (IPSNAPT). This mechanism also allows a mobile node of a WMN to change gateway transparently (e.g., when node moves out of the reach of the initial gateway due to the mobility of the associated user), without suffering service disruption. To correctly manage the handoff process, CNAPT and SNAPT exchange handshaking packets with each other using a control socket.

In a generalized setting, mobile devices have CNAPT installed on them, while an Internet server or any node in a network (as in the scenario previously described) have SNAPT installed on them.

A. WiOptiMo Architecture for a WMN

In [18], we present a general configuration of our WiOptiMo for a WMN. We exploit the flexibility of location where a SNAPT can be installed to address scalability issues that might arise in a WMN. In this scenario, multiple SNAPTs can be deployed on mesh routers or on Internet nodes to avoid network congestion in a single spot. Every mobile wireless device has CNAPT installed on it to provide independent mobility support. We use a combination of network status monitoring and user configurable policy to enable every CNAPT to choose a suitable SNAPT that will relay its application flows. At start-up, each CNAPT connects to a fixed SNAPT specified in a configuration file. Then, it receives a list of other available SNAPTs from the currently connected SNAPT, and measures the delay towards them by means of passive and active monitoring of the control connection towards the SNAPTs, used for handshaking. CNAPTs also take into account the bandwidth used by applications in order to make a more wise SNAPT choice. The CNAPTs select a SNAPT to relay their data depending on the measured delay and estimated remaining throughput (based on the application's bandwidth requirements). This selection policy also helps in reducing the overload on any single SNAPT. Figure 2 shows WiOptiMo's architecture for a WMN. The SNAPTs can be managed by private administrators (otherwise called mobility service providers), who may require a fee for the use of their mobility service. This circumstance might foster the competition between mobility service providers, forcing them to increase the quality of provided service and benefit the entire WMN.

B. Implementation changes

We adapted WiOptiMo's implementation (both CNAPT and SNAPT) for low profile devices and to provide a fast handoff procedure. Figure 3 shows the changes to the basic

 ISBN: 978-1-61208-366-7

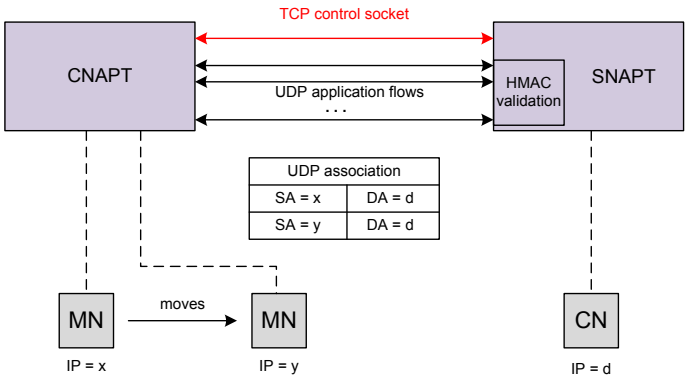


Figure 3: WiOptiMo adaptation for a WMN.

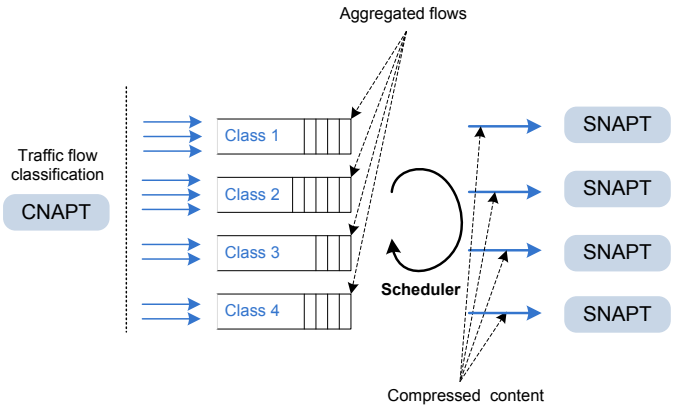


Figure 4: Software architecture of the aggregation scheme.

implementation of WiOptiMo. A TCP control socket still manages the communication between a CNAPT and a SNAPT. It provides network configuration parameters (e.g., the MTU of the underlying network) and also transmits data packets in a fall-back mode when middle-boxes, such as firewalls and/or NATs, block UDP packets. Further, the control socket is used to authenticate the CNAPT and to exchange a session key for providing data authenticity and integrity during a handoff. The CNAPT relays data packets to SNAPTs (and vice versa) using UDP sockets—this solution increases performance during handoffs, because UDP does not need to retransmit lost packets nor does it perform any connection setup. When a SNAPT receives a UDP data packet, it validates it using HMAC [19] and tests it against replay attacks using a sequence number. During handoffs (i.e., when the source IP address of data packets changes), the SNAPT updates the return IP address for the flow and transmits a keep-alive request to the CNAPT, which will reset the control connection or hasten the detection of a timeout. This event will then trigger the re-establishment of the control socket connection to the SNAPT.

IV. QoS SUPPORT IN WIOPTIMO

We need an efficient delivery of heterogeneous traffic to meet the QoS requirements of applications. Since WiOptiMo relays each outgoing data flow from a client to a server application (through the link between CNAPT and SNAPT), every flow from a mobile device to its intended destination can be managed separately, according to its characteristics. In this section, we present the improvements to the WiOptiMo framework that enable it to efficiently deal with QoS, while still providing mobility support.

A. Flow classification

To meet the QoS requirements of applications, data flows are relayed to different SNAPTs based on their delay and throughput needs. In this regard, we identified four different flow classes according to the minimum throughput and maximum delay requirements of applications: *High Throughput and High Delay* (HT & HD), *High Throughput and Low Delay* (HT & LD), *Low Throughput and High Delay* (LT & HD), *Low Throughput and Low Delay* (LT & LD). In terms of throughput, the minimum threshold for classifying HT flow classes is 64kbit/s. We set the maximum delay for LD classes to 1s. As previously stated, during the normal workflow, a CNAPT periodically measures delay (one-trip time) and throughput (amount of received data over a time period) towards the different SNAPTs. Then, for each application flow, it detects the class type on the basis of process name, protocol and port number. Every class has an assigned delay and throughput requirements and data flows get relayed to a SNAPT that meets their delay and throughput requirements.

While our solution for flow classification is conceptually similar to DiffServ [20], it doesn't have its drawbacks. First, flow classification is performed dynamically per SNAPT, so that new flows are allocated depending on the current network performance statistics (e.g., the increase of the delay with the increase of the load). Second, our framework might refuse to serve a flow if its QoS requirements cannot be met, hence avoiding to disrupt the traffic already allocated. Moreover, the routing layer, as explained in [18], knows which traffic is managed by WiOptiMo. In this way, a QoS-aware routing mechanism can be executed whenever needed. In particular, network statistics about each single flow are reported to the routing layer so that there is no loss of granularity in the traffic management.

V. FLOW AGGREGATION MECHANISM

WiOptiMo allocates a UDP socket for each application flow (i.e., TCP connection). This behaviour does not favor the efficient handling of application flows with short frequent sessions (e.g., DNS requests), because useless computational overhead can be generated. It is also inefficient in terms of performance because the wireless link can be under-utilized. Furthermore, major unfairness may occur between competing flows—a major drawback when wireless links have high latency [21]. A naive solution would be to aggregate all data into a single flow, however applications with high bandwidth requirements would delay low latency applications. To overcome these issues, we designed a class based aggregation technique. Classified flows that belong to the same class are treated as a single aggregate and transmitted to a SNAPT using the same UDP socket. Our objective is to maximize the utilization of the available link bandwidth and reduce network overhead, thereby increasing the achieved throughput without significantly impacting the latency requirements.

Figure 4 presents the details of our aggregation mechanism.

 ISBN: 978-1-61208-366-7

TABLE I: Different Parameters of the Experiment.

Application Class	Packet Size (Bytes)	Range of bit-rate (bit/s)	Range of Flows
HT & HD	1360	1M - 20M	1 - 5
HT & LD	576	128k - 2M	1 - 5
LT & HD	1360	15k - 1M	1 - 5
LT & LD	100	15k - 128k	1 - 5

HT - High Throughput LT - Low Throughput
HD - High Delay LD - Low Delay

We implemented four connection queues, one for each of the application classes defined in Section IV-A. The queues feed into a scheduler, which uses a connection strategy based on flows' priority: the scheduler sends classes with more stringent requirements in terms first of delay and then of bandwidth. To reduce the amount of exchanged data, we enabled compression of the aggregated flows—packets are appended to the aggregated compound until their cumulative compressed size does not exceed the 70% of the underlying network's MTU. We chose this threshold to maximize the effectiveness of aggregation without having to resort to a slower algorithm.

VI. Experimental Results

In this section, we present the experiments conducted to assess the performance and QoS support of WiOptiMo with flow aggregation.

A. Performance of WiOptiMo with flow aggregation

We conducted experiments in three different scenarios:

1) Baseline: without WiOptiMo.
2) WiOptiMo basic.
3) WiOptiMo with flow aggregation mechanism.

Measurements showed that the performance of the baseline and WiOptiMo basic configurations are comparable (the degradation on throughput and the additional end-to-end delay introduced by the WiOptiMo solution are negligible, as presented also in [6]). For this reason, we report only the results for the baseline and WiOptiMo with flow aggregation scenarios.
In the next paragraphs, we show that our flow aggregation scheme achieves a better link utilization and reduces the amount of bytes exchanged in the network.

Experiment setup: We installed the WiOptiMo SNAPT on a Dell Optiplex 760 (server) and WiOptiMo CNAPT on a Dell Precision M4300 (client) with LinkSys Dual-Band Wireless A+G PCI Card. To avoid interference with nearby 802.11 access points operating on the 2.4 GHz band, we connected the client and server through a Netgear WNDR3800 wireless router (with OpenWRT 12.09 and only 802.11a networking enabled). Both client and server operated on a Linux distribution (Ubuntu 12.04 with Linux kernel 3.11).

We used the *Iperf* [22] network testing tool to send a stream of UDP packets (at a specific bit-rate) to server and measured the number of bytes sent between client and server using the *dumpcap* utility [23]. Instead of using the default UDP packets generated by Iperf—all packets contain same data—we configured the Iperf utility to generate UDP packets containing random text stored in a file. We performed experiments under the four different classes described in Section IV-A. For each flow class, we fixed the size of data in every UDP packet transmitted by the Iperf utility. We repeated experiments 10 times, to get more reliable results. Table I shows the characteristics of every flow generated by Iperf to measure the performance of WiOptiMo (for each application class).

We measured the performance of WiOptiMo by varying the number of flows and bit-rate of each flow, and observing their impact on the percentage of bytes saved on the link, due to flow aggregation and compression. It is calculated by subtracting pre-aggregation (and compression) bytes and post-aggregation (and compression) bytes, and dividing this difference by the pre-aggregation (and compression) bytes. This metric measures the bytes saved in the packet transfer between the client and server with the flow aggregation configuration, compared to the baseline configuration. It captures the energy spent to transfer data to the server. Since WiOptiMo performs flow aggregation and compression, this metric will enable us to measure the amount of energy that could be saved without impacting the QoS of applications.

Results: Figure 5 shows the percentage of bytes saved for applications with high throughput and high delay network requirements. We observe that for bit rates lower than 10Mbit/s, the percentage of bytes saved increases as the number of flows increases. Even for a single application flow, WiOptiMo with flow classification and aggregation helps in reducing, on average, the 60% of data sent between client and server. For bit-rates higher than 10Mbit/s, the percentage of bytes saved is still high but its relationship with the number of flows is no longer linear. This behaviour is due to the saturation of the system's modules capacity (wireless card, aggregation and compression mechanisms).

In Figure 6, we observe that when applications have high throughput and low delay requirements, savings by WiOptiMo increase from 38% for single flow to a maximum of 82.5% for applications with 5 flows. For all flows, the percentage of bytes saved increases until the bit-rate reaches about 400kbit/s. For much higher rates we observe that the percentage of bytes saved remains constant.

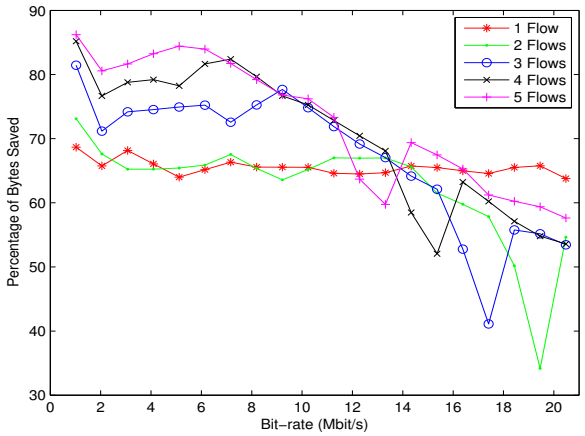


Figure 5: Percentage of bytes saved due to flow aggregation in HT & HD applications.

 ISBN: 978-1-61208-366-7

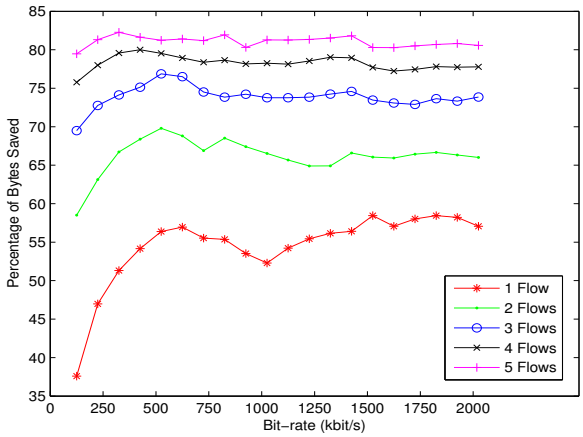


Figure 6: Percentage of bytes saved due to flow aggregation in HT & LD applications.

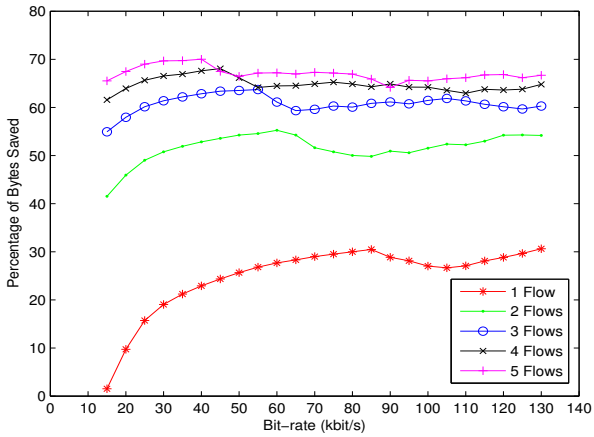


Figure 8: Percentage of bytes saved due to flow aggregation in LT & LD applications.

For low throughput and high delay tolerant applications (see Figure 7), we observe that for low bit-rates (∼125kbit/s), the percentage of bytes saved is not significant because no additional savings could be achieved by compressing and aggregating data packets arriving at long intervals of time. For higher bit rates (that is after the size of the aggregated packets allows better compression), savings increase and then stay constants (we can achieve a maximum savings of around 90%). In Figure 7, we also observe that savings achieved by WiOptiMo increase as the number of application flows increases.

Finally, for applications with low throughput and low delay requirements, we could achieve a maximum saving of 70% (see Figure8). Even at very low bit-rate (∼20kbit/s), WiOptiMo is able to save 10% of the data transferred between client and server.

B. QoS support by WiOptiMo

In the second set of experiments, we tested the capability of the WiOptiMo with an aggregation schema to provide QoS support. We used Iperf and measured the throughput between client and server using two different flow classes (HT & LD and HT & HD), in two distinct configurations: with a single SNAPT and with two SNAPTs. We show that a software configuration with multiple SNAPTs increases the network throughput and then helps preserving the QoS of applications.

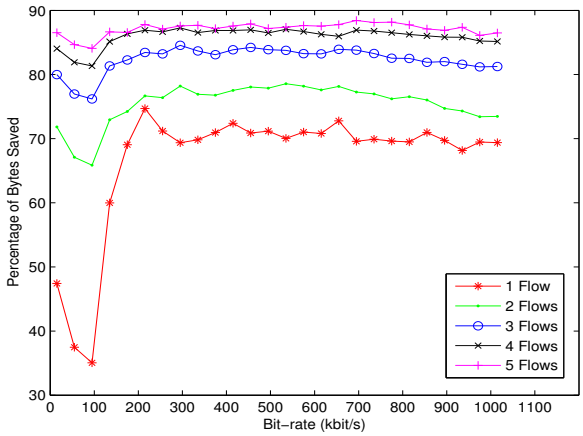


Figure 7: Percentage of bytes saved due to flow aggregation in LT & HD applications.

Experiment setup: We setup a wireless mesh network testbed to measure the QoS offered by WiOptiMo. The testbed consists of three static Internet-sharing nodes and two wireless mobile nodes. Each static node consists of an ALIX.2D2 system board, which supports two mini-PCI radios. We used one Wistron DNMA92 miniPCI card for each board, which is in turn connected to two 802.11n antennas. Each board mounts a 500 MHz AMD Geode LX800 processor and 256 MB DDR DRAM, runs Debian Wheezy 7.0 with Linux Kernel 3.12.6, and uses an ath9k driver for WiFi.
We used two ASUS EeePC 900 (with a Atheros 5008 Wireless Card, a 900MHz Celeron Processor and 1GB DDR RAM) as mobile nodes in our experiments. They operated on Debian Wheezy 7.0 with an ath5k WiFi driver.

To complete the hardware set-up, we installed WiOptiMo SNAPT on two Dell Optiplex 760 (servers) and a Lenovo ThinkPad T410a had WiOptiMo CNAPT installed on it. Both the machines operated on a Linux distribution (Ubuntu 12.04 with Linux kernel 3.11). Two static nodes (gateways) and

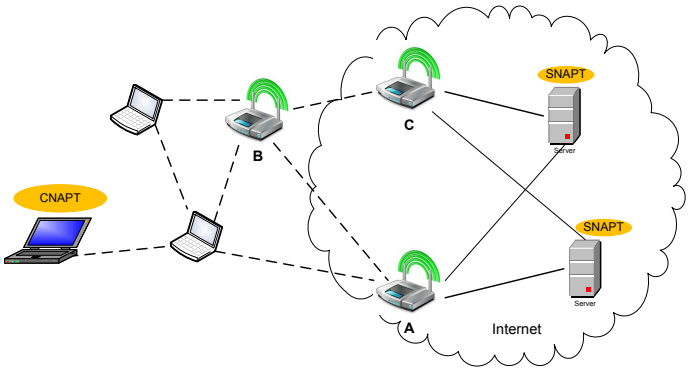


Figure 9: Testbed mesh network architecture.

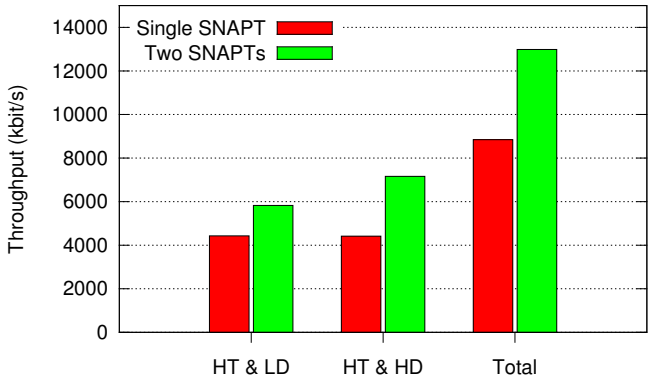


Figure 10: Throughput with multiple SNAPTs.

two servers were connected to the Internet with an Ethernet connection, while the rest of the nodes participate in the mesh network. We set the bandwidth of Ethernet connection to 10Mbit/s. The gateways performed NAT between the mesh network and the Internet. We ran the Optimised Link State Routing Protocol daemon (OLSRd, version 0.6.2) [24] on each node for network path resolution and configured the network to ensure that the two SNAPTs could be reached by separate gateways. The final testbed architecture is shown in Figure 9.

Results: Figure 10 shows the throughput comparison for two scenarios: with single SNAPT and with two SNAPTs (with different network delays) that could be reached from separate gateways. The results clearly show that in the first scenario the available bandwidth gets divided equally between the two application classes. In the second scenario, the HT & HD class achieves on average higher throughput compared to HT & LD class because the data of HT & LD class always gets routed to the SNAPT with lowest delay. Specifically, in the two SNAPT scenario, we observe a higher throughput compared to the bandwidth available towards each single gateway. Finally, we did not observe any significant additional delay in the network due to the introduction of WiOptiMo.

VII. Conclusion

In this paper, we have proposed a flow classification and aggregation scheme for enabling the WiOptiMo framework to manage multiple applications with different QoS requirements in a wireless mesh networking environment. We evaluated the proposed scheme on a Linux-based wireless mesh network testbed. Experimental results show that the aggregation mechanism we designed improves network performance in terms of link utilization and QoS, while still providing mobility support, without requiring any changes to be made to the network protocol stacks of either the mobile or fixed end systems. In the future, we would like to define and integrate into WiOptiMo, a requirements based policy that optimizes the use of mobility services and rewards users who do not waste network resources.

Acknowledgment

This work is supported by a grant from Switzerland through the Swiss Contribution to the enlarged European Union (PSPB-146/2010, CARMNET).

References

[1] S. Jakubczak, D. Andersen, M. Kaminsky, K. Papagiannaki, and S. Seshan, "Link-alike: using wireless to share network resources in a neighborhood," pp. 1–14, October 2008.

[2] C. Middleton and A. Potter, "Is it good to share? a case study of fon and meraki approaches to broadband provision," in Proceedings of International Telecommunications Society 17th Biennial Conference, 2008.

[3] M. Glabowski and A. Szwabe, "Carrier-grade internet access sharing in wireless mesh networks: the vision of the carmnet project," in Proceedings of The Ninth Advanced International Conference on Telecommunications, June 2013, (in press).

[4] P. Walkowiak, R. Szalski, S. Vanini, and A. Walt, "Integrating carmnet system with public wireless networks," ICN 2014, The Thirteenth International Conference on Networks, feb 2014, pp. 172–177.

[5] D. Johnson, C. Perkins, and J. Arkko, "Mobility support in ipv6," RFC 3775, June 2004.

[6] G. A. D. C. et al., "Wioptimo: A cross-layering and autonomic approach to optimized internetwork roaming," in AHSWN Journal, May 2007, pp. 104–113.

[7] T. Ernst and L. H., "Network mobility support goals and requirements," in RFC 4886, July 2007.

[8] M. Buddhikot, A. Hari, K. Singh, and S. Miller, "Mobilenat: A new technique for mobility across heterogeneous address spaces," in ACM Mobile Networks Apps, vol. 10, no. 3, June 2005, pp. 289–302.

[9] I. Ramani and S. Savage, "Syncscan: Practical fast handoff 802.11 infrastructure networks," in 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2005), vol. 1, 2005, pp. 675–684.

[10] V. Navda, A. Kashyap, and S. R. Das, "Design and evaluation of imesh: an infrastructure-mode wireless mesh network," in IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks (WOWMOM), Italy, June 2005, pp. 164–170.

[11] Y. He and D. Perkins, "Bash: A backhaul-aided seamless handoff scheme forwireless mesh networks," in International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM 2008). IEEE, June 2008, pp. 1–8.

[12] R. Huang, C. Zhang, and Y. Fang, "A mobility management scheme for wireless mesh networks," in IEEE GLOBECOM 2007, Washington DC, USA, November 2007, pp. 5092–5096.

[13] D. Johnson, C. Perkins, and J. Arkko, "Mobility support in ipv6," in RFC 3775, June 2004.

[14] M. Sabeur, G. A. Sukhar, B. Jouaber, D. Zeghlache, and H. Afifi, "Mobile party: A mobility management solution for wireless mesh network," in 3rd IEEE Int. Conf. Wireless and Mobile Comp., Networking, and Commun. (WiMob), October 2007.

[15] S. Speicher and C. H. Cap, "Fast layer 3 handoffs in aodv-based ieee 802.11 wireless mesh networks," in 3rd Int. Symp. Wireless Commun. Syst. (ISWCS), 2006, pp. 233–237.

[16] Y. Amir, C. Danilov, R. Musaloiu-Elefteri, and N. Rivera, "The smesh wireless mesh network," ACM Transactions on Computer Systems, vol. 28, no. 3, September 2010, pp. 6:1–6:49.

[17] D. Gallucci, S. Giordano, D. Puccinelli, N. Tejawsi, and S. Vanini, "Fixed mobile convergence: The quest for seamless mobility," in Fixed/Mobile Convergence Handbook. CRC Press, 2010, pp. 185–196.

[18] S. Vanini, D. Gallucci, S. Giordano, and A. Szwabe, "A delay-aware num-driven framework with terminal-based mobility support for heterogeneous wireless multi-hop networks," in ICTF 2013 Information and Communication Technology Forum, 2013, (in press).

[19] H. Krawczyk, M. Bellare, and R. Canetti, "Hmac: Keyed-hashing for message authentication," RFC 2014, February 1997.

[20] "Ietf diffserv working group page," http://datatracker.ietf.org/wg/diffserv/charter, 2014.

[21] R. Chakravorty, S. Katti, J. Crowcroft, and I. Pratt, "Flow aggregation for enhanced tcp over wide-area wireless," in IEEE Conference on Computers and Communications, 2003, pp. 1754–1764.

[22] [Online]. Available: https://github.com/esnet/iperf (2014)

[23] [Online]. Available: http://www.wireshark.org/docs/man-pages/dumpcap.html [retrieved: May, 2014]

[24] "An ad-hoc wireless mesh routing daemon," http://www.olsr.org, 2014.

 ISBN: 978-1-61208-366-7

Self-organizing Mobile Medium Ad hoc Network

Nada Alsalmi, John DeDourek, Przemyslaw Pochec
Faculty of Computer Science
University of New Brunswick
Fredericton, Canada
e-mail: {nada.alsalmi, dedourek, pochec}@unb.ca

***Abstract*— MANETs are mobile networks of wireless mobile devices capable of communicating with one another without any reliance on a fixed infrastructure. A Mobile Medium Ad hoc Network (M2ANET) is a set of mobile nodes forming a Mobile Medium and functioning as relays for facilitating communication between the users of this Mobile Medium. Movement of the nodes affects the performance of a M2ANET. We propose a scheme for controlling the movement of mobile nodes in a M2ANET based on an attraction/repulsion paradigm. The new node movement has an advantage over a random movement in keeping the nodes in an unbounded region in a sufficient density to allow for an efficient transfer of data over the Mobile Medium. Simulation results show tripling of the delivery ratio in a self-organizing M2ANET compared to a mobile network with all nodes moving randomly, in one experimental scenario.**

Keywords-mobility models; self-organizing mobile network; M2ANET; Mobile Medium; MANET; NS-2; AODV

I. INTRODUCTION

A Mobile Ad Hoc Network (MANET) is a set of mobile devices that cooperate with each other by exchanging messages and forwarding data [1][2]. Mobile devices are linked together through wireless connections without infrastructure and can change locations and reconfigure network connections. During the lifetime of the network, nodes are free to move around within the network and node mobility plays a very important role in mobile ad hoc network performance. Mobility of mobile nodes significantly affects the performance of a MANET [2].

A Mobile Medium Ad Hoc Network (M2ANET) is a particular configuration of a typical MANET proposed in [3], where mobile nodes are divided into two categories: (i) the forwarding only nodes (shown in black in Fig. 1) forming the so called Mobile Medium, and (ii) the communicating nodes (shown in red in Fig. 1), mobile or otherwise, that send data and use this Mobile Medium for communication. The advantage of this M2ANET model is that the performance of such a network is based on how well the Mobile Medium can carry the messages between the communicating nodes and not based on whether all mobile nodes form a fully connected network. An example of a M2ANET is a cloud of drones released over an area of interest facilitating communication in this area. Recently, a number of projects that match the M2ANET model have been announced; they include Google Loon stratospheric balloons [4] and Facebook high altitude solar powered planes [5] for providing Internet services to remote areas, and the Swarming Micro Air Vehicle Network (SMAVNET) project where remote controlled planes are used for create an emergency network [6].

Controlling the movement of all forwarding nodes forming a Mobile Medium is a problem in deploying M2ANETs in real world scenarios like emergency or disaster recovery. While movement of each node is most easily directed independently there is a need for keeping the nodes in relative proximity to maintain their connectivity one with another. In practical terms, the nodes may move on closed paths (e.g., circular), or at random. With randomly selected trajectories maintaining the nodes in one area becomes a problem.

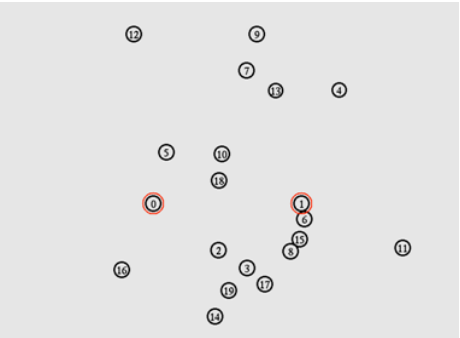


Figure 1. ns2 simulation screen of a M2ANET

The problem is simple to handle in MANET simulation: simulators typically allow setting the simulation area defined as a bounded region which guarantees that the nodes do not disperse any further. If any node tries to move too far away it hits the boundary and then moves in another direction, but still in the same area together with the other nodes. The same cannot be said about the real world scenarios.

In this paper, we propose a solution for controlling the Mobile Medium nodes for M2ANET deployments in an unbounded region. The mechanism is based on an attraction/repulsion paradigm for controlling the movement of mobile nodes in a region without boundaries while providing means for maintaining all nodes in the same area. In principle, when a node moves too far away from other

 ISBN: 978-1-61208-366-7

nodes it should detect the separation and turn back. While the decision making in our simulation is based on the actual distance between the nodes, in a practical deployment the same can be done based on the radio signal strength.

In Section II, we present background on MANETs and mobility patterns. The new movement pattern based on the attraction/repulsion principle for MANETs is discussed in Section III. Simulation experiments of this movement under different scenarios are in Section IV. Finally, we present the experimental results in Section V, followed by the conclusion and future work.

II. STATE OF THE ART

A MANET is comprised of interconnected mobile nodes, which make use of wireless communication links for multi-hop transmission of data. They offer distinct advantages over infrastructure based networks and are versatile for some particular applications and environments. There are no fixed or prerequisite base stations or infrastructures; therefore, their set up is not time consuming and can be done at any time and in any place. MANETs exhibit a fault-resilient nature, given that they are not operating a single point of failure and are very flexible. The deletion and addition of new nodes, forming new links are a normal part of operation of a MANET [1][7][8]. A group of nodes can facilitate communication between distant stations by forming a Mobile Medium, as introduced in [3].

Many mobility models have been proposed for recreating the real world application scenarios of MANETs. A mobility model attempts to mimic the movement of real mobile nodes that change speed and direction with time. There are two main types of mobility models currently used in simulation of MANETs [2][9]: trace and synthetic. A trace uses actual node movements that have been observed in a real system. In the absence of traces, synthetic mobility models can be used. The synthetic models attempt to realistically mimic the movements of mobile nodes in mobile networks [2]. The categorization of synthetic models is based on interactions between the nodes and the environment in a mobile network [2]: we can distinguish between individual node movements and group node movements. Based on specific mobility characteristics these models can be further classified into four categories: models with temporal dependency, models with spatial dependency, models with geographic restriction, and random models [2]. In the mobility model with temporal dependency the movement of a mobile node is affected by its movement history. A node's current movement is affected by past movement such as in the Gauss Markov Model and the Smooth Random Mobility model [2]. In mobility models with spatial dependency, the mobile nodes tend to travel into a group and are interdependent one on another. The movement of a node is affected by surrounding nodes in group mobility such as in the Reference Point Group Model [2]. Another class is the mobility models with geographic restriction. The mobile node movement is limited to certain geographical areas such as streets or freeways as for example in the Pathway Mobility Model and the Obstacle Mobility Model [2].

In simulation, a random mobility is often used as a reference case scenario, mostly because of the relative ease of implementing it in a simulator. One of these popular models is the Random Way Point (RWP) model available in ns2 [10]. Nodes are moved in a piecewise linear fashion, with each linear segment pointing to a randomly selected destination and the node moving at a constant, but randomly selected speed.

III. ATTRACTION/REPULSION MOVEMENT

One of the most incredible sights in nature happens when animals form a group and move together in a flock. How exactly do these individuals do it? A group, such as a herd of land animals or flock of birds, consists of individuals but exhibits some characteristics of team collaboration in the population. While it seems that the group is under a centralized control, in reality what is observed is an aggregated behavioral performance of independent individuals, each of which is acting on the basis of its own local perception [11].

Similar principles can be applied to controlling node movement in our self-organizing M2ANET. The objective of the proposed approach is to control the collective movement of locally interacting nodes similar to the behavior observed in flocks of birds or swarms of insects. Our goal is to keep randomly moving nodes (similar to RWP model) in a limited area without imposing a hard constraint of an external boundary. Our approach is based on an attraction principle to keep the nodes together in a flock (we use the name "flock" or a "cloud" when referring to a number of mobile nodes moving together) and on a repulsion mechanism to keep them sufficiently far apart so that they cover a large area. Though the actual simulation we conducted is based on the distance calculation, in practice the attraction/repulsion principles can be implemented based on the received signal strength at each node.

A. Attraction

The main deficiency of the RWP model for controlling the movement of nodes in a MANET is that, aside from the border effect [12], the nodes tend to fill the entire available space. If there is a boundary limiting the node movement, like in the case of most simulation environments, ns2 included, the nodes tend to disperse approximately evenly resulting in the node density and the average distance between the nodes determined by the available area and the number of nodes in the network. The situation becomes worse in an environment with no boundaries where nodes would disperse completely and lose any connectivity over time.

 ISBN: 978-1-61208-366-7

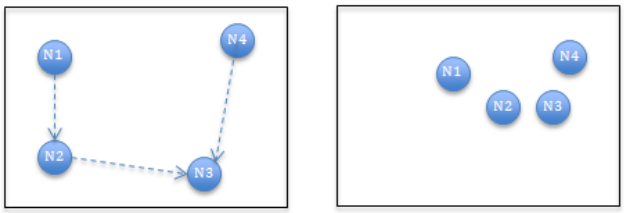


Figure 2. Attraction keeps nodes in a flock.

Attraction between the nodes, when used in addition to the RWP model, can remedy this problem. In our proposed approach, nodes normally move following the RWP model, but when the distance to the nearest neighbor becomes too large they turn towards the nearest neighbor (Fig. 2) rather than choosing a random direction.

B. Repulsion

While the attraction mechanism would be sufficient for a set of randomly moving nodes to form a flock (or a cloud) and remain connected and stay over a limited area without imposition of a hard boundary, the network coverage could be improved with an added mechanism, also based on watching the distance to the nearest neighbor. The coverage of a M2ANET is where the Mobile Medium nodes are, so keeping the nodes apart assures a larger area of coverage by preventing the nodes from congregating in only one place.

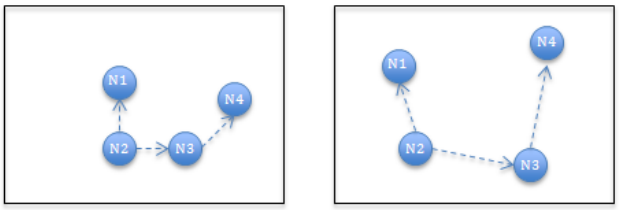


Figure 3. Repulsion prevents the nodes from collapsing into one point.

In our proposed approach, nodes normally move following the RWP model, but when the distance to the nearest neighbor becomes too small they move away from the nearest neighbor (Fig. 3) rather than choosing a random direction.

C. Implementation

Nodes normally follow RWP model movement pattern, with the next move direction determined by parameters stored locally at each node. Attraction and repulsion mechanisms can be implemented based on the received signal strength at each node. We could assume that each node periodically sends a beacon signal (possibly as a part of functioning routing mechanism like in the Destination-Sequenced Distance Routing (DSDV) protocol [13]). The received signal strength determines the identity, and possibly the direction towards, the nearest neighbor. Alternatively, the direction towards the nearest neighbor could be determined by querying the nearest neighbour for the location information (assuming it has a Global Positioning System (GPS), or similar, built in).

In ns2 simulation, nodes move piecewise linearly with each movement of a node specified with the *setdest* command [10]. In our simulation experiments we use the distance between the current node and its nearest neighbor D and define two thresholds: Th_1 to mark when nodes are too far apart, and Th_2 when nodes are to close. The next move is specified:

i. *towards* the nearest node, when $D > Th_1$,
ii. *away* from the nearest node, when $D < Th_2$, and
iii. in a *random* direction, when $Th_1 > D > Th_2$.

The distance covered is chosen randomly (in cases (i) and (ii), uniform distribution $U(0,D)$), but within the bounds of the simulated area.

D. Simulation environment

Each simulation of a network consists of a different number of nodes roaming in a square 1000 x 1000 meters with a reflecting boundary. The transmission range is 250m. The link data rate is 1 Mbps. Every packet has a size of 512 bytes. The buffer size at each node is 50 packets. Data packets are generated following a Constant Bit Rate (CBR) process [10]. The source and destination nodes are stationary and located at coordinates (300, 500) and (700, 500). The summary of the simulation parameters used in ns2 is shown in Table 1.

TABLE I. SIMULATION PARAMETERS

Parameters	
Simulator	NS-2.34
Channel Type	Channel / Wireless Channel
Network Interface Type	Phy/WirelessPhy
Mac Type	Mac/802.11
Radio-Propagation Type	Propagation/Two-ray ground
Interface Queue Type	Queue/Drop Tail
Link Layer Type	LL
Antenna	Antenna/Omni Antenna
Maximum Packet in ifq	50
Area (n * n)	1000 x 1000
Source Type	(UDP) CBR
Simulation Time	900 sec
Routing Protocol	AODV

The forwarding nodes are mobile and move according to the attraction/repulsion algorithm. In each experiment, the designated source node transmits to one designated destination node for 900 seconds.

V. RESULTS

Four sets of simulation experiments were conducted: one set with all forwarding Mobile Medium nodes moving randomly, and three sets with the forwarding nodes moving based on the attraction/repulsion principle using three

 ISBN: 978-1-61208-366-7

different threshold levels:

i. Low threshold: $Th_1 = 60, Th_2 = 30,$
ii. Medium threshold: $Th_1 = 120, Th_2 = 60,$
iii. High threshold: $Th_1 = 200, Th_2 = 120.$

In each experiment, data regarding the node location and the delivery ratio were collected.

A. *Node movement behavior*

Topologically, the purpose of the attraction/repulsion mechanism is to keep the nodes together while allowing them to move independently. To measure the togetherness of the nodes we collected samples of node coordinates (every 10s) over the duration of each experiment and calculated the standard deviation of all X coordinates, for all the samples.

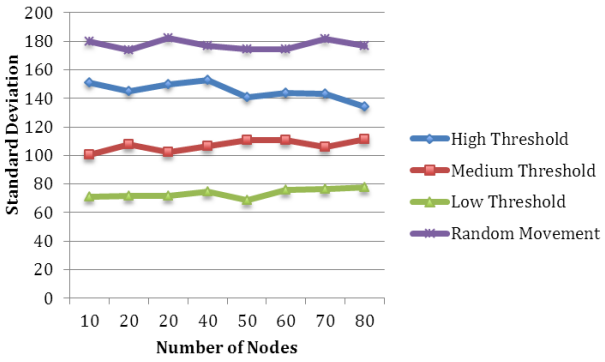


Figure 4. Node location standard deviation: X axis

Fig. 4 shows a measure (standard deviation of *X* coordinates of all mobile nodes, sampled every 10 seconds) of the spread of all the nodes in four sets of experiments. The results show that the lower the threshold the tighter the flock (cloud) formed by the mobile nodes. Also, the nodes of the proposed self-organizing M2ANET stayed closer together than they normally would if all the nodes just moved randomly over the 1000 by 1000 m simulation area.

B. *Network delivery ratio*

The main goal of a self-organizing M2ANET is to avoid node dispersion and to provide enhanced communication over the area covered by the Mobile Medium (forwarding nodes). Fig. 5 shows the comparisin between the delivery ratios in a self-organizing M2ANET versus a M2ANET with Mobile Medium nodes moving randomly over the entire simulation area. The graph shows that decreasing the treshold values and thus keeping the Mobile Medium nodes closer together improves the delivery ratio. In our experiments, all self-organizing networks do better than a network with nodes moving totally randomly. The improvement is most signifincat for experiments with small number of nodes: in a M2ANET with only 10 nodes in an area 1000 by 1000 m the delivery ratio of 9% for a random movement scenario was improved threefold to almost 30% in a self-organizing M2ANET when a low threshold settings of $Th_1 = 60, Th_2 = 30$ were used.

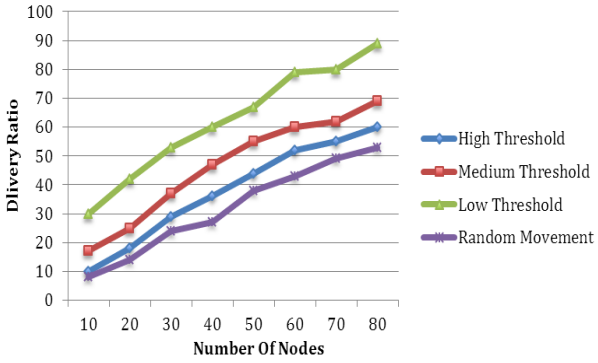


Figure 5. Delivery ratio.

The improved performace is due to keeping the nodes closer together (Fig. 1), which increases a likelihood of forming a route from the source to the destination.

VI. Conclusion and Future Work

In this paper, we proposed a control paradigm for a self-organizing MANET network. The approach is particularly attractive for M2ANETs where the goal is to create a Mobile Medium out of mobile forwarding nodes, and use this Mobile Medium to facilitate data communication between other users.

The new mobility control mechanism is based on an attraction/repulsion principle: the Mobile Medium nodes normally move randomly, but they turn back when they get too far from their neighbors. This mechanism keeps all the nodes in a "flock", with the flock (or cloud) density controlled by two thresholds, and thus allowing the M2ANET creator to control the performance of the Mobile Medium: the lower the attraction/repulsion thresholds the closer the nodes of the Mobile Medium remain and the higher the delivery ration of the resulting M2ANET network.

Based on our results, we suggest further testing self-organizing M2ANET networks using different routing algorithms. Also the role of the lower threshold Th_2 needs to be investigated: it is not clear which protocols might benefit form maintaining the minimum distance between the mobile nodes.

ACKNOWLEDGMENT

This work is sponsored and funded by the Ministry of Higher Education of Saudi Arabia through the Saudi Arabian Cultural Bureau in Canada.

References

[1] S. Basagni, M. Conti, S. Giordano, and I. Stojmenovic (Eds.), Mobile

Ad Hoc Networking. New York: Wiley-IEEE Press. 2001.
[2] F. Bei and A. Helmy, A survey of mobility models in wireless Ad hoc Networks, University of California, USA,2004.
[3] J. DeDourek and P. Pochec, "M2ANET: a Mobile Medium Ad Hoc Network", Wireless Sensor Networks: Theory and Practice, WSN 2011, Paris, France, Feb. 2011, pp. 1-4.
[4] H. Hodson, "Google's Project Loon to float the internet on balloons", New Scientist, October 2013.
[5] J. Brustein, "Facebook's Flying Internet Service, Brought to You by Drones", Bloomberg Businessweek, March 4, 2014.
[6] A. Jimenez Pacheco, et al., "Implementation of a Wireless Mesh Network of Ultra Light MAVs with Dynamic Routing", IEEE GLOBECOM 2012, 3rd International IEEE Workshop on Wireless Networking & Control for Unmanned Autonomous Vehicles 2012, Anaheim, California, USA, 2012.
[7] D. P. Agrawal and Q. A. Zeng, Introduction to Wireless and Mobile Systems, Thomson Engineering, 2010.
[8] S. K. Sarkar, T. G. Basavaraju, and C. Puttamadappa, Ad Hoc Mobile Wireless Networks, Principles, Protocols, and Applications, Auerbach Publications Taylor & Francis Group, 2007.
[9] N. Aschenbruck, E. G. Padilla, and P. Martini, "A survey on mobility models for performance analysis in tactical mobile networks", Journal of Telecommunications and Information Technology, vol. 2, 2008, pp. 54-61.
[10] H. Ekram and T. Issariyakul, Introduction to Network Simulator NS2, Springer, 2009.
[11] M. Dorigo, V. Maniezzo, and A. Colorni, "The Ant System: Optimization by a colony of cooperating agents", Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 26 (1), 1996, pp. 29-41.
[12] R. Alghamdi, J. DeDourek, and P. Pochec, "Avoiding Border Effect in Mobile Network Simulation", The Twelfth International Conference on Networks ICN 2013, Seville, Spain, Jan 27 - Feb 1, 2013, pp. 184-189,
[13] C. E. Perkins and P. Bhagwat, "Highly dynamic destination-sequenced distance-vector routing (DSDV) for mobile computers", ACM SIGCOMM Computer Communication Review, Oxford University Press, 24(4), 1994, pp. 234-244.

 ISBN: 978-1-61208-366-7

COmpAsS: A Context-Aware, User-Oriented Radio Access Technology Selection Mechanism in Heterogeneous Wireless Networks

Sokratis Barmpounakis,
Panagiotis Spapis, Nancy Alonistioti
Dept. of Informatics and Telecommunications
University of Athens, Greece
{sokbar,pspapis,nancy}@di.uoa.gr

Alexandros Kaloxylos
Department of Informatics and Telecommunications,
University of Peloponnese, Greece
kaloxyl@uop.gr

***Abstract*—5G networks will have to cope with an increase of data traffic, as well as a vast number of devices, which already transpires in the wireless/mobile communication environments. Several on-going efforts both from 3rd Generation Partnership Project (3GPP) and several proposals from the literature as well, attempt to overcome the existing barriers by enabling the use of Wi-Fis and femto-cells. The evolution of Access Network Discovery and Selection function (ANDSF) in Evolved Packet Core (EPC) networks, as well as the Hotspot 2.0 approach, can be used to facilitate a seamless integration of WiFis with the cellular networks. Although this integration clearly presents benefits, a handover mechanism that will capitalize on the new standards is still missing. This paper acts in a two-fold way. We design and evaluate a novel context aware selection mechanism that is using fuzzy logic to select the most appropriate Radio Access Technology (RAT). To this end, we propose network extensions that allow the ANDSF entity to be aware and provide up to date information to end devices about the network status. Extensive simulation results illustrate the advantages of our approach.**

Keywords-RAT selection;ANDSF;LTE;handover.

I. INTRODUCTION

Traffic analysis clearly indicates that 5G networks will have to cope with a huge increase of data traffic and the number of the end devices (e.g., smartphones, tablets, sensors etc). To address this issue the research community designs solutions to improve the spectral efficiency, to increase the network cell density and to exploit the underutilized radio spectrum resources [1]. Such approaches suggest the exploitation of the available femto-cells or Wi-Fi Access Points (APs) to reduce the network load of an operator in a particular area [2].

Integrating Wi-Fi access points with cellular networks has been a hot topic for over a decade. However, apart from limited deployment examples, this approach has not been widely adopted by the network operators. This is because of a number of reasons. Wi-Fi suffers from interference issues since it operates on the unlicensed spectrum. Typically, the installed access points in homes, offices, public spots do not belong to the cellular operator. Also, up to now, switching from a cellular network to a Wi-Fi access point was not a transparent process for the end users (e.g., authentication). Finally, there was not a clear business case for the operators on how to increase their revenues by supporting Wi-Fi access points.

Some new technological solutions may change the landscape. ANDSF and Hotspot 2.0 if combined together may prove the right solution for simplifying the access of end users among RATs. Also, roaming among cellular operators and wireless internet service providers may also be supported. The new business case for cellular operators would be the support of the same QoE for their services among different RATs that may even belong to another operator. Thus, the integration of cellular networks with Wi-Fi APs needs to be revised not only due to the new business cases that arise, but also because the new protocols can be exploited to design more efficient RAT selection mechanisms.

3GPP has already specified how Wi-Fi access points may be integrated with the Evolved Packet Core (EPC) architecture [3][4]. Also, a new network entity, which takes account of policy rules and security requirements was introduced, namely Access Network Discovery and Selection Function (ANDSF) [5]. Closely coupled with the Policy and Charging Rules Function (PCRF) [6], ANDSF implements dynamic data offload for the User Equipment (UE) in a structured method. The ANDSF is a cellular technology standard, which enables the operator to store its policies for discovery and selection of RATs on a server. The UEs are updated with these polices either via push (network-initiated information to the UE) or pull (UE-initiated request) methods by the server. The policies within ANDSF contain information on which of the available Wi-Fi hotspots are preferable during specific a specific time or day, and at a specific location as well, based on indications from past measurements.

The ANDSF information is represented by the ANDSF Management Object (MO) and may contain information with regard to the UE location, Inter-System Mobility Policies (ISMPs) and Inter-System Routing Policies (ISRPs) ([7]). The ISRPs are available for UEs, which support IP Flow Mobility (IFOM), multiple-access Packet Data Network (PDN) connectivity (MAPCON), or non-seamless offload [8] - [10]. MAPCON enabled UEs may establish different PDN connections through different RATs. IFOM enabled terminals may establish a single PDN connection via multiple access networks, for instance 3G/LTE and Wireless

 ISBN: 978-1-61208-366-7

Local Area Network (WLAN). For such UEs, IFOM enables to move individual IP flows from one access network to another with session continuity. The ANDSF prioritized rules in the case of MAPCON apply per PDN connections, while in IFOM and non-seamless offload cases per flow. ANDSF communicates with the UE over the S14 reference point.

Hotspot 2.0 Wi-Fi technology standard from Wi-Fi Alliance acts in a complementary way to ANDSF as it improves the ability of WLAN devices to discover and connect in a secure way to public Wi-Fi APs. Hotspot 2.0 builds on 802.11u specifications that enable devices to discover information about the available roaming partners using query mechanisms. The query and response protocol, which supports Hotspot 2.0, is the Access Network Query Protocol (ANQP) [11]. ANQP is used to collect the following: the operator's domain name, the accessible roaming partners, the IP address type availability, the type of the access point (private, public free, public chargeable, etc.), and most significantly load information (i.e., total number of currently associated devices to the AP, channel utilization percentage and an estimate of the remaining available admission capacity).

The WLAN_NS working item of 3GPP ([12]) is working to Enhance 3GPP solutions for WLAN and access network selection based on Hotspot 2.0 and ensure that data, i.e., Management Objects (MO) and policies provided via HotSpot 2.0 and ANDSF are consistent. This alignment of ANDSF and HotSpot 2.0 provides an excellent basis for the complementarity of ANDSF and Hotspot 2.0, as well a number of multi-operator scenarios that can be supported. In [2], a rather exhaustive list of possible scenarios is presented.

From the above description it is clear that several efforts have already taken place to address the interworking between cellular networks and WiFi. In the new landscape it is imperative to design new mechanisms for the RAT selection for every terminal. The reason is that UEs will have to choose among typical macro-cells, femto-cells and APs. Due to the diverse set of parameters that have to be evaluated by a UE and the network we adopt the use of fuzzy logic [13] that can handle multi-criteria problems.

The rest of the paper is organized as follows. In Section II, we present related work from the literature, which attempts to deal with the aforementioned challenges. Section III is split into two main parts: the first presents a proposed extension of the ANDSF entity to collect information from HeNBs and APs to support the RAT selection process; the second part goes through a comprehensive description of our mechanism, which we call *COmpAsS*. In Section IV, simulation results based on a realistic business case are presented. In Section V, we describe the conclusions, which are derived from the overall work and we discuss our future steps.

II. RELATED WORK

There has been a lot of effort into further optimizing the standardized mechanisms, and plenty of proposals and algorithmic solutions to improve the handover procedure.

The survey in [12] provides an overview of the main handover (HO) decision criteria in the current literature and presents a classification of existing HO decision algorithms for femto-cells. According to this, some researchers focus on evaluating the Reference Signal Received Power (RSRP), the user location or speed, the mobility patterns, the battery level, the mean UE transmit power and the UE power consumption, the load of the cell and the service type. Apart from the case of RSRP, typically researchers are using multiple criteria (e.g., battery lifetime, traffic type, cell load, speed) and are using different tools (e.g., cost based functions, fuzzy logic, etc.) to reach a decision.

Xenakis et al. [14] present an overview of the vertical handover (VHO). Initially, a categorization of the information parameters of the VHO processes into layers is made: application (e.g., user preferences), transport (e.g., network load), network (e.g., network configuration, topology), data-link (e.g., link status) and physical (i.e., available access media). From the network perspective the ones highlighted are: latency, coverage, RSS, RTT, number of retransmissions, BER, SINR, packet loss, throughput, bandwidth, network jitter and the number of connected users. From the UE perspective, the parameters that are presented are user monetary budget, preferred network (user choice), location, movement (change of direction), velocity, technologies available in the device, as well as battery consumption. Many of the proposed mechanisms that this survey presents attempt to create an overall context-aware mechanism, by combining several of the aforementioned parameters for the VHO decision outcome.

Several other existing surveys attempt to present a unifying perspective with regard to HO mechanisms. Rao et al. [15] deal with the network selection concept as a perspective approach to the always best connected and served paradigm in heterogeneous wireless environment. From the origin point of view, they classify them in four categories: network-related criteria, terminal-related, service-related and finally, user-related. In addition, in [16]-[18], several efforts are described, which aim to improve the selection mechanisms, which support heterogeneous RATs. In principle, all mechanisms combine parameters like RSS, bandwidth, mobility, power consumption of the UE, security, monetary cost and user preferences.

In all the above a cases, the researchers are using for the most advanced schemes a number of parameters. However, very rarely they clearly state how this information is collected and from which network entities. Such information is necessary because the hypothesis that a value (e.g., the location of terminal) can be collected may require extensive signaling exchange among the network components. Also, in most cases solutions target either handovers for macro-femto cells or vertical handovers among different RATs. In this paper, we attempt to clearly indicate how the information required for our solution is collected and from which network entities. We also examine the possibility of UE to handover among macro-femto and Wi-Fi APs.

When dealing with diverse parameters in order to reach a decision, in the literature many authors have proposed Fuzzy Logic (FL) Inference Systems. Indicatively, Xia et al. [19]

 ISBN: 978-1-61208-366-7

propose a scheme taking into consideration the actual RSS, as well a predicted RSS, and they combine it with the speed of the UE in order to determine if a handover should be made or not. Moreover, they estimate the suitability of a RAT for handover, taking as input the current RSS, the estimated RSS, as well as the available bandwidth. In [20], FL is also used for estimating the output suitability of a network based on the inputs of the environment (bandwidth, delay, charging, power consumption). In addition, Ma and Liao use GPS, in order to adapt the monitoring rate of the afore-mentioned values. For our solution we have also chosen to use a FL scheme to support the decision making process.

III. THE PROPOSED SOLUTION

A. COmpAsS mechanism

The aim of the COmpAsS is to enable a UE at selecting in an intelligent way the most suitable RAT to perform a per-flow handover. COmpAsS is a user-oriented, context-aware scheme, which takes into account the mobility of the UE, the Received Signal Strength (e.g., RSRQ for 3GPP access networks), the load of the (Home) eNodeBs ((H)eNBs) and WLAN APs, the backhaul load of the network, as well the sensitivity to latency for each of the candidate flows for handover (Fig. 1). Based on FL, the five inputs are assessed using a Fuzzy Inference process, which resides in the UE and calculates the suitability of the available RATs for each one of the flows of the UE. The calculation inside the FL Inference Engine is based on pre-defined rules regarding all the possible combinations of the different inputs. According to the rules, in principle, it is assumed that a RAT is more attractive to the UE when it is characterized by low (backhaul) load and high RSS. In addition, the higher the sensitivity to latencies, the more important is the mobility of the UE; high mobile UEs prefer larger cells to avoid unnecessary handovers. In the proposed scheme, the information is obtained from an extended ANDSF network entity, which is described in detail in the following section.

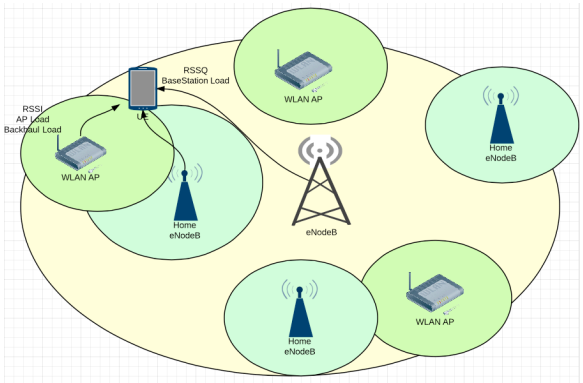


Figure 1. Context-aware RAT selection by COmpAsS

Although the FL computational requirements are minimum, in order to further optimize the energy consumption of COmpAsS inside the UE, as well as to minimize the unnecessary handovers, additional mechanisms are used (Fig. 2), i.e., a) a suitability threshold: no FL computation is performed if the current RAT's suitability is higher than 90%, b) a suitability hysteresis value, i.e.: neighbor RAT's suitability must be at least 10% higher than the current RAT's (if a neighbor RAT is a macro cell) or at least 1% higher than the current RAT (if neighbor RAT is a femto-cell) in order to trigger a handover. The higher hysteresis in the case of macro neighbor RAT is chosen aiming to impel the handover to smaller RATs for offloading reasons.

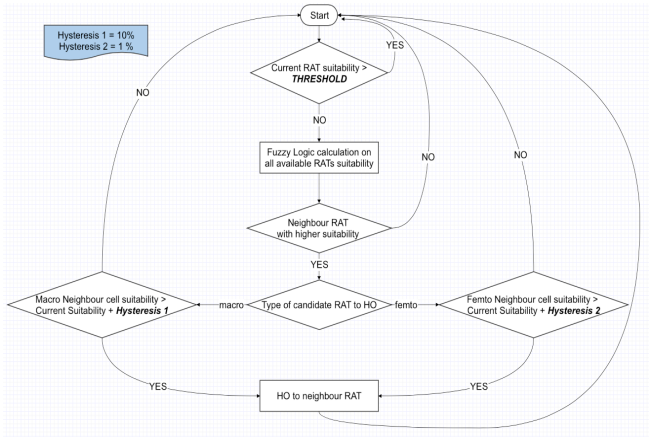


Figure 2. RAT suitability Hysteresis and Margin for minimizing unnecessary handovers

B. Extension of Access Network Discovery and Selection Function (ANDSF) functionality

As described earlier in this paper, ANDSF is a cellular technology standard, which implements dynamic data offloading for the UEs in a structured way. However, the purpose of ANDSF is currently limited to provide the UE with policies with regard to access networks. Moreover, one of the most crucial aspects in relation to offloading and handover mechanisms, that the ANDSF MO is missing, is real-time network conditions, such as the load of a Base Station. This type of information, as well as additional features, which are not provided by the ANDSF, may be provided by the Hotspot 2.0 standard described earlier, supported by the ANQP protocol.

On the contrary, ANDSF provides WLAN AP location information, supports UE location reporting, as well as may provide a list of preferred or restricted access networks, -features, which are not provided by Hotspot 2.0 -.

It becomes clear that ANDSF and Hotspot 2.0 could act in a supplementary way to maximize the available information to the UE, resulting in more efficient offloading mechanisms. In this paper, we propose an enhanced version of the ANDSF server capable of:

a) collecting real-time load information regarding the available 3GPP access networks, based on a new logical interface (e.g., between the (H)eNB and the ANDSF entity). This information is evaluated in a coarse manner (i.e., low, medium, high).

b) supporting queries to Hotspot 2.0 enabled WLAN APs using the ANQP protocol

 ISBN: 978-1-61208-366-7

c) gathering information from the UE measurements regarding RSRQ measurements

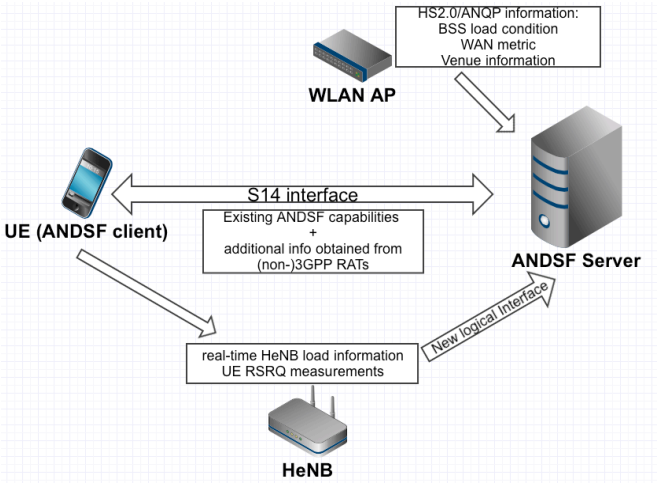


Figure 3. Extended ANDSF architecture

As a result, the UE will be capable of assessing both 3GPP and non-3GPP available RATs using the same input parameters and ultimately take the optimal decision for handover. S14 existing interface between the UE and the ANDSF component will provide to the UE already-supported information, as well as the additional information obtained from the available (non-) 3GPP RATs. A high-level description of the above architecture is demonstrated in Fig.3.

IV. SIMULATION RESULTS

In order to evaluate the performance of COmpAsS mechanism advanced topology simulations were carried out using the *ns-3* simulator [21]. The *fuzzylite* C++ Fuzzy Logic library is also integrated inside the custom *NS-3.19* build. The figure, which follows, presents a realistic business case scenario of a shopping mall comprising 3 floors (ground floor, 1st and 2nd floor), and 20 shops per floor (Fig. 4). The UEs are either static or moving, and are roaming around the shopping mall rooms (shops, cafes, etc.). Several HeNBs are deployed in the three floors. In addition, two macro cells (eNBs) exist outside the mall area in a distance of 200m to different directions. Due to the fact that COmpAsS handles Wi-Fi APs and HeNBs in a similar way, with regard to the pre-defined rules of the Fuzzy Inference Engine, for the sake of simplicity, in the simulations only macro and femto-cells are deployed.

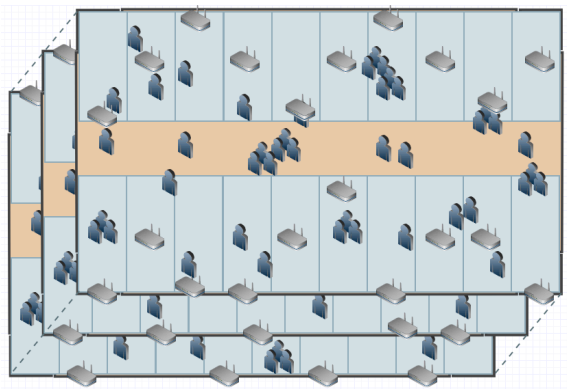


Figure 4. Shopping Mall with 3 floors and 20 shops per floor (simulation environment)

Besides the several UEs, which are roaming inside the mall area and creating respective traffic to the HeNBs, we use one "test UE", in which COmpAsS is deployed. Different simulations were carried out to test the UE at different velocities (low, medium, high), in each one of the scenarios in order to evaluate the proposed scheme for varying UE mobility, as mobility is one of the inputs, which are taken into consideration for the decision. The test UE is moving with linear velocity between the rows of the shops, on the 1st floor. An overview of the simulation details is presented in the following table:

Table 1 SIMULATION DETAILS.

Environment	Shopping mall: 3 floors, 100 x 200 meters per floor, 20 rooms per floor (2 rows of 10 equal rooms)
Number of UEs	Variable (UEs connecting/disconnecting)
Number of (H)eNBs	2 eNBs, 9 HeNBs
Carrier frequency (MHz)	Downlink: 2120.0, Uplink: 1930
Channel bandwidth	50 RBs for eNBs, 15 RBs for HeNBs
Transmit power	35.0 dBm (eNBs) , 23.0 dBm (HeNBs)
Simulation time	100 s
Time unit	0.1 s
UE mobility	0.4 m/s, 0.8 ms, 1.4 m/s (linear constant velocity)
HeNB load	Varying depending on the number of associated UEs (very low, low, medium, high, very high)
Traffic sensitivity to latency	High (0.7/1.0)

The proposed scheme is evaluated against A2A4 RSRQ mechanism –a well-established handover algorithm found often in the literature-. A2-A4-RSRQ may be triggered by the two events; Event A2 is defined as the situation during the serving cell's RSRQ becomes worse than a *threshold*. A4 event describes the situation when a neighbor cell's RSRQ becomes better than a *threshold*.

The following figures illustrate the measured Key Performance Indicators (KPIs), which resulted from the two mechanisms with regard to the number of overall handovers which took place during the simulation, the throughput of the test UE, the experienced delays, as well as the packet loss during the measurements.

Variable load of the femto-cells of the shopping mall was tested, calculated in relation to the overall associated users per base station and traffic that is generated. In particular, the load of the base stations varies from 10% up to 90% of their available resources (horizontal axis in Fig. 5-11).

 ISBN: 978-1-61208-366-7

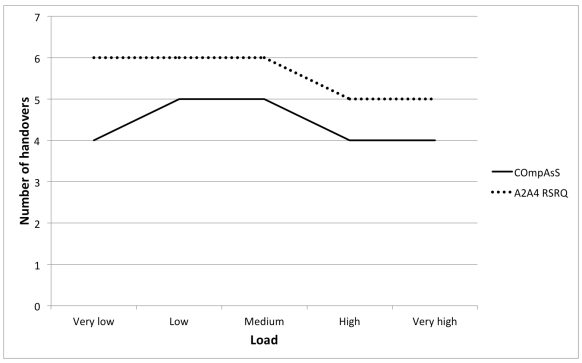


Figure 5. Number of handovers

In Fig. 5, the overall number of handovers is shown. According to the graph, the proposed mechanism tends to minimize the number of handovers as it realizes less handovers than A2A4 RSRQ in all load situations.

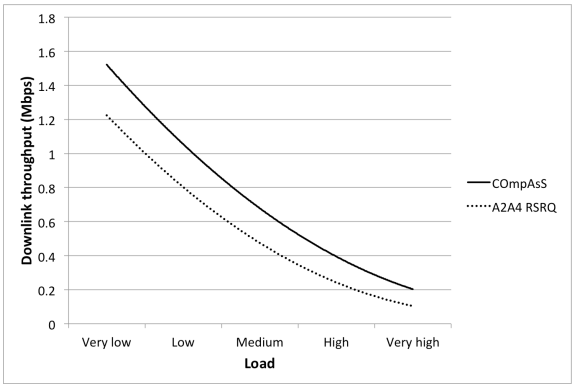


Figure 6. Downlink throughput

In Figs. 6-8, the results of the downlink are illustrated: throughput, delay and packet loss. With regard to the throughput (Fig. 6), COmpAsS outperforms the A2A4 RSRQ algorithm in all load scenarios by 10-20 %. In the case of the proposed scheme, the high interference, which results from the tested environment retains the UE from handing over to the femto-cells, which suffer more; instead, the UE tends to stay more time attached to the eNBs, achieving finally a higher throughput. Moreover, the UE mobility is taken into consideration from COmpAsS, in contrast to A2A4 RSRQ; for high mobile users femto-cells are less attractive, particularly if the load of them increases as well, which makes them even more unattractive. In the case of the delay (Fig. 7), a significant difference between the two mechanisms is observed throughout the measurements. Similarly, the packet loss (Fig. 8) that experiences the UE, which uses the COmpAsS mechanism, is by 20% lower than the other scheme, no matter how high the load of the network –and as a result the experienced interference as well- is.

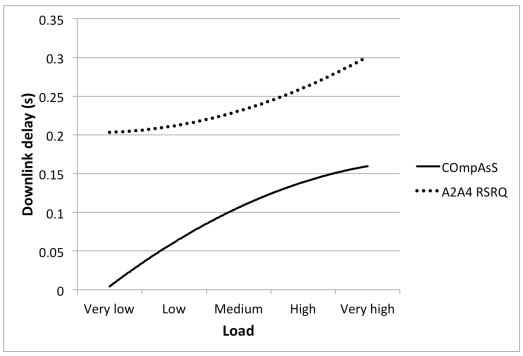


Figure 7. Downlink delay

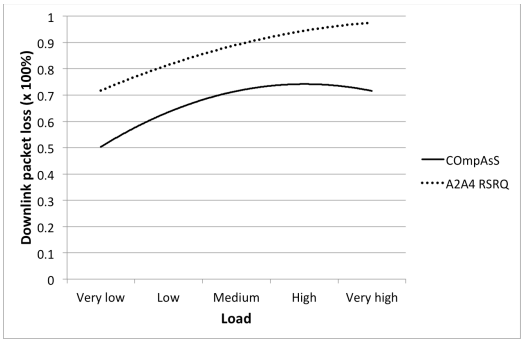


Figure 8. Downlink packet-loss

Figs. 9-11 illustrate the measured KPIs of the uplink. Noticeably, the difference of the throughputs of the two schemes is even higher than in the case of the downlink, i.e., 200 – 400 Kbps (Fig. 9).

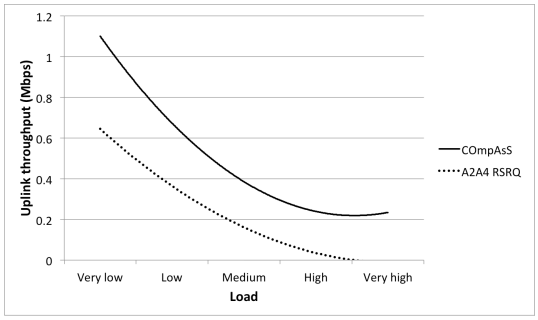


Figure 9. Uplink throughput

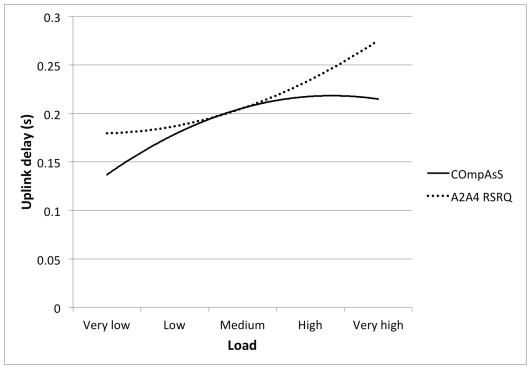


Figure 10. Uplink delay

With regard to the uplink delay (Fig. 10), it is shown that, although at medium load the two algorithms have almost identical results, as the load increases further, COmpAsS's performance is significantly better –roughly 50ms-, maintaining constant delay. In contrast, A2A4 RSRQ's delay is increasing further. This is explained by the fact that, the suitability by COmpAsS during the load increase of the femto RATs, reduces radically, particularly for faster users.

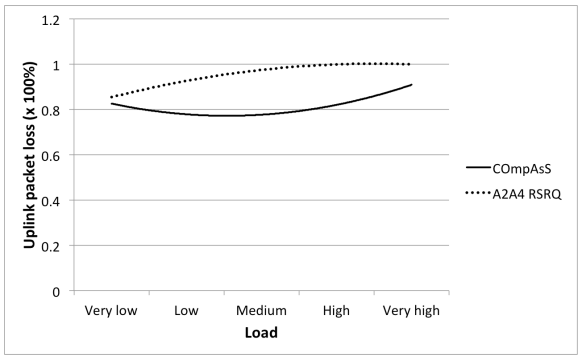


Figure 11. Uplink packet-loss

The packet-loss in the uplink case (Fig.11), similarly with the previous figures confirms the superior performance of the proposed mechanism.

V. CONCLUSION AND FUTURE STEPS

This paper proposed COmpAsS, a context-aware RAT selection mechanism, based on Fuzzy-Logic. The proposed solution emphasizes on the actual way of obtaining the different types of information, which ultimately lead to the handover decision, via an extension of the current solutions such as ANDSF and Hotspot 2.0. The realistic business case scenario, which was simulated, and the extensive results confirm the high performance of COmpAsS in challenging environments of several mobile users and different co-existing RATs, while at the same prove that it can be broadly applicable, in simpler, less demanding use cases as well.

The proposed mechanism, on the one hand avoids the unnecessary handovers minimizing the redundant signaling overhead; on the other hand, the context awareness of the UE remarkably improves the handover decisions resulting at the end in higher service quality and -eventually- higher quality of experience for the end-user.

Future steps will be: (a) define an adaptive sampling rate of the mechanism, in order to further optimize the battery consumption of the UE and minimize the unnecessary signaling, and (b) carry out more simulation scenarios with more users and diverse service types.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Commission's 7th Framework program FP7-ICT-2012 under grant agreement Nº 317669 also referred to as METIS (Mobile and Wireless Communications Enablers for the 2020 Information Society).

REFERENCES

[1] J. Wannstrom, "LTE-Advanced (2012)". May 10, 2012, http://www.3gpp.org/IMG/pdf/lte_advanced_v2.pdf [retrieved: April, 2014].

[2] BT & Alcatel Lucent White paper, Wi-Fi Roaming building on ANDSF and HOTSPOT 2.0, October 2012.

[3] 3GPP TS 23.401, V12.1.0, "GPRS enhancements for E-UTRAN access, (Release 12)", June 2013.

[4] 3GPP TS 23.402, "Architecture enhancements for non-3GPP accesses, (Release 12)", June 2013.

[5] 3GPP TS 24.312, "Access Network Discovery and Selection Function (ANDSF) Management Objects (MO), (release 12)", June 2013.

[6] 3GPP TS 23.203 V12.2.0, "Policy and charging control architecture", September 2013.

[7] Wi-Fi Roaming Building on ANDSF and Hotspot 2.0", Alcatel Lucent – BT White Paper .

[8] 3GPP 24.237 V10.1.0, "Mobility between 3GPP Wireless Local Area Network (WLAN) interworking (I-WLAN) and 3GPP systems", March 2013.

[9] 3GPP TS 24.302 V10.5.0, "Access to the 3GPP Evolved Packet Core (EPC) via non-3GPP access networks", Release 10, September 2011.

[10] 3GPP TS 23.261 V10.0.0, "IP Flow Mobility and Seamless Wireless Local Area Network (WLAN) offload", Release 10, June 2010.

[11] IEEE 802.11u-2011 IEEE Standard for Information Technology-Telecommunications and information exchange between systems-Local and Metropolitan networks-specific requirements - Amendment 9: Interworking with External Networks.

[12] 3GPP, TR 23.865 "Study on Wireless Local Area Network (WLAN) network selection for 3GPP terminals; Stage 2" September 2013.

[13] T. J. Ross, "Fuzzy Logic with Engineering Applications", 2nd Edition, October 2004.

[14] D. Xenakis, N. Passas, L. Merakos, and C. Verikoukis, "Mobility Management for Femtocells in LTE-Advanced: Key Aspects and Survey of Handover Decision Algorithms", Communications Surveys & Tutorials, IEEE, vol. 16, iss. 1, 1st Quarter 2014, pp. 64-91.

[15] K. R. Rao, Z. S. Bojkovic, and B. M. Bakmaz, "Network Selection in Heterogeneous Environment: A Step toward Always Best Connected and Served", Telsiks 2013, October 2013, pp 83-92.

[16] A. Ahmed, L. Merghem Boulahia, and D. Gaıti, "Enabling Vertical Handover Decisions in Heterogeneous Wireless Networks: A State-of-the-Art and a Classification" , IEEE Communications Surveys and Tutorials, accepted for publication. Available: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6587998 [retrieved April, 2014].

[17] P. Bellavista, A. Corradi, and C. Gianneli, "A Unifying Perspective on Context-Aware Evaluation and Management of Heterogeneous Wireless Connectivity", IEEE Communications Surveys and Tutorials, vol. 13, No. 3, 3rd quarter 2011, pp. 337 - 357.

[18] X. Yan, Y.A. Sekercioglu, and S. Narayanan, "A survey of vertical handover decision algorithms in Fourth Generation heterogeneous wireless networks", Computer Networks vol. 54, 2010, pp. 1848-1863.

[19] L. Xia, L Jiang, and C. He, "A novel fuzzy logic vertical handoff algorithm with aid of differential prediction and pre-decision method", IEEE ICC, June 2007, pp. 5665 - 5670.

[20] B. Ma and X. Liao, "Vertical Handoff Algorithm Based on Type-2 Fuzzy Logic in Heterogeneous Networks", Journal of Software, vol. 8, No 11, November 2013, pp. 2936-2942.

[21] Ns-3 simulator, http://www.nsnam.org/overview/what-is-ns-3/ [retrieved: May, 2014].

 ISBN: 978-1-61208-366-7

Use of Bluetooth Technology on Mobile Phones for Optimal Traffic Signal Timing

Hyoshin Park, Ali Haghani
Department of Civil and Environmental Engineering
University of Maryland
College Park, Maryland, USA
hspark@umd.edu, haghani@umd.edu

***Abstract*— Optimizing traffic signal timing is an effective and economical way to improve mobility in an urban area and reduce traffic congestion. The objective of the proposed algorithm is to enable traffic to traverse through the maximum number of downstream intersections without a stop. In this study, Bluetooth technology, to measure travel times on arterial roads, is used as input for an optimal bandwidth progression algorithm. The trajectories of vehicle platoons are tracked and decomposed into link-based samples using adaptive smoothing method, and paired with signal timing on each signalized intersection. Predicted travel time, a value representing the travel time between signalized intersections, is obtained by Support Vector Regression (SVR) model. According to bandwidth efficiency and attainability, the signal timing generated by the proposed model yields lower delays than the current signal planning. The applicability of the proposed model has been validated.**

Keywords-Bluetooth technology; bandwidth optimization; adaptive smoothing; support vector machine

I. INTRODUCTION

Mobility is a key performance area, the enhancement of which supports the economy and the community by facilitating the movement of people and goods [1]. It is critical to maintain reliable traffic flow: travelers can plan and execute their journeys seamlessly using available software applications, and vehicles will flow more freely through existing infrastructure. To overcome the increasing congestion of arterial roads, investments on infrastructure have increased. However, expanding infrastructure is not the only way to improve mobility. Making better use of existing roads can also increase transport capacity.

Arterial street signal systems must coordinate timing of adjacent intersections to improve mobility of platoons. Vehicles in the platoons could encounter fewer red lights, shortening the travel time, decreasing number of stops, and reducing time delays. Bandwidth efficiency and attainability are major criteria for judging the quality of a coordinated signal timing plan. Bandwidth-based solutions, the most visible indicator to individual drivers, generally outperform delay-based solutions [2]. To provide an optimal bandwidth progression, traffic engineers are faced with problems of providing accurate travel time between intersections. Various models have been developed to accurately estimate arterial travel times or delays. However, collecting reliable traveling time data on signalized intersections is challenging. Previous sensor technologies have issues with privacy protection, quality of data, and cost of dedicated hardware.

The number of mobile phones used worldwide has grown, and more than half of those were smartphones in 2013 [3]. Wireless communications are considered enablers of innovation in the field of smart mobility in smart cities [4]. Therefore, it will be worthwhile to identify vehicles carrying mobile phones. One of the latest technologies using wireless communication is the Bluetooth detector, becoming more common to enable real-time continuous traffic monitoring. This paper introduces Bluetooth technology as an effective means of data collection of ground truth travel time. Measured travel time data is used as input for an optimal bandwidth progression algorithm. Compared to traditional method depending on point speed at their fixed locations, Bluetooth technology provide point-to-point travel time over the segments. A new traffic light based on traffic data collected by Bluetooth technology also make traffic flow more smooth and fast [5]. By placing sensors along roads, tracking Bluetooth devices in passing vehicles, the solution is able to accurately detect and record how long it takes a car to drive along a corridor, segment by segment. Fig. 1 presents a well-configured signal coordination system using Bluetooth technology.

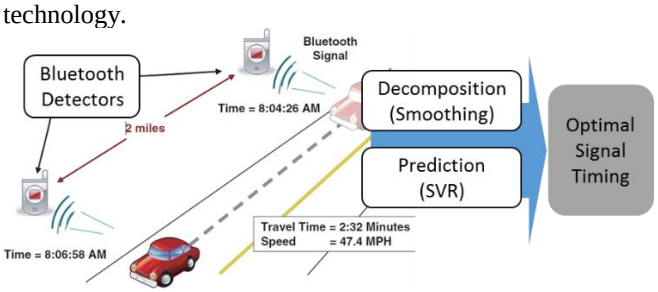


Figure 1. A signal coordination system for arterial roads

The trajectories of vehicle platoons are tracked and decomposed into link-based samples in this study. However, Bluetooth sensors, collecting the data in a point-to-point format, may not be used directly for real-time purposes. It takes time for the actual trip to be realized and for the travel time to become available. In most urban networks, the actual travel time is not available within one discrete time step, as traffic congestion increases. Therefore, we use predicted values of link travel time. The performance of travel time prediction varies with many variables, such as day-to-day traffic demand and other factors. Considering the complex and dynamic nature of traffic flows in the system, traditional models cannot capture the complete characteristics of the stochastic traffic data and may not predict the traffic under

 ISBN: 978-1-61208-366-7

variation with high accuracy. We develop Support Vector Regression (SVR) to predict link travel time.

The Bluetooth technology is introduced for travel time data collection in Section II. Usages of travel time for optimal signal timing is proposed in Section III. We present decomposition and prediction of travel time in Section IV. The signal timing plan is evaluated and future work is presented in Section V and Section VI, respectively.

II. STATE-OF-THE-ART

A. Sensor Technologies

Accurate travel time information between two intersections can be essential to get optimal signal coordination, yet reliable methods for travel surveillances are slow in coming.

1) Until recently, inductive *loop* detectors [6] were the most common traffic data collection for arterial streets even though they are not always reliable. These sensors disrupt traffic during installation and repair, and therefore have high installation and maintenance costs.

2) License plate matching has been used to travel time data collection purpose. However, this system have high equipment costs and their accuracy depends on environmental conditions.

3) Acoustic sensors are attractive especially for their low cost and simple and non-intrusive installation. However they require a sophisticated post-processing algorithm for extracting useful information [7]. These sensors depend on measurements at a point that will over-represent the number of fast vehicles and under-represent the slow ones, and hence give a higher average speed than the true average.

4) Recently, mobile phones have been used as primary source of floating car data. A camera-based traffic signal detection algorithm was used to learn traffic signal schedule patterns and predict their future schedule. Based on when the signal ahead will turn green, drivers can then *adjust speed* so as to avoid coming to a complete halt [8]. It is also possible to accurately infer traffic volume through *Global Positioning Systems* (GPS) collected from a roving sensor network of taxi probes that log their locations and speeds at regular intervals [9]. However, energy consumption of GPS on some phones can be a challenge, then less energy-hungry but noisier sensors like WiFi can be used to estimate both a user's trajectory and travel time along the route [10].

B. Bluetooth Technology

The traditional floating car method is very costly and produces a sparse amount of data [11]. As a result, a new data collection methodology was developed that receives anonymous emissions from Bluetooth equipped accessories in passing vehicles that have been activated in the discovery mode. Bluetooth technology is good for collecting high-quality travel time data that can be used as ground truth for evaluating other sources of travel time [12]. This method proved to be more cost-efficient than floating car method. Various application of Bluetooth technology can be found in [13].

Bluetooth is a telecommunications industry specification that defines the manner in which mobile phones, computers, personal digital assistants, car radios, and other digital devices can be easily interconnected using short-range wireless communications. One example of the use of this technology is the interconnection of a mobile phone with a wireless earpiece to permit hands-free operation. Bluetooth enabled devices can communicate with other Bluetooth-enabled devices anywhere from 1 m to about 100 m (300 ft). This variability in the communications capability depends on the power rating of the Bluetooth sub-systems in the devices. The Bluetooth protocol uses a 48-bit electronic identifier, or tag, in each device called a Machine Access Control (MAC) address. Bluetooth transceivers transmit their MAC ID for the purpose of identifying a device with which to communicate. This "inquiry mode" is used to establish a link with the "responding devices." Inquiries are made by a Bluetooth transceiver, even while it is already engaged in communication with another device. The continuous nature of this process facilitates the identification of passing vehicles containing Bluetooth devices, since all equipped and activated devices will be transmitting inquiries as long as they have their discovery mode enabled.

The main purpose of this paper is a *temporary* (i.e., two weeks) installation of sensors on a small network. Our future study also includes dynamic installations of sensors so that sensors can cover the entire network in several stages [14]. However, there is a case for *permanent* installation of Bluetooth sensors running on solar energy with an overall power budget of less than 5 watts. For details on the technology, readers can refer to [15].

The Bluetooth traffic monitoring system calculates travel times by matching public Bluetooth wireless network IDs at successive detection stations. The time difference of the ID matches provides a measure of travel time and space mean speed based on the distance between the successive stations. Each vehicle at the same signal timing in a different time of day is categorized to represent vehicle platoon.

C. Data Quality Issues

Although Bluetooth has been demonstrated as a promising technology, there remain problems which affect the accuracy of the estimation such as difficulty of distinguishing between multiple transportation modes (e.g., passenger cars, buses, bicycles, or pedestrians.). For example, the probability of multiple Bluetooth travel time records from a bus was analyzed [16]. It is observed that bus is overrepresented in the BMS dataset and it is rare to have overrepresentation by more than six travel time points. The chances of observing more than three travel time records for a bus, is less than 20 %. Nevertheless, in our study, there is no data suspected to be other mode than motor vehicles. However, an effort to distinguish different transportation mode may need in more congested urban area.

The Bluetooth receiver can pick up signals within a 300-ft radius around the sensor. Having two sensors at both ends of an arterial segment implies that in the resulting travel time samples obtained using this technology, one might expect to

 ISBN: 978-1-61208-366-7

see errors caused by a maximum of 600-ft error in the length traveled. Since Bluetooth devices might be detected at any point within the detection zone, this study used the first detection in a group to calculate travel times from the MAC address data.

The Bluetooth traffic detectors sample only a fraction of the vehicles in the traffic stream. To approximate the sampling ratio of the new technology, actual traffic volume in a roadway segment is needed. Traffic volume data are available where other sources of traffic surveillance systems are in place. The average Bluetooth hourly sampling rate is between 2.0% and 3.4%.

D. Privacy Concerns

The anonymous nature of this technique is due to the use of MAC addresses as identifiers. MAC addresses are not directly associated with any specific user account (as is the case with cell phone geo-location techniques) or any specific vehicle (as is the case with deriving travel time from automated toll tags). The MAC address of a cell phone, camera, or other electronic devices, though unique, is not linked to a specific person through any type of central database thus minimizing privacy concerns. Additionally, users concerned with privacy can set options in their device (referred to as "Discovery Mode" or "Visibility") so that the device is not detectable.

III. BANDWIDTH OPTIMIZATION ALGORITHMS

Bandwidth optimization algorithm using three signals with simple two-phase operations is illustrated (Fig. 2). An intersection with the minimum arterial green split, G_{min}, is called the critical intersection (e.g., the middle intersection). The arterial green times for the other intersections in the system are all greater than G_{min}. This minimum green time, G_{min}, determines the largest possible bandwidth progression that can be achieved for the system.

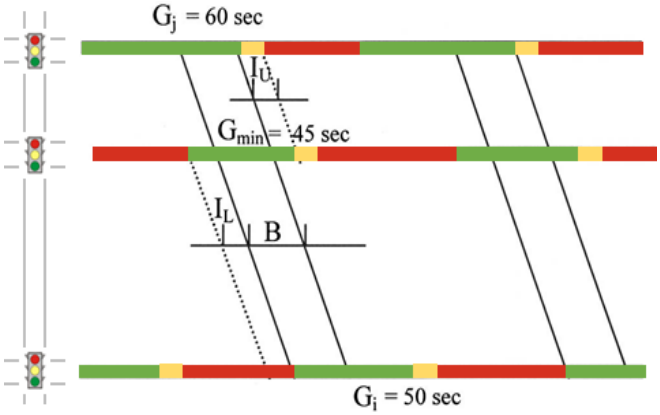


Figure 2. Optimization of bandwidth progression

The system bandwidth is reduced if the progression band encounters interference from other signals in the system. Only one type of interference, either an upper interference, I_U, or a lower interference, I_L, can occur at each signal. The final system bandwidth, B, is determined by G_{min} minus the minimum possible combination of the upper interference and the lower interference,

$$B = G_{min} - \min\{\max(I_{U,i}) + \max(I_{L,j})\} \quad (1)$$

where B=bandwidth(s); $I_{U,i}$=upper interference at intersection i (s); $I_{l,j}$ lower interference at intersection j (s); $\max(I_{U,i})$=maximum value from all signals producing upper interference and max maximum value from all signals producing lower interferences.

The enhanced Brook's algorithms [17], such as those in PASSER II [18], search for the best phasing sequences and offsets at each signal location to minimize the combined interference. The optimization process simultaneously considers progression in both directions.

To maximize the progression bandwidths for both directions, the offset and phasing of each signal should be carefully designed. For an intersection j with multi-phases (e.g., the option of a leading left turn phase or a lagging left turn phase), the interference for one direction is also related to the timing parameters for the other direction. Equations (2) and (3) show how the upper interference or the lower interference can be calculated for intersection j with respect to a master intersection m for one of the directions

$$I_{U,j}(p) = [G_{min} - T_{mj} + T_{jm} - O_m(n) + O_j(p) + G_j] \bmod C \quad (2)$$

$$I_{L,j}(p) = [T_{mj} + T_{jm} - O_m(n) + O_j(p) - S_j] \bmod C \quad (3)$$

where $I_{U,i}(p)$, $I_{L,j}(p)$=upper interference and lower interference at intersection j with phase sequence p (only one phase sequence could occur) (s); T_{mj}, T_{jm}=travel times between intersections m and j (s); $O_m(n)$=relative offset between direction a green time and direction b green time at signal m with phase sequence n (s); $O_j(p)$=relative offset between direction a green time and direction b green time at signal j with phase sequence p (s); G_j=direction a green time at signal j (s); S_j=difference between green times of intersections j and m in direction b (s); and C=cycle length (s).

The interference (either upper or lower) is largely affected by the signal spacing as reflected by the travel times, T_{mj} and T_{jm}. Representative travel time has been predicted by using decomposition and SVR in travel time prediction section of this paper. With the increase of the number of signals in a system, the chances of having larger interference values also increase. For example, there might be a signal whose spacing may actually produce maximum interference, which equals to G_{min}, the green time of the reference intersection. In this case, the bandwidth would be zero.

The arrival sequence of green time at each intersection presents four scenarios.

- Scenario 1. upstream bands projected to arrive after downstream queue discharges
- Scenario 2. upstream bands projected to not arrive after downstream queue discharges
- Scenario 3. upstream bands projected to arrive before queue discharges

 ISBN: 978-1-61208-366-7

- Scenario 4. upstream bands projected to arrive after downstream queue discharges

To evaluate the proposed algorithm, delay is calculated for each vehicle compare to free flow traffic condition. Calculated delay for each signal cycle at specific time of day is aggregated for forty seven days. Efficiency and attainability measure the quality of through progression provided by a timing plan. Efficiency for a direction is the percent of cycle used for progression. Attainability is the percent of bandwidth in a direction in relation to the minimum green split in the same direction. When attainability is at 100%, the bandwidth is at its maximum. Theoretically, the maximum bandwidth in a direction can be no more than the smallest through green split in that direction. We calculate efficiency and attainability for the two arterial directions (4) and (5).

$$\text{Efficiency}(\%) = \frac{(\mathrm{B_U} + \mathrm{B_L})}{2 \times \text{Cycle length}} \times 100 \qquad (4)$$

$$\text{Attainability}(\%) = \frac{(\mathrm{B_U} + \mathrm{B_L})}{G_{\min,\mathrm{U}} + G_{\min,\mathrm{L}}} \times 100 \qquad (5)$$

IV. ARTERIAL TRAVEL TIME

While license plate matching techniques are several miles apart due to associated costs, Bluetooth sensors are deployed 0.7-1 miles apart. A normal segment between Bluetooth sensors has two or three intermediate intersections. Proposed decomposition method reconstructs the trajectory of point-to-point path into intersection to intersection link data. Accurate prediction of travel time provides inputs for optimal bandwidth progression.

A. Decomposition of Path Travel Time

For a vehicle traveling from an origin point *A* to a destination point *B* through *x* intersections, we decompose the travel time as the sum of travel times on each link (Fig. 3).

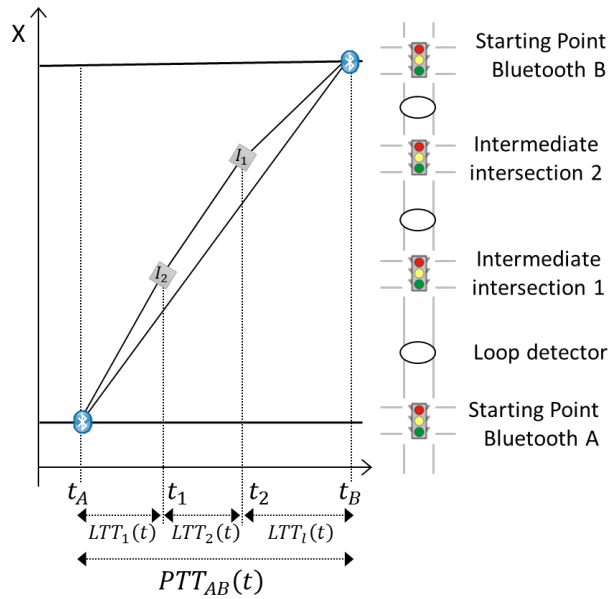


Figure 3. Reconstruction of travel time

We use link-based travel time, in which traffic conditions (speed reported by loop detectors) are assumed to be constant. The vehicle speed is considered linearly increasing or decreasing between intersections.

A serious challenge in traffic data is that the typical scale of some traffic patterns, such as the wavelength of stop-and-go waves, is similar to the spacing of stationary detectors. Consequently, important dynamical features may be lost in the interpolation process, and even entirely spurious patterns may be reconstructed [19].

The switch between free and congested traffic is then managed by an adaptive speed filter. A smoothing function performs two-dimensional interpolation to reconstruct the spatiotemporal traffic state from discrete traffic data. The adaptive weight factor $0 < W_s(t) < 1$ controls the superposition of the free and congested velocity fields and can be estimated as

$$\mathrm{W_s(t)} = \frac{\sum_1^{\mathrm{l}} \mathrm{LTT_x(t)} - \sum_1^{\mathrm{l}} \mathrm{LTT_x^f}}{\sum_1^{\mathrm{l}} \mathrm{LTT_x^j} - \sum_1^{\mathrm{l}} \mathrm{LTT_x^f}} \qquad (6)$$

where LTT_l^j denotes congested traffic operations and LTT_l^f denotes free flow conditions from historical data between intersections. $LTT_x^s(t)$ is smoothed $LTT(t)$ of detector *x* at time interval *t*, estimated by combining the values for free and congested traffic:

$$\mathrm{LTT_x^s(t)} = \mathrm{W_s(t)LTT_x^j} + \big(1 - \mathrm{W_s(t)}\big)\mathrm{LTT_x^f} \qquad (7)$$

The ratio of $LTT_x^s(t)$ is used to generate piece-wise link travel times $LTT_x(t)$:

$$\mathrm{LTT_x(t)} = PTT_{AB}(t) \times \frac{\mathrm{LTT_l^s(t)}}{\sum_1^{\mathrm{l}} \mathrm{LTT_l^s(t)}} \qquad (8)$$

B. Travel Time Prediction

Support Vector Machines (SVMs), learning machines implementing the structural risk minimization inductive principle, is used to obtain good generalization on a limited number of learning patterns in travel time prediction. SVMs work by solving a constrained quadratic problem where the convex objective function for minimization is given by the combination of a loss function with a regularization term [20].

Traditional regression procedures are often stated as the processes deriving a function *f*(*x*) that has the least deviation between predicted and experimentally observed responses for all training examples. One of the main characteristics of Support Vector Regression (SVR) is that instead of minimizing the observed training error, SVR attempts to minimize the generalized error bound to achieve generalized performance. This generalization error bound is the combination of the training error and a regularization term that controls the complexity of the hypothesis space.

The approximate function is determined by a small subset of training samples called Support Vectors (SVs). A specific loss function is developed to make a sparseness property for SVR. In order to learn the non-linear relations by linear machines, selecting a set of non-linear features and rewriting

 ISBN: 978-1-61208-366-7

the data in the new representation are needed, equivalent to applying a fixed non-linear mapping of the input space to a feature space in which the linear machine can be used. In SVR, the input *x* is first mapped onto a *m*-dimensional feature space using some fixed (nonlinear) mapping. Then, a linear model is constructed in this feature space. Using mathematical notation, the linear model in the feature space, $f(\mathbf{x},\omega)$, is given by

$$f(\mathbf{x},\omega)=\sum_{j=1}^{m}\omega_j g_j(\mathbf{x})+b \tag{9}$$

where $g_j(\mathbf{x}), j=1,\ldots,m$ denotes a set of nonlinear transformations, and *b* is the "bias" term. Often the data are assumed to be zero mean, so the bias term in (9) is dropped.

The quality of estimation is measured by the loss function $L(y, f(\mathbf{x},\omega))$. The SVR uses a new type of loss function called $\mathcal{E}$-insensitive loss function

$$L_\varepsilon(y, f(\mathbf{x},\omega))=\begin{cases}0 & if\ |y-f(\mathbf{x},\omega)|\le\varepsilon\\ |y-f(\mathbf{x},\omega)|-\varepsilon & otherwise\end{cases} \tag{10}$$

The empirical risk is

$$R_{emp}(\omega)=\frac{1}{n}\sum_{i=1}^{n}L_\varepsilon(y_i, f(\mathbf{x}_i,\omega)) \tag{11}$$

Note that $\mathcal{E}$-insensitive loss coincides with least-modulus loss and with a special case of robust loss function when ε =0. Hence, we shall compare prediction performance of SVM (with proposed chosen ε) with regression estimates obtained using least-modulus loss (ε =0) for various noise densities.

SVM regression performs linear regression in the high-dimension feature space using $\mathcal{E}$-insensitive loss and, at the same time, tries to reduce model complexity by minimizing $||\omega||^2$. This can be described by introducing (non-negative) slack variables ξ_i, ξ_i^* ($i=1,..n$), to measure the deviation of training samples outside $\mathcal{E}$ -insensitive zone. The SVR is formulated as minimization of the following objective function:

$$\min\quad \frac{1}{2}||\omega||^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{12}$$

$$\text{s.t.}\quad \begin{cases} y_i-f(\mathbf{x}_i,\omega)\le\varepsilon+\xi_i^*\\ f(\mathbf{x}_i,\omega)-y_i\le\varepsilon+\xi_i\\ \xi_i,\xi_i^*\ge 0, i=1,\ldots,n\end{cases}$$

This optimization problem can transformed into the dual problem, and its solution is given by

$$f(\mathbf{x})=\sum_{i=1}^{n_{SV}}(\alpha_i-\alpha_i^*)K(\mathbf{x}_i,\mathbf{x}) \tag{13}$$

$$\text{s.t. } 0\le\alpha_i^*\le C,\ 0\le\alpha_i\le C$$

where n_{SV} is the number of Support Vectors (SVs) and the kernel function

$$K(x,z)=\left(c+\langle x,z\rangle\right)^d \tag{14}$$

In this study, polynomial kernel with *c* = 1, and *d* = 2 is used for prediction of link travel time (Fig. 3).

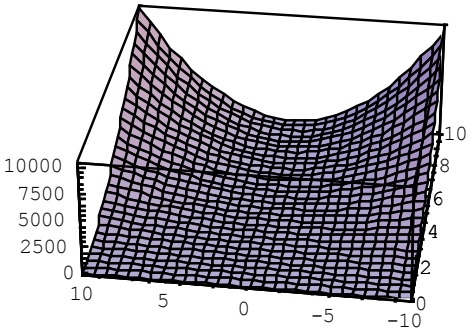


Figure 4. Instantaneous and Predicted Travel Times.

To handle fully dense quadratic optimizations in SVR, decomposition methods are designed. Unlike most optimization methods that update the whole vector α in each step of an iterative process, the decomposition method modifies only a subset of α per iteration. This subset, denoted as the working set B, leads to a small sub-problem to be minimized in each iteration. An extreme case is the Sequential Minimal Optimization (SMO) [21], which restricts B to have only two elements. Then, in each iteration one does not require any optimization software in order to solve a simple two-variable problem.

The model is applied to real world transportation network from the FHWA test data set [22]. The network consist of 4 intersections on 82nd street for afternoon peak hours (4:00 PM–6:00 PM) from 9/15/2012 to 11/14/2012. The data include following information:

- Phase and timing data consists of active calls and phasing information for four signals.
- Bluetooth data consists of travel times derived from matching MAC addresses that are captured by the Bluetooth readers between a pair of locations
- Loop detector data consists of speed on link between upstream and downstream.

In order to compare the performance measures before and after the field implementation, forty seven weekdays travel time need to be trained, after paring signal timing and reconstructing link travel time. The polynomial kernel function is used for SVR travel time prediction model. We examine the travel time of three links from the intersections of 82nd and Woodward to 82nd and Foster. Relative Mean

Errors (RME), the ratio of difference between predicted error and actual travel time to the quantity, is calculated to evaluate prediction performance of the model, for 60 seconds interval. The results in Table I show the RME and RMSE of SVR for different travel distances over all the data points of the testing set. They show that the SVR predictor represent each temporal and spatial vehicle platoon in a feasible range. However, if penetration rate is higher, shorter interval with higher frequency of detection will be available and we can provide more accurate inputs for signal optimizations.

TABLE I. PREDICTION RESULTS

	RME	**RMSE**
Link 1 (0.5mi)	10.52%	19.56%
Link 2 (0.6mi)	9.84%	17.94%
Link 3 (0.6mi)	12.32%	22.54%

The optimized offset values were implemented on afternoon peak hours on 11/15/2012. The arterial outbound bandwidth is 29 seconds, and 25 seconds for the inbound. Arterial bandwidth efficiency is 21.16%, and bandwidth attainability is 63.74%, which means a fair progression according to the guidelines of bandwidth efficiency [23].

The existing field offset setting is {0, -24.9, -21.6, 4.6}, and its weighted total delay per cycle for each intersection is 183.8 seconds. In comparison, the optimized offset values were implemented and the best offset result is {0, -21.4, -21, -20.9} for four intersections, and the weighted total delay per cycle is 30.4 seconds. We should note that the above offset values are computed under the transformed time coordinates.

Table II compares the calculated travel time delays of both eastbound (from stop line of Boone to stop line of TH100) and westbound (from stop line of TH100 to stop line of Boone) based on different offset settings.

TABLE II. AVERAGE DELAY COMPARISION BEFORE AND AFTER

	Original (Before)	**Optimized (After)**	**Change percentage**
Northbound average delay (seconds)	784.8	678.4	13.6%
Southbound average delay (seconds)	119.8	107.4	10.4%

As we can see, both eastbound and westbound travel time delays are substantially reduced after the offset adjustment. On average, the northbound travel time delay with original offset (9/3/2009) is 784.8 seconds and it decreases to 678.4 seconds after optimization (9/14/2009), which is a 13.6% reduction. For southbound, average travel time delay with original offset is 119.8 seconds and it decreases to 107.4 seconds after optimization, which indicates a 10.4% reduction. As traffic condition is more congested (northbound), reduction of travel time delay is higher. Considering that the original offset setting was already optimized, the improvement is significant.

V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an application of Bluetooth technology to signal control by improving the quality of travel time prediction. Proposed method presents fair bandwidth progression efficiency and attainability, and lower delays than the current signal planning.

A number of possible future research directions exist. For example, the applicability of the proposed model is currently limited to through movement of traffic. A worthwhile research effort would be to generalize the model to the network level to obtain network-wide movements into signal controls. A challenging part of Bluetooth data is the small number of sample data. By using distribution of travel time data collected from each loop detector, Bluetooth data can be augmented. Accurate estimation of queue and arrival dynamics can be integrated for the optimal signal timing.

ACKNOWLEDGMENTS

Parts of the research were funded by the I-95 Corridor Coalition and the Center for Integrated Transportation Systems Management at the University of Maryland, College Park.

REFERENCES

[1] S. Mahapatra, M. Wolniak, K. F. Sadabadi, E. Beckett, and T. Jacobs, "2013 Maryland State Highway Mobility Report," A Joint Effort between the University of Maryland Center for Advanced Transportation Technology, the Maryland State Highway Administration, and Johnson, Mirmiran & Thompson, Available: http://www.roads.maryland.gov/OPPEN/2013_Maryland__Mobility.pdf [retrieved: June 2014]

[2] X. K. Yang, "Comparison among Computer Packages in Providing Timing Plans for Iowa Arterial in Lawrence, Kansas," Journal of Transportation Engineering, 1274, pp. 314–318, 2001.

[3] D. L. Gilstrap, "2013 Ericsson Mobility Report: On the Pulse of the Networked Society," Ericsson SE-126 25 Stockholm, Sweden, Available: http://www.ericsson.com/res/docs/2013/ericsson-mobility-report-november-2013.pdf [retrieved: June 2014]

[4] G. Pasolini, A. Bazzi, B. Masini, and O. Andrisano, "Smart Navigation in Intelligent Transportation Systems: Service Performance and Impact on Wireless Networks," IARIA International Journal on Advances in Telecommunications (ISSN: 1942-2601), vol 6, no 1&2, 2013, pp. 57-70. Available: http://www.iariajournals.org/telecommunications/tocv6n12.html [retrieved: June 2014]

[5] Blip Systems, "How Traffic is Optimized by Using Bleutooth Technology," Available: http://www.youtube.com/watch?v=QqZEjA8XAXs [retrieved: June 2014]

[6] A. Somov, C. Dupont, and R. Giaffreda, "Supporting Smart-City Mobility with Cognitive Internet of Things," Future Network and Mobile Summit (FutureNetworkSummit), 2013 , pp.1-10.

[7] B. Barbagli, L. Bencini, I. Magrini, G. Manes, and A. Manes, "A Traffic Monitoring and Queue Detection System Based on

 ISBN: 978-1-61208-366-7

an Acoustic Sensor Network," International Journal on Advances in Networks and Services, vol. 4, pp. 27-37, 2011.

[8] E. Koukoumidis, L. Peh, and M. Martonosi. "Signalguru: Leveraging mobile phones for collaborative traffic signal schedule advisory," Proc. The 9th International Conference on Mobile Systems, Applications, and Services, Washington, D.C., July 2011.

[9] J. Aslam, S. Lim, X. Pan, and D. Rus, "City-scale Traffic Estimation from a Roving Densor Network," Proc. The 10th ACM Conference on Embedded Network Sensor Systems, Toronto, Canada, November 2012.

[10] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo, and J. Eriksson. "Vtrack: Accurate, energy-aware road traffic delay estimation using mobile phones," Proc. The 7th ACM Conference on Embedded Network Sensor Systems, California, November 2012.

[11] A. Haghani, S. Yang, and M. Hamedi, "Cellular Probe Data Evaluation, Case Study: The Baltimore Multimodal Traveler Information System," Maryland Department of Transportation, Hanover, January. 2007.

[12] A. Haghani, M. Hamedi, H. Park, Y. Aliari, and X. Zhang, "I-95 Corridor Coalition Vehicle Probe Project: Validation of INRIX Data," Civil Engineering Department, University of Maryland College Park, August 2013, Available: http://www.i95coalition.org/i95/Portals/0/Public_Files/upload up/Vehicle-Probe/I-95%20CC%20Valid%20Report-Aug% 202013-data%20May%202013-GA.PDF [retrieved: June 2014]

[13] A. Haghani, M. Hamedi, K. F. Sadabadi, S. Young, and P. Tarnoff, "Freeway Travel Time Ground Truth Data Collection Using Bluetooth Sensors," in Transportation Research Record: Journal of the Transportation Research Board, No. 2160, Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 60-68.

[14] H. Park and A. Haghani, "Optimal Number and Location of Bluetooth Sensors on Arterial Roads" International Conference on Engineering and Applied Sciences Optimization (OPTI 2014), Kos Island, Greece, June 2014

[15] S. Young, "Bluetooth Traffic Detectors for Use as Permanently Installed Travel Time Instruments," Maryland State Highway Administration Research Report, Febuary 2012. Available: http://www.roads.maryland.gov/OPR_Research/MD-12-SP909B4D-Bluetooth-Traffic-Detectors_Report.pdf [retrieved: June 2014]

[16] A. Bhaskar, L. M. Kieu, M. Qu, A. Nantes, M. Miska, and E. Chung, "Is Bus Overrepresented in Bluetooth MAC Scanner data? Is MAC-ID Really Unique?" International Journal of Intelligent Transportation Systems Research, April 2014.

[17] W. D. Brooks, "Vehicular Traffic Control: Designing Traffic Progression Using A Digital Computer," IBM-Data Processing Division, Kingston, N.Y., 1965.

[18] S. Venglar, P. Koonce, and T. Urbanik. "PASSER III-98 Application and User's Guide," Texas Transportation Institute, Texas A&M University System, College Station, Texas, 1998.

[19] Treiber and D. Helbing, "Reconstructing the Spatio-temporal Traffic Dynamics from Stationary Detector Data," Cooper@tive Tr@nsport@tion Dyn@mics 1, 3.1–3.24, 2002, Internet Journal, Available: http://www.TrafficForum.org/ journal [retrieved: June 2014]

[20] V. Vapnik and A. Lerner, "Pattern Recognition using Generalized Portrait Method," Automation and Remote Control, 24, 1963.

[21] J. C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization," in B. Sch¨olkopf, C. J. C. Burges, A. J. Smola, editors, Advances in Kernel Methods-Support Vector Learning, Cambridge, M.A., 1998.MIT Press.

[22] Research Data Exchange (RDE), "Multimodal Data Set Clean-up for Portland Oregon Metropolitan Region", Federal Highway Administration, 1200 New Jersey Avenue, SE | Washington, D.C., Available: https://www.its-rde.net/home [retrieved: June 2014]

[23] Federal Highway Administration and USDOT, "Traffic Signal Timing Manual, United States Department of Transportation," Washington, D.C., USA, 2008.

 ISBN: 978-1-61208-366-7

Empowering Mobile Users: Applications in Mobile Data Collection

Arlindo F. Conceição
Institute of Science and Technology
Federal University of São Paulo (UNIFESP)
São J. dos Campos, Brazil
Email: arlindo.conceicao@unifesp.br

Dario Vieira
French School of Electronics and Computer Science (Efrei)
Paris, France
Email: dario.vieira@efrei.fr

***Abstract*—This paper presents an architecture for collecting and analyzing mobile data. The system offers simple and intuitive interfaces to create mobile applications (Apps). It allows the collection of conventional data, such as numbers and text, and also non-conventional data, such as multimedia files, location information, and barcodes. The collected data can be shared among users on a social network. In addition, we propose a pipeline architecture to data analyzing.**

Keywords–Mobile services; Smartphones; Data Collection.

I. INTRODUCTION

The mobile communication market has evolved fast. This evolution is mainly characterized by three factors: reduced smartphone prices, launch of mobile devices with high processing capability, and emergence of new technologies for the development of Mobile Applications (Apps). These factors have created conditions for the large-scale usage of Apps.

However, despite the advances in hardware and software, the creation of mobile applications continues to demand programming efforts and involvement of programmers and IT professionals. In our opinion, this is the main limitation to wider usage of mobile solutions and applications. The resources (money or/and programmers) to develop these mobile applications are not always available. In general, the end user cannot pay for the development of customized applications.

In order to mitigate this problem, we are developing an open cloud infrastructure for data collection and automatic creation of mobile Apps [1]. The platform allows the user to create and customize their own mobile applications using simple interfaces. The user does not need to know how to program.

By providing new tools to the users, we are opening opportunities for new applications, services and usage of mobile devices. We refer to this concept as Mobile User Empowering, which includes the following goals:

- To allow customization of mobile software requirements using simple interfaces.
- To host data and applications, transparently, in the cloud.
- The service must be free and the data must be that of the user.

To create a proof of concept of Mobile User Empowering, we focused on applications for Mobile Data Collection (MDC) and mobile surveys. These applications usually have the format of a questionnaire and contain a pre-established number of objective questions [2]. There are several good reasons to use mobile applications instead of traditional methods. First, we can reduce or even eliminate the usage of paper. Second, mobile applications may enhance the reliability of the collected data by implementing validation procedures. Finally, if we collect data using electronic devices, the information does not need to be manually moved from paper to an information system.

The project was called **Maritaca** (*MARitaca Is a Tool to creAte Cellular phone Applications*). Furthermore, *Maritaca* is also the name of a bird in Brazil. The project is open source and was designed to be highly scalable. The tool is available for evaluation [3].

The remainder of this paper is structured as follows: Section II presents the related works, Section III describes the distributed architecture, Section IV briefly shows the interfaces and features of the platform and integration model of the project and Section V proposes a pipeline architecture to data analysis. Section VI explores applications of the platform. Finally, we present future work and our final considerations.

II. RELATED WORK

There are several projects with similar purposes to Maritaca. For example, *App Inventor* [4] allows to build applications for Android visually. It focuses on the drawing interface components, step-by-step, connecting their respective events. The advantage of Maritaca over *App Inventor* is that it allows a simpler and more intuitive design of interfaces. This is possible because it focuses on MDC applications.

Nokia Data Gathering [5] is a system that allows questionnaires to be built which can be accessed by mobile devices with connection to the Internet. Data is gathered and stored in the mobile devices and can be transmitted to a server. However, it is a proprietary solution.

Another tool is Open Data Kit (ODK) [6], which consists of a set of open source tools that help to create and manage mobile data collection. It is composed of three tools: *Build*, *Collect* and *Aggregate*. *Build* is used for modeling the forms. *Collect* is used to start data collection on mobile devices that run the Android operating system, and also for sending the data to the server. Finally, *Aggregate* gathers the collected data on the server and converts to standard formats. It should be noted that the tool *Collect* generates the interfaces of the mobile application from a file format XForm, a standard formatting of forms, specified in XML. However, ODK does not offer an infrastructure, but sets standards and APIs.

The product DoForms [7] allows the creation of multiplatform and mobile questionnaires. It is similar to Maritaca, however, it is not open source.

The Mafuta Go project [8] is a specific mobile application designed to find the nearest gas stations in Uganda (located in

 ISBN: 978-1-61208-366-7

East Africa). The App also compares the prices of gas. The Maritaca could be used to create the same application.

Fulcrum [9] allows the creation of Android and iOS applications for data collection. It also collects location information, so that the collected data can be viewed on maps. It is quite similar to Maritaca. On the site of the project it is possible to see several Apps created using the system; the Apps are organized in several categories, such as tourism, utilities, financial tools, etc. However, Fulcrum is not a free platform.

All these products have similarities to Maritaca, but none of them implements all proposed features, and none of them allows flexible data analysis.

III. System Architecture

The Maritaca project was developed as a cloud application [10]. The data collected using Android devices are stored in the cloud and can be visualized using standard web browsers.

There are two main components:

Mobile component: this is an Android application that interprets the XML file (questionnaire descriptor) and generates the interfaces automatically. In fact, the mobile component is an engine, based on the design pattern Interpreter [11]. The mobile component design was the key factor for allowing to create mobile applications automatically.

Server component: the server side was written in Java, using the application server *JBoss* [12] and the framework *Spring* [13]. All web services were implemented based on the *RESTFul* approach [14]. The server also integrates the following products: Form Editor, Analytics Editor, Cassandra database, Hadoop file system, Solr search engine, and MongoDB.

- **Form Editor**: this is an independent Web application, written using HTML5 and Ajax. It allows the quick and intuitive development of questionnaires by implementing drag-and-drop interfaces. As a result, this component generates a questionnaire descriptor, which is persisted in XML format, and is parsed by the Mobile component.
- **Analytics Editor**: it is also an independent Web application used to create queries about the collected data.
- **Cassandra database**: it is used for scalable storage of information. It is based on the paradigm *NoSQL* [15], [16].
- **Hadoop file system**: this is a distributed file system [17], [18] used to store non structured data, such as Apps and multimedia files.
- **Solr Engine**: it is a distributed search engine [19], [20] used to enable searching of Apps. Each App has a description; we used Solr to index the keywords of this description, so that it is possible to search for specific Apps.
- **MongoDB**: this is a scalable distributed database [21] used to analyze collected data. MongoDB is a NoSQL data repository that implements data queries using a semantic similar to SQL [22].

The Figure 1 illustrates the system architecture and the relation among server components.

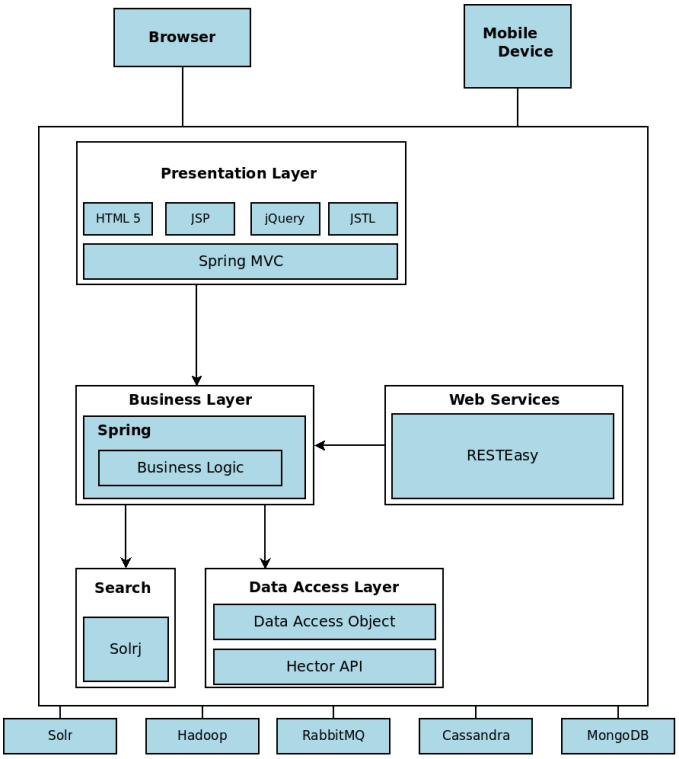


Figure 1. System architecture and main components of the server side.

A. System integration

The interaction between mobile devices and the system is always done with RESTFul services. The web layer and the RESTFul services interact with the business layer using the framework Spring.

The architecture predicts the usage of many instances from the system components (JBoss, Cassandra, MongoDB, and Hadoop) in one computational *cluster*. The load balance of the requisitions will be implemented using nginx [23].

The project also includes one additional component, the RabbitMQ [24], that is used to enqueue messages, mainly to send email messages. For example, it is used to send new passwords (if they have been forgotten), send invitations to new users, etc.

B. Funcionalities of Mobile component

The mobile component is an engine that translates the descriptor of the questionnaire (represented in XML format) into a hierarchy of instantiated objects. These objects are responsible for rendering the interfaces and implementing data validation. The computational model used to represent the questionnaires is sophisticated, and it was this technological innovation that makes the solution possible.

The technique of mapping XML into a list of objects was based on the design pattern *Interpreter* [11], where we can use a hierarchy of classes to simplify message protocol and data interpretation. In our case, see Figure 2, the mobile application is a Context Manager, which always points to the object currently in use (question being answered). Each subclass of type *Question* [25] can implement its own policies, such as, field validation, interfaces, and data storage methods.

 ISBN: 978-1-61208-366-7

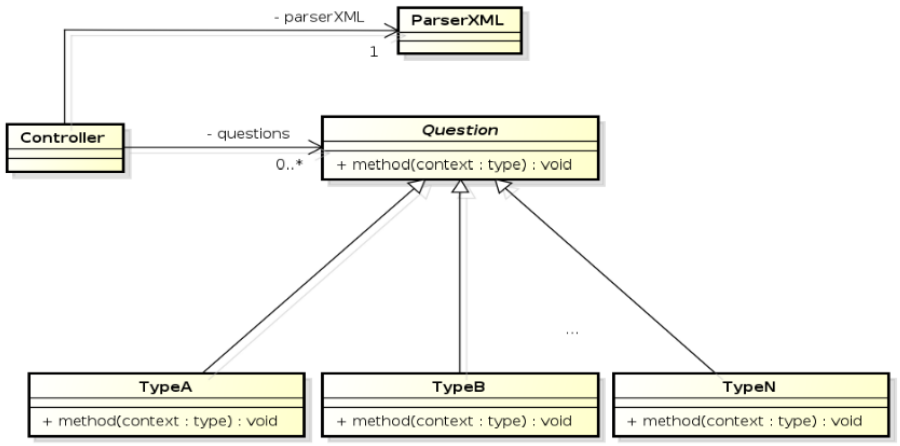


Figure 2. Application controller and XML mapping into objects using *Interpreter* design pattern.

1) Hierarchy of objects: We create a hierarchy of classes in which the questions are mapped. The interfaces for manipulating objects allows, for example, to validate the value collected (using method *validate*), to render the interface for each type of question (method *getLayout*) and to control navigation between each of the questions (methods *getNext* and *getPrevious*).

The *validate* method allows, for example, to validate minimum and maximum limits for the inputs of numeric data. Thus, if a form contains a question about the interviewee's age, the minimum and maximum value for the answers can be defined, for example, respectively as 0 and 105. Data validation is implemented in the class that defines the type of question and is a very effective method to prevent incorrect data collection.

2) Techniques for XML interpretation: The project uses the Simple framework [26] for the serialization and deserialization of the XML files. That is, the framework directly converts XML files into objects and vice versa. This technique is simpler to implement than a XML parsing, which simplifies code maintenance and extension.

3) XML file format: The mobile component interprets the XML file generated by the form editor. In this file, each question is represented as a XML tag with the following basic attributes: *id*, *next*, *previous*, *required*, *label*, *help*, *type* and *default*.

The following fields: *id*, *next* and *previous* are numeric type; *id* identifies the number of the current question, *next* points to the id of the next question and *previous* points to the previous question.

The field *required* determines whether the question is mandatory; its value is defined as: *true* or *false*. The field *label* must contain a question text to be shown in the questionnaire. The attribute *help* is not mandatory and contains a clarification about the question. The attribute *type* defines the type of the data, and can assume, for example, the following values: *text*, *number*, *radiobutton*, *combobox*, *video*, *gps*, etc. The attribute *default* contains a default value of the current question.

Furthermore, some types of questions may have a *conditions* structure, which is used to define the conditional navigation between questions. By using this tag, the response to the current question is used to determine which question will be displayed. For example, consider the following question: "*How old are you?*". If the answer is a value under 18 years, the next question might be: "*What is the name of your parent?*"; otherwise, this question could be omitted.

4) Authentication: Before the first data collection, the mobile application user must authenticate their identity on the server; this guarantees that the user has permission to collect data for that form. This process is done using the OAuth authorization framework [27] that enables third-party applications to obtain limited access to an HTTP service. All data transfer is implemented using RESTFul services and JSON message format.

C. Capturing unusual data

In addition to collecting usual data, such as texts and numbers, the solution also allows collection of unusual data, such as multimedia (audio, video and images) [28], geolocation [29], drawings, barcodes, etc. In summary, the questionnaire can include questions such as: *What is your current location? Take a picture! Record an audio message!*

The implementation of new types of data captured can be easily performed. To do this, simply extend the class *Question* and make the appropriate changes in the parsing class.

D. Automatic generation of Apps

Every time a form is saved in the Form Editor, the system generates a new Android App (executable APK file format) and stores it in the Hadoop distributed file system. This was not the first approach adopted. Initially, we planned to create a single Android application, where the XML descriptors would be loaded.

However, it was difficult to maintain several versions of Apps, thus we implemented this new process of form publication. The process of compilation takes a few seconds, but it is done in the background, and thus does not affect the perceived usability of the server system.

IV. SYSTEM USAGE

This section briefly describes how the system can be used [3], [30].

 ISBN: 978-1-61208-366-7

First, the user can access the system and create an App (mobile application) for data collection; the user must be registred in the platform. The App can be installed onto any compatible mobile device that runs Android 2.2, or above. The installation of the mobile App is simple and straightforward.

On the mobile device, the application allows data to be collected using user-friendly interfaces. Figure 3(a) shows the main interface of the application. Figure 3(b) shows a screenshot of a data collection interface. By default, it uses a Wizard interface design pattern, i.e. each question is shown on a screen. The user can also choose to see a complete list of questions, it is useful in large surveys.

(a) Screenshot of the home interface of the Mobile component.

(b) A data collection interface for a question of type Text.

Figure 3. Interfaces automatically created.

To carry out the data collection, it is not necessary to be connected to the Internet. After collection, data is stored on the mobile device. An Internet connection is needed only for authentication and data upload.

The user can use the web interface to visualize and manage forms, see Figure 4. The list of forms is organized in two panels, forms created by the user (top) and shared forms (bottom).

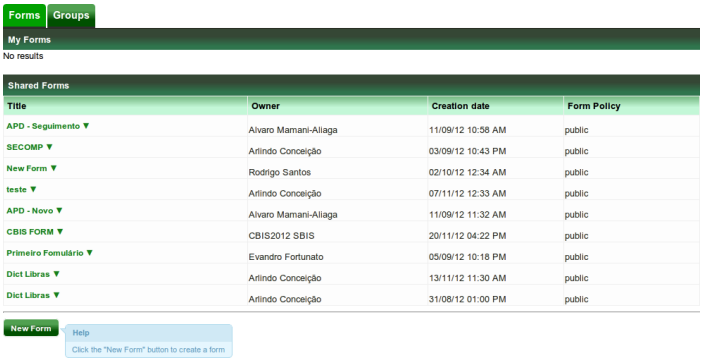


Figure 4. Form management interface.

Currently, the system allows the creation of three types of forms: private, public and shared. In Table I, these policies are summarized. As the names indicate, the **private** form can be only used by its owner and the **public** form can be viewed by any user of the platform.

For **shared** forms, the owner can invite other users, and the owner can see all data collected. Furthermore, the shared forms can be divided into two subtypes: hierarchical and social. For **shared-hierarchical** forms, the owner can invite users, for example, users A and B, but the data gathered by user A is not visible by user B, and vice versa. In turn, for **shared-social** forms, all invited users can visualize the data collected by one another.

TABLE I. DATA SHARING POLICIES.

		Private	Shared Hierarchical	Shared Social	Public
Forms	Read	Owner	Owner and List	Owner and List	All
	Update	Owner	Owner	Owner	Owner
Answer	Read	Owner	Owner	Owner and List	All
	Collect	Owner	Owner and List	Owner and List	All

V. PIPELINE DESIGN FOR DATA ANALYSIS

In addition, we recently created a solution that allows to analyze the data. The user can configure a pipeline for data processing, as illustrated in Figure 5. The user can: filter the data (time, specific fields, or geographical region), apply data transformations similar to SQL commands (*order by, sum, average*, etc.), and, finally, choose a data visualization mode, such as table or map.

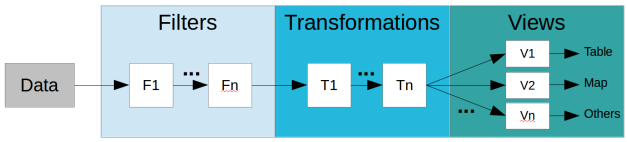


Figure 5. Analytics module, based in a pipeline of information

For example, suppose that a user creates a App to collect the price and location of restaurants. Then, the user can define a query using the pipeline architecture. A filter can be applied to restrict the results to restaurants near the user. A transformation can be applied to order the restaurants by price. Finally, the user can choose to see the results as a map. Figure 6 shows the result of a query created using the *Analytics* module. In fact, every time a user makes a query request, the data is imported from the Cassandra database to the MongoDB, where the filters and transformations are applied using a map reduce strategy [22].

VI. APPLICATIONS

The project allows:

- **The creation of your own mobile data application.** The user can create and modify their own app for data collection. A salesman can coordinate a customer's orders, students can share pictures of a party, and parents can visualize where their children are. There are no costs and no need for programming skills.
- **Your own Social Network.** The user can create an application for data collection and define three different models of data sharing. Thus, it is possible to create social networks for specific interests.

 ISBN: 978-1-61208-366-7

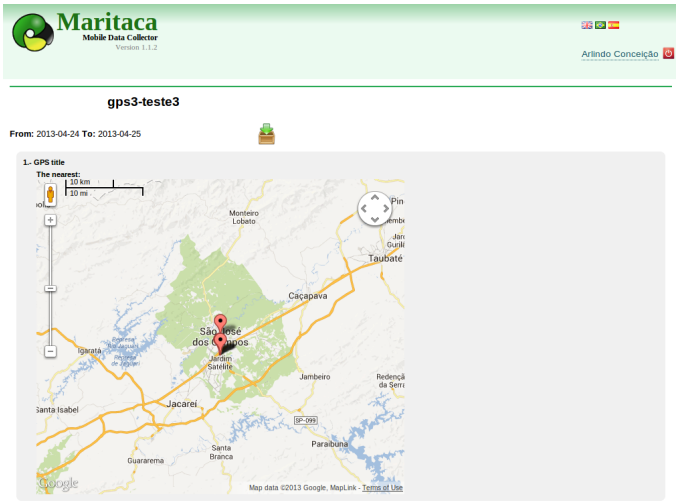


Figure 6. A result of *Analytics* module.

- **Extracting Collective Intelligence.** By using the *Analytics* module, the user can define special views of the data, offering services with high aggregated value.

Therefore, the user can create an effective and complete mobile solution using simple interfaces. There are many possibilities. Currently, we are using the system to collect data in homecare health services and nutritional monitoring.

VII. Conclusion and Future Work

We explored the concept of Mobile User Empowering, that provides to the user the power to create, modify and use his own mobile applications. The project offers tools to collect, share and analyze mobile data, allowing users total customization of software requirements using simple interfaces, without needing knowledge of programming languages or IT infrastructure. The architecture has been developed to cover most mobile applications based in questionnaires, storing both conventional (number, text, etc.) and non-conventional data (video, pictures).

In addition, we proposed a pipeline architecture and its cloud implementation that can be used to data analysis.

Currently, the project can create mobile applications on Android platform. We are developing a multi-platform version using the Phonegap technology. In addition, we are deploying the solution in a private cloud with high processing power.

The latest version of the project is available for evaluation [3]. The source code and additional documentation can be found at the code repository [30].

Acknowledgment

We thank FINEP [31] by funding this research project.

References

[1] A. F. da Conceição, J. V. Sánchez, T. Barabasz, A. H. Mamani-Aliaga, B. G. dos Santos, and M. F. Mendonça, "Open architecture for mobile data collection using cloud computing," in International Workshop on Mobile Cloud Computing: Data, Management & Security (mCloud). In conjunction with 14th IEEE International Conference on Mobile Data Management (IEEE MDM). Milan, Italy., 2013.

[2] R. Ghiglione, B. Matalon, C. Pires, and A. de Saint-Maurice, O inquérito: teoria e prática. Lisboa: Editora Celta, 1997.

[3] A. F. da Conceição et. al, "Maritaca," http://maritaca.unifesp.br, accessed: 2014-06-07.

[4] MIT, "Mit App Inventor," http://appinventor.mit.edu, accessed: 2014-06-07.

[5] Nokia Corp., "Nokia Data Gathering," https://nokiadatagathering.net, accessed: 2014-06-07.

[6] ODK Community, "Open Data Kit," http://opendatakit.org, accessed: 2014-06-07.

[7] doForms Inc., "DoForms," http://www.doforms.com, accessed: 2014-06-07.

[8] MafutaGo, "Official Mafuta Go website," http://www.mafutago.com, accessed: 2014-06-07.

[9] Fulcrum, "Gather data anywhere, anytime with fulcrum," http://fulcrumapp.com, accessed: 2014-06-07.

[10] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica et al., "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, 2010, pp. 50–58.

[11] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, Design Patterns. Addison-Wesley Professional, 1994.

[12] S. Davis and T. Marrs, "JBoss at work: A practical guide," 2005.

[13] B. Tate and J. Gehtland, Spring: a developer's notebook. O'Reilly Media, Incorporated, 2005.

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

[15] E. Hewitt, Cassandra: the definitive guide. O'Reilly Media, Incorporated, 2010.

[16] A. Lakshman and P. Malik, "Cassandra: a decentralized structured storage system," SIGOPS Oper. Syst. Rev., vol. 44, no. 2, 2010, pp. 35–40.

[17] T. White, Hadoop: The definitive guide. O'Reilly Media, 2012.

[18] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST). Washington, DC, USA: IEEE Computer Society, 2010, pp. 1–10.

[19] D. Smiley and E. Pugh, Solr 1. 4 Enterprise Search Server: Enhance Your Search with Faceted Navigation, Result Highlighting, Fuzzy Queries, Ranked Scoring, and More. Packt Publishing, 2009.

[20] O. Gospodnetic and E. Hatcher, Lucene. Manning, 2005.

[21] K. Chodorow, MongoDB: the definitive guide. O'Reilly, 2013.

[22] MongoDB Aggregation, http://docs.mongodb.org/manual/aggregation, accessed: 2014-06-09.

[23] nginx, http://nginx.org, accessed: 2014-06-07.

[24] "RabbitMQ Messaging," http://www.rabbitmq.com/, accessed: 2014-06-07.

[25] A. Durham, E. Sussumu, and A. da Conceição, "A framework for building language interpreters," in Companion of the 18th annual ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications (OOPSLA). Educators Symposium. ACM, 2003, p. 196.

[26] Simple Framework, http://simple.sourceforge.net, accessed: 2014-06-07.

[27] A. Tassanaviboon and G. Gong, "OAuth and ABE based authorization in semi-trusted cloud computing: aauth," in Proceedings of the second international workshop on Data intensive computing in the clouds, ser. DataCloud-SC '11. New York, NY, USA: ACM, 2011, pp. 41–50.

[28] A. da Conceição, R. Pereira, J. Rezende, B. Silva, R. Correia, H. Domingues, R. Kon, and F. Kon, "Projeto Borboleta: Ferramentas Móveis e Multimídia para Atenção Básica Domiciliar," in Congresso Brasileiro de Informática em Saúde. Artigo curto, 2008.

[29] A. El-Rabbany, Introduction to GPS: The Global Positioning System. Artech House Publishers, 2002.

[30] Maritaca Team, "Maritaca Source Code," http://sourceforge.net/p/maritaca, accessed: 2014-06-07.

[31] Financiadora de Estudos e Projetos (FINEP), http://www.finep.gov.br, accessed: 2014-06-07.

 ISBN: 978-1-61208-366-7

Car Ride Classification for Drive Context Recognition

Stefan Haas
Institute for Informatics
LMU Munich
Munich, Germany
Email: haasst@cip.ifi.lmu.de

Kevin Wiesner
Institute for Informatics
LMU Munich
Munich, Germany
Email: kevin.wiesner@ifi.lmu.de

Thomas Christian Stone
BMW Group
Munich, Germany
Email: thomas.stone@bmw.de

***Abstract*—The automotive domain, with its more and more increasing number of comfort and infotainment functions, offers a field of opportunities for learning and context-sensitive functions. In this respect, personal and frequent trips of drivers provide very promising and interesting contexts. To identify frequent driving contexts in a set of recorded GPS tracks, this paper presents two different clustering algorithms: First, a hierarchical *Drive-Clustering*, which combines drives based on their number of common GPS points. Second, a *Start-Stop-Clustering*, which combines trips with the same start- and stop-cluster utilizing density based clustering. Especially the *Start-Stop-Clustering* showed particularly good results, as it does not depend on the concrete routes taken to a stop position and it is able to detect more trip clusters. To predict these trip contexts, a Bayesian network is presented and evaluated, with logged trip data of 21 drivers. The Bayes classifier uses context information such as the time, weekday and the number of persons in the car, to predict the most likely trip-context and thus achieves a good accuracy in the prediction of the different trip contexts.**

Keywords–Context-aware Vehicle; Spatial Clustering; Drive Context Prediction

I. INTRODUCTION

Context-awareness is an important building block in the development of intelligent systems as it can significantly improve the interaction between a user and a system. Any information that enables a system to provide the user with useful, context-related information or intelligent behavior, can be considered a context. Knowledge about a specific context is normally gathered by sensor readings and their interpretation [1][2].

With its steadily increasing number of comfort and infotainment functions, the automotive domain offers a unique field of opportunities for learning and context-sensitive functions. In recent years, many different context-aware advanced driver assistance systems (ADAS) have already been introduced. They are based on information which is provided by dedicated sensor systems, especially in the areas of safety and comfort, like the lane departure warning system (LDW), adaptive cruise control (ACC) or intelligent speed adaption (ISA).

Another interesting and promising context to advance vehicle personalization is the drive itself. Above all, the repeated drives of a person offer a lot of potential for finding consistent usage patterns and subsequently the possibility of automating recorded user behavior after a certain learning period. For example, if a driver usually checks his mail on the way to work or likes to listen to the news, the vehicle could adapt to his preferences by recognizing the drive context as a regularly drive to work and by automating the desired functions. This automation of functions could improve safety as well as comfort because the driver is no longer forced to adjust his personal settings by himself.

In the following, we will describe and evaluate different methods for the detection and prediction of repeated drives of individual drivers. To develop and evaluate our proposed methods, we had the possibility of utilizing recorded vehicle sensor data of 21 drivers collected over several months by a data logger. The collected data included many different sensor signals exchanged between the different in-car electronic control units (ECU) over the Controller Area Network (CAN) bus, ranging from Global Positioning System (GPS) position to seat belt status.

The contributions of our paper are two novel clustering methods for detecting repeated trips of individual drivers, a novel distance measure based on the Jaccard distance for comparing GPS tracks and a hybrid Bayesian network for predicting frequent drive contexts right away from the start of the trip based on contextual information like the time of the day or the number of passengers in the car.

The paper is structured as follows. Section II gives an overview on existing work in the fields of route prediction, route recognition, destination prediction and place mining. Section III outlines two new spatial clustering methods for detecting the frequent drive contexts of a particular driver. In Section IV, we present a hybrid Bayesian network to predict the frequent drive contexts of an individual driver right away from the start of the trip. The results we obtained running the before presented algorithms individually on the collected drive data of every single driver are described in Section V. We close our work in Section VI with a summary and an outlook on possible future work.

II. RELATED WORK

Route recognition and prediction systems have been proposed in many different works [3][4][5][6][7]. In the majority of these publications, the general way to predict respectively recognize the current route is based on the comparison of the current driving trajectory against previously recorded trajectories using a distance measure. As comparing GPS tracks can not be done with classic L_p metrics due to their length related inequality, dimension and noise, novel more elastic distance measures are needed. Already proposed distance measures, were for example, based on the longest common sub-sequence (LCSS) algorithm [3][8][9], the Hausdorff distance [4] or the Jaccard distance [10]. In [8], this simple instance based learning approach of comparing the current route to already recorded routes is further enhanced by the inclusion of contextual information (e.g. time of the day) to better differentiate overlapping routes.

 ISBN: 978-1-61208-366-7

Probabilistic approaches for route and destination prediction have been presented amongst others in [10][11][12] and [13]. The investigated prediction methods hereby often underlie a Bayesian approach and include additional contextual information like the time of the day, the particular weekday or even background information about locations to infer the most likely route or destination [13]. In [12], a Markov model is used instead of a Bayesian approach to predict the next location of a user.

Identifying personally important places of users in recorded GPS data has for example been investigated in [14][15][16][12] and [7]. Density based clustering hereby proved more efficient than classic partitioning algorithms like k-means [17][18][14][15], as the final clusters only consist of dense regions in the data space. Regions of low object density are not included in the final clusters and are considered as noise.

Our work differs from existing publications, as we focus on the personal repeated drives of individual drivers and their prediction. We thereby consider a set of similar drives included in a repeated drive cluster as a certain drive context and as a basis for learning and automating user settings to advance comfort and safety.

III. Detecting frequent drives

To detect frequent drive clusters of an individual driver, we present and evaluate two different spatial clustering methods explained in the following two Subsections. *Drive-Clustering* is based on the Jaccard distance and compares whole trajectories using hierarchical clustering, whereas *Start-Stop-Clustering* focuses on semantically similar routes based on the before determination of frequent start and stop positions of the particular driver. The goal of both algorithms is to identify repeated patterns in the set of recorded GPS tracks in order to detect repeatedly occurring drive contexts, e.g., drives from home to work. In Section V, we compare the obtained results of both algorithms applied to our test data set.

A. Drive-Clustering

An important factor in cluster analysis is a distance measure to determine the distances between elements contained in the data, for the purpose of grouping similar elements together in clusters. In trajectory data the standard way for identifying patterns is to compare whole trajectories. In our case, the trajectory data of each drive is stored as a sequence of GPS points $S_i = \{p_{i,1}, p_{i,2}, ..., p_{i,n}\}$, with $p_{i,1}$ being the start point of the drive and $p_{i,n}$ being the end or stop point.

To compare two point sequences we use a dissimilarity measure based on the well known Jaccard distance, which measures dissimilarity between sample sets [19] (see equation 1):

$$d(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}. \quad (1)$$

Our dissimilarity measure thereby calculates the intersection of the two GPS sequences S_i and S_j by counting the number of common points $NOCP(S_i, S_j)$ contained in both sequences starting from the shorter sequence (see equation 2). This number of common points value is then divided by the number of points contained in the shorter sequence $min(S_i, S_j)$. In order to obtain a dissimilarity measure the whole term is subtracted from 1, so that a result of 0 signifies maximum similarity and a value of 1 maximum dissimilarity.

$$d(S_i, S_j) = 1 - \frac{NOCP(S_i, S_j)}{min(S_i, S_j)}. \quad (2)$$

GPS points of two geometrically similar trajectories are very unlikely to have the exact same coordinates, due to different driving speeds and other noise. Hence it is necessary to define a threshold Θ from which two points are considered as equal or contained in both sequences (common points), e.g., 50 meters. The threshold needs to be defined dependent on the logging frequency. In our case the logging frequency is $f = 1Hz$. So when we for example consider 135 km/h as the maximum vehicle speed, the maximum distance between two succeeding points will be $(135*1000)m/3600s = 37.5m$. In the evaluation we set the threshold to 50 meters, which is sufficient for driving speeds up to 180 km/h with a logging frequency of $f = 1Hz$.

The number of common points (NOCP) algorithm iterates over all points $p_{i,k} \in S_i$ included in the shorter sequence and tries to find at least one point in the other sequence $p_{j,l} \in S_j$ whose distance is less or equal than the defined threshold distance Θ. If the set of found points in range is not empty, the number of common points counter is increased. Consequently, the presented distance measure is more elastic than distance measures based on dynamic programming, like the longest common sub-sequence (LCSS) or dynamic time warping (DTW), as it is able to match several elements of one sequence to just one element of the other sequence. This behavior is important in our case to handle traffic jams and different driving speeds. The implementation of the number of common points (NOCP) function can be significantly sped up by storing the queried sequences' points in a *k-d tree* [20].

To calculate the distance between two-dimensional GPS points we use a simplification of the *haversine* formula [21] based on the euclidean distance, which in contrast to the standard euclidean distance allows metric parametrization of our algorithms (ϕ latitude, λ longitude) (see equation 3).

$$\begin{aligned} dist(\phi_1, \lambda_1, \phi_2, \lambda_2) = (((111.3 * \cos(\frac{\phi_1 + \phi_2}{2}) * \\ (\lambda_1 - \lambda_2)^2) + (111.3 * (\phi_1 - \phi_2)^2))^{\frac{1}{2}} * 1000. \end{aligned} \quad (3)$$

In order to avoid the problem of a very much shorter sequence being contained in a longer sequence and to speed up the comparison, the number of common points in the two sequences is only calculated, when the start and stop points of the two sequences are sufficiently similar, e.g., their respective distances do not exceed 250 meters ($p_{i,1} \sim p_{j,1}$ and $p_{i,n} \sim p_{j,m}$). Otherwise the maximum dissimilarity value 1 is returned without any further calculation (see equation 4).

$$d_{opt}(S_i, S_j) = \begin{cases} 1 - \frac{NOCP(S_i,S_j)}{min(S_i,S_j)}, & \text{if } p_{i,1} \sim p_{j,1} \\ & \wedge\ p_{i,n} \sim p_{j,m} \\ 1, & \text{otherwise} \end{cases}. \quad (4)$$

To group similar drive contexts in clusters, we use agglomerative hierarchical clustering, starting from single GPS sequences. To stop the calculation when no sequence anymore undercuts a distance ε to another sequence we need to define

 ISBN: 978-1-61208-366-7

a similarity threshold, e.g., $\varepsilon = 0.05$. The smaller the value ε the more similar are the trips contained in a cluster. This threshold will cut the *dendrogram* at a certain level and lead to the final drive clusters. To predefine the minimum cluster size we use another parameter $MinDrives$, referring to the $MinPoints$ parameter in density based clustering [18].

B. Start-Stop-Clustering

Another way of determining frequent drive contexts of a certain driver is based on his frequent start and stop positions. In contrast to the above presented trajectory clustering method this method rather focuses on semantically similar drives with the same start and stop positions than on geometrically similar drives or routes.

As the vehicle is typically not parked at the exact same coordinates, it is necessary to merge similar parking positions to *start-stop-clusters*. To obtain these frequent start and stop position clusters of a particular driver, we use density based clustering, to be exact the DJ-Cluster algorithm presented in [14], which is a simplification of DBSCAN [18] [22]. Density based clustering has the advantage of explicitly eliminating outlier points compared with partitioning clustering, e.g., k-means [17] [22]. As we are only interested in dense regions included in the set of start and stop positions of an individual driver in order to identify frequent drive contexts, density based clustering is suitable for our task.

Consequently, the first step in Start-Stop-Clustering is to calculate dense regions of start and stop positions in the set of GPS sequences and to store the cluster IDs of every GPS sequences' start and stop points. Therefore, it is necessary to specify the two parameters $MinPoints$ and ε, representing the minimum cluster size and search radius respectively. Figure 1 shows an example of a dense point cluster found in the drive data of a particular driver with $\varepsilon = 100$m.

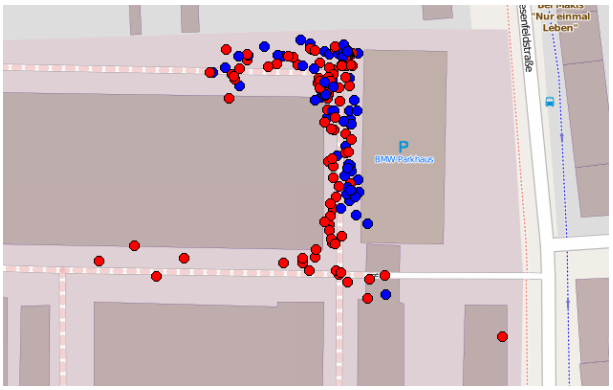


Figure 1. Visualization of the start (red) and stop points (blue) of a driver. All shown points are included in the same point cluster.

The binary dissimilarity measure for Start-Stop-Clustering then looks as follows (see equation 5):

$$d(S_i, S_j) = \begin{cases} 0, & \text{if } C_s(p_{i,1}) = C_s(p_{j,1}) \\ & \wedge\ C_e(p_{i,n}) = C_e(p_{j,m})\ . \\ 1, & \text{otherwise} \end{cases} \quad (5)$$

Two GPS sequences S_i and S_j are considered as equal, when their corresponding start ($p_{i,1}$, $p_{j,1}$) and stop points ($p_{i,n}$, $p_{j,m}$) lie in the same start C_s respectively end cluster C_e.

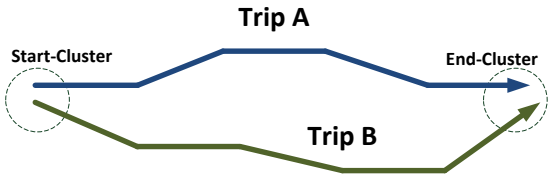


Figure 2. Illustration of a route-independent *Start-Stop-Cluster*.

Hence, the final drive clusters are comprised of GPS sequences whose start and stop points lie in the same dense region or point cluster and therefore have the same cluster IDs. The found frequent drive contexts are direction-dependent just like those obtained with the above presented Drive-Clustering approach. However, the drives included in a *Start-Stop-Clustering* drive context cluster do not necessarily follow the same routes. In contrast to *Drive-Clustering* they are route-independent (see Figure 2). To predefine the minimum cluster size we also use the $MinDrives$ parameter.

IV. Predicting Frequent Drive Contexts

To predict frequent drive contexts that have been identified with one of the above presented methods, we propose a hybrid Bayesian network. The structure of the network is shown in Figure 3.

The goal is to predict a present frequent driving context, e.g., a drive to work, as early as possible during the drive. Therefore we make use of contextual information associated with a certain drive context cluster. The contextual information used to infer the current drive context includes the start point of the drive, the number of passengers in the car, the weekday, the start time and the fuel level.

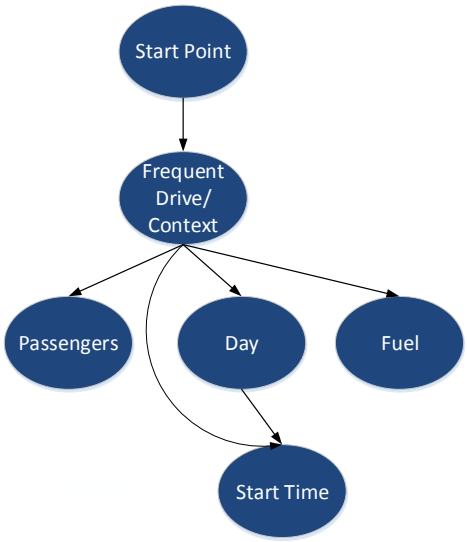


Figure 3. Topology of the hybrid Bayesian network fo predicting the most likely frequent drive context.

Using the start point of the drive we are able to eliminate impossible contexts, e.g., a drive from work to home if the start point is home, which significantly reduces the possible

 ISBN: 978-1-61208-366-7

contexts, prevents false positives and speeds up the implementation. The variable *Frequent Drive/Context* represents the *a priori* probability distribution over the set of identified drive contexts, already constrained by the current start point. The variables *Day*, *Passengers* and *Fuel* are conditionally independent of each other given the *class* variable *Frequent Drive/Context*. The variables described so far all underlie a discrete probability distribution.

In contrast to the other probability variables, we model the variable *Start Time* as continuous. By the edges between *Frequent Drive/Context*, *Day* and *Start Time* we receive a drive context dependent start time *probability density function* (PDF) for every single day. This enables a stronger differentiation between the drive contexts, as the start time probabilities for the different contexts are also day dependent.

To approximate the probability density function for the start times associated with a certain drive context we use *kernel density estimation* (KDE) (equation 6) with a Gaussian kernel (equation 7) and Scott's *rule of thumb* (equation 8) for bandwidth selection h [23]:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}). \quad (6)$$

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}). \quad (7)$$

$$h_{scott} = n^{-1/(d+4)}. \quad (8)$$

By using *kernel density estimation* we receive continuous day and context dependent probability density functions for the start times, with high probabilities during day times the drive context normally occurs (see figure 4).

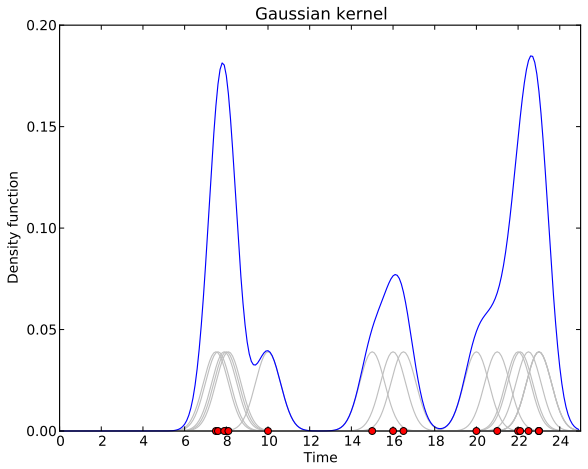


Figure 4. Example of a probability density function for the *Start Time* variable of a particular drive context.

We deliberately do not use *Laplacian correction* to deal with zero probabilities. When a drive context has not occurred before, at a certain day or time, the probability for the whole context will be zero. This helps in preventing false positives.

The probability for a certain context C, given the start point s, the weekday d, the time t, the number of persons in the car p and the fuel level f, can then be calculated with the following formula:

$$P(C|s,d,t,p,f) \propto P(C|s)P(d|C)P(t|d,C)P(p|C)P(f|C). \quad (9)$$

The context C_i leading to the highest probability value $P(C_i|s,d,t,p,f)$ is then assumed to be the present context:

$$\arg\max_{C_i}\{P(C_i|s,d,t,p,f)\}. \quad (10)$$

V. Evaluation

To evaluate the described methods, we had access to a data set collected by 21 drivers over several months. The logger used for collecting the data records all kinds of data bus traffic, also when the car is not moved, e.g., when the electronic key is pressed. To filter out this unwanted noise, we only used recorded data for our evaluation where the vehicle was at least moved 1 kilometer (air-line distance). The minimum number of filtered drives of one driver was 216, the maximum number 986. The majority of the probands ranged between 400 to 600 recorded drives.

A. Drive clustering

Figures 5 and 6 show the results obtained applying Start-Stop-Clustering and Drive-Clustering to the data set. Figure 5 illustrates the average number of found clusters for different minimum cluster sizes (MinDrives={3,5,10}). Figure 6 presents the average share of repeated drives of the total quantity of drives.

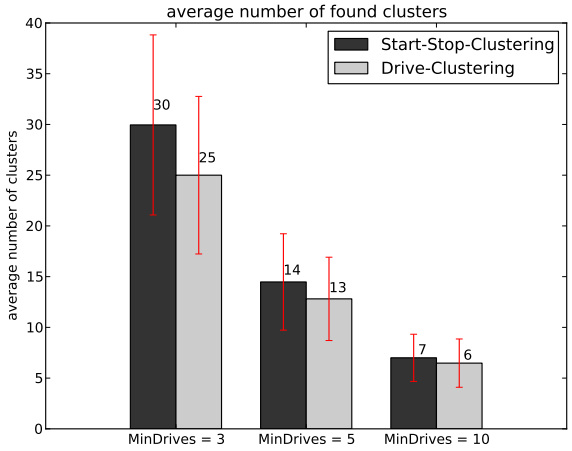


Figure 5. Average number of found clusters with Start-Stop- and Drive-Clustering dependent on the minimum number of drives contained in the clusters (MinDrives).

As one can see, Start-Stop-Clustering is on average able to identify more clusters than Drive-Clustering (see Figure 5). However, with increasing the minimum cluster size, the difference between the average number of found clusters by Start-Stop-Clustering and Drive-Clustering decreases. This leads to the assumption that for frequent drives (MinDrives=10), drivers usually have a preferred route that they normally take, whereas

 ISBN: 978-1-61208-366-7

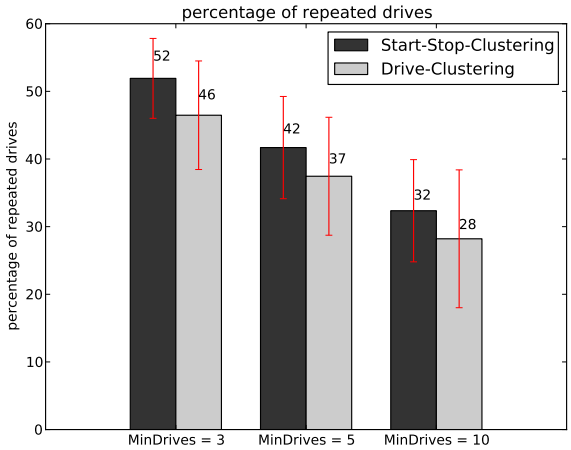


Figure 6. Percentage of repeated drives identified with Start-Stop- and Drive-Clustering dependent on the minimum number of drives contained in the clusters (MinDrives).

for less frequent drives (MinDrives=3) they also take different routes to the same destination. In addition to the number of found clusters, Start-Stop-Clustering is on average able to assign a larger fraction of the overall number of drives to a repeated drive cluster compared to Drive-Clustering, as it also includes all route alternatives (see Figure 6).

As we are rather interested in detecting frequent drive contexts than the frequent routes taken by a driver, Start-Stop-Clustering is more appropriate for our use case. Especially large clusters (MinDrives $\geq$ 10) may provide promising and interesting contexts, on the basis of which usage patterns may possibly be learned and automated. The average fraction of trips repeated at least 10 times by the participants during the survey amounts to approximately 30% of the overall trips (see Figure 6).

To keep the set of frequent driving contexts up-to-date one could use a shifting time frame and only consider drives for the cluster calculation that for example occurred during the last 6 months. This would lead to a slow exclusion of no longer appearing driving contexts over time and also limit the amount of data used for the context identification.

B. Prediction

To evaluate our proposed Bayesian inference system for predicting frequent drive contexts, we made use of cross-validation and focused on clusters identified by Start-Stop-Clustering with a cluster size larger than 10 drives.

Figure 7 shows the overall prediction result for all drives, including also non-frequent drives, as well as the prediction result for solely frequent drives belonging to a cluster. The prediction result improves significantly, to almost 100% ($\sim$97%), when a prediction result is considered correct when lying within the top 3 predictions.

The differentiation between the different drive contexts is relatively accurate ($\sim$ 89% respectively $\sim$97% for top 3 matches). Moreover, in Figure 8 one can see that, when considering all drives, the main share in false predictions not lying within the top 3 matches is produced by false positives. A large fraction of false positives could be detected correctly

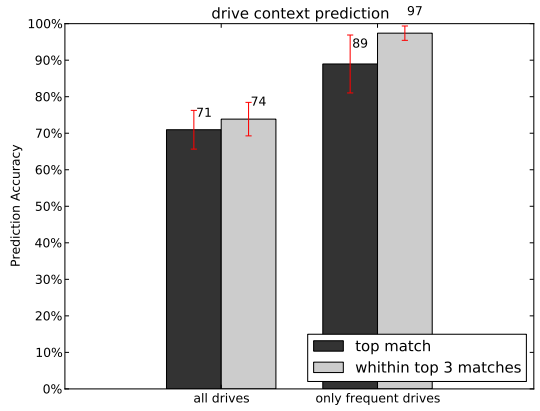


Figure 7. Prediction result for all drives and only frequent drive contexts (MinDrives=10).

($\sim$60%), but as there might be highly frequented start and stop positions like home, with overlapping context information, e.g., time and weekday, some infrequent drives were predicted as belonging to a frequent drive context.

In the evaluation we used a binary probability distribution for the day variable (workday, weekend) due to the relatively small minimal cluster size of 10 drives. It might be possible to achieve a better recognition of infrequent drives by assuming a discrete probability distribution for every day (Monday, Tuesday, Wednesday, etc.), which would also lead to time probabilities for every day for each drive context. However, this would only make sense with a higher minimal cluster size, in order to get representative probability distributions for every day.

Figure 8. Overall prediction error rate and the share of false positives at the overall error rate.

Compared to the rate of false positives the rate of true negatives is extremely low and underlines the accuracy of our inference system related to the prediction of frequent drive contexts (see Figure 8). However, eliminating false positives is crucial in order to not annoy the driver with unwanted function automation and might only be solvable with little driver interaction. A solution could be providing the driver with the top 3 most likely contexts and letting the driver decide

 ISBN: 978-1-61208-366-7

if one is appropriate for him in the current situation. If none is selected by the driver after a certain driving time the system assumes that in the current situation no function automation is wanted by the driver.

VI. Conclusion

In this paper, we investigated the detection and prediction of frequent drive contexts as an important building block for vehicle personalization. We proposed two different spatial clustering approaches for identifying frequent drive patterns in a GPS data set. Especially the route independent Start-Stop-Clustering is promising, as it is able to detect frequent drive patterns independently of the chosen route. The presented Bayesian inference systems accuracy in differentiating frequent drive contexts was about 89% respectively 97% for a top 3 match. Future work will consist of linking context information and adaptive function automation together, as well as in in-car field and acceptance tests.

Acknowledgment

The authors would like to thank the participants of the study and BMW Group for providing the drive data. All location and drive data was anonymized to ensure the participants' privacy.

References

[1] G. D. Abowd et al., "Towards a better understanding of context and context-awareness," in Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing, ser. HUC '99. London, UK, UK: Springer-Verlag, 1999, pp. 304–307.

[2] A. Schmidt, "Ubiquitous Computing - Computing in Context," Ph.D. dissertation, Lancaster University, November 2002.

[3] O. Mazhelis, "Real-time recognition of personal routes using instance-based learning," in IEEE Intelligent Vehicles Symposium (IV 2011), 2011, pp. 619–624.

[4] J. Froehlich and J. Krumm, "Route prediction from trip observations," in Proceedings of the Society of Automotive Engineers (SAE) 2008 World Congress, SAE Technical Paper 2008-01-0201, April 2008, pp. 1–13.

[5] D. Tiesyte and C. S. Jensen, "Similarity-based prediction of travel times for vehicles traveling on known routes," in Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ser. GIS '08. New York, NY, USA: ACM, 2008, pp. 14:1–14:10.

[6] A. Brilingaite and C. S. Jensen, "Online Route Prediction for Automotive Applications," in Proceedings of The 13th World Congress and Exhibition on Intelligent Transport Systems and Services (ITS 2006), London, October 2006, pp. 1–8.

[7] K. Torkkola, K. Zhang, H. Li, H. Zhang, C. Schreiner, and M. Gardner, "Traffic Advisories Based on Route Prediction," in Proceedings of Workshop on Mobile Interaction with the Real World, 2007, pp. 33–36.

[8] O. Mazhelis, I. Žliobaite, and M. Pechenizkiy, "Context-aware personal route recognition," in Proceedings of the 14th international conference on Discovery science, ser. DS'11. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 221–235.

[9] M. Vlachos, G. Kollios, and D. Gunopulos, "Discovering similar multidimensional trajectories," in Data Engineering, 2002. Proceedings. 18th International Conference on, 2002, pp. 673–684.

[10] K. Laasonen, "Route Prediction from Cellular Data," in Proceedings of the Workshop on Context-Awareness for Proactive Systems (CAPS). Helsinki, Finland: University Press, 2005, pp. 147–158.

[11] K. Tanaka, Y. Kishino, T. Terada, and S. Nishio, "A destination prediction method using driving contexts and trajectory for car navigation systems," in Proceedings of the 2009 ACM Symposium on Applied Computing, ser. SAC '09. New York, NY, USA: ACM, 2009, pp. 190–195.

[12] D. Ashbrook and T. Starner, "Using gps to learn significant locations and predict movement across multiple users," Personal Ubiquitous Comput., vol. 7, no. 5, Oct. 2003, pp. 275–286.

[13] J. Krumm and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," in In Ubicomp, 2006, pp. 243–260.

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

[15] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering Personally Meaningful Places: An Interactive Clustering Approach," ACM Trans. Inf. Syst., vol. 25, no. 3, July 2007.

[16] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," in Proceedings of the 2Nd ACM International Workshop on Wireless Mobile Applications and Services on WLAN Hotspots, ser. WMASH '04. New York, NY, USA: ACM, 2004, pp. 110–118.

[17] J. B. Macqueen, "Some methods of classification and analysis of multivariate observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[18] M. Ester, H.-P. Kriegel, J. S, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." AAAI Press, 1996, pp. 226–231.

[19] M. Lewandowsky and D. Winter, "Distance between sets," in Letters to nature. nature publishing group, 1971, pp. 34–35.

[20] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," ACM Transactions on Mathematics Software, vol. 3, no. 3, September 1977, pp. 209–226.

[21] R. W. Sinnott, "Virtues of the Haversine," Sky and Telescope, vol. 68, no. 2, 1984, pp. 159+.

[22] J. Han, M. Kamber, and A. K. H. Tung, "Spatial clustering methods in data mining: A survey," in Geographic Data Mining and Knowledge Discovery, Research Monographs in GIS, H. J. Miller and J. Han, Eds. Taylor and Francis, 2001, pp. 201–231.

[23] D. W. Scott and S. R. Sain, "Multi-Dimensional Density Estimation". Amsterdam: Elsevier, 2004, pp. 229–263.

 ISBN: 978-1-61208-366-7

What am I Doing Now? Pythia: A Mobile Service for Spatial Behavior Analysis

Amnon Dekel, Tomer Weller, Hanny Bar,
Cadan Ojalvo
Department of Software Engineering
Shenkar: Engineering, Design, Art
amnoid@gmail.com, tomer.weller@gmail.com,
barhanny@gmail.com, cadan85@gmail.com

Scott Kirkpatrick, Benjamin Kessler
Selim Benin School of Engineering
and Computer Science
The Hebrew University, Jerusalem, Israel
kirk@cs.huji.ac.il, benjy.kessler@gmail.com

***Abstract*—Pythia is a prototype hybrid Mobile/Cloud service for ascertaining what the user of a mobile phone is currently doing. The service continuously captures and uploads context data to an analysis engine in the cloud. Early field-testing showed that the service can categorize activities into working, at home, traveling, or shopping.**

Keywords-Mobile Context Awareness; Mobile Context Capture; Mobile Applications; Mobile Services; Cloud Services.

I. INTRODUCTION

In the last few years, the use of context awareness as a way of enabling a mobile application to react and perform in relevant ways to the needs and expectations of users has picked up [12][14]. With mobile smart phones containing multiple sensors, the platform has become more capable of capturing data that can be used to identify context. A number of applications have emerged in the last few years that attempted to use this data in order to release the user from having to specifically tell the application what they are doing [2][5][11][13]. Being able to discover the user context is valuable and can enhance and streamline the service being offered.

The work in the area has generated a number of specialized mobile applications on the major smart phone platforms (Android, IOS, Windows phone) that claim to capture signals and recognize user context in an effort to streamline the user experience within specific use cases. The major use cases that have been implemented are driving versus parking in the physical domain (and in the process remembering where the user last parked their car) [2], the personalization of a user's home screen depending on the currently understood context (i.e., presenting a specific set of application icons on the phone desktop when at work versus a different set of application icons when at home [5]), and the presentation of search results that are relevant for the user's current context [11].

We tested some of the available services and found them to be an interesting start, but also lacking in consistent context recognition or transparency. For example, the AGENT [2] application suffered from too many false negatives when parking, leaving the phone in a driving context after we had left the car and thus was not able to remember where the car was last parked. The personalized desktop application COVER [5] that presents a differing set of home screen application icons that it deems to be relevant to the current user context, caused us to feel confused about where the needed applications icons were located on the home screen. This was probably the result of a drastic change in context without the user being an active participant in the change.

Two current services which provide a more consistent and clear experience are MOVES [13] and Google Now [11]. The MOVES application is an activity tracking application that tracks it's users "everyday life" [13]. By gathering sensor based data, it can track a user's activities such as walking, cycling and running, while identifying where these activities take place. It then visualizes its user's day. Google NOW is the most ubiquitous context based mobile service today since it is part of the ubiquitous Android operating system. It is a sophisticated perpetual search service on android that continuously analyses a user's searches, email, calendar and location and uses these signals to present contextually relevant information cards at appropriate times. For example, in the morning it presents a card showing how long it will take to get to work in the current traffic situation (see Figure 1).

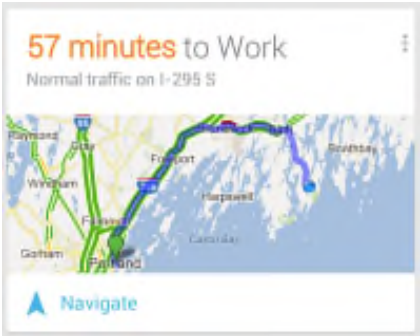


Figure 1. Google Now Context Relevant Card

In both the MOVES application and Google NOW, the actual methodologies for data capture and analysis are proprietary and therefore remain the property of the companies developing them. Although Google makes some of these capabilities available to developers via Application

 ISBN: 978-1-61208-366-7

Programming Interfaces that it opens every once in a while [8], the algorithms remain proprietary, secret and under the control of their owners.

In this paper we present the Pythia system. We start with defining the objectives of the project and then describe the Pythia system. We continue with a description of the methods we used to test the system and end with conclusions and a description of future work to be carried out.

II. OBJECTIVES

With Pythia, we are striving to develop a mobile and cloud based service for capturing contextual signals and ascertaining the current user context. Because of a lack of context capture consistency and transparency in existing mobile applications, the goal of the project is to learn about and improve the capability of a mobile-based service to arrive at better context classifications.

Pythia was developed as part of a research effort to tap the intimate relationship that people have with their phones in order to learn about their activities and to use this knowledge in order to make the phone a smarter companion in our everyday lives.

III. THE PYTHIA SYSTEM

Pythia is a hybrid mobile/cloud service that captures packages and uploads data to a web-based repository where the data is parsed, cleaned, normalized and classified. All Pythia data is stored on a MongoDB instance. The Pythia server is a lightweight Node.js application running in a hosted Heroku service. Data is hosted on MongoHQ (see Figure 2).

The phone client includes a management User Interface (see Figure 3) to enable the user to set up the data capture resolution and upload dynamics (i.e., what is the minimal distance traveled that the system will save and upload to the online service), and a background service (see Figure 4) that gathers location data and uploads it to the data repository according to the application settings. The background service shows the amount of events it has captured and is ready to upload to the server on the next synchronization process (”events in pipe”).

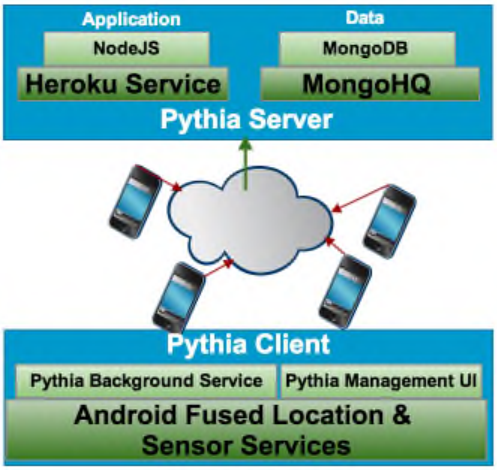


Figure 2. Pythia System Overview

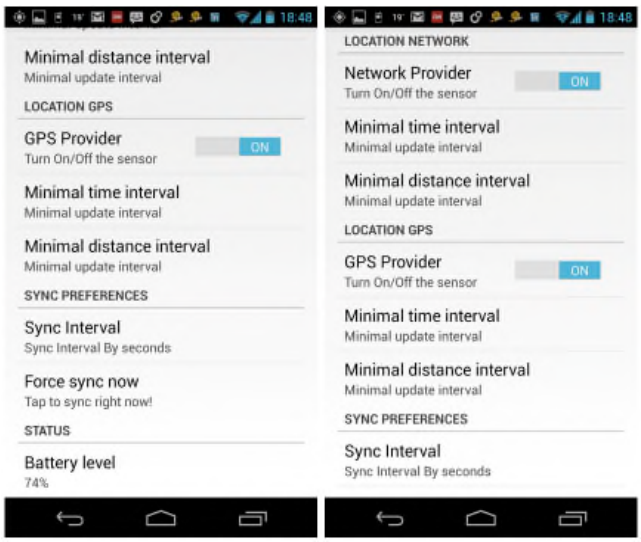


Figure 3. Pythia Management UI

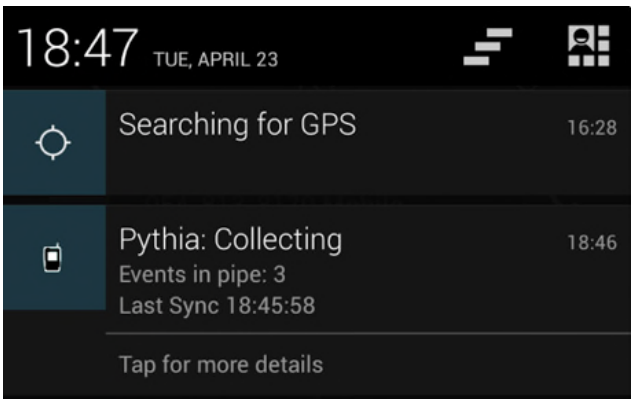


Figure 4. Pythia background Service

Note that the management UI was designed for our initial technical field trial and is not designed for normal users. Future iterations of the service will be released to larger numbers of users and will include a more streamlined and simple user interface that will be easy to use by anyone.

The Pythia server receives the sensor data (GPS or Cell based location), from the participating mobile clients and performs a spatial classification process on the data. This is achieved with a series of three map/reduce [6] computations:

1. **Pre-processing:** Similar location events are grouped by their bounding timestamp and geobox and then these aggregated events are indexed using geohashing- a method of transforming a 2d point into a 1d hash, which allows for easy indexing and simple geoboxing of locations. This process reduces the size of the data while filtering out noise caused due to the instability of the mobile device location reads.

2. **"HotSpot" aggregation:** Grouping of aggregated location events by bounding geobox, and

 ISBN: 978-1-61208-366-7

calculating the sum of timeframes per geobox and then sorting by this sum. Geoboxes with highest timeframe count represent the users most visited locations.

3. **"HotSpot-by-hour-slot" aggregation:** Same as "HotSpots" but also assigning time of day to geoboxes. Allows us to see the user's most visited locations per hour.

Once the sensor data per phone was processed as described, the next step was to ascertain a context for each location and timeslot. For this version, we kept the context to the following: *Home*, *Travel*, *Work* and *Shopping*. It is clear that this classification is very rough, but we think that being able to identify these prototypical contexts is a valuable step in the process of being able to identify more refined contexts later on.

IV. METHOD

An initial version of the Pythia data capture service was installed on two Motorola Razr's and two Nexus 4 phones. The devices continued to be used as active phones for a week while capturing and uploading the data to the Pythia server.

In this version, the service was only used to gather location and sensor data. Apart from being able to turn the service on or off, or to tweak the sensor gathering resolution, the user did not get any information from the application itself. The uploaded data was then processed at the server end with the goal of ascertaining what the user was doing throughout the period of data capture.

The goal of the service was to establish if it was feasible to use the sensor information in order to conclude what context the user is in, while keeping within a viable power consumption envelope. It is clear to us that a service that can conclude a context correctly, while using too much power, will not be useful, and at the same time, a service that is very power efficient but cannot ascertain the current context is not useful either. The sweet spot is to be power efficient while correctly ascertaining context. We define power efficiency as a power consumption envelope that does not visibly lower the service life of a smart phone in normal use to below a full day of use. A power consumption envelope that lowers the daily service life of a phone will simply be discarded by most users.

In this version, the process of ascertaining user activity context was semi-automatic and necessitated human supervision. The service would automatically aggregate and run through the list of locations and timeslots and then receive human supervision as to the naming of the current context. Once enough of these were seen by the system it was able to classify contexts. Thus, after seeing that a specific location was always classified as Home by the supervisor, the service would be able to reach the same conclusion going forward.

V. INITIAL RESULTS

The first version of the service was quickly seen to be overly power hungry and severely shortened the daily service life of the phone by 50%. This was unacceptable. Google then released a new sensor capture framework (called Fused Location Provider [10]) that was purported to be more power efficient. After implementing it in the Pythia client we were able to lengthen the service time to the full day threshold we expected and enabled us to continue into a 4-week field test with the 4 phones that continued to be used in a normal fashion by their owners.

Figure 5 shows the distribution of classified activity contexts over a representative 24-hour period. For each of the participants, we compared the resulting activity map to their calendar and notes as well as interviews. The comparison showed us that the classifications were correct.

Figure 6 shows the total averaged activity distribution as identified over the course of the process. In this case, we averaged identified activities over the time of the trial and the most frequent activity per timeslot became the selected classification. These were mapped onto a 24-hour representative period. Note that we analyzed only normal working days, ignoring weekends and vacations.

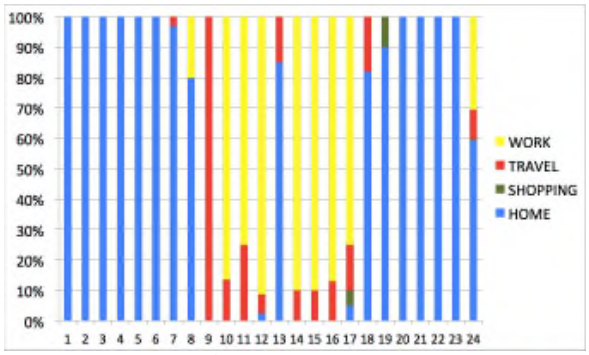


Figure 5. Classified Activity Context over 24 hours
Y-axis: the percentage of each activity during a specific hour
X-axis: The hour in the day

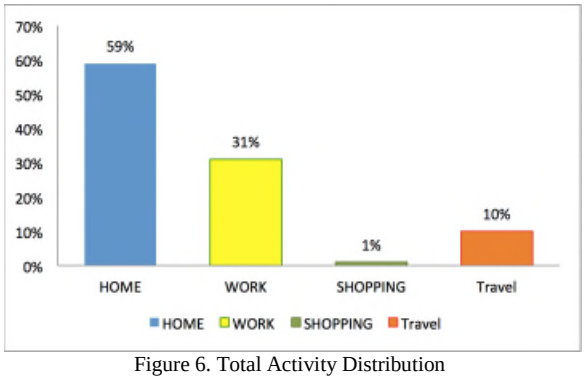


Figure 6. Total Activity Distribution

Analyzing the results presented above, we conclude that within the constraints described, the system was able to classify the major activities across time.

VI. CONCLUSION AND FUTURE WORK

The Pythia system analyzed an aggregated database of user locations over time and used this in order to classify what the user was doing at each location during each time slot. We believe that being able to semi-automatically identify what a user is doing (their personal context) is an important signal that can help in making our mobile phones more useful and helpful in our daily lives.

While we have shown that the Pythia system is able to classify activity contexts related to location and time, we believe that ascertaining context via location alone may be too coarse as a signifier in many cases. Because of this, we are now in the processes of improving the system in the following ways:

1. Widening the net of possible activity context classifications that the service can identify.
2. Minimizing the need for human supervision in the classification.

We believe that both of these improvements can be achieved by supporting additional sensors in the system. Additional *physical sensors* can help to an extent, i.e., using physical activity recognition to identify if the user is stationary, walking, running, riding a bike or travelling in a vehicle, is valuable, but has a limited capability for refined context discovery. Similarly, the use of audio analysis can help us identify how many people are around the user, if they are in a meeting or in a social or sporting event. This can also be achieved by identifying the wireless fingerprints of radio devices around a person (wireless hotspots, mobile phones, etc.). But we believe that fusing sensors of different types will create the most value. For example, by adding *semantic sensors* such as the phone calendar or various text-messaging services, we gain both a wider classification capability, and also a method that can minimize the human supervision needed. When a person adds a calendar entry (for example "Meeting with Joe" or "Pick up Anna from school"), the descriptive text can serve as a semantic data point that helps identify a more refined context (i.e., “Work Meeting” or “Family Task”) while at the same time serving as an automatic supervision entry. Similarly, a calendar entry named “workout” combined with physical sensors that identify the user as in a running activity out of doors will help in identifying the context as “Jogging”.

An updated version that we are now working on (Pythia Occursum) will include the calendar as our first semantic sensor. Additional sematic sensors that will be added are the contacts database, the text-messaging database(s), and the user's email.

ACKNOWLEDGMENT

This research was funded in part by the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI).

REFERENCES

[1] A. Abecker, A. Bernardi, K. Hinkelmann, O. Kühn, and M. Sintek, “Context-aware, proactive delivery of task-specific information: The KnowMore project,” Information Systems Frontiers, 2(3), 2000, pp. 253–276.

[2] “Agent Mobile Application,” On the Google Play store: [https://play.google.com/store/apps/details?id=com.tryagent]

[3] P. Bellavista, A. Corradi, M. Fanelli, and L. Foschini, “A survey of context data distribution for mobile ubiquitous systems,” ACM Comput. Surv. 44, 4, Article 24, August 2012.

[4] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni, “A survey of context modelling and reasoning techniques,” Pervasive and Mobile Computing 6 (2), 2010, pp. 161–180.

[5] “Cover Mobile Application,” On the Google Play store: [https://play.google.com/store/apps/details?id=com.coverscreen.cover]

[6] D. Jeffrey and S. Ghemawat, "MapReduce: simplified data processing on large clusters," Communications of the ACM 51.1, 2008, pp. 107-113.

[7] AK. Dey and GD. Abowd, “Towards a better understanding of context and context-awareness,” CHI'2000 Workshop on the What, Who, Where, When, and How of Context-Awareness, 2000, pp. 304-307 [https://smartech.gatech.edu/ bitstream/ handle/ 1853/ 3389/ 99-22.pdf]

[8] A. Duvander, “Google Now for Android Full of APIs,” On the Web at [http://blog.programmableweb.com/2012/06/27/ google-now-for-android-full-of-apis/]

[9] P. Fahy and C. Siobhan, "CASS–a middleware for mobile context-aware applications," Workshop on Context Awareness, MobiSys, 2004.

[10] “Google Fused Location Provider Application Programming Interface,” on the Android Developer web site: [http://developer.android.com/google/play-services/ location.html]

[11] Google Inc, “Google Now,” On the web at [http://www.google.co.il/landing/now/]

[12] R. Kyle, “The Future Of Mobile Is Figuring Out What You Want Before You Do,” On the Web at Business Insider: [http://www.businessinsider.com/the-future-of-mobile-is-context-2014-2?op=1]

[13] “Moves Mobile Application,” On the Google Play store : [https://play.google.com/store/apps/details?id=com.protogeo.moves]

[14] S. Perez, “T3 Reveals Scout, A Mobile “Context Engine” That Knows Where You Are & What You’re Doing, To Personalize Apps & Sites,” On the web at Techcrunch [http://techcrunch.com/2013/08/06/t3-reveals-scout-a-mobile-context-engine-that-knows-where-you-are-what-youre-doing-to-personalize-apps-sites/]

[15] A. Schmidt, M. Beigl, and H. Gellersen, "There is more to context than location," Computers & Graphics 23.6, 1999, pp. 893-901.

[16] Y. Tingxin, D. Chu, D. Ganesan, A. Kansal, and J. Liu, “Fast App Launching for Mobile Devices Using Predictive User Context,” MobiSys 2012, Low Wood Bay, UK, Jun 25-29, 2012, pp. 113-126.

[17] K. Wan, “A Brief History of Context,” IJCSI International Journal of Computer Science Issues, Vol. 6, No. 2, 2009, pp. 33-42

 ISBN: 978-1-61208-366-7

Design and Evaluation of a Mobile Payment System for Public Transport: the MobiPag STCP Prototype

Marta Campos Ferreira, Teresa Galvão Dias, João Falcão e Cunha
Department of Industrial Engineering and Management
Universidade do Porto, Faculdade de Engenharia
Porto, Portugal
Emails: {mferreira, tgalvão, jfcunha}@fe.up.pt

***Abstract*— The general adoption of mobile devices and their increasing functionality allow their use to make payments. This wide-spreading reality is being applied to several sectors, including public transport. In fact, there are several advantages of mobile payment and ticketing over traditional systems, such as queue avoidance, ubiquitous and remote access to payment, and the lack of need to carry physical money. This paper presents a prototype of a mobile payment system for public transport using customers' smartphones with Internet connection. The purchase and validation of tickets is made Over-The-Air (OTA), and location providers are used to locate the traveller and reduce the number of options when it comes to purchasing or validating a ticket. This system was tested in the city of Porto, by real travellers of Sociedade de Transportes Colectivos do Porto (STCP), the main bus transport company, during their normal use of public transport services. The users considered the system extremely useful, since it is more convenient than traditional systems, improving the travelling process and experience. They also felt secure to pay with their mobile phones, and valued the fact they could access information about their journeys, tickets, and account. The ticket validation process revealed to be one of the main challenges that any payment system for public transport should address, as compared to the simplicity of traditional systems.**

Keywords–mobile payments; public transport; mobile ticketing; user experience; field trial.

I. INTRODUCTION

Mobile payment can be defined as the use of a mobile device (mobile phone, Personal Digital Assistant (PDA), wireless tablet) "to initiate, authorize and confirm an exchange of financial value in return for goods and services." [1]. For more than a decade now, several attempts have been done to use mobile phones for payment transactions. In fact, there are several advantages of mobile payments over traditional systems, such as queue avoidance, ubiquitous and remote access to payment, and lack of need to carry coins and cash [2]. For instance, the users can pay for transport tickets without the need to visit an Automated Teller Machine (ATM) or a ticketing machine [3].

In this paper, we present a mobile payment system for public transport based on customers' mobile devices that only need to have Internet connection. The purchase and validation of tickets is made OTA, and location providers are used to locate the traveller and reduce the number of options when it comes to purchasing or validating a ticket, making the system easier to use. Since the system is totally based on customers' mobile devices, Public Transport Operators (PTOs) do not need to adapt or buy new infrastructures, such as gates, ticket vending machines or ticket readers.

The system was tested in real environment in the city of Porto, by real travellers, during their normal use of public transport services. Twenty-six users tested the system during 2 weeks and were accompanied by a Facebook group created for this purpose. This evaluation method was a success, since it allowed users to report in real time their difficulties, opinions and improvement suggestions. After the experiments, individual interviews were carried out, being useful to explore additional questions related with travelling habits, security perception and mobile payment business models.

The outline of the current paper is as follows: the next section characterizes mobile payment systems and traditional ticketing systems in public transport sector. Section 3 describes the proposed mobile payment system and Section 4 details the evaluation procedure and the main results. Finally, Section 5 presents the conclusions and future research

II. RELATED WORK

PTOs already used basic mobile phone features, like making phone calls and sending text messages, to allow travel tickets purchase. For instance, Paybox in Austria allows the Austrian railway OBB customers to purchase travel tickets via Short Message Service (SMS) or through the Vodafone live! Portal [4], and enable the customers to pay through their monthly phone bills. Proximus SMS-Pay in Belgium, Mobipay in Spain and AvantixMetro in UK are other examples of implemented mobile ticketing systems based on SMS.

While SMS can be considered a simple and easy to use technology, it has limitations when used to make payments. SMS uses store and forward technology, does not use any encryption method and there is no proof of delivery within the SMS protocol [5]. Most SMS-based mobile payment models do provide a proof of delivery, requiring a second separate message to be sent, which increases the costs of a transaction. This problem is particularly pertinent when small payments are at stake.

 ISBN: 978-1-61208-366-7

The evolution of mobile phones to smartphones has broadened the range of payment possibilities [6]. Also, when contactless technologies like Near Field Communication (NFC) were added to smart phones, more functionality became possible. Tickets can be purchased, downloaded, and accessed on the phone, and when in contact with NFC-enabled readers, the tickets are redeemed and a receipt is sent [7]. Several pilots of NFC-enabled phones have been launched in the public transport area. For instance, the Touch&Travel service in Germany allows passengers to make payments with their mobile phones. Travellers have to tap their NFC-enabled mobile phone to the Touchpoint device at the departing station and at the destination. The length of the journey and the ticket price are calculated at the end of the journey, and the customer receives, each month, a statement with all travel data and an attached invoice [8].

A NFC pilot was also launched in London [7], where 500 customers were given Nokia handsets with Oyster functionality. Passengers could top up their Oyster by touching their handset on Oyster ticket machines in tube stations or at Oyster tickets shops. Key findings of the research were that customers maintained high levels of interest and satisfaction throughout the trial and that the main customer benefits were convenience, ease of use, and status.

NFC was considered a good choice for mobile payments in terms of speed, security and usability when compared with traditional mobile payment service concepts, such as Interactive Voice Response, SMS, Wireless Application Protocol and One Time Password Generator [9][10]. In fact, NFC allows two-way contactless communication, offers faster connection between devices, less chance of interference, and has a shorter range, making it more secure for use in crowded places. However, NFC is failing in get critical mass, since it requires service providers to invest in new POS and NFC-reading systems and enough number of customers with NFC-enabled phones and wanting to use them to use them for payment purposes.

Bohm et al. [11] and Ferreira et al. [12] propose further mobile ticketing models for public transport based on Global Positioning System (GPS). According to these models, apart from having a smartphone with Wi-Fi and GPS technologies, the user only needs to check-in when starting a trip and check-out at the end. The customer is also located by the service provider during his trip at defined intervals. At the end of the journey, the system determines the route within the public transport network and calculates the price, which is then debited from the customers' account. This kind of system is really convenient and easy to use for customers, as they are not required to have any particular knowledge about tariffs or ticketing machines [12].

Mobile phones' features make them unique and suitable to be used to make payments and to offer additional services. Mobile phones are network-connected, have easy-to-use sound and text interfaces and provide anytime-anywhere access to information. When applied to the public transport sector, mobile ticketing systems allow PTOs to reduce operational and maintenance costs, acquire better knowledge about customers' travel behaviour, and shorten the interaction with the customers.

III. PROPOSED MOBILE PAYMENT SYSTEM

The proposed mobile payment system is the result of a project involving the main bus transport company in Porto – STCP – and potential customers. The mobile payment system was designed taking into consideration this specific service provider and its characteristics. Nevertheless, the concept and design of the system are scalable and adaptable to other realities. In the next subsection we describe the background and the challenges beyond this development. In Subsection B, we present the architecture of the system and finally, in Subsection C, we present the system itself.

A. Background and challenges

The Metropolitan Area of Porto (AMP) is served by an extensive public transport network which includes buses (from STCP), light rail (Metro do Porto) and trains (CP – Portuguese Railways). The electronic ticketing system in AMP is an open (ungated) system that required a significant technological investment, such as card readers along the platforms at each metro station and at each bus vehicle, and handheld devices for conductors.

The pricing policy implemented in the AMP is based on two types of price discrimination: journey-based and passenger-based price discrimination. The price for the journey-based perspective was settled based on a zone concept. The AMP network is divided into zones, with a flat rate within each zone, and the price is determined according to the number of zones crossed by the passenger. Once the ticket is validated, the passenger can travel, within a certain period of time, in the zone he chose. After that period of time, the traveller must validate the ticket again. The price of the tickets also depends on the characteristics of the passenger (child, student, senior or pensioner).

Tickets are available in several types: zonal single ticket, season ticket and multi-journey ticket. The ticketing system adopted in AMP is the contactless card, Andante, based on RFID technology [13]. Travellers can buy the contactless cards or recharge them at ticket vending machines, Service Provider Stores and spots, Third Party Agents and inside the vehicle (bus). Each Andante card can only contain one type of ticket at the same time (e.g., it cannot have a Zone2 ticket and also a Zone3 ticket), but it can contain several tickets of the same type (for instance, 10 Zone2 tickets).

In order to travel along the AMP network, passengers must buy the Andante contactless card and charge it with zone tickets. Then, they must validate the travel card in the reader at the beginning of the journey, and the ticket is redeemed. There is no need to validate the Andante at the end of the journey, but travellers must validate the travel card every time they change vehicle.

One of the main challenges of this work was to propose a ticketing solution for public transport services requiring the minimum investment cost from PTOs point-of-view, achieving at the same time the maximum consumer acceptance. The proposed system is based on customers' mobile devices, which are widely available and offer numerous functionalities, and is based on wireless communication technologies (3G and/or Wi-Fi) and on location providers, such as GPS and network triangulation.

 ISBN: 978-1-61208-366-7

SMS and NFC technologies were not considered a viable option, due to different reasons. SMS has several limitations (already referred) and have a premium price associated. NFC technology would require huge investments to convert existing infrastructures into NFC readers and would not represent a ubiquitous solution.

Another challenge we had to face was to guarantee the supervision of valid tickets by conductors. Since AMP is an ungated system, Porto PTO must have a way to confirm that a certain ticket is still valid for a specific journey. This confirmation is visual and uses security symbols and sequence numbers. This process is described in detail in the Subsection C. Such information can also be confirmed by assessing the backend system, in case of mobile phone's dead battery.

Another concern has to do with customers' information and data gathering. It is true that PTOs have heavy infrastructures installed and incur in maintenance costs every month, but these infrastructures are powerful data collectors that helps PTOs to know customers' travel patterns and to adjust service offerings. With the proposed system we do not lose this precious information, rather we enhance it. PTOs have access to individual customers' travel behaviour and preferences. This may represent a shift in public transport service delivery, since SP may direct recommendations, services and institutional and operational information particularly suited to a specific customer. We move from mass communication to one-to-one communication.

B. System Architecture

The system architecture comprises three main components: server, client and conductor (see Fig. 1). The client component allows customers to interact directly with the services. This interaction is achieved through the use of a mobile phone, tablet, or any other mobile device running the Android operating system. This component allows buying, store and validating travel tickets, as well as checking tickets balance, account movements, validation history, check prices and maps, and find near stations.

The conductor component allows conductors to verify if a traveller has a valid ticket for the journey or not. The client and conductor components are integrated in the same application, in the customers' mobile phone, which removes the need of an additional device to be carried by the conductors.

The server may be considered the heart of the system, since it provides the services to the other components of the architecture. It comprises three subcomponents:

a) Database: all information is stored in a database, only accessible by server-side business logic.

b) Webpage: acts as a control panel through which the platform manager can manage all aspects of the system and access to the customers' travelling information.

c) Web service: most of the logic of the system will be processed by this component, which function as an intermediary between the customer/conductor component and the central database.

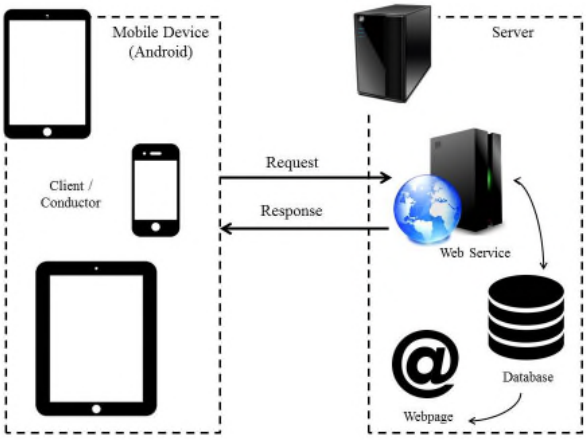


Figure 1. MobiPag STCP architecture.

C. MobiPag STCP Application

According to the proposed mobile payment system, the purchase and validation of tickets is made OTA and location providers are used to locate the traveller and reduce the number of options when it comes to purchase or validate a ticket, making the system easier to use. Before choosing which ticket the user wants to buy, a list with the tickets already stored in his wallet is presented to him, preventing the user to buy tickets he already has. The user can buy the travel tickets in two ways: he chooses the zone ticket he wants to buy and selects the number of tickets for each zone (see Fig. 2 (a)), or alternatively, he chooses the departure and the arrival station and the system automatically converts this information into zones.

To validate a ticket, location providers (GPS and network triangulation) are used to identify customers' location. From the given list of near stops the user chooses which stop he is going to enter and then he selects the ticket he wants to redeem from those stored in his virtual wallet. In order gather valuable information, users must also select the bus line they are entering (see Fig. 2 (b)). This requirement allows the Porto PTO to know exactly which vehicle the user is entering, by crossing this information with time and vehicles on the road. Once the ticket is validated, the

Figure 2. Mobile payment system screens: (a) buy ticket by choosing the type (zone); (b) validate ticket; (c) active ticket (interfaces in Portuguese).

passenger can travel in the zone he chose for a certain time, and check the remaining time on the display. The user is also warned when the journey time expires.

If a conductor wants to verify if a traveller as a valid ticket for that journey, the user only needs to show the active ticket screen on his mobile phone (see Fig. 2 (c)). This screen has information about the ticket (stop, date and type of ticket), a security symbol, and a sequence number. The security symbol, represented by the watermark image, will act as a secure element to prevent users from creating false tickets images. This symbol changes every day, and the conductor has access to it in order to know what he expects to see on customers' mobile devices. The sequence number acts also as a secure element. Each validation corresponds to a different sequence number. So, the conductor will be able to verify a pattern (sequence numbers very close) inside a bus. If he identifies a sequence number very different from others, this acts as a warning sign for the conductor to check carefully the other information to see if the title is valid. In Fig. 2 (c), this number is represented by the number 746321.

The proposed mobile payment system also comprises several additional services beyond payments in order to attract potential consumers. For instance, the user can check tickets balance, account movements, validation history, check prices and maps, and find near stations.

IV. EVALUATION

In this section, we explain how we evaluated the MobiPag STCP system. Our goals were to understand users' perception about the concept of buying and validating travel tickets with the mobile phone, and to analyse the usability of the application, identify major problems and potential improvements. The next subsection explains the evaluation procedure that was used. Subsection B details the sample characteristics. The test phase is described in Subsection C, and finally, Subsection D presents the major results.

A. Procedure

The experiments were conducted in real environment, by real travellers, during their normal use of public transport services in the city of Porto. The recruitment and selection of the participants was carried out by STCP, who solicited participation through their website and information inside the buses. In order to get as much heterogeneity in terms of various demographic factors (gender, age, occupation), 37 travellers were selected, from which 26 participated in the tests. The users tested the application for two weeks.

There were some prerequisites that the participants had to meet in order to participate: owning a mobile phone with Android operating system, being a frequent user of public transport (at least 5 validations per week), and have Internet connection via mobile phone.

The experiments were divided in three phases:

1) Pre-test phase: Explanation of the evaluation process (by email and in person) and administration of a questionnaire to characterize the sample.

2) Test phase: The users tested the application in real environment and in the context of use of public transport services. During this phase users were accompanied by a Facebook group created for this purpose and by email, which allowed for a very detailed review in real time.

3) Post-test phase: In depth interviews in order to gather additional information about the experiments.

B. Sample Characterization

Before starting the experiments, the participants had to fill in a questionnaire that was applied online using the google drive platform. The main objective was to characterize the sample in terms of socio-demographic characteristics, and smartphone and public transport usage. Participants had also to rate, according to five-point Likert scale [14] ranging from strongly disagree to strongly agree, several statements related with the purchase and validation of tickets through traditional methods and through mobile devices.

From the total of 26 participants, 16 are male and 10 female, aged between 21 and 68 years and average age of around 34 years (see Table I). Most users (23) have a smartphone for 6 months or more, which indicates some familiarity with the use of smartphones, making it easier to adapt to the application. They are all frequent users of public transport services and about half of the users perform intermodal transhipment (bus-subway; bus-train; etc.).

Most users buy their tickets at vending machines (18) or through third party agents (11) and ATM network (9) and use both debit card (15) and money (11) to make payment. The most used additional services are checking timetables (25) and transport network maps (11), being the access to this information mostly done via the website of the operators (23) but also in stops and stations (18).

In some situations, users do not know what kind of ticket to buy to perform a certain trip. Despite considering easy to buy tickets in vending machines, it is frequent not to have change to make the purchase. The need to go to a physical store to purchase the monthly pass is considered inconvenient, especially at the end or beginning of each month because of the long queues. Users also stated that it is rather more likely to leave the Andante card at home than the mobile phone.

Regarding the purchase and validation of tickets with the mobile phones, users revealed very receptive to it, considering this payment method useful and secure, and compatible with their lifestyle and normal use of the phone. However they showed some concern about connectivity problems and short battery life that may jeopardize the completion of the payment operations.

TABLE I - SAMPLE CHARACTERIZATION

Characteristics	Number of Participants
Sample number (n)	26
Age	20-29y (11), 30-39y (8), 40-49y (4), 50-59y (3)
Gender	Male (16), Female (10)
Smartphone Ownership	More than six months (23), less than six months (3)

 ISBN: 978-1-61208-366-7

C. Test Phase

The Mobipag STCP application was made available via Google Play two days before the beginning of the experiments, and only the authorized participants could download it. This allowed users to get familiar with the application and clarify doubts.

During this test phase, the users had to buy and validate travel tickets through the mobile application during their normal use of public transport. Since the mobile tickets had no legal value, users had also to buy and validate physical tickets at the same time.

In two weeks, the users made 723 validations with their mobile phone, in 234 different stops and 111 different routes and 36 transhipments. Analyses of Fig. 3 indicate a sharp decline of validations during the weekends and an average of 50 validations per day during the week. The users bought 24 monthly passes and made 63 purchases of single tickets.

To promote the communication among participants, it was created a group on the social network Facebook, where users were encouraged to share their experience, doubts and questions. This method was a success, since it allowed users to report in real time their difficulties, opinions and improvement suggestions. They interacted with each other by sharing their experiences and trying to solve common problems. It also allowed us to correct, in real time, any bug they reported and to gather a lot of information regarding the experience.

Every comment on the Facebook group was analysed and coded. These codes were then aggregated into categories, according to the relationships between them. After the testing phase, the participants were interviewed individually. Each interview was recorded and lasted about 50 minutes. The interviews were useful to explore additional questions related with travelling habits, security perception and mobile payment business model. The interviews content was then analysed and coded. Some codes were related to the ones founded in the Facebook content analysis, while others have led to the emergence of new categories. Regarding the problems of the application and improvement suggestions, the Facebook group comments revealed to be much richer, since they were reported in real time. The main conclusions regarding the analysis of the Facebook comments and interviews are presented in the next subsection.

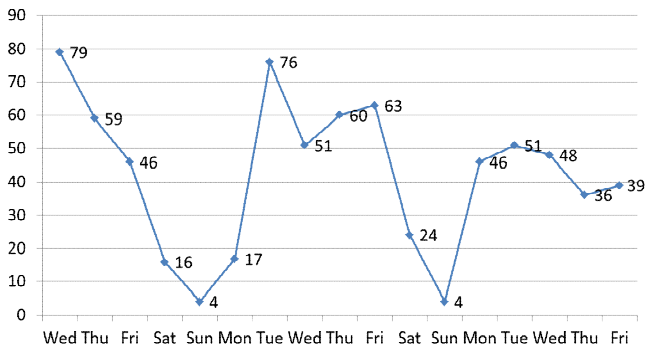


Figure 3. Number of validations per day.

D. Discussion

The comments on Facebook group and the individual interviews were analysed, coded and grouped in five main categories: perceived value, suggestions, concerns and issues, security and fraud, and business model. These categories were summarized in Table II and are detailed below.

1) Perceived value

The users liked the concept of buying and validating tickets with the mobile phone and considered the application very intuitive and easy to use. The possibility of buying tickets everywhere and anytime was greatly valued. They also found very useful the additional information about the journeys and users' account provided by the application. Such customized information, like tickets balance, remaining time of the journey, and journey details, is not possible to provide through a contactless card with no screen.

TABLE II - PERCENTAGE OF PARTICIPANTS THAT MENTIONED EACH TOPIC RELATED TO THE MOBIPAG STCP APPLICATION

Categories	% of participants mentioning the argument (n=26)	Argument in favour or against the proposed system
Perceived Value		
- Satisfaction (e.g., easy-to-use; intuitive; great functionalities (historic, purchase, remaining time))	86%	+
Suggestions		
- Improvement suggestions (e.g., PIN and password; colours; method of selecting the stops; storage and upload of personal photo)	71%	+ -
- New functionalities (e.g., languages; maps; historic of most used stops; alerts)	57%	+ -
- Application Bugs (e.g., aplicattion crashes; unknown characters; wrong alert about ending station; wrong price)	67%	-
Concerns and Issues		
- Technology (e.g., GPS bad performance)	29%	-
- Insatisfaction (e.g., takes too long to complete the validation process)	33%	-
Security and Fraud		
- Security (e.g., security when paying with the mobile phone)	71%	+
- Fraud (e.g., people may not validate tickets; use of the same account in different mobile phones)	24%	-
Business Model		
- Payment method (e.g., pre-paid account for travelling purposes)	86%	+
- Pay for the mobile ticketing service (e.g., willing to pay a modest fee for the application)	76%	+

 ISBN: 978-1-61208-366-7

2) Suggestions

The participants were very active in identifying problems and bugs related to the application. The use of the social network Facebook in the evaluation process was a major advantage, since users were able to communicate in real time the bugs they were finding, and the developing team was able to fix those errors immediately.

A lot of improvement suggestions were also identified by users, such as the design and colours of the application, PIN and password procedures, method of selecting the stop (by alphabetical order, by most used, etc.). New functionalities were also suggested, such as adding new languages to the application, historic of most used stops, alerts about the expiration of a monthly pass. In Table II the topics improvement suggestions and new functionalities have the "+" and "-" signs simultaneously. These topics are a negative argument against the application because those functionalities were not (well) covered by the application, but at the same time are in favour because some of those were implemented during the tests.

These inputs were fundamental to improve the application and to set new ideas for future versions of the system.

3) Concerns and issues

The validation process was considered more complex when compared with the traditional one. Users had to choose the stop, the route and the ticket before the validation. They proposed several ideas to simplify the process: provide the last used stops, create favourite stops, and use other technologies, such as NFC or QR codes. The validation process is a major challenge in the design of a ticketing solution.

In order to facilitate the validation process, the system locates the user through triangulation or GPS to indicate the departure stop where he is. However, GPS location takes some time to locate the user and the location through mobile networks proved in many cases to be inaccurate. This meant that in most cases it was necessary to resort to a manual selection of the stops (which was thought to be used only in special cases), requiring an extra step in the validation process.

4) Security and fraud

The users felt secure to pay with the mobile phones. They even compared this system with mobile banking systems. The participants that were already familiar with the use of mobile banking applications, they felt equally safe to purchase tickets with the phone.

From service providers' point-of-view some concerns regarding security and fraud may emerge due to the inspection process. The process for conductors to check the validity of a ticket is mainly visual, which may require adding further security to the process, such as providing reading devices to the conductors.

In addition to these, there will always be risks associated with the behaviour of the people itself, regardless of the ticketing system used. For instance, some participants raised some concerns about what prevented people from validate the ticket only when they saw the conductor approaching.

5) Business Model

Most of users stated they prefer to have a pre-paid account for travelling purposes instead of having the ticketing application linked with their bank account or mobile phone bill. This is important for PTOs, since it means that users are willing to pay before they travel. This lag between the payment for the service and its provision, functions as a way of funding for PTOs.

When questioned about how much they were willing to pay for the mobile ticketing service, most users stated that they were willing to pay a modest fee for the application, but not an additional amount per ticket purchased.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented the design and evaluation of a prototype for a mobile payment system for public transport. This system is based on customers' mobile devices that only need to have Internet connection. The purchase and validation of tickets is made OTA, and location providers are used to locate the traveller and reduce the number of options when it comes to purchase or validate a ticket, making the system easier to use.

The system was tested in real environment, by real travellers, during their normal use of public transport services in the city of Porto. The 26 users tested the system during 2 weeks and were accompanied through a Facebook group created for this purpose. After the experiments individual interviews were carried out, in order to explore additional information related with travelling habits, security perception and mobile payment business model.

The users liked the concept of buying and validating tickets with the mobile phone and considered the application very intuitive and easy to use. They also felt secure to pay with the mobile phone and valued the fact they could access to personal information about their journeys, tickets and account. They also stated they prefer to have a pre-paid account for travelling purposes instead of having the ticketing application linked with their bank account or mobile phone bill. The ticket validation process revealed to be one of the main challenges in the design of mobile ticketing systems, since the validation through traditional physical systems is very simple.

This field trial allowed corroborating the great potential that mobile ticketing systems have over traditional systems. They are more convenient (tickets can be purchased everywhere, anytime), users have access to more information about their journeys and PTOs can interact more closely with their customers, opening doors to a one-to-one communication. In order to maximize the potential of such solutions, the validation process should be as simple as possible and additional and complementary services should be integrated.

As future work, we want to improve this payment system and add additional and complementary services. It is our intention to involve further service providers beyond public transport operators, since the travellers' value constellation is composed by other players.

 ISBN: 978-1-61208-366-7

ACKNOWLEDGMENT

This work was supported by TICE - MOBIPAG project 13847, Mobile Payments National Initiative. This project also involved Universidade do Minho, CEDT, Cardmobili, Creative Systems, and Wintouch. Funding is provided under the COMPETE, QREN programme, managed by AdI, in the context of European Union FEDER. OPT (www.opt.pt), IBM CAS Portugal, INEGI and IDMEC Pólo FEUP are also supporting the project at FEUP (www.fe.up.pt/IBM-CAS-Portugal).

REFERENCES

[1] Y. Au and R. Kauffman, "The economics of mobile payments: Understanding stakeholder issues for an emerging financial technology application," *Electron. Commer. Res. Appl.*, vol. 7, no. 2, pp. 141–164, 2008.

[2] N. Mallat, "Exploring consumer adoption of mobile payments – A qualitative study," *J. Strateg. Inf. Syst.*, vol. 16, no. 4, pp. 413–432, Nov. 2007.

[3] M. Mut-Puigserver, M. M. Payeras-Capellà, J.-L. Ferrer-Gomila, A. Vives-Guasch, and J. Castellà-Roca, "A survey of electronic ticketing applied to transport," *Comput. Secur.*, vol. 31, no. 8, pp. 925–939, Nov. 2012.

[4] "Paybox." [Online]. Available from: http://www.paybox.at/ 2014.05.22

[5] R. Boer and T. de Boer, "Mobile payments 2010: Market Analysis and Overview," Chiel Liezenber (Innopay) and Ed Achterberg (Telecompaper), 2009.

[6] A. Becker, A. Mladenow, N. Kryvinska, and C. Strauss, "Aggregated survey of sustainable business models for agile mobile service delivery platforms," *J. Serv. Sci. Res.*, vol. 4, no. 1, pp. 97–121, 2012.

[7] NFC Forum, "NFC in Public Transport," 2011.

[8] "Touch&Travel." [Online]. Available from: http://www.touchandtravel.de/ 2014.05.22

[9] M. Massoth and T. Bingel, "Performance of Different Mobile Payment Service Concepts Compared with a NFC-Based Solution," *ICIW 2009, The Fourth Int. Conf. Internet Web Appl. Serv.*, 2009, pp. 205–210.

[10] K.-Y. Chen and M.-L. Chang, "User acceptance of 'near field communication' mobile phone service: an investigation based on the 'unified theory of acceptance and use of technology' model," *Serv. Ind. J.*, vol. 33, no. 6, pp. 609–623, May 2013.

[11] A. Bohm, B. Murtz, C. Sommer, and M. Wermuth, "Location-based ticketing in public transport," in *Proceedings of the 8th International IEEE Conference on Intelligent Transportation Systems 2005.*, pp. 837–840.

[12] M. C. Ferreira, A. Cunha, J. F. e Cunha, H. Nóvoa, T. Galvão, and M. Moniz, "A survey of current trends in smartphone based payment and validation services for public transport users," in *The Art & Science of Service Conference, Maastricht*, 2012, pp. 1–28.

[13] "Andante." [Online]. Available from: http://www.linhandante.com/ 2014.05.22

[14] M. S. Matell and J. Jacoby, "Is There an Optimal Number of Alternative for Likert-Scale Items?," *J. Appl. Psychol.*, vol. 56, no. 6, pp. 506–509, 1972.

 ISBN: 978-1-61208-366-7

Expected Penetration Rate of 5G Mobile Users by 2020: A Case Study

Andrey Krendzel
5G Networks
Huawei Technologies
Helsinki, Finland
andrey.krendzel@huawei.com

Philip Ginzboorg
5G Networks/Communications Systems and Networking
Huawei Technologies/Aalto University
Helsinki, Finland
philip.ginzboorg@ huawei.com

***Abstract*—The next 5th generation mobile network, or 5G key concepts, scenarios and requirements are actively debated in the research community. In this context, it is interesting to estimate a tentative 5G penetration rate, i.e., the mobile community of 5G users. In this paper, we focus on the initial 5G penetration, i.e., proportion of people who are willing to use the 5G networks, when the first 5G equipment is projected to be deployed, around the year 2020. The 5G penetration rate can be used as an input parameter for business viability, traffic estimation and network planning/dimensioning related to the 5G network infrastructure. The 5G penetration level will be country-specific; it may be different in different countries. We assume that the initial 5G penetration will depend on the penetration rate of the previous wave of the mobile wireless technology, which is called "fourth generation" (4G), or Long Term Evolution (LTE). Finland, currently, has the highest LTE penetration rate in Western Europe. As a case study, we estimate a number of potential users of the 5G network in Finland by 2020. In our approach we use relationships between mobile penetration rate, Gross Domestic Product (GDP) per capita, inequality of income distribution within population, and the Pareto law.**

Keywords-penetration rate; 5G; 4G; Pareto distribution; Logistic function; Lorenz curve; Gini coefficient; GDP per capita.

I. INTRODUCTION

So far, new technologies for mobile wireless networking have been deployed once in about ten years. The appearances of a new generation equipment on the market happened roughly in: 1981 (termed 1G, or first generation), 1991 (2G), 2001 (3G) and 2011 (4G). The fifth generation (5G) is expected to emerge around 2020-2021.

Even though the 5G mobile network is not defined yet in any official specification or standard, issues related to advanced 5G network infrastructure provoke intense interest in research community, e.g., within the framework of the Horizon 2020 European programme [19] for research and innovation. Please note that the exact time when the 5G technology will be introduced is still uncertain. We shall assume that it is the year 2020 in this paper.

The main factors/drivers towards 5G that should be taken into account are: (i) demand for services/applications from different groups of end-users in the 2020 time frame, i.e., competitive market impact; (ii) Gain/cost ratio related to new innovations/technologies/solutions/business models; (iii) existing limitations of frequency bands and spectral bandwidth; (iv) political factors that can impose some restrictions on innovative solutions.

The first 5G networks are projected to be deployed around the year 2020. For the first three factors, it can be useful to estimate the potential number of people that are ready to become 5G subscribers by that time. The 5G penetration rate will be different for each country. In this paper, we select Finland as a country for our case study. Finland has the highest LTE penetration level between countries of Western Europe in 2013 [1]. We estimate a potential number of 5G users in the country and the 5G density in its two largest urban areas by 2020. This is the main contribution of the short paper. To the best to our knowledge, there has been no prior work in literature showing how to estimate 5G penetration rate by 2020.

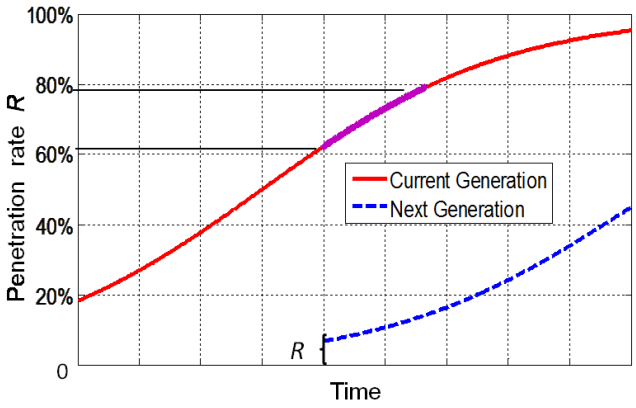


Figure 1. Penetration process R of current and next generation technology/service among the affluent part of the population. When 60 % to 80 % of the wealthy have "current generation" networking technology, it is time to introduce the "next generation." Then the whole group of affluent "current generation" users becomes the initial group of potential "next generation" users.

Our main premises are that first, the potential 5G subscribers initially come from the affluent subgroup of the population where the 4G penetration is much higher than the average. Second, the penetration of technology/service including telecommunication services follows the logistic function [2]-[4]. (Please note that qualitatively this function has three distinct phases: initial exponential growth up to 20%, almost linear growth up to 80%, and the final saturation stage [9].) We shall assume that when the new 5G technology is introduced, the penetration level of the previous, 4G technology among that wealthy (affluent)

 ISBN: 978-1-61208-366-7

subgroup will be between 60 % and 80 % that corresponds to the second part of the linear phase of the logistic curve (i.e., the greatest demand for the current technology/service [2] [10]). These assumptions are schematically illustrated in Fig. 1. The idea is that the wealthy (affluent) subgroup of people in the general population will be ready for the new 5G service when most of their wealthy peers have adopted the existing 4G technology.

We estimate the size R of that subgroup based on relationships between the Pareto law [5], Gross Domestic Product (GDP) per capita, inequality of income distribution within population, and the 4G penetration rate.

It is worthwhile also to note that methods similar to ours were applied to estimate a demand for telecommunication services in the past. For instance, the number of Integrated Services Digital Network (ISDN) users in several developed countries was estimated in [3][4]. Today, the results of these estimations seem quite plausible. In particular, very low demand for ISDN was predicted. It was estimated that the number of ISDN users even in developed countries had to be around 5-6% from the number of Public Switched Telephone Network (PSTN) users that corresponded with the real situation those years.

The rest of the paper is divided into three sections. The next section describes briefly the approach to evaluate the expected number of 5G users. Section III presents the case study related to estimating 5G penetration level in Finland by 2020. Finally, Section IV concludes the paper.

II. APPROACH FOR ESTIMATING 5G PENETRATION RATE

In this section, we estimate the proportion of wealthy people in the population (potential 5G users) from the penetration level μ of 4G users and the Pareto parameter α. Then we show how α itself can be derived from μ.

It is argued in [2]-[4] that the demand for services depends on both GDP and its distribution within society and there is the relationship between a demand for telecommunication services, labour productivity, distribution of incomes between individuals, and GDP per capita. In particular, it is shown in [4] that the relationship between a telecommunication demand and income distribution is close to the Pareto law [5].

If X is a random variable with a Pareto (Type I) distribution, then the probability that X is greater than some number x is given by [5]

$$R(x) = \Pr(X > x) = x^{-\alpha}, \quad 1 < \alpha < \infty, x \geq 1, \qquad (1)$$

where $R(x)$ can be expressed as the proportion of individuals who have income more than x, and α is the distribution parameter, called the Pareto parameter or the tail index. (Please note that a range of small incomes has very small influence on statistical characteristics of income distribution [3]-[5].)

The income x in (1) is a normalized value that is equal to the ratio g/g_{min}, where g is one of income values, g_{min} is the minimum income value in a population.

The minimum income value may be expressed as $g_{min}(\alpha)=g_0/L(\alpha)$, where g_0 is the average value of personal annual income, that is GDP per capita, and $L(\alpha)=\alpha/(\alpha-1)$ is the average value of the normalized income. Thus, the expression (1) for estimating the number of individuals who have income more than $x = g/g_{min}$ takes the following form:

$$R(g / g_{\min}, \alpha) = \left(\frac{g\alpha}{g_0(\alpha - 1)} \right)^{-\alpha}. \qquad (2)$$

It is shown by Varakin [3] that the linear dependence takes place between the average amount of produced information generated by society per an individual in a country and its GDP per capita. Mobile telecommunications (as a part of society and economical infrastructure) impact on economic development [6]. Conversely, the economic development of a country determines its level of mobile telecommunications [7]. As a result, there is a relationship between a telecommunication/mobile penetration level and GDP per capita, that is generally assumed to be linear [3][4][6][7]. In our case, we also suppose that there is linear dependence between the penetration level of 4G users and GDP per capita g_0 (in the first approximation).

Mathematically, this relationship may be presented as

$$\mu = A \cdot g_0, \qquad (3)$$

where $\mu = N_{4G}/100$ is the penetration level of 4G users, N_{4G} is the average number of users that have subscriptions for 4G per 100 individuals, and A is the normalizing dimension factor that is country-specific.

Since parameters μ and g_0 are the average values obtained by averaging many input data, in general case, it is also plausible that the penetration level T of 4G technology in the subgroup of the affluent individuals follows the relation $T = Ag_1$, where is g_1 is some income within the affluent group, and A is the same normalizing factor as in (3).

Thus, the ratio between the parameters is

$$k = \frac{T}{\mu} = \frac{g_1}{g_0}, k \geq 1, \qquad (4)$$

where the coefficient k determines the excess of the penetration level of 4G subscribers in the affluent subgroup above the average value of the 4G penetration level in the total population.

As mentioned in Introduction, we assume that T is between 60 % and 80 % (i.e., k is between $0.6/\mu$ and $0.8/\mu$.) when 5G is introduced. Then, the whole group of affluent 4G users becomes the group of potential 5G users.

Based on equation (2) and the above assumption, the expression to determine the relative number of individuals who have the 4G penetration level more or equal than the parameter μ, or, in other words, the expression to estimate the relative number of the affluent 4G users or the potential 5G users (R) by the time when new generation is launched has the following form:

$$R\left(\frac{g_1}{g_0}, \alpha\right) = R(k, \alpha) = \left(\frac{k\alpha}{\alpha - 1} \right)^{-\alpha}. \qquad (5)$$

Recall that we assume that k is between $0.6/\mu$ and $0.8/\mu$ when the new, 5G technology is introduced. As a rule,

 ISBN: 978-1-61208-366-7

forecasts of 4G penetration level μ in a region may be found in statistical literature. For instance, in the European Union report [1], the forecast related to the 4G penetration level in different Western European countries is presented up to 2020. But, to compute R according to equation (5), we also need to estimate somehow the Pareto parameter α at the time when the new, 5G technology is introduced.

In the rest of this section, we will show a way to estimate α from μ.

Generally, the parameter α depends on the inequality of income distribution between individuals. In our case, this parameter depends on the inequality of distribution of the number of 4G subscribers between individuals. The inequality of income distribution in a subgroup of individuals is described by the Lorenz curves [8]. In Fig. 2, the Lorenz curves show relationships between the current average income value in a subgroup of population Q and the number of individuals in the subgroup F for several values of α [18].

In particular, the set of the Lorenz curves illustrates that with increasing the parameter value α income in a subgroup is becoming more evenly distributed.

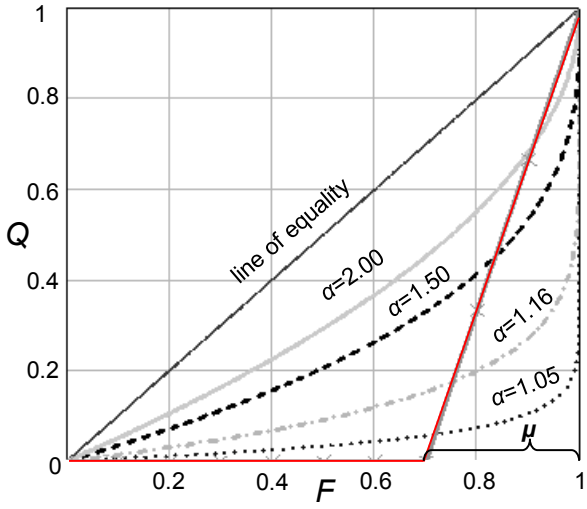


Figure 2. Set of Lorenz curves

The analytical function $Q(F)$ that allows assigning a set of the Lorenz curves has the following form [5]:

$$Q(\alpha, x) = 1 - (1 - F(x))^{\frac{\alpha - 1}{\alpha}} \quad (6)$$

To estimate the Pareto parameter α for the expression (5), it is needed to approximate the function describing the broken line $Q(\mu,F)$ by the function (6) $Q(\alpha,F)$ corresponding to the Pareto distribution with the parameter α. It can be done by means of the Gini coefficient [8] W related to each of these functions. It is equal twice the area between the Lorenz curve and the line of equality, i.e., $W = 1 - \mu$ and $W = (2\alpha\text{-}1)^{-1}$ for $Q(\mu,F)$ and $Q(\alpha,F)$, respectively [18].

As a result of this approximation, the Pareto parameter α can be expressed as function of the penetration level μ:

$$\alpha \approx \frac{0.5(2 - \mu)}{1 - \mu}. \quad (7)$$

Then, the absolute value of the number of 5G potential users is estimated as

$$N_{5G} \approx R(k)N \quad (8)$$

where N is the population size in a region.

Thus, the expressions (4), (5), (7), and (8) give a basis to estimate the number of potential 5G users in a region.

III. CASE STUDY

In this case study, we estimate a number of the potential subscribers of 5G networks in Finland by 2020 using the presented approach.

According to statistical information, 5.44 million people live in Finland (2013) [11]. The projection of the population growth in years 2010-2060 [12] predicts a number of inhabitants in the country by 2020 as 5.64 million (N).

Fig. 3 shows the forecast of LTE residential penetration of Western Europe up to 2020 presented in the EC report [1] (based on the Analysys Mason research [16]).

In accordance with it the penetration level of 4G (LTE) users in Finland (μ) is estimated as 32% by 2020. Then, applying the expressions (4), (5), (7), (8) we can evaluate the relative and absolute number of potential 5G users for Finland in 2020.

If T = 0.8 (conservative value), the relative number of potential 5G users in the country may be estimated as $R(k, \alpha) = R$ (2.5, 1.23) = 0.04. It means that just **4%** of people in Finland will be willing in 2020 to use the 5G network infrastructure to get their services. The absolute number of 5G subscribers in this case is around 0.22 million.

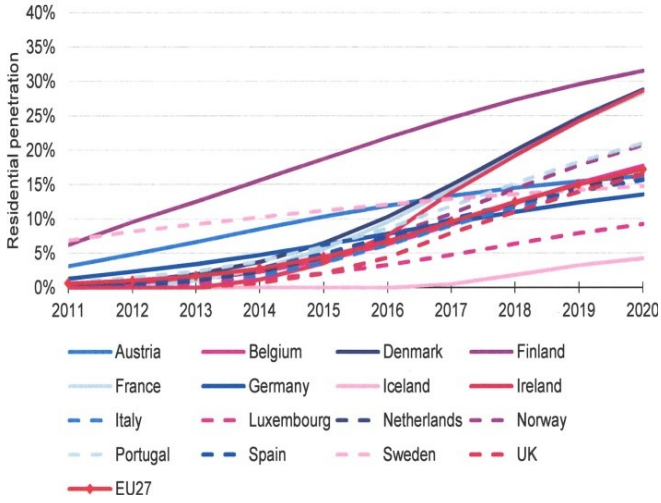


Figure 3. Residential penetration of LTE in Western Europe [Source: Analysys Mason]

If T = 0.6 (optimistic value), the relative number of tentative 5G users in Finland is equal to $R(k, \alpha) = R$ (1.9, 1.23) = 0.06. In this case, **6%** of Finish citizenships can be ready to become the 5G subscribers. It corresponds to 0.34 million people. Note that these values are initial numbers of potential 5G subscribers by 2020 and can be considered as lower limits.

It should be taken into account that population distribution in Finland throughout of the country is not

 ISBN: 978-1-61208-366-7

uniform. People are mainly concentrated in the large urban areas. If we take the Helsinki metropolitan area (Helsinki, Vantaa, Espoo and Kauniainen) [13] and the Tampere urban area, the population of these two regions equals 1.4 million (1.1 and 0.3 million, correspondingly [14]). This is around 26% of the current 5.44 million Finland's population. The Gini coefficient value for Finland is very low (W is 25.9 [15]). It means more or less equal income distribution between country residents. Then, we can roughly assume that there is no big difference in income distribution level between these two regions and the rest of the country. Thus, we can suppose that 26% of all potential 5G users are concentrated in the Helsinki and Tampere urban regions.

As a rule, people in large cities have larger income and a share of 5G users for these two areas can be taken even a bit higher than 26%. But, we focus on a lower bound of the 5G penetration rate in this paper. In the absolute values, the number of potential 5G users in Helsinki and Tampere urban areas by 2020 is forecasted in accordance with the proposed approach to be between 60 thousand (conservative value) and 90 thousand (optimistic value). These two urban areas cover an area of about 1000 km^2 [13][14]. That is, the density of 5G users by 2020 in this territory is expected around **60~90 users/km^2**.

Definitely, only the future can confirm or disprove the estimations. However, it is interesting what happens if we would apply this approach to the past statistical information to estimate the number of 3G users based on 2G penetration level. On the one hand, the penetration level of 2G mobile phones in Finland in 2000 was 72% [17]. The first network equipment of UMTS (3G) was deployed in the beginning of 2000s. The initial penetration percent of mobile Internet phone (3G) in Finland was 22% [17]. On the other hand, if we use the presented methodology, then the relative number of 3G users had to be $R(k, \alpha) = R\ (1.11, 2.29) = 0.21$ (T=0.8), i.e., around 21%, which is very close to the actual 3G penetration rate at the time.

Note that the 4G penetration forecast (Fig. 3) ends in 2020. If the 4G prediction would be known also for later years (for instance, up to 2025), then using the proposed methodology it may be possible to get the long-term dynamic forecast of 5G user growth.

IV. CONCLUSION

In this paper, we have presented an approach to estimate a tentative 5G penetration rate by 2020 when the first 5G network equipment is planned to be deployed. As a case study, we evaluated this parameter for Finland, but the approach can be applied also for other countries if corresponding statistical information is available. Though Finland is one of more promising countries in this context (it is predicted to have the highest LTE penetration in Western Europe), initial level of 5G penetration rate by 2020 is expected to be only around 4~6 % of the total number of inhabitants. The density of 5G users in two largest urban areas (Helsinki and Tampere) is also forecasted to be quite low, 60~90 users/km^2. These estimated values indicate a starting point of 5G penetration process.

To conclude, it is not worthwhile to expect an initial demand for 5G services at a level of 20-30% as it was when the first 3G network services became available. It is needed also to be cautious with regard to the density of 5G users per km^2 even in urban environment in the 2020 time frame. It is reasonable to support concentrating initial 5G deployment in "strategic" places like city centres and shopping malls.

The presented approach can help in issues related to traffic load estimations in 5G networks, network planning and network dimensioning aspects, in assessing the potential revenue from 5G subscriptions at the first stages of 5G network deployment.

REFERENCES

[1] "The social economical impact of bandwidth", the European Commission final report, p.198, 2010.

[2] Paul A Samuelson, William D Nordhaus, "Economics", 19th Edition, 2010.

[3] L.E. Varakin, "Economics, telecommunications and development of the society: macroeconomic mechanisms of telecommunications development," Electrosvyaz Journal (in Russian), no. 1, 1994.

[4] L.E. Varakin, "The Pareto law and the rule 20/80: distribution of incomes and telecommunication services," MAC proceedings (in Russian), pp. 3-10, no. 1, 1997.

[5] Barry C. Arnold, "Pareto Distributions", International Co-operative Publishing House, 1983.

[6] "GSMA. The Mobile Economy 2013", ATKearney, 2013.

[7] H. Gruber, P. Koutroumpis, "Mobile telecommunications and the impact on economic development", Economic Policy Panel, October 2010.

[8] C. Dagum. "The generation and distribution of income, the Lorenz curve and the Gini ratio," Econ. Appl. No. 33, 327-367, 1980.

[9] J.S. Cramer "The origins and development of the logit model", University of Amsterdam, August 2003.

[10] C.V. Brown, P.M. Jackson, "Public sector economics", Oxford, Blackwell, 1990.

[11] Statistics Finland. Population, http://tilastokeskus.fi/til/vrm_en.html [retrieved June 2014].

[12] Statistics Finland, years 2010 to 2060: projection, http://tilastokeskus.fi/til/vaenn/2009/vaenn_2009_2009-09-30_tau_001_en.html [updated September 2009]

[13] Helsinki Region Information and Statistics, http://www.helsinginseutu.fi/hki/HS/The+Region+of+Helsinki/City+information+and+statistics [retrieved June 2014]

[14] Population Register Center of Finland, http://vrk.fi/default.aspx?docid=7809&site=3&id=0 [retrieved February 2014].

[15] Eurostat. Gini coefficient of equivalised disposable income (source: SILC), November 2013. http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=ilc_di12& [retrieved June 2014].

[16] Analysys Mason, http://www.analysysmason.com/

[17] Statistics Finland. Science, Technology and Information Society, http://www.stat.fi/tup/suoluk/suoluk_tiede_en.html [retrieved April 2014].

[18] A. Krendzel, "Network Planning Aspects for 3G/4G Mobile Systems", TUT publication, Tampere, 2005.

[19] The EU Framework Programme for Research and Innovation, Horizon 2020, http://ec.europa.eu/programmes/horizon2020/

 ISBN: 978-1-61208-366-7

Design and Implementation of Co-Presence Transportation for Physical Objects

Lars Fischer

Research Group IT-Security
Business and Information Systems Engineering
University of Siegen. Germany
Email: fischer@wiwi.uni-siegen.de

Julia Dauwe

Operating Systems and Distributed Systems
Electrical Engineering and Computer Science
University of Siegen, Germany
Email: julia.dauwe@uni-siegen.de

***Abstract*—This work introduces a prototype for negotiation-based routing in co-presence networks. The Physical Object Sneaker Transport (P.O.S.T.) is aimed at forwarding physical objects towards their destination using local wireless communication devices (i. e., smartphones) for opportunistic route negotiation. The combination of existing data communication technology with the physical world provides not only interesting challenges, but may also provide novel methods for distributed authentication, payment systems and social networking. The prototype provides the basic functionality for further research and development of protocols and concepts related to spatial, distributed networking of physical and digital objects.**

Keywords–Co-Presence; Opportunistic Routing; Emergence.

I. INTRODUCTION

We have developed an early prototype to support transportation of physical objects in a purely distributed manner using techniques from co-presence networking [1]. The general idea of co-presence networking is to exploit spatial movement of individuals and the occasional contacts between them to transport objects. A *co-presence network* is inherently a distributed network of contact events and unidirectional, ephemeral links between all participants. The scenario motivates research on a wide array of challenges and opportunities that emerge from the combination of distributed networking and corporeal co-presence.

This paper describes the ongoing work on protocol and prototype development for co-presence based transportation of physical objects. At this state of the project, the main objective is to show that routing and communication methods can be practically developed to enable peer-to-peer transportation of physical objects. The underlying objective is to provide a first platform that is based on spatial closeness relations and can prospectively be used to research authentication and payment in distributed, co-presence networks.

Network infrastructure has reduced the effects of and the requirements for spatial closeness in many forms of social interaction. There is, nonetheless, a tight relation between humans and their spatial and temporal location. Corporeal co-location generally is still the fundamental mode of interaction and the key source of social relations. One of the main objectives is to create an application which utilises spatial closeness in the digital domain.

The current version of the prototype is a very fundamental solution for peer discovery, co-presence opportunistic routing and integration of spatial attributes into the digital domain. The prototype is able to connect to instances of the prototype on other devices and negotiate an estimation of the best carrier, based on manually configured data about destinations of participants and objects. It already implements a basic model to protect location privacy by reducing the amount of disclosed information on destinations to a necessary minimum.

This paper is organised as follows: Section II introduces the scenario in the context of related work. In Section III, the protocol scheme is introduced. Section IV describes the architecture of our. prototype. A brief test of the functionality is summarised in Section V. Future work is discussed in Section VI. The paper is concluded in Section VII.

II. SCENARIO AND RELATED WORK

The objective of P.O.S.T. is to transport physical — as opposed to digital — objects on a network of contact events and movements of physical entities, i. e., transport books, letters and other small goods by passing it to the next person moving in the right direction. The idea is that mobile devices establish a co-presence network, discover peers, negotiate routes and handle security protocols. The concept is comparable to Software Defined Networking, with the distinction that [2] human participants have to be involved to handle objects similar to the forwarding layer.

The idea has first been published in [3], where we undertook a first simulation to show general feasibility of the idea. But, co-presence networks are well known in the domain of transportation of digital objects. Transportation of physical objects otherwise is only found in centralised systems, e. g., ride-sharing agencies. Also, analysis shows that the network formed by encounters is scale-free, and thus can provide comparatively short paths for transportation [4].

Distributed transportation of physical objects is related to Delay-Tolerant Networks (DTN) [5]. Research on DTN is striving to engineer *data mules*. The objective of P.O.S.T. is aiming at *mules*, i. e., the transportation of physical — not only digital — objects. Known addressing and routing schemes, for example opportunistic routing [6], can be used to forward objects towards spatial locations. But also, direct addressing

 ISBN: 978-1-61208-366-7

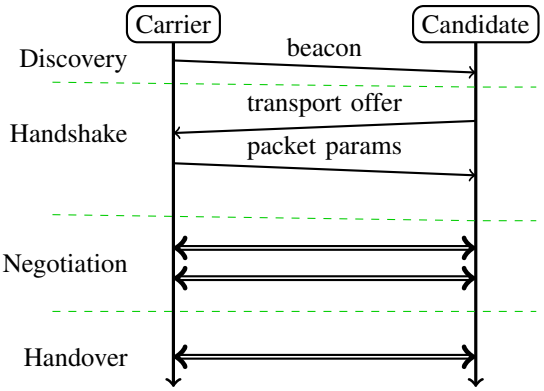


Fig. 1. Protocol Overview

of individual recipients, using *probabilistic routing* [7] can be imagined. While research on DTN for digital data is very well established, transportation of physical objects is, albeit comparable, rarely considered [8].

Routing decisions in this network are based on predictions of future location and contacts. This means that personal information — especially regarding location — has to be disclosed to peers. Formal models of location privacy provide techniques to establish a balance between privacy and efficiency [9]. Recent application concepts for contact-based communication, e. g., MoP-2-MoP [10], already address the topics of effectiveness and privacy for digital communication.

III. Protocol Scheme

This section briefly introduces the four parts of communication of the P.O.S.T.-prototype: Discovery, Handshake, Negotiation and Handover. Participants then either take the role of *carrier* if the considered physical object is in their custody, or of *candidate* if not. This section describes the primary intentions and attributes of the protocol. The final communication standard is still under development.

The *Discovery* of devices that are P.O.S.T.-enabled is implemented by beacon messages. During the initial *Handshake* the candidate decides whether the attributes of an object, e. g., weight, dimensions, safety and security requirements, are agreeable by the user. The *Routing Negotiation* generates a decision about the optimal next carrier based on commitments to itineraries or destinations of participants and objects. The final *Handover* signifies the transfer of custody of physical objects. (See Figure 1).

A. Discovery

The main obstacle for device discovery is to synchronise remote devices within a brief physical link duration [11]. Discovery of P.O.S.T.-enabled devices — in its current implementation — makes use of the Bluetooth Service Discovery Protocol (SDP). Bluetooth has been chosen, because it is wide available in smartphones. Every device subsequently alternates between actively scanning for devices or waiting for incoming connections. Waiting time is randomized by a parametrized amount. A P.O.S.T.-service is recognised if a connection on a common Universally Unique IDentifier (UUID) is established.

An established connection is then handed over to the Handshake protocol below.

B. Handshake

An existing connection, i. e., the communication socket, is used for the *handshake protocol*. Both communication partners take on the role of carrier for all carried objects and the role of candidate for the objects carried by the partner. Starting with the client of the connection at first taking the role of carrier, a carrier sends descriptive data about carried object to the candidate. The data currently includes dimensions and weight of objects. This data explicitly excludes information on the objects or carriers destination which is exchanged only during negotiation.

The candidate then decides whether it is generally acceptable to carry this object. A negative decision ends the protocol for the current object. A positive decision lets both partners enter the route negotiation, described below, for this object.

C. Route Negotiation

Route Negotiation describes the part of the protocol where the decision which of two partners will carry a given physical object onwards from the contact event. The protocol adheres to three principles to thwart attacks. The first principle is that the partner not holding the object must disclose his destination before the destination of the object is revealed. This order of disclosure makes sinkhole attacks more difficult. The second principle states that the final decision is made by the current carrier of an object. The current carrier is entrusted with custody of the object and already in the position to misuse this trust. The third principle demands that no partner will reveal movement predictions and destinations with a higher precision than the partner.

In the first message of the negotiation, the candidate discloses a current prediction of future movement to the current carrier. The precision of this spatial information is reduced to a defined degree to protect the user's privacy. The current prototype implements a prediction of the direction of movement which is cloaked by calculating a cone with a user-defined opening angle. The orientation of the cone is selected uniformly at random from all orientations that contain the original direction within the cone.

The current carrier of an object then decides whether the partner is moving closer to the destination than himself, not moving closer to the destination, or whether he cannot decide definitely. The first two outcomes lead to the termination of the routing negotiation with either a handover of the object following, or not. The third outcome of the decision leads to a request for higher precision sent to the partner not holding the object. The partner may now decide whether he accepts lower location privacy by calculating a new cone, contained within the previous cone which has a smaller opening angle. If a partner is not willing to increase precision of his movement prediction, the negotiation is terminated with a negative result.

D. Handover

In the event of a positive routing decision, the object is handed over to the candidate. *Handover* is the only part

 ISBN: 978-1-61208-366-7

of the protocol where participants have to be involved. The current handover consist of the transfer of data describing the physical object to the new carrier, i. e., dimensions, weight and precise destination. The transfer has to be acknowledged by both involved users. Only if the acknowledgements are communicated, the respective database entries are updated, meaning the former carrier deletes his entry on the object while the new carrier adds an entry to his database.

E. Location Privacy Protection Mechanism

One objective in the route negotiation protocol is to reduce the amount of disclosed information of future destinations of participants to a necessary minimum. The Location Privacy Protection Mechanism (LPPM) [9] utilised here is obfuscation by reducing the precision of destination predictions exchanged with the communication partner. The current prototype implements destination as geodetic location, which is used to calculate the direction of predicted movement from the location of the current contact event.

To reduce the amount of disclosed information the protocol does not exchange the precise direction but a cone opening towards the direction of predicted movement. The opening angle of the cone is used as a measure of the precision of the disclosed information, i. e., the LPPM privacy parameter. The direction of the cone is chosen uniformly at random in a way that guarantees that the precise direction is included within the cone. It is further specified, that both participants disclose information with the same precision, emphasising the equality of both partners in the exchange.

Negotiation is initiated with a user-determined wide opening angle and successively reduced if, based on the disclosed cone, no decision can be reached. The prototype ensures that the opening angle is never reduced below a minimum privacy level as defined by the user. The result of a routing decision, based on a disclosed cone thus has a ternary result. If, finally, the minimum privacy level of either of the participants is reached with no definite result who the best carrier for a given object is, the object is left with the current carrier.

As the chosen privacy level is controlled by the user, the project is not yet able to determine to what degree the LPPM is reduces the quality of routing decisions. This topic is an interesting part of future analysis of the whole socio-technical system.

IV. PROTOTYPE

The project has implemented a prototype for the android operating system that provides the core functionality of P.O.S.T., namely peer discovery, routing negotiation, handover, as well as a related database and a user interface.

The architecture is an extension of the model-view-controller (MVC) primitives [12], introducing a separate user service. The common primitives are represented by database, main service (dæmon) and User Interface (UI). Figure 2provides a structured view on the main components of the prototype.

User Service and *Database* are combined into the model-component and provide the complete state of the local system,

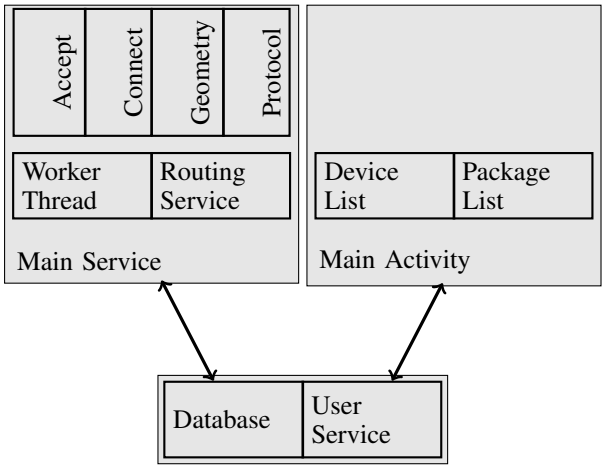


Fig. 2. Prototype Architecture

which contain the user's intentions, including privacy configurations. Within the database currently carried objects and data on the communication status of known devices are stored. In the future, the database will be extended to contain data to authenticate users and handle handover receipts. The idea is to utilise semantic vocabularies and methods to improve the integration into the physical world.

Main Service and *Main Activity* are structured following android implementation recommendations. A *Worker Thread* within the Main Service handles discovery of devices and runs through necessary handshakes alternately using *Accept* and *Connect* subroutines to establish communication. The routing process is then handled by an additional *Routing Service* in order to allow for concurrency, supported by necessary *Geometry* and *Protocol* instances. The *Main Activity* provides views on known devices and carried objects to the user, as well as providing the configuration interface to the *User Service*.

V. TEST RESULTS

The project executed some primary tests that verified that the prototype provides the intended functionality. Tests where executed as a small field test with five individuals carrying P.O.S.T.-enabled smartphones simulating multiple contact-events at a crossing. We tested that communication between devices takes place and that routing negotiation produces the expected results. It could not yet be tested whether objects are indeed propagated over multiple hops.

For the current state of the prototype, the main problem seems to be peer discovery. During the tests with multiple devices, only a fraction of the present devices were positively identified as P.O.S.T.-enabled. Repeated laboratory tests identified mismatched accept- and connect-phases of the discovery protocol. Results from Nayebi and Karlson [11] suggest that the accept-time must adhere to the physical link duration in pedestrian situations.

The tests further identified usability of the prototype as one of the main obstacles for its success. The participants were dissatisfied with manual enrolment of objects. The suggestions hinted towards better automated support for selection of participant destination, object destination and identification of

 ISBN: 978-1-61208-366-7

objects during handover. It was further deemed unsatisfactory, that the objects where handed over to "strangers" without any receipt.

VI. Future Work

The main contribution of the prototype is to provide the context to a rich selection of challenges. Distributed networking in a co-presence world may be natural for human interaction, but it poses very fundamental problems for digital devices. On the other hand, the research community does not seem to have used the advantages and attributes of spatial closeness for security related operations, e. g., validation of identities for authentication.

Among the base communication problems, the peer-discovery is the most urgent problem to solve. The Bluetooth-solution seems to be a dead-end in this respect, but common mobile devices lack a dedicated channel usable for peer discovery. There seem to be two different approaches to be followed here: improving multi-channel communications, i. e., protocols to manage hand-overs from discovery to communication mediums and the exploration of unusual communication mediums, e. g., utilisation of ultrasound or light, depending on circumstances.

The routing scheme must be extended to realistic routing based on road maps. In the next iteration, routing on combined address-spaces, e. g., symbolic, personal and geodetic, will be incorporated. Symbolic representations, i. e., names of places, are better known to users, but the actual destination of an object usually is an individual person or organisation. Distributed transportation may provide a way to reach even mobile destinations of physical objects, the so-called *probabilistic-routing* might be utilised within P.O.S.T..

To enable the routing in this context, a precise prediction of future movements of an individual are fundamental. The main hindrance probably is the computational cost attached to predictive heuristics. It is nowadays common for individuals to provide precise predictions while using navigation software, but a person rarely uses navigation under every day circumstances on known territory. The project intends to use or develop algorithms that detect and exploit regularly visited locations for predictions.

The remaining area of research where the prototype provides a motivating and enabling platform is the wide field of physical security. We identify a need to physically secure the objects against damage and theft. We further require secure proof-of-work schemes to allow for payment or reputation systems. Further, without authenticity, there can be no accountability for lost objects and no penalty for stolen goods.

The project intends to use the prototype to research ways to exploit spatial closeness for authentication and the establishment of reputation. It has been shown by others, that spatial closeness and social relation are in correlation to each other [13]–[16].

VII. Conclusion

This paper introduced a protocol and prototype for transportation of physical objects on co-presence networks that are still under development. The P.O.S.T. prototype provides the basic functionality for minimum-angle routing on geodetic coordinates. This paper discussed a list of areas for research that are opened up by the scenario and whose development is supported by the prototype.

Acknowledgment

The authors would like to thank the student team that did most of the implementation work, in alphabetical order: Patrick Brooks, Johannes Hees, Heinz K. Hiekman, Julian Huperts, Marius Müller, and Niels Stahlhut.

References

[1] S. Zhao, "Toward a taxonomy of copresence," Presence: Teleoper. Virtual Environ., vol. 12, no. 5, October 2003, pp. 445–455.

[2] H. Kim and N. Feamster, "Improving network management with software defined networking," Communications Magazine, IEEE, vol. 51, no. 2, February 2013, pp. 114–119.

[3] L. Fischer, M. Heupel, and D. Kesdogan, "Evolving logistics: Physical-objects sneaker transport (post)," in 8. GI/KuVS-Fachgespräch Ortsbezogene Anwendungen und Dienste, M. Werner and J. Roth, Eds., 2011, pp. pp. 192ff.

[4] V. Kostakos, E. O'Neill, A. Penn, G. Roussos, and D. Papadogkonas, "Brief encounters: Sensing, modeling and visualizing urban mobility and copresence networks," ACM Trans. Comput.-Hum. Interact., vol. 17, no. 1, 2010.

[5] Y. Zhu, B. Xu, X. Shi, and Y. Wang, "A survey of social-based routing in delay tolerant networks: Positive and negative social effects," Communications Surveys Tutorials, IEEE, vol. 15, no. 1, First 2013, pp. 387–401.

[6] L. Song and D. F. Kotz, "Evaluating opportunistic routing protocols with large realistic contact traces," in In Proc. ACM 2nd Workshop on Challenged Networks (CHANTS '07), pp 35–42, 2007.

[7] A. Lindgren, A. Doria, and O. Schelén, "Probabilistic routing in intermittently connected networks," LNCS Service Assurance with Partial and Intermittent Resources, vol. 3126, 2004, pp. pp. 239–254.

[8] A. Voyiatzis, "A survey of delay-and disruption-tolerant networking applications," Journal of Internet engineering, vol. 5, no. 1, 2012.

[9] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux, "Quantifying location privacy," in IEEE Symposium on Security and Privacy, Oakland, CA, USA, 2011.

[10] M. Senftleben, M. Bucicoiu, E. Tews, F. Armknecht, S. Katzenbeisser, and A.-R. Sadeghi, "Mop-2-mop – mobile private microblogging," in Proceedings of Financial Cryptography and Data Security 2014, 2014.

[11] A. Nayebi and G. Karlsson, "Beaconing in wireless mobile networks," in proceeding of: Wireless Communications and Networking Conference, 2009, ser. IEEE Xplore, 05 2009.

[12] T. Reenskaug, "The model-view-controller (mvc) its past and present," University of Oslo Draft, 2003.

[13] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh, "Bridging the gap between physical location and online social networks," in Proceedings of the 12th ACM international conference on Ubiquitous computing, ser. Ubicomp '10. New York, NY, USA: ACM, 2010, pp. pp. 119–128.

[14] S. Pan, D. Boston, and C. Borcea, "Analysis of fusing online and co-presence social networks," in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, March 2011, pp. pp. 496–501.

[15] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: Improving geographical prediction with social and spatial proximity," in Proceedings of the 19th international conference on World wide web, ser. WWW '10. New York, NY, USA: ACM, April 26–30 2010, pp. pp. 61–70.

[16] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins, "Geographic routing in social networks," PNAS, vol. 102, no. 33, August 2005, pp. pp. 11 623–11 628.

 ISBN: 978-1-61208-366-7

The Connectivity Control Framework: Enabling Session Continuity in Multi-Domain Environments

Michelle Wetterwald
HeNetBot
Sophia Antipolis, France
e-mail: michelle.wetterwald@henetbot.fr

Christian Bonnet
EURECOM
Sophia Antipolis, France
e-mail: christian.bonnet@eurecom.fr

***Abstract*— Last decades have witnessed a massive evolution of mobile communications. When no agreement between the network providers exists, changing the attached network still means breaking the session and relying on the application to recover the lost data. A large set of mobility solutions has been proposed, which impact the network architecture or the applications communications methods. To cope with this issue, this paper presents an innovative technological framework, which applies changes to the terminal only and ensures the continuity of the session when roaming through independent wireless access networks. This framework is based on abstract interfaces hiding the specificities of technologies, a shared knowledge base constantly improved by system learning, and generic service enablers dedicated to specific connectivity tasks, such as a socket session handler. A simulated model based on a heterogeneous wireless playground is used to prove the benefits of this distributed system, which is easily suitable for deployment.**

Keywords-heterogeneous networks; IEEE 802.21; session continuity; multi-domain mobility; generic service enablers; device abstraction; autonomous systems.

I. INTRODUCTION

The evolution of mobile communications has generated new challenges for the design of the connectivity functions in future terminals. The trend has been the conception of multimode devices, with an increasing number of interfaces, and able to connect to any available network. A multimode Mobile Terminal (MT), such as a laptop, smartphone, tablet or car device, is equipped with several network interfaces and able to support communications through one or several of these interfaces at a given time. In parallel, users' requirements in terms of Quality of Experience (QoE) have been soaring, triggering a massive effort from network designers and the conception of devices more and more complex. Because of the turn up of various wireless standards and technologies with different properties, mobile networks have become heterogeneous, incorporating several types of access technologies under the same administrative domain. These accesses provide different connectivity characteristics to the user applications and protocols and require additional adaptability and system control at higher levels of the protocol stack to allow seamless roaming through the different accesses. Roaming and mobility across heterogeneous networks are thus part of the critical operations under study in mobile communications. Currently, when no federation exists between two mobile network providers and no mobility-specific mechanism is deployed in the wireless network, roaming very often means that the session hosting the running application is broken and must be restarted manually, at the cost of lost data, except if the application is designed to recover by itself.

Most of the existing popular user applications use Transmission Control Protocol (TCP) as basis or for the control of their data transfer. TCP has been designed as a stationary protocol, so when the identifier of one of its endpoints, i.e., its Internet Protocol (IP) address and socket port, changes, the TCP connection fails and is terminated. Some of the applications freeze or stop their execution while others are set to establish a new TCP connection and resume their activity. But at the end, it all depends on the way the application itself was developed.

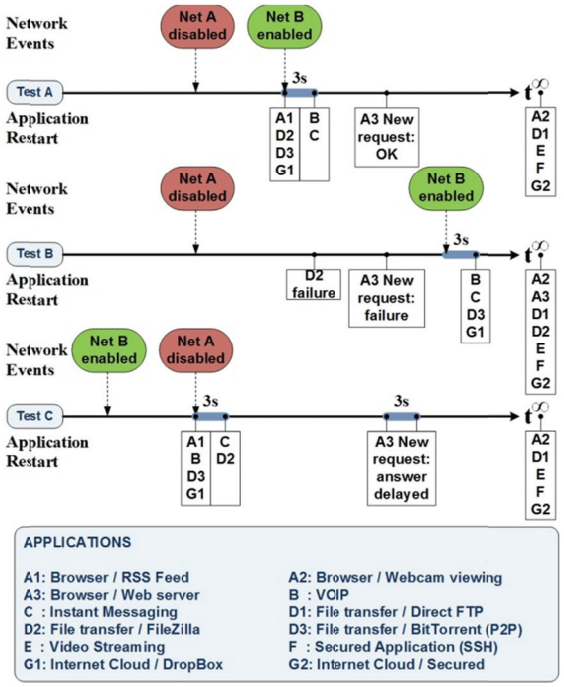


Figure 1. Application Recovery Timeline

 ISBN: 978-1-61208-366-7

Fig. 1 shows the result of experiments that have been performed using a real terminal and applications to frame the issue of application session continuity. The terminal was equipped with two network interfaces accessing independent networks. For each test, the application was started on NetA, then Net A was disabled and NetB started with different delays, as shown in the figure. We could observe that, while some of the popular user applications tested are able to recover by themselves thanks to their smart design, others are frozen or completely stopped, as shown with an infinite time recovery on the right of the figure.

So, the primary objective of this paper is to tackle the problem of the session failure when changing to an access network that does not support mobility. The result will allow an individual using a mobile device to roam seamlessly across non-federated heterogeneous wireless environments. Such environments can be a mobile operator network, a campus hotspot, a road operator communication network or the user private home network.

Several networking techniques such as Media Independent Services, mobility management, or autonomous systems can contribute to achieve our objective, but none of them provides the solution on its own. By enhancing and combining efficiently these mechanisms, the target scheme involves a strong level of cross-layer design and enables many generic services such as handovers, broadcast services, session mobility, battery saving or security.

In this contribution, we propose an innovative framework to resolve at MT level the problem of session continuity between independent domains and leave the network totally unaffected. The rest of the paper is organised as follows. Section II discusses the existing technologies available to address this type of issue, analysing their potential contributions to an integrated framework and their limitations. After this, in Section III, we propose our integrated framework and describe its internal components. This is followed in Section IV by its evaluation with a simulation model implementing the proposed system, and includes its main results. The document is closed in Section V with an assessment of the contribution and the indication of direction for future research topics.

II. ANALYSING EXISTING TECHNOLOGIES

In this section, the existing technologies and challenges lying in the path of the target architecture are identified and reviewed.

A. Media Independent Services

Operating multimode devices in heterogeneous networks can become very complex if each access technology has to be controlled directly and separately by the upper layer entities. This has led to the emergence of a strategy based on a shared abstraction layer above the access layer. In this direction, the IEEE 802.21 standard proposes three different Media Independent Handover (MIH) Services [1], which offer to the upper layer management protocols some generic triggers, information acquisition and the tools needed to perform handovers. The Event Service (MIES) provides the framework needed to manage the classification, filtering and triggering of network events, and to report dynamically the status of the different links. The Command Service (MICS) allows the upper layer management entities to control the behaviour of the links. The Information Service (MIIS) is distributes the topology-related information and policies from a repository located in the network. They result in a cross-layer architecture where the Media Independent Handover Function (MIHF) operates as a relay between the media-specific Link layer entities and the media-agnostic upper layer entities, or MIH-Users. In existing solutions, the MIH-User is represented by a Connection Manager (CMGR) whose main role is to decide which path if best suited to reach the application server or the Correspondent Node (CN) located across the Internet [2].

Currently, the IEEE 802.21 standard provides valuable mechanisms to control the network interfaces of a multimode terminal in a media-independent and abstracted way. However, it involves a few strong limitations. It currently only enables handover services and deals exclusively with the control of wireless network interfaces. It does not consider the information from other devices, such as battery consumption or positioning, in the terminal. It thus offers the possibility to be developed to support an extended set of services and devices in the terminal. This extension will be a main axis for the design of the target solution.

B. Handling Mobility

The most recent mobile devices are expected to be usable while walking on the streets, carried in road vehicles or even in fast speed trains. However, when a device moves out of its original routing area, it cannot continue using the same IP address and the executing session is broken. Incoming packets are still forwarded along the former route and are not able to reach the mobile anymore. To solve this problem, the IETF groups address the issue of mobility at various levels of the protocol stack. The solution coverage is wide spread as well: at device, transport, session, application, or even more recently, at flow level. The objective is to design protocols able to survive the change of the terminal environment context or discontinuities of its connectivity. A large set of mechanisms and protocols has been proposed to solve this issue. Mobile IP and its enhancements, Fast Mobile IP or Proxy Mobile IP (PMIP), operate at the network layer level. Other protocols like mobile Stream Control Transmission Protocol (mSCTP) or Session Initiation Protocol (SIP) address the transport layer level or above [3]. All these solutions thus affect the network or transport layer. They depend on control entities located in the network that must be owned and maintained by specific organizations. They most often infer heavy changes to the network architecture, including in the anchor point, or to the communication interface at the application, and thus face strong unwillingness for their deployment. For their part, the cellular systems handle mobility with 3rd Generation Partnership Project (3GPP) proprietary protocols and procedures [4], sometimes adapted from the previous ones.

Beside these continuous mechanisms, an interesting technique named Delay Tolerant Networks (DTN) allows the mobile nodes to survive long connectivity disruptions.

 ISBN: 978-1-61208-366-7

Intermittent connectivity is overcome by using store-and-forward message switching [5]. Whole or pieces of a specific message are moved between persistent storage nodes (called DTN nodes), which buffer the message pieces for long periods of time until they are able to forward them to the next DTN node. This functionality is provided by an end-to-end message-oriented overlay, called the "Bundle layer", which is inserted between the application and the transport layer. Since the DTN nodes terminate transport protocols at the Bundle layer, it makes this architecture tolerant to delay and connectivity problems. However, these techniques are over-sized compared to the requirements of a short handover

C. *Automating the System*

Recent conceptual studies of future network architectures introduce a totally new cognitive plane, where the environment is sensed and observed, leading to the acquisition of knowledge. This is exploited in a novel capability of self-management [6]. The system operates by undertaking intelligent control loops [7]. It senses its operating environment, works with models that analyse its own behaviour in that environment, and, based on existing policies and learned knowledge, derives the appropriate actions to adapt and change the environment, its own state or its operation. A basic knowledge source is installed at setup and further enhanced by self-learning in an evolutionary process through progressive steps. These self-management architectures have been designed in a layered fashion with a hierarchy of decision modules monitoring the information retrieved from sensors and actively coordinating the action of executors, while maintaining a common cross-layer knowledge base. They are currently used for cognitive radio or the management of network infrastructures. By mirroring their functionalities, it sounds interesting to apply the same concept to the self-configuration and self-healing of the MT connectivity in order to optimize its operation.

The individual technologies that have been analysed in this section can be adapted to obtain a better optimized solution. Their efficiency can be improved by combining them into a single framework. Their common factor is the mobile terminal, which is the only node that the end user controls. Accordingly, in the remainder of this study, an innovative approach has been adopted, choosing to apply the designed changes to the mobile terminal only and leaving the network totally unaffected. The connectivity has to be maintained efficiently while remaining transparent to the applications. The system should capitalize on the layered architecture introduced in autonomous systems.

III. THE CONNECTIVITY CONTROL FRAMEWORK

Accordingly, this section proposes a solution based on a cross-layer architecture that leverages the optimization of dedicated generic services. These services operate in close relationship with an abstraction layer, which hides and takes care of the specificities of the embedded devices. Different services, such as access network selection, connectivity and mobility management, and application session management are combined and enhanced to reach the objective of seamless connectivity. Moreover, the new cognitive capabilities of autonomous systems are involved to bring autonomy to the roaming and support a faster decentralized operation.

Following the requirements defined above, the resulting layered system, the Connectivity Control Framework (CCF), pictured in Fig. 2, modifies only the MT, leaving the network unchanged. It revolves around three main principles that guarantee a simple and flexible architecture, and which could be summed up in a simple modification of the terminal operating system. The first principle is to share the knowledge about the terminal context and its environment in a cross-layer fashion. This is achieved by the Cross-Layer Agent (CLA). It stores the configuration, policies and status of the whole framework in a Local Information Base (LIB). The second principle is to hide the heterogeneity and diversity of the internal devices and access networks behind an abstract interface, which facilitates a range of services wider than handover management. This is achieved by the Media Independent Services Function (MISF) and the Link Interfaces, inspired from the MIH model. The third principle is to provide coordinated generic service enablers that can take care of dedicated operations. They ensure the terminal seamless and optimized connectivity and its operational behaviour, coping with dynamic changes and events in the network environment, while preserving the application data transfer continuity. This is achieved by the Network Access Generic Service Enabler (NAGSE), the Mobility Generic Service Enabler (MGSE) and the Session Generic Service Enabler (SGSE). To enhance their efficiency, the Cognitive Manager (CM) has been introduced. It coordinates autonomously the actions of the GSEs and relies on human interaction (component User Interactions Application (UIA)) only when the level of confidence of its self-management algorithms is too low. The GSEs and the CM are integrated in the Connectivity Agent (CA) sub-system.

The components defined in the CCF combine their individual actions in order to bring the whole framework to its expected level of resilience and efficiency.

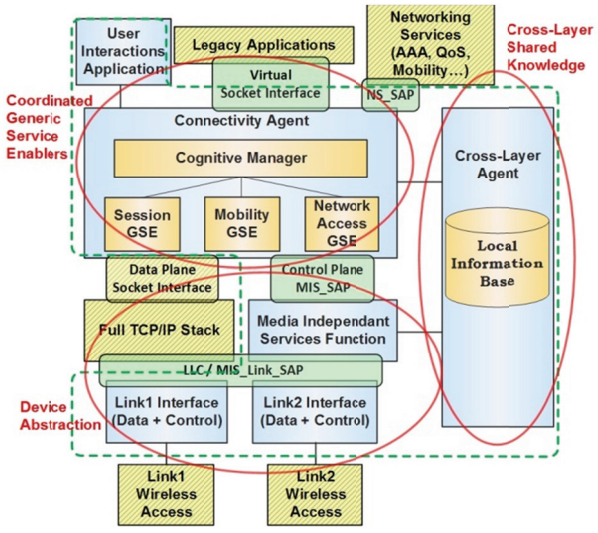


Figure 2. Global architecture of the CCF

A. Shared Knowledge

Implementing the first CCF concept, the CLA gathers the data from the devices and Link Layer technologies and from the upper layers, aggregates them in the LIB and provides them on request when needed to the other CCF components.

The cross-layer approach adopted here is a hybrid of two common solutions, in order to integrate their benefits. A cross-layer engine, the CLA, is introduced that works as a local Information Server. It manages a local storage, the LIB, making it accessible to the other components in the framework, namely the MISF and the sub-components of the CA, while preserving its integrity. In parallel, direct interactions between the adjacent components (CM, GSEs, MISF, and Link Interfaces) are maintained to transfer related events and commands, according to the layered model. This scheme distributes the complexity and ensures a quick response of the overall framework to changes in the external environment.

The LIB is the shared knowledge source for the whole system. It contains all the data relevant for an optimized operation of the framework. These data are classified in three types: (i) pre-defined information stored at configuration time, either by the user or by accessing remote databases at the network operator servers (e.g., MIIS) or in the cloud, (ii) status information about the mobile and its environment reported by the other CCF components, (iii) policies and utility functions resulting from the learning process.

B. Device Abstraction

The MISF is an enhanced abstraction layer responsible for dealing with the wireless multimodality of the terminal. It is a key component of the connectivity optimization process, as it also provides the means for the abstracted interaction between the radio access and the upper layers. It is based on the IEEE 802.21 MIH model, keeping only its local components, but is not restricted to handover; it fully manages the wireless accesses and the other devices in the terminal, hiding their specificities to the upper layers. It provides a whole set of additional services, including system statistics and status retrieving, resource configuration to comply with a certain level of Quality of Service (QoS), setting and getting identities, handling power sources, positioning, etc. Moreover, it also provides an abstract interface to the CLA component, directly forwarding to the local storage the network or device information received, and contributing to the system learning. The MISF is later able to retrieve the link parameters when requested to issue a command to the lower layers, avoiding that the upper layers get involved with the device details.

The Link Interface components make the link between the MISF and the technology drivers. There is one Link Interface per type of device, completely specific to its implementation. Its main function is to translate the MIS commands and forward them to their target destination. It acts as the endpoint for parameters retrieval from the device. Its location at the edge of the CCF minimizes the overall energy and processing power consumed by the framework. Before generating events, it smooths the values of retrieved dynamic parameters and applies hysteresis thresholds to avoid unnecessary and too frequent reactions. When related to the mobile connectivity, internal devices (positioning systems, power supplies or other sensors) can be integrated and coordinated through the CCF and the MISF, provided the availability of a Link Interface component that translates the abstract MIS primitives into the corresponding set of commands.

C. Coordinated Generic Service Enablers

The Generic Service Enablers, or GSEs, are the key elements of this framework. They allow the legacy services to benefit from the technology-agnostic framework. They complement at service level the abstraction introduced by the MISF. These functional blocks are called generic because each of them provides a set of specialized functionalities; they take care of the specificities of the applications and legacy Network Services (NS). They act as MIS-Users and hide the MISF interface to their own users. They can query the LIB for aggregated relevant cross-layer metrics and to provide them to the upper layer services.

The NAGSE deals with aspects related to the monitoring of the networks availability, learning the characteristics of the unknown accesses and selecting the best access network by running its algorithm on a set of parameters retrieved from the CLA. The reader is referred to [8] for more details on its functionality.

The MGSE is the most commonly developed part of the CA. Its role is generally included in the CMGR. It takes care of connectivity related services, including network interfaces and link management, networking aspects, reception and filtering of network and device events, keeping track of current location and connectivity. The MGSE enables the capability to be connected to different types of networks using abstracted processes and mechanisms. Moreover, it smartly filters and dynamically reports changes of the MT context, whether internal or in the external network environment. The process of the MGSE is completely independent from the existing mobility protocols, such as Mobile IP, that may be available in the network and/or the terminal and would run as part of the NS. It thus provides an enhancement to these mechanisms.

The SGSE deals with aspects related to the management of the data sessions opened by the applications. It also takes care of the availability of resources according to the application QoS requirements. For each user application, two related addresses are used. A Personal Address (PA) is attributed to the application when it starts. This address is kept identical throughout the whole session [9]. The Local IP Address (LA) depends on the user network location. It is the global address seen by the external network nodes. In case Mobile IP is used, the LA is equal to the Care of Address (CoA). In the data plane, the SGSE performs the address translation between the PA seen by the application and the LA seen by the network. As will be demonstrated in the validation part in Section IV, this action has a very low impact of the performance of the whole system. For the

 ISBN: 978-1-61208-366-7

duration of handovers, it executes a buffering mechanism inspired from DTN techniques, storing the packets received from the application in a temporary local storage until a new connection is safely established. It re-starts the TCP connectivity when it has been broken due to the change of network connection. When an application starts, it opens a socket on a Virtual Socket interface, providing the address and port number of the destination. From the application, this interface is seen as the standard socket Application Programming Interface (API) unchanged. Then, the SGSE opens a real socket on the TCP/IP stack, with the same properties as the virtual one, providing the LA as source address. During an inter-domain handover, the SGSE copes with the break of the session by automatically re-establishing a fresh one through the new access network, using the same parameters. It then updates the binding between the PA and the new LA in order to transfer the flow to the new session. In case TCP is used, it makes sure that the TCP congestion window is set to the same value as it was with the previous session before it failed, in order to avoid the slow start mechanism and reduce the impact of the handover to the local application.

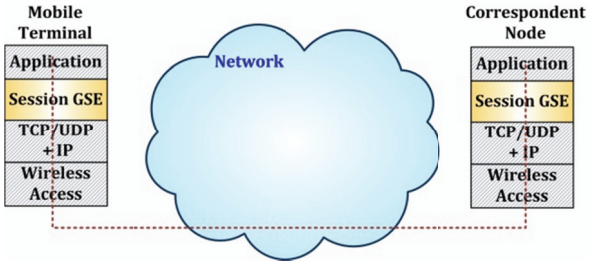


Figure 3. Communication with the correspondent node

The session continuity is ensured at the CN as well, whether it is a server or another mobile terminal, and despite the break of the TCP connection. A simplified version of the CCF, including only the SGSE, is present in the CN and executes the same address translation and packets buffering as in the MT. It is thus able to hide the change to its endpoint of the application. This solution is scalable, as it does not require any intermediate node. The mobility is handled above the transport layer; it is compatible with existing mechanism and protocols, thus opening the path to fast acceptability. The SGSE brings to the framework the mechanisms that allow it to recover when a terminal movement has endangered the operation of a running application.

The GSEs are directly interacting with the MISF, each of them dedicated to a specific role. They need to be coordinated to provide an integrated autonomic behaviour. The CM plays the role of the system controller, orchestrating the self-management functions in the MT to increase the level of efficiency of the global framework. It is assisted by the UIA, a simple user interface to the human owner of the terminal. The CM obtains information and triggers actions from the GSEs. It coordinates their actions according to its own state and the events received. It makes the operational decisions using the knowledge source in the CLA, and triggers their execution. The CM brings to the framework the capability to operate in a smart and autonomous manner, hiding the complexity of maintaining the network connectivity from the mobile user. The overall terminal operation benefits from increased robustness, adaptability to internal or external events and enhanced effectiveness.

The UIA establishes the link with the mobile user and is the component which allows him to control the operation of his mobile according to his needs and requirements. It obtains the user preferences at configuration time or is used to validate the CCF decisions at runtime, i.e., every time some human-originated information is needed. Such an interaction makes sense and is expected to become less tedious with the apparition of more intuitive user interfaces based on voice rather than screen pop-ups.

IV. MODEL VALIDATION AND RESULTS

This section aims at validating the chosen technological scheme. The test displayed in Fig. 1 could be reproduced with a simulation model that was developed with the objective to demonstrate the capabilities and the benefits of the framework approach. The model focuses on adding the CCF components to a wireless multimode terminal and moving randomly that terminal in a sample wireless network. By doing so, it evaluates the impact of the framework in the context of the scenario described above, using a selection of common applications (file transfer, Web browsing). The framework here is assessed on its efficiency rather than traffic throughput perspective. The evaluation criteria are the number of data bytes that did not arrive at destination, the recovery from broken TCP sessions and the time between two handovers to control the Ping-Pong effect between the two access networks.

The prototype has been developed as a simulation running under the OMNET++ tool [10]. All the components of the CCF framework shown in Fig. 2 are implemented and tested in the simulator. A small part of the features are streamlined due to the wide range of services described in Section III. The following statistics are collected during each simulation run: number of bytes transmitted and received by the applications, number of TCP connections opened / broken during the test, number of handovers (HO) and time between two handovers, connection time on each technology and in total, usage of the DTN buffers and DTN process duration, time when the last packet was received by the application in the mobile. Each test was performed under steady state conditions, with a fixed duration of 2000s of simulation time and is reproduced 50 times to allow a reasonable simulation time, while providing sufficient data for results aggregation, as shown in Fig. 4b and Fig. 5b.

The results obtained with the CCF prototype-enabled terminal moving randomly across the playground are compared with two other similar use cases that do not involve any change in the network and transport layers or in the network infrastructure: standard multimode terminal with no connectivity control and mobile terminal equipped with a CMGR equivalent to current smartphones. The results

 ISBN: 978-1-61208-366-7

obtained from the simulation runs, shown in Fig. 4 and Fig. 5, focus on the critical issue of the TCP session continuity and recovery.

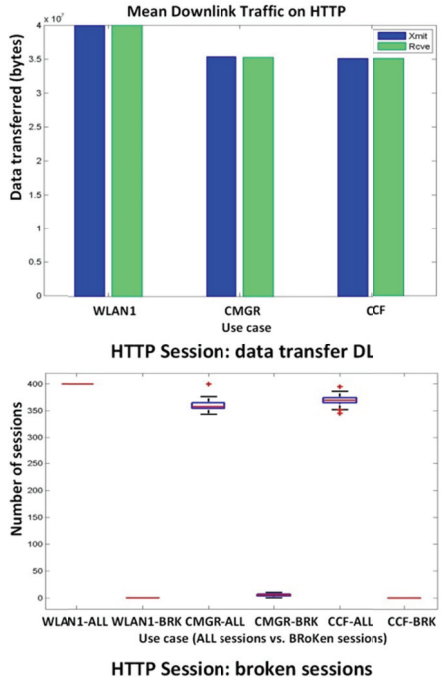


Figure 4. Simulation results for web browsing

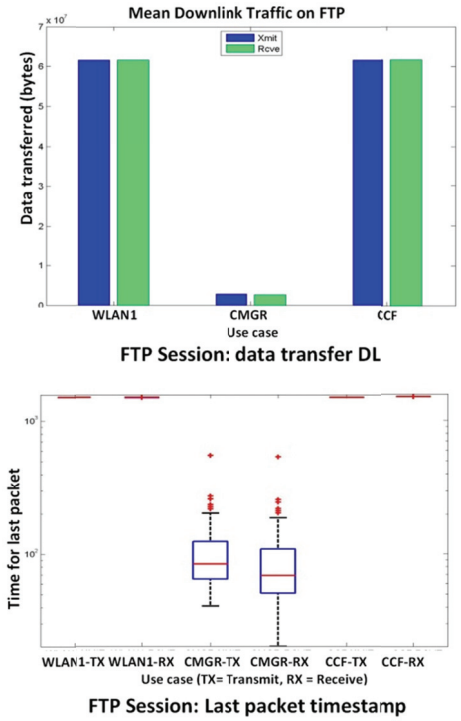


Figure 5. Simulation results for file transfer

When executing the web browsing application, no session is broken anymore midway to its completion (CCF-BRK box in Fig. 4b), as happened with the CMGR (CMGR-BRK box in Fig. 4b). The amount of traffic transferred is slightly reduced, because the application is interactive and a larger part of the data traffic is transferred through the safer cellular access which, on the other hand, offers a reduced bandwidth.

These results show that when executing a file transfer application, the virtual TCP connection is not broken anymore. With the CMGR (result in the middle of each graph), the traffic exchange stopped after the first handover and resulted in a very low quantity of data exchanged. With the CCF, the application can complete its task until the end and transfer all its packets, as in the reference case on the left. Some of them require the additional support of the buffering queue technique, as can be seen in Table 1.

Table 1 shows the usage of the buffer queue in the SGSE. A few packets are saved during the connectivity break and sent when it is complete, which fully prevents packet loss. We could observe that their number during a specific HO remains very low, at a level quite acceptable compared to the amount of memory available in a mobile terminal.

TABLE I. MEASURES FOR BUFFERING MECHANISM

Parameter	Use Case	
	FTP	***HTTP***
Average Number of Handovers	15.65	15.47
Total packets queued - max	36.00	10.00
Total packets queued - min	2.00	0.00
Max packets queued during 1 HO	3.00	1.00
Min packets queued during 1 HO	2.00	0.00

Finally, statistics have been collected that demonstrate the low fingerprint of the CCF operation on the MT processing power and the usage of the buffering mechanism, as it is measured smaller than 0.2%. On a specific test where the MT executes the file transfer application and performs 12 handovers, only 1233 events (over 761839 for the full simulation run) did involve one of the CCF components. It represents less than 0.2% of the total operations.

V. CONCLUSION AND FUTURE WORK

This paper has presented the design and evaluation of a cross-layer, integrated and coordinated framework, the Connectivity Control Framework, whose target is to ensure session continuity across independent wireless networks. The CCF is restricted to the mobile terminal and has no impact on the mobile network infrastructure, while maintaining full compatibility with existing networking standard. Its layout is based on three main principles: (i) a cross-layer agent, which maintains and shares the knowledge acquired by the other components of the framework; (ii) an abstraction layer, which hides the network specificities to the rest of the framework, including as well the support of other hardware devices such as positioning systems or diverse sensors; (iii)

 ISBN: 978-1-61208-366-7

coordinated generic service enablers responsible of dedicated tasks and taking care of the various functions necessary to handle the terminal connectivity. To assess its benefits, a simulation model has been developed, that compares the framework behaviours in a testing heterogeneous environment to other solutions that do not affect the network. The concept of keeping the changes in the mobile device distributes the effort in the global system, reduces the risk of bottleneck functions in the network and improves its scalability. No additional network entity has to be deployed and maintained by the operators, the system installation and configuration are simplified. This study has investigated comprehensively the distribution of functions and the integration and coordination of the system proposed. For each individual component, existing solutions that satisfied the main requirements have been selected and necessary enhancements described. A future continuation of this topic will perform a more precise analysis and definition of the generic service enablers introduced in the CCF.

REFERENCES

[1] E. Piri and K. Pentikousis, “IEEE 802.21”, the Internet Protocol Journal, Volume 12, No.2, June 2009, pp. 7-27.

[2] M. Kassar, B. Kervella, and G. Pujolle, “An overview of vertical handover decision strategies in heterogeneous wireless networks”, Computer Communications, Volume 31, Issue 10, June 2008, pp 2607-2620.

[3] G. Lampropoulos, N. Passas, L. Merakos, and A. Kaloxylos, "Handover management architectures in integrated WLAN/cellular networks", IEEE Communications Surveys & Tutorials, February 2006, pp 30-44.

[4] 3GPP TS 23.401, "General Packet Radio Service (GPRS) enhancements for Evolved Universal Terrestrial Radio Access Network (E-UTRAN) access", Release 10, June 2012, available at ftp://ftp.3gpp.org, last accessed on 05/2014.

[5] F. Warthman, "Delay-Tolerant Networks (DTNs): A Tutorial; v1.1", Wartham Associates, 2003.

[6] J. Larsen, "Cognitive Systems", tutorial presented at IEEE Workshop on Machine Learning for Signal Processing, Cancun, Mexico, October 2008.

[7] IBM Corporation. “An Architectural Blueprint for Autonomic Computing,” White Paper, 4th Edition, June 2006.

[8] M. Wetterwald and C. Bonnet, "Devices and wireless interface control in vehicular communications: an autonomous approach", Ambi-sys 2013 , Athens, March 2013, pp 91-103.

[9] R. Bolla, R. Rapuzzi, and M. Repetto, “An Integrated Mobility Framework for Pervasive Communications”, Proc. of IEEE Globecom 2009 Next-Generation Networking and Internet Symposium (GC’09 NGNI), Honolulu, Hawaii, USA, Nov.-Dec. 2009, pp 1-6.

[10] OMNET++ simulator, available at http://www.omnetpp.org/, last accessed on 05/ 2014.

 ISBN: 978-1-61208-366-7

Performance of Novel Target Detection in Radar Network Systems with a 3D Vehicle Model

Hiroyuki Hatano
Department of Advanced Interdisciplinary
Graduate School of Engineering,
Utsunomiya University
Tochigi, Japan
e-mail: hatano@is.utsunomiya-u.ac.jp

Masahiro Fujii, Atsushi Ito, Yu Watanabe
Department of Information Sciences
Graduate School of Engineering,
Utsunomiya University
Tochigi, Japan
e-mail: {fujii, at.ito, yu}@is.utsunomiya-u.ac.jp

Yusuke Yoshida, Takayoshi Nakai
Department of Electrical and Electronic Engineering
Graduate School of Engineering,
Shizuoka University
Shizuoka, Japan
e-mail:yoshida@hatanolab.eng.shizuoka.ac.jp
e-mail:tdtnaka@ipc.shizuoka.ac.jp

***Abstract*—We focus on forward-looking systems with automotive radar network systems. By using multiple radars, the radar network systems will achieve reliable detection and wide observation area. The forward-looking systems by cameras are famous, but not all-around system. In order to realize more reliable safety, the cameras had better be used with other sensing devices such as the radar network. In the radar network, processing of the data derived from the multiple receivers is important because the processing decides the detection performance. In this paper, we will introduce our data processing and detection algorithm. Finally, the performance will be evaluated via a 3D target model. From results of computer simulations, we can confirm that our proposal can achieve stable detection even if the target positions differ.**

Keywords-Radar Network; forward-looking radar; multiple radars; Wide detection

I. INTRODUCTION

By applying intelligent devices, more safety and comfortable driving is desired. Intelligent Transportation Systems (ITS) is considered to solve some transportation problems such as an accident, a traffic jam and an environmental pollution. The forward-looking alert or braking system is one of the elemental technologies for the realization of ITS world as described by Rasshofer and Gresser [1]. For the forward-looking, various devices are now researched and some systems are realized. Examples are shown in the researches by Meinecke et al. [2] and Sakamoto [3]. Especially, image processing technologies with cameras are famous. Sensing by the image processing can detect targets in wide area. Such system can alert sudden pedestrians from blind spots. In such case, the wide detection for wide are is needed.

However, the image processing has fatal weaknesses. The poplar cameras cannot achieve the adequate performance under optical disturbances, such as bad weather. In order to realize more reliable safety, the cameras had better be used with other sensing devices. In this paper, we focus on radar sensors as other devices.

For achievement of wide and reliable detection, we focus on radar network systems which have multiple receivers. Fig. 1 shows the example structure of the radar network. By using multiple distributed radars, wide observation area and detection will be realized. So, these systems have been researched as the forward-looking systems in automotive usages. The similar structures of the radar network are also researched by Klotz et al. [4] and Folster et al. [5].

In the radar network systems, it is important to process the data derived the from multiple receivers because the process decides the detection performance. In the articles [6] [7], we have proposed some algorithm for position estimation in the forward-looking radar network systems for automotive. In order to estimate target positions precisely, our methods regard the distances to the targets as stochastic variables. Then, the target positions are derived from the calculated probability, which means “target existence”.

As our past works, we have discussed our novel estimation algorithm, Existence Probability Estimation Method (EPEM) and Existence Probability Estimation Method using Reflected Signals (EPEMR) in our articles [6] [7]. However, the evaluation is simple simulator as the target is single point. In this paper, we will evaluate the detection performance by 3D target models.

This paper is organized as follows. In Section II, we introduce the position estimation algorithm briefly. In Section III, we present the simulation settings for the evaluation. Especially, the simulation tools, data processing, simulated cases and results are described. Finally, Section IV summarizes

 ISBN: 978-1-61208-366-7

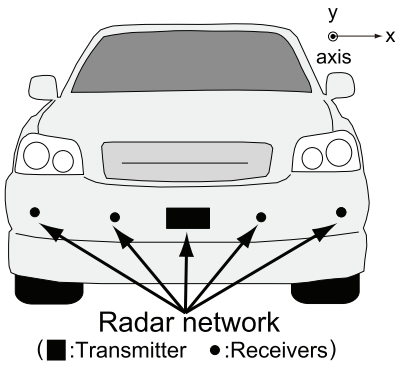


Figure 1. Example of radar network structure

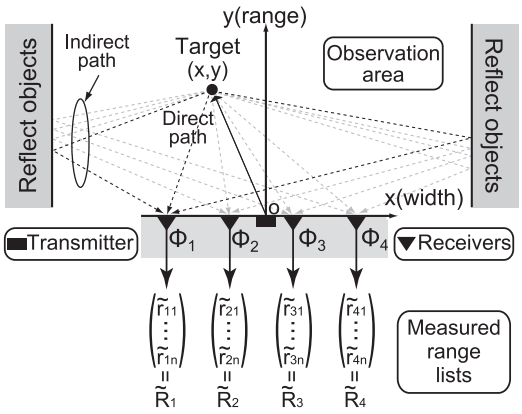


Figure 2. System model of radars and target

the paper.

II. Position estimate algorithm

A. System model

In this section, we will present our system model. The radar network is constructed with a transmitter and multiple receivers (Fig. 2). We assume four receivers. The transmitter is set up at a origin of x-axis. The four receivers are set up in equal interval. The center of the receivers is also the origin of x-axis (see Fig. 2). x-position of the receivers are $\phi_1, \phi_2, \phi_3, \phi_4$ [m] respectively. The position of a target is (x, y). The kth receiver outputs measured ranges. We obtain measured range lists $\tilde{R}_k = \{\tilde{r}_{k1}, \tilde{r}_{k2}, ..., \tilde{r}_{kN_k}\}$ from the kth receiver. The variable N_k refers to the number of measured ranges in the measured range list of the kth receiver's. The variable $\tilde{r}_{kn}$ refers to the distance of wave propagation, that is, the sum of the distances from the transmitter to the nth reflection point and from the reflection point to the kth receiver. The N_k reflection points include, of course the target and other objects such as walls.

The measured range $\tilde{r}_{kn}$ has the measurement error. The error is modeled as follows.

$$\tilde{r}_{kn} = r_{kn} + \epsilon_k \tag{1}$$

The variable r_{kn} means the real distance between the nth reflection points, the transmitter and the kth receiver. The variable ϵ_k means the measurement error which is modeled as a random variable with variance σ^2. Also, the notation "˜" means measured values. The position of the target has to be estimated by the above measured range lists $\tilde{R}_k$ of all receivers.

B. EPEM

EPEM is used as the position estimation method. Our final estimation method EPEMR is based on EPEM. In this section, we will introduce EPEM briefly. The detailed algorithm of EPEM is introduced in our past article [6].

EPEM estimates the target position by calculating the existence probability of the targets, which is explained in this section. In this method, the measured ranges, that is $\tilde{r}_{kn}$, is regard as the random variables.

In order to estimate the target position, EPEM calculates the following conditional probability.

$$P(\hat{x}, \hat{y} \mid \tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \tilde{R}_4). \tag{2}$$

The above probability means that the target exists on the coordinate (x, y) when the measured range lists $\tilde{R}_1, \tilde{R}_2, \tilde{R}_3, \tilde{R}_4$ are obtained. The notation "ˆ" means estimated values.

By using Bayes'theorem and assumptions, (2) can be transformed. The following equation has the same distribution shape of (2).

$$\prod_{k=1}^{4} \sum_{n=1}^{N_k} P(\tilde{r}_{kn} \mid \hat{x}, \hat{y}) \tag{3}$$

$$= \prod_{k=1}^{4} \sum_{n=1}^{N_k} P(\tilde{r}_{kn} \mid \hat{r}_k) \tag{4}$$

where is $\hat{r}_k = \sqrt{(\hat{x} - \phi_k)^2 + \hat{y}^2} + \sqrt{\hat{x} + \hat{y}}$. The probability $P(\tilde{r}_{kn} \mid \hat{r}_k)$ means the probability of getting the measured range $\tilde{r}_{kn}$ when the target exists in the range $\hat{r}_k$. This means the measurement characteristic which each radar has. The measurement characteristic means the error ϵ_k in (1).

From the measured range lists and (4), we can calculate the distribution of the probability which means the target exists on the coordinate (x, y). The distribution of (4) is called as "existence probability distribution". The high probability in the above distribution indicates the target position. An example of the existence probability distribution is shown in Fig. 3.

C. EPEMR

EPEM does not have enough accuracy. EPEM tends to generate large error in the same direction to the receiver layout. In order to improve estimation accuracy, we also construct novel estimation algorithm "EPEMR". EPEMR uses not only direct path from the target but also indirect path which is reflected other objects (see Fig. 2). By using indirect paths, EPEMR can observe the target as the target is surrounded by both the real and virtual receivers. EPEMR

 ISBN: 978-1-61208-366-7

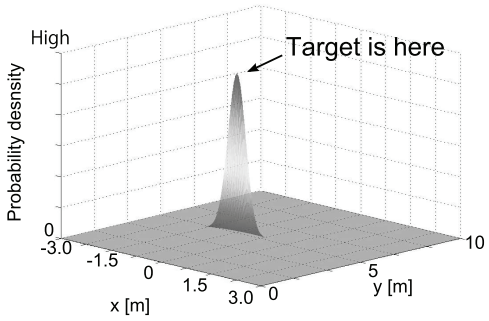


Figure 3. Example of existence probability distribution

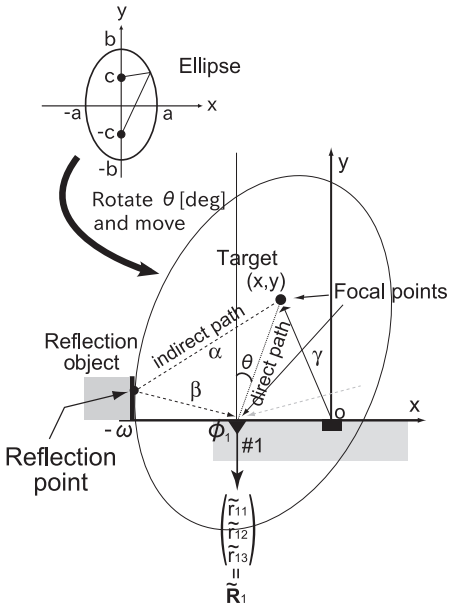


Figure 4. Ellipse image (focuses, target, receiver)

is expanded from the above mentioned EPEM. From now, we will introduce EPEMR briefly. The detailed procedure is introduced in our past work [7].

We estimate according to the following procedure.

Step(a) Estimating the position of the target by EPEM

EPEMR is based on EPEM. First, we estimate the positions of the target by using EPEM as described in Sec. II-B. EPEM can estimate the rough positions under multipath environment.

Step(b) Estimating the reflected points

Next, we estimate reflecting points on other objects except the target. In order to estimate reflecting points, we focus on the properties of a ellipse.

The distance of direct path is presented as the distance from the target to each receiver (Fig.4). The indirect path means the path which is reached by reflecting at a kind of objects, such as walls. We can derive the distance (length) of the direct path from the estimated position of the target at Step(a). By comparing the distance value of the direct path, we eliminate the close value in the measured range lists. After eliminating, only the measured distances of indirect path remain in the lists.

By using the derived distances of the indirect paths, we construct a virtual ellipse. From now, we prepare the ellipse which is illustrated in Fig. 4. The ellipse, which we will prepare, has two focuses at the positions of both the target and each receiver.

Generally, an elliptical equation is expressed as follows.

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1 \quad (a < b) \tag{5}$$

The sum of the distance from a arbitrary point to 2 focus points, that is $\alpha+\beta$ in Fig. 4, is constant. This distance is equal to the measured value $\tilde{r}_{kn}$ of the indirect path. As mentioned before, the distance of the indirect path can be gotten from the measured range lists after removing the direct distance. Then, the relation among the ellipse parameters and the distance $\tilde{r}_{kn}$ is:

$$2b = \tilde{r}_{kn} \tag{6}$$

$$a^2 = b^2 - c^2 \tag{7}$$

where the parameters a, b are coordinates of the intercept of long/short axis on the ellipse. The focal distance of the ellipse is denoted as c. These parameters are also illustrated in Fig. 4.

To fit the ellipse in the geometric relation between the receiver and the target, we rotate and move the above ellipse. The rotation angle θ and the amount of the movement are decided as the two focal points of the ellipse are placed at the receiver and the estimated target respectively. The detailed procedure is explained in our past work [7].

The wall's positions are known. Then, by using both the derived ellipse equation and the wall's position, we can compute the candidate positions of the reflections on the wall.

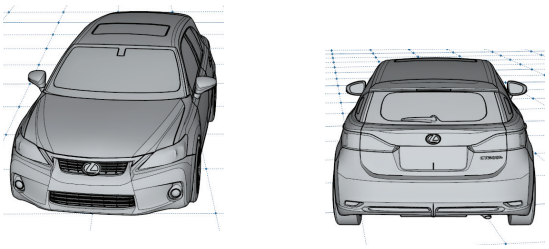
Step(c) Set up virtual receivers

We set the virtual receivers at the reflection points. We also calculate the distance α, β from the reflection points. Then, we prepare new measured range lists R'_k. These new lists means the range lists of the virtual receivers. The distance in the lists R'_k is α. We have prepared the virtual receivers which has own virtual measured range list. We estimate the target position by EPEM with the all measured range list, that is both the virtual receivers and the real receivers again.

III. Estimation performance by simulations with 3D vehicle model

A. 3D vehicle model

We introduce the characteristic of our estimation algorithm by computation simulations. In our past works, we



(a) Front view (b) Rear view

Figure 5. Target model

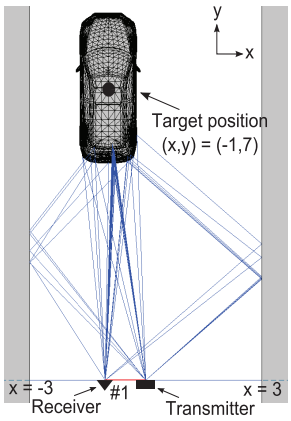


Figure 6. Example of the propagation path

considered and evaluated the position estimation algorithm using the target which is modeled as a single point. The single point model is important to evaluate a performance, in term of comparing the estimated position to the true one. However, it is also important to evaluate the estimation performance of using the 3D modeled object. So, we consider a 3D simulation model and re-construct data processing. From the new evaluations, we can grasp the performance in case of the surface which the real target has, not a single point.

We prepare the realistic 3D vehicle model as the target. We download the 3D vehicle model via Trimble 3D gallery (Fig. 5). In Trimble 3D gallery, we can download modeled files which can be imported to our simulation software, the file type is .skp. Size of the vehicle model is 3.6 meters long, 1.4 meters wide and 1.2 meters height. For calculating the measured ranges, we use software “Raplab” which is an analysis tool of radio propagation by 3D ray tracing. This tool can simulate the propagation path like Fig. 6. In Fig. 6, the receiver is denoted as a triangle mark and the transmitter as a rectangular mark. We set the target position (x, y) as the center of the vehicle like Fig. 6. By using this 3D model, the measured ranges at each receiver become more reality than before.

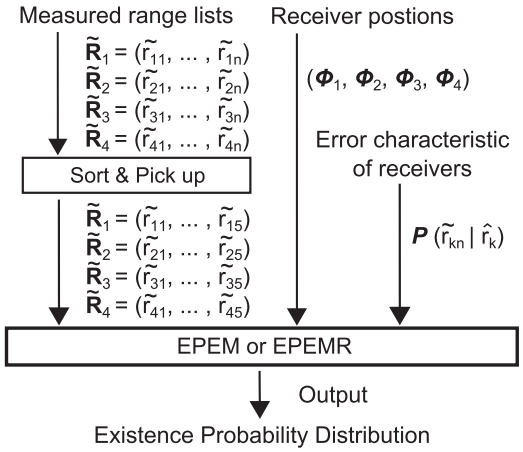


Figure 7. Data flow

B. Data processing

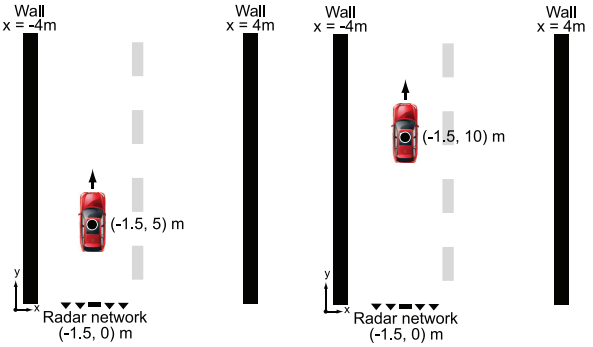
Fig. 7 shows data flow. In our algorithm, we use the propagation distance calculated by the above-mentioned tool as the true value of the measured range. The measured range lists of each receiver have many measured ranges because there are about 100 propagation paths. Some of them are unnecessary information such as multipath. Moreover, some of the reflection points of these paths are not on the same plane which expends at the same ground level to the radar network. In the position estimation, we should get the direct path which propagates via the way of the transmitter - the target - the each receiver. That is the shortest path. For applying to EPEM, we sort the measured ranges of the lists $\tilde{R}_k$ in ascending order, and pick up s ranges from the smallest. The variable s means the number of the selected ranges. In this paper, we set the variable $s = 5$, this is not optimal but experimental. The parameter s affects the calculation time and the detection performance. The shorter time is desired. Moreover, in case of selecting larger s, we sense unnecessary part of the body such as the side of the target. This results in confusing detection. The most important part is the nearest part of the target. So, we pick up the shorter $s = 5$ ranges. Although the situation of multiple targets is not scope of this paper, the parameter s may be set larger value if there are multiple targets which we want to detect.

We derive the existence probability distribution of the target by EPEM with the measured range lists, the receiver position and the error characteristic of each of receivers. Finally, we get the result of the position estimation by choosing the high probability in the existence probability distribution.

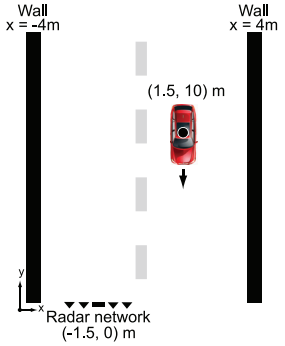
C. Simulated cases and results

We simulate the following cases. In these cases, there is one target of the vehicle. We note that the coordinate (x, y) of the target position is denoted as the center of the vehicle; see Fig. 6. The important detected part is the rear of the

 ISBN: 978-1-61208-366-7



(a) Case I-(i), Same lane, Near (b) Case I-(ii), Same lane, Far



(c) Case II, Opposite lane, Far

Figure 8. Simulated Cases (Target Layout)

vehicle. In the simulation, we assume two cases in different layout of the vehicle; see Fig. 8.

1) Case I: the target is arranged at the same lane in the front of own car.
   (i) Set on the target at near area $(x, y) = (-1.5, 5)m$.
   (ii) Set on the target at far area $(x, y) = (-1.5, 10)m$.
2) Case II: the target is arranged at opposite lane and far area $(x, y) = (1.5, 10)m$.

For comparison, we also simulate the conventional EPEM algorithm.

In 3D ray tracing, we set the maximum times of reflection and diffraction as 1, respectively. Simulation parameters are summarized in Table I. The measured ranges are modeled as (1). The distribution of the error ϵ_k in (1) is assumed as Gaussian distribution with standard deviation σ. The standard deviation is decided by the error range of the measured range at each receiver. The error range of the direct path is set as 0.3m. The amount 4σ means including more than 90 % in Gaussian distribution. So, we set $4\sigma = 0.3$ [m].

As a result, we summarize the existence probability distribution in Figs. 9, 10, and 11. The presented figures indicate the existence probability at $x - y$ plane. Each figure has the color bar which distributes from red color to white color.

TABLE I
SIMULATION PARAMETERS

Number of radars: k	4
Position of receivers: ϕ_k	-2.4,-1.8,-1.2,-0.6 [m]
x-coordinate of wall: ω	-4,4 [m]
Distribution of error ϵ_k [m]:	Gaussian
Measurement error of the radar [m]	0.3 ($4\sigma = 0.3$)
Resolution of x-y plane	$x = 0.05$, $y = 0.05$ [m]
Height of receivers and transmitter	0.3 [m]

The red color means high probability. So, the place of the red color has possibility of the target existence. Figures 9 and 10 show the results of Case I-(i) and -(ii), respectively. Figure 11 shows the results of Case II. We note that the coordinate of the each target is the center of the car. So, the detected areas of Figs. 9, 10 and 11 are the nearest part of the car, that is outside of the body. From Figs. 9, 10 and 11, it results that the proposed EPEMR can reduce the error compared to EPEM. Especially, in EPEM, the farther the distance between the target and the radar is, the larger the error in the x-direction is. This is the typical problem of multiple sensing system such as radar network. The multiple sensing from the same side generates large error in the same direction of the sensor arrangement. For example, from Figs. 10 and 11, the high probability in EPEM can be found along about 2.31m in x-direction. It results that the target exists over the width of the car. On the other hand, EPEMR can suppress the error and the detected area becomes within the car width. This improvement can be confirmed in all cases.

IV. CONCLUSION

In the realization of ITS world, we research the forward-looking radar network. Especially, we focused on the combination between the imaging detection and the radar network detection. In order to be more accurate radar sensing, we regarded the measured ranges to targets as the random variables. We have proposed and evaluated some position estimation algorithms. In this paper, we introduced our proposals EPEMR, the data processing and the estimation performance with new 3D target model. EPEMR estimates the reflection points on the surrounding structures with the results by the imaging devices. EPEMR sets the virtual receivers on the estimated reflection points. By adding the virtual receivers, the target can be observed from various directions. From the computer simulations with 3D target model, we confirmed that the EPEMR can reduce the positioning error. We also confirmed the advantage and robustness of the proposal by different situations.

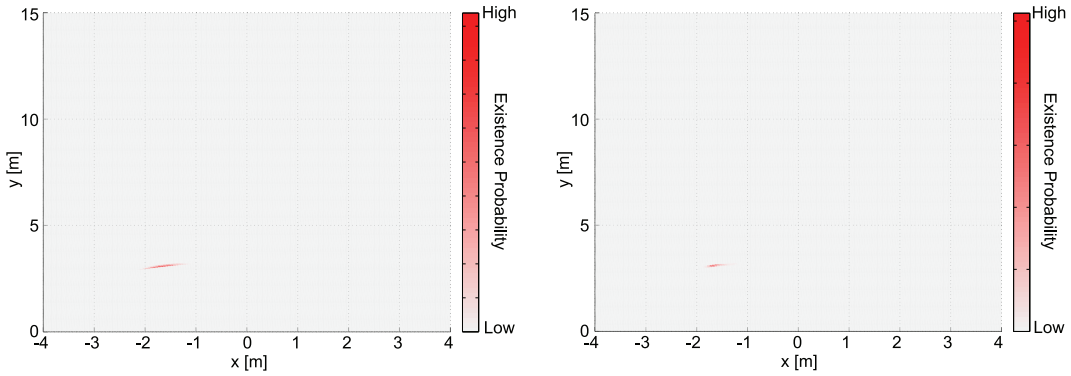
ACKNOWLEDGEMENT

The part of this work is supported by Grant-in-Aid for Young Scientists (B), Japan Society for the Promotion of Science (JSPS). We are also particularly grateful for the 3D model made by username:brecht on Trimble 3D gallery.

REFERENCES

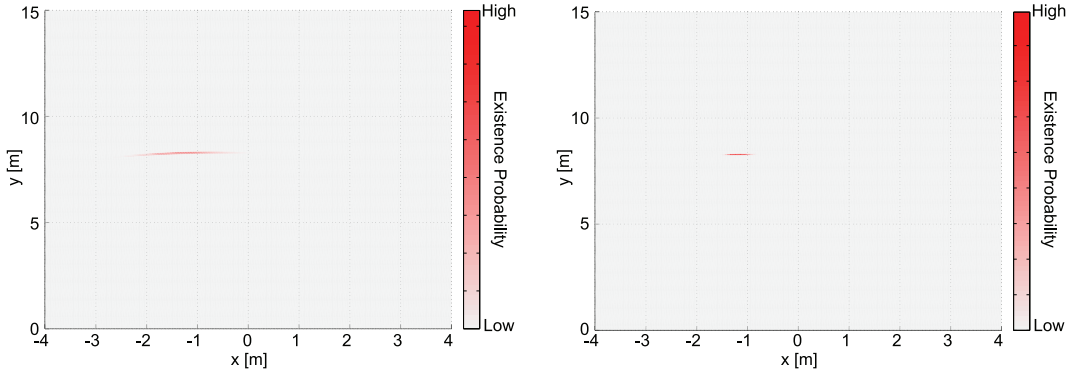
[1] R. H. Rasshofer and K. Gresser, “Automotive radar and lidar systems for next generation driver assistance functions,” Advances in Radio Science, vol.3, May 2005, pp. 205–209.

 ISBN: 978-1-61208-366-7



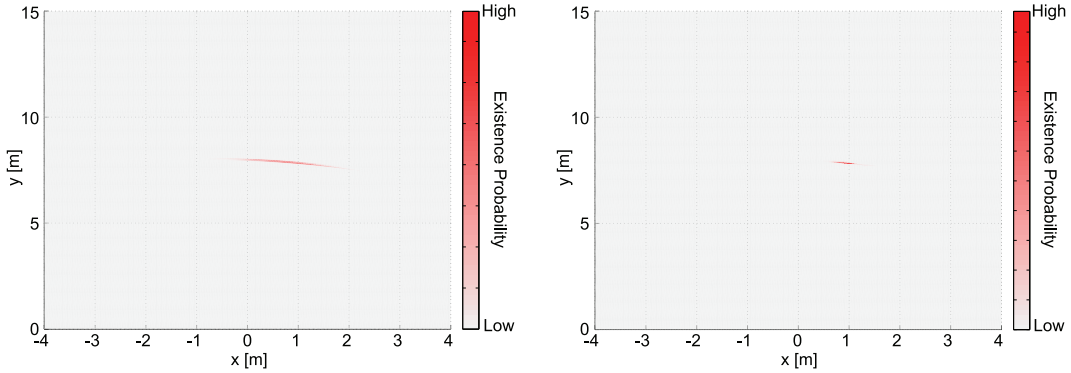
(a) EPEM (b) EPEMR

Figure 9. Target existence probability(Case I-(i), 5m, EPEM vs EPEMR)



(a) EPEM (b) EPEMR

Figure 10. Target existence probability(Case I-(ii), 10m, EPEM vs EPEMR)



(a) EPEM (b) EPEMR

Figure 11. Target existence probability(Case II, Oncoming, 10m, EPEM vs EPEMR)

[2] M. M. Meinecke et al., “Approach for protection of vulnerable road users using sensor fusion techniques,” International Radar Symposium, Sept. 2003, pp. 125-130.

[3] T. Sakamoto, “A fast algorithm for 3-dimensional imaging with UWB pulse radar systems,” IEICE Trans. on Commun., vol.E90-B, no.3, Mar. 2007, pp. 636–644.

[4] M. Klotz and H. Rohling, “24 GHz radar sensors for automotive applications,” International Conference on Microwaves, Radar and Wireless Communications, May 2000. pp. 359–362.

[5] F. Folster, H. Rohling, and U. Lubbert, “An automotive radar network based on 77 GHz FMCW sensors,” IEEE International Radar Conference, May 2005, pp. 871–876.

[6] H. Hatano, T. Mizutani, K. Sugiyama, and Y. Kuwahara, “Target position estimation algorithm under corrupted measurement data for radar network systems,” IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences, vol.E95-A, no.1, Jan. 2012, pp. 317–323.

[7] Y. Yoshida, H. Hatano, and K.Sugiyama, “An evaluation of error reduction by indirect path in forward-looking radar network systems,” 12th International Conference on ITS Telecommunications, Nov. 2012, pp. 269–274.

 ISBN: 978-1-61208-366-7

TV Content Delivery to PC, Tablet, Smartphone
From the Accessibility Vision into Market Reality

Hadmut Holken
Market Insights and Take-Up
Holken Consultants & Partners
Paris, France
holken@holkenconsultants.com

Pilar Orero
Catalan Centre for Research in Ambient Intelligence and Accessibility, Universitat Autònoma de Barcelona, UAB
Barcelona, Spain
pilar.orero@uab.cat

***Abstract*— The article addresses media accessibility for all citizens in the connected TV environment. HbbTV (Hybrid Broadcast Broadband TV) is an European standard increasingly adopted by European broadcasters. One of the challenges in the coming years will be the delivery of multi-platform audiovisual content (anytime, anywhere, any device) and making this content accessible for all. The elderly and people with various disabilities rely on subtitles, audio description or sign language. Customizing accessibility services through options for personal preferences is only one example of future possibilities. This essay highlights accessibility issues in the connected TV and smart phone environment. It gives insights on the genesis and future modules of the European project Hbb4All (Hybrid Broadband Broadcasting for All), which started in December 2013 for a 3-years-period. It addresses a wide range of interactivity, interoperability and personalized accessibility features based on the HbbTV(hybrid broadcast-broadband) concept and will be user trial oriented. Given the recent start of the project, first results are expected after a one-year-running-time up from 2015.**

Keywords-interoperability; accessibility; connectivity; multi-platform devices; market take-up.

I. INTRODUCTION

Is Michel Hazanavicius' silent black and white film *The Artist*, against the current, at the time of stereoscopic 3D, UHD 4K and interactivity? This French film is brimming with inventiveness and finds a wonderful support in black and white.

Blancanieves is a 2012 Spanish-French black and white silent fantasy drama master film written and directed by Pablo Berger [1]. Based on the fairy tale "Snow White" by the Brothers Grimm, the story sets a romantic vision of 1920s Andalusia and is a love letter to European silent cinema.

Neither *The Artist* nor *Blancanieves* was specially invented for the deaf. And yet, those concerned by some form of deafness are very sensitive and may consider the film as an advantage. Thus, there have been comments about the film as "*being deaf, I finally saw my first French film in the cinema and I laughed to read on the lips of the actors. I can already tell you that they do not jabber anything and respect the dialogues. For once I understand better than a hearing person ^ ^*" [2]. *Blancacieves* has been audio-described in the French language by author Paul Memmi for blind people. His creation is also a standalone oeuvre, to which one can listen to without watching the film.

Inventiveness and creativity, when associated with intelligence and talent, even against the current, can add value to the greatest number. The same applies for accessibility: first addressing people with disabilities, it is a process facilitator that brings ease of use for all, thus significant (mass) market developments.

Connected TV and the second screen allow new user experiences with personalized user centric content delivery. In terms of accessibility, user studies reported in the EU project DTV4ALL [3] that a solution with the interpreter in a small window is not optimal, as the picture-in-picture does not contain enough detail. Other solutions were preferred, and today the average TV screen is becoming bigger [4] allowing new user experiences. One of the major challenges of the coming years will be the multi-platform delivery of audio-visual content (anytime, anywhere, any device) [5], be it a broadcast or a (future) Internet IPTV service. Hybrid delivery platforms such as connected TVs and two screen solutions will be ubiquitous.

In the following, this article describes the relevance of the project with regard to mobility, gives an idea about the connected market environment, where accessibility services may be deployed, focuses in the following part on HbbTV and second screens (understood mobile devices), a section that explains the 4 projected pilots considering needs for people with disabilities. As the project started only some months ago, it concludes with expected results.

II. RELEVEANCE WITH REGARD TO MOBILITY

Since access to information was officially declared by the United Nations a Human Right in 2003, legislation, policy and regulations have been introduced and standards were drafted to assure einclusion. The legal framework of the European Commission is the "European i2010 initiative on e-Inclusion – to be part of the information society" (2007); this called on the ICT industry to work to help disabled people access digital TV and electronic communications products. It adopted the Audiovisual Media Services

 ISBN: 978-1-61208-366-7

Directive in 2010 [6]. A toolkit was set up for the Transposition of the Audiovisual Media Services Directive into National European Member States Law in 2008 [7]. Further initiatives fostered einclusion like the Web accessibility Initiative [8] and the possibility for hearing impaired and blind to access movies in theatres [9]. The roadmap to the European Disability Act [10] is a legislative initiative to improve accessibility of goods and services within the European internal market, and studies report on various projects and initiatives [11]. However, "content" processes - from conception, production, translation, exchange and archiving to distribution and use - are still complex procedures, both technologically and commercially. All access services, be they for the elderly or for people with disabilities, are language dependent. To turn the accessibility vision into reality, the active participation of multiple stakeholders is required in the value chain. This is the objective of HBB4ALL project [12], that builds on HbbTV, as the major European standard, for converged services and looks at both the production and service sides.

The project is co-funded by the European Commission under the Competitiveness and Innovation Framework (CIP) [13] and let by 12 partners from several complementary fields: universities, TV channels/broadcasters, research institutes, and SMEs (small and medium sized companies) [14].

This young major European project started in December 2013 for a 3 years period. It addresses technical, research, societal and social issues. It is not R&D (Research and Development) financed as such, but fosters market take-up of innovations. For example the partners define common technical components through different existing applications and products to create new services. Furthermore the project intends to accelerate the "go-to-market" in building on (existing) applications ready to come to the market. Therefore, technical adjustments need to be done within the connected TV and mobile devices environment, for example synchronization. The project partners intend to reach market take-off through large scaled user tests in at least 3 European countries.

This project is dealing with connected TV, which means interactivity coming on TV, mobile devices using TV services and accessibility (for people with impairments like non seeing or non/hard hearing people). The second screen (mobile and tablets) will be used for accessibility services. The project addresses all those screens within the connected environment, because the heterogeneity allows creating meaningful new applications. Technical working groups among the partners have been set up. No public project information is yet available on technical issues, and first results are expected beginning 2015.

The following highlights some market considerations, technical prerequisites, and focuses on accessibility with the HBB4ALL all project, integrating mobility use.

III. CONNECTED MARKET ENVIRONMENT

EDF (European Disability Forum) [15] counts 80 million people with disabilities in Europe, Age Platform AGE Platform Europe [16] refers to 100 million ageing people throughout Europe. They describe themselves as "European network of around 167 organisations of and for people aged 50+ which aims to voice and promote the interests of the 30 million senior citizens in the European Union and to raise awareness on the issues that concern them most." As a mix, this represents an estimated 40%+ of the European population. These given figures alone represent already mass market potentials, only within the field of concerned users.

Beyond the fact of producing "personalized services for all", originally invented for specific populations, it is all about moving from classic accessibility mechanisms to personalised media systems that allow to make life and access easier for all users. Tablets or touchscreen devices expressly show new ways for innovative interactive TV content handling, especially with input from the content industries.

Such social innovation topics are discussed for example among players from the value chain within the French public-private Media4D Think Tank initiative [17]. Media4D is a public-private brand initiative from Holken Consultants and co-funded by the French State (Direccte), the cities agglomeration Plaine Commune (North Paris Region) and private partners, among which France Televisions, group La Poste, Icade (real estate group), SMEs, user associations and creative people. Members of the Think Tank are working to set-up a very first multi-device (4 screens) and multi-accessible (audio-description, subtitling, sign language) user experience in different public places in Northern Paris territories. User tests, cross sectorial awareness creation, deployment of accessibility applications and services are major objectives of the envisaged experiences.

In the meantime, the Media4D Think Tank is the place to get access to concrete examples, discuss the state of the art in terms of accessibility in the audiovisual and digital content world: R&D, content, audiences' needs, but also financing and funding for content providers as well as potential new business models are among the topics. Creative people who include accessibility from the scratch within their content production process, story telling or scenario writing will probably create very original œuvres and therefore also interest larger audiences.

Imagine a huge film success winning an Oscar – which does not exist yet - especially conceived for attention to accessibility, that the filmmaker would have included directly into the content production process. This Oscar would go around the world, inform implicitly and explicitly all people, public and professional audiences, about accessibility and e-inclusion. Once "evangelized", coincidentally market shares for connected TV sets and applications for mobile platforms would increase immediately, and time to market for innovative research projects on accessibility would be consequently reduced.

With this thought, Media4D put together researchers and content providers, TV channels and digital equipment providers, regulators and legislators … to start discussions and exchange between stakeholders in the value chain with the goal to find ways to boost or create corresponding meaningful markets for universal accessibility on the different existing devices/screens (cinema, TV, PC and

 ISBN: 978-1-61208-366-7

mobile devices). With this background, the social innovation platform integrated the Hbb4All project in bringing social innovation strengths to supports dissemination activities for the HBB4ALL project, which deals with connected TV accessibility, thus interactive service access using among others mobiles devices.

This approach will complement technical skills and user trials, as outlined in the following parts.

IV. HBBTV SECOND SCREENS FOR ALL – INCLUDING PEOPLE WITH DISABILITIES

In November 2012, ETSI published version 1.2.1 of the HbbTV specification to include progressive streaming (MPEG-DASH) and some DRM (Digital Right Management) support [18].

A Connected (or Smart) TV set is not necessarily apt for truly hybrid interactive viewing experiences. Most often it is merely a multi-purpose device that *just* allows the viewing of broadcast television content *or* using separated and limited add-on functionalities through the Internet connection on the same screen.

For the truly hybrid services as offered by HbbTV, an "engine" is required that links the broadcast content (offered via satellite, terrestrial over-the-air and CATV or IPTV networks) and the Internet content (provided by any IP connection be it via DSL, CATV, or via mobile broadband networks). HbbTV provides such an engine that is activated via appropriate signaling within the broadcast transport stream. HbbTV is also being used as technology platform for portals from network operators, manufacturers and even for independent TV Apps, thus benefitting the entire eco-system. In principle, HbbTV can be used to provide any access service required: EPG, video on demand, enhanced text services. For the consumption of television, connected TV represents a prime means to help the elderly and people with disabilities (but also minorities) to improve their access to the TV content. Access services such as sign language, subtitles, audio description, clean audio, etc., can be made available via the IP link and can be displayed on either the main screen (or main loudspeakers, respectively) or on a second screen. The services can be made adjustable to the individual needs of the users. Especially, the second screen allows such tailoring as this is a personal device. The second screen application was primarily developed within the EC project FI-CONNECT [19]. The technical challenge (not yet standardised) is to time-synchronize the displaying of the broadcast and the IP delivered content. The EC project HBB-NEXT [20] is working on a solution that could be part of the HbbTV 2.0 specifications.

These specifications are, are among others, the basis to develop user applications and large scale tests within the HBB4ALL project prefiguring future deployments.

V. ACCESSIBILITY TO DIGITAL SOCIETY WITH HBB4ALL

The European HBB4ALL (Hybrid Broadcast Broadband for All) project addresses media accessibility in the connected TV, namely the new HbbTV (Hybrid Broadcast Broadband) environment. One of the challenges for the coming years will be the delivery of multi-platform audiovisual content (anytime, anywhere, any device), it will be a program or Internet service. Platforms' hybrid delivery as connected TVs and solutions to two screens allows a cost effective and convenient delivery access for those who need those services. The elderly and people with various disabilities rely on subtitles, audio description, improving dialogue or sign interpretation. Customizing personal preferences will be possible within predetermined limits.

The Hbb4All project aims at:

- Advancing solutions to future accessibility problems, when HbbTV becomes widespread in Europe;
- Understanding interoperability in a multiplatform and multi-language communication to test easy solutions for media accessibility;
- Benchmarking quality of access services from a user-centric approach, and promoting accessibility as an added value for education and social inclusion;
- Becoming a major platform/player in the e-Inclusion economy currently taking place, fostering the future market take-up of exiting innovations in conceiving universal accessibility tools and concepts to satisfy the diverse interests of all societal groups.

How to watch TV content in PC, tablets, smart phones and TVs with an array of communication solutions, such as subtitling, audio description, clean audio, and many customizable features? Multiple European languages, large and small, sign language, and language situation – monolingual, bilingual - will be taken into consideration and also the three translation modes: dubbing, subtitling and voice-over.

For this purpose, the project will test access services in various pilot implementations (from the definition to the operational phase) and gather implicit and explicit user feedback to assess the acceptance and the achievable quality of service in the various delivery scenarios (broadcasting, hybrid, full IP). Four interlinked sub-pilots will be implemented in the HBB4ALL project:

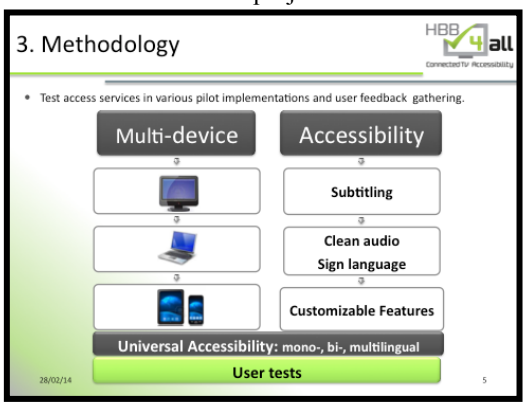


Figure 1. Hbb4All Methodology

A. Pilot-A: Multi-platform subtitle workflow chain

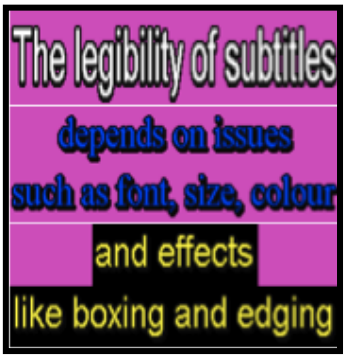


Figure 2. Subtitling effects

Pilot A deals with Multi-Platform Subtitle Services. Across Europe, broadcasters are working to provide subtitles on multiple platforms for individuals who are deaf and hard-of-hearing, or do not have sufficient language skills to understand the content without textual support either in the original or foreign languages. The main challenge is to provide subtitles tailored to the specific needs of the end-users in terms of channels, platforms and consumption requirements. This requires a well-conceived production and distribution strategy that allows for the exchange of subtitles and their automatic re-purposing producing quality and impact-driven access services for multiple platforms.

B. Pilot-B: Alternative audio production and distribution

Pilot B deals with alternative audio production and distribution. Given EU citizen mobility, TV content is not only seen by nationals, but also by large communities living away from home. There is also a need to broadcast same content in different languages synchronically (e.g., Swiss TV or Brussels TV), but the content is not the same across languages.

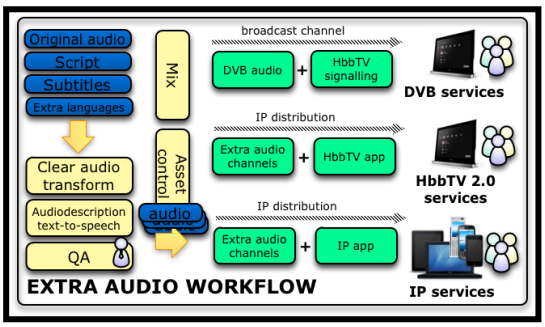


Figure 3. Extra Audio Workflow Scheme

C. Pilot-C: Automatic UI (User Interface) adaptation – accessible Smart TV applications

Pilot C looks at automatic User Interaction (UI) adaptation, and smart TV applications. During the last years digital TV as a media platform has increasingly turned from a simple receiver and presenter of broadcast signals to an interactive and personalised media terminal with access to traditional broadcast as well as web-based services.

The accessibility features of such a service will make use of the UI adaptation framework that was developed within the European project GUIDE (Gentle user interfaces for elderly people) [21].

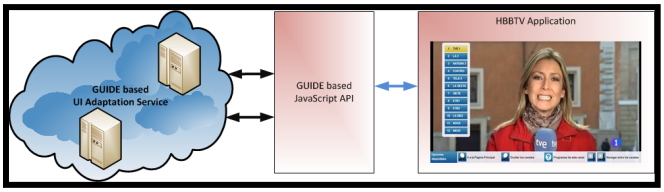


Figure 4. Example UI adaptation framework

D. Pilot-D: Sign-language translation service

The last pilot is related to sign language translation. Visual signing for audiovisual media such as film and television was shown for the first time in1929 as a means to make such content accessible to individuals whose mother tongue is a sign language and not an oral language. Users of sign language are often born deaf. In many European countries, there are constitutional and legal provisions to assure the provision of sign language for such citizens who, in numerical terms, account for less than 1% of the population.

Figure 5. Signing in Belgium – RTBF

Figure 6. Signing in Portugal - RTP1

Broadcasters dependent on advertising express concerns that an obligation to offer signing would lead to a noticeable reduction in advertising revenue, since audiences dislike "screen contamination" with the interpreter. Offering closed signing (where the viewer can choose to see or not to see the interpreter) requires much more bandwidth than closed

 ISBN: 978-1-61208-366-7

subtitles or audio description. Signing is important not only for mainstream programming and TV programming specifically for the signing communities in Europe and elsewhere but also emergency alerts on TV.

On this basis, HBB4ALL is elaborating pertinent guidelines, guides of good practice, metrics, and recommendations. It will initiate campaigns to promote the project results and thus raise awareness not only on the necessity of access services but also on the technical solutions available. For that purpose, and to transform accessibility vision into reality, Hbb4All targets all relevant stakeholders of the value chain.

VI. EXPECTED RESULTS

Being an ETSI standard, HbbTV is currently linked with the DVB TV system family but can, in principle, be used in conjunction with any digital TV service in the world. DVB is widely used throughout all continents. Sooner or later, all countries in the world will have completed their analogue-to-digital switch-over. As a consequence, the results of HBB4ALL will be of worldwide relevance and will, through standardisation bodies such as the ITU and ISO, also be publicised on a world-wide level. Given the impact in close fields such as eHealth and eEducation for example, the results from this project will have important results and direct impact. On its basis, HBB4ALL is elaborating pertinent guidelines, guides of good practice, metrics, and recommendations and will initiate campaigns to promote the project results, and thus raise awareness not only on the necessity of access and interaction services but also on the technical solutions available with interoperability. For that purpose, all relevant stakeholders, from content providers to user associations, will be addressed. The overall objective of HBB4ALL is to become a major platform/player in the e-Inclusion economy currently taking place, fostering the future market take-up of exiting innovations in conceiving universal accessibility tools and concepts to satisfy the diverse interests of all societal groups.

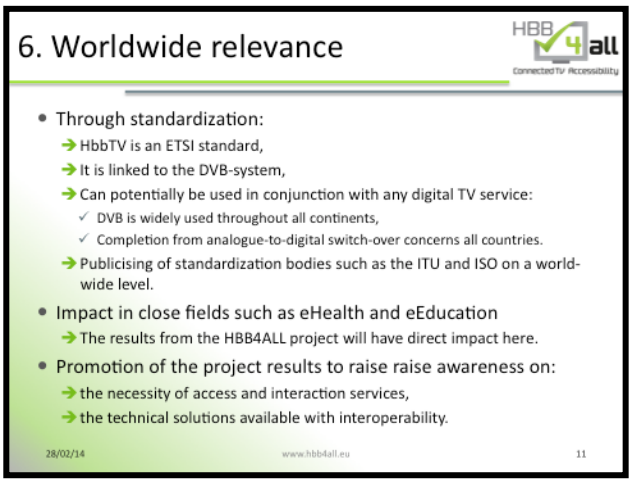


Figure 7. Worldwide relevance of the European Hbb4All project

References

[1] Biography: http://en.wikipedia.org/wiki/Pablo_Berger

[2] Translated in English from the following original French text http://marvelll.fr/critique-the-artist/ [retrieved February 2014]

[3] http://www.psp-dtv4all.org/

[4] CES – Consumer Electronic Show, January 2014, Las Vegas, www.cesweb.org/

[5] NEM Position Paper on Connected TV, December 2012 http://nem-initiative.org/wp-content/uploads/2013/12/NEM-PP-015.pdf [retrieved July 2014]

[6] http://ec.europa.eu/smart-regulation/impact/ia_carried_out/docs/ia_2007/sec_2007_1469_en.pdf [retrieved July 2014]

[7] EDF/European Disability Forum, Toolkit for the Transposition of the Audiovisual Media Services Directive into National EU Member States Law, September 2008: http://cms.horus.be/files/99909/MediaArchive/EDF_Toolkit_for_the_Transposition_of_AVMS_Directive.pdf [retrieved November 2013]. See also: European Disability Forum, EDF Answer to the European Commission Consultation on the Green Paper on the Online Distribution of Audiovisual Works in the European Union: Opportunities and Challenges towards a Digital Single Market, November 2011 http://ec.europa.eu/internal_market/consultations/2011/audiovisual/non-registered-organisations/european-disability-forum_en.pdf [retrieved June 2014]

[8] W3C (2008) Web Accessibility Initiative http://www.w3.org/WAI/ [retrieved March 2014]

[9] Olivier Hillaire, Solutions for Visually and Hearing Impaired people to access Cinema Theatres (Les solutions permettant aux handicapés visuels et auditifs d'accéder aux salles de cinéma), Study of FNCF – Fédération Nationale du Cinéma Français/National Federation of French Cinemas, April 2012, in: http://www.fncf.org/updir/3/etude_olivier_hillaire_accessibilite_def.pdf [retrieved February 2014], www.fncf.org/ [retrieved February 2014]

[10] Roadmap to the European Accessibility Act: legislative initiative to improve accessibility of goods and services in the Internal Market (2012), http://ec.europa.eu/smart-regulation/impact/planned_ia/docs/2012_just_025_european_accessibiliy_act_en.pdf [retrieved February 2014], available also at http://www.edf-feph.org/Page.asp?docid=29753&langue=EN [retrieved June 2014]

[11] eAccessibility-related EC studies and similar projects: http://www.e-accessibility2020.eu/portal/index.php?option=com_content&view=article&id=1424:eaccessibility-related-ec-

 ISBN: 978-1-61208-366-7

studies-a-similar-projects&catid=109 [retrieved January 2014]

[12] http://www.hbb4all.eu [retrieved April 2014]

[13] http://ec.europa.eu/cip/ [retrieved April 2014]

[14] Consortium: http://www.hbb4all.eu/partners/consortium [retrieved April 2014]

[15] http://www.edf-feph.org/ [retrieved April 2014]

[16] http://www.age-platform.eu/[retrieved April 2014]

[17] www.socialmedia4D.com

[18] http://www.etsi.org/deliver/etsi_ts/102700_102799/102796/01.02.01_60/ts_102796v010201p.pdf [retrieved June 2014]

[19] http://www.fi-content.eu/ [retrieved June 2014]

[20] http://www.hbb-next.eu [retrieved June 2014]

[21] http://www.sumat-project.eu/ [retrieved 16.11.2013] – see: http://www.guide-project.eu/[retrieved November 2013]

 ISBN: 978-1-61208-366-7