JYVÄSKYLÄ STUDIES IN COMPUTING

271

Riaz Uddin Mondal

Radio Frequency Fingerprinting for Outdoor User Equipment Localization

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 271

Riaz Uddin Mondal

Radio Frequency Fingerprinting for Outdoor User Equipment Localization

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen Gamma-salissa
joulukuun 13. päivänä 2017 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, Gamma hall, on December 13, 2017 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Radio Frequency Fingerprinting for Outdoor User Equipment Localization

JYVÄSKYLÄ STUDIES IN COMPUTING 271

Riaz Uddin Mondal

Radio Frequency Fingerprinting for Outdoor User Equipment Localization

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2017

Editors
Timo Männikkö
Faculty of Information Technology, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-7285-1

URN:ISBN:978-951-39-7285-1
ISBN 978-951-39-7285-1 (PDF)

ISBN 978-951-39-7284-4 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä 2017

ABSTRACT

Mondal, Riaz Uddin
Radio Frequency Fingerprinting for Outdoor User Equipment Localization
Jyväskylä: University of Jyväskylä, 2017, 68 p. (+included articles)
(Jyväskylä Studies in Computing
ISSN 1456-5390; 271)
ISBN 978-951-39-7284-4 (nid.)
ISBN 978-951-39-7285-1 (PDF)
Finnish summary
Diss.

The recent advancements in cellular mobile technology and smart phone usage have opened opportunities for researchers and commercial companies to develop ubiquitous low cost localization systems. Radio frequency (RF) fingerprinting is a popular positioning technique which uses radio signal strength (RSS) values from already existing infrastructures to provide satisfactory user positioning accuracy in indoor and densely built outdoor urban areas where Global Navigation Satellite System (GNSS) signal is poor and hard to reach. However a major requirement for the RF fingerprinting to maintain good localization accuracy is the collection and updating of large training database. The Minimization of Drive Tests (MDT) functionality proposed by 3GPP LTE Release 10 & 11 has enabled cellular operators to autonomously gather and update necessary amount of RF fingerprint samples by utilizing their subscriber user equipments (UEs). The main objective of this thesis is to propose a framework for RF fingerprint positioning (RFFP) of outdoor UEs using MDT data and to further improve its performance capability to provide better localization. In the first part only LTE base-station (BS) RSS values were used to improve grid-based RF fingerprint positioning (G-RFFP) by using novel approaches: using overlapped grid-cell layouts (GCL), weighting based grid-cell unit selection and Artificial Intelligence based G-RFFP method. In the second part real measurement RSS values from LTE BS and WLAN access points (APs) were utilized and a generic measurement method referred to as GMDT was proposed to correlate WLAN RSS to LTE RSS measurements and its significance to RFFP was studied using a partial fingerprint matching technique. To remove the computational cost associated with training data preprocessing a new cluster-based RF fingerprint positioning (C-RFFP) method was proposed. This thesis provides a good source of information and novel techniques for cellular operators to build a low cost RF fingerprint positioning system which can deliver acceptable results in emergency user localization.

Keywords: RF fingerprinting, LTE, WLAN, Mahalanobis distance, Kullback-Leibler divergence, K-nearest Neighbor, K-means clustering, Hierarchical clustering, Fuzzy C-means clustering

Author Riaz Uddin Mondal
Faculty of Information Technology
University of Jyväskylä
Finland

Supervisor Professor Dr. Tapani Ristaniemi
Faculty of Information Technology
University of Jyväskylä
Finland

Reviewers Professor Peter Chong
Auckland University of Technology
New Zealand

Professor Yevgeni Kucheryavy
Tampere University of Technology
Finland

Opponent Professor Mohammed Elmusrati
University of Vaasa
Finland

ACKNOWLEDGEMENTS

I am forever grateful to ALLAH for without His permission this research work would not have been possible.
I would like to express my deepest gratitude to my supervisor Professor Tapani Ristaniemi for all his guidance, support, encouragement and valuable suggestions which has been absolute necessary for my progress throughout this doctoral study. I wish to express my sincere appreciation to the Faculty of Information Technology, University of Jyväskylä for the financial support during the past years. I humbly thank to all the teachers, colleagues and friends of the University of Jyväskylä how has been beside me and helped me to increase my knowledge and experience.
I would like to give special thanks to Dr. Jussi Taneli Turkka, his presence and cooperative support helped me a lot in learning and developing new ideas. I would also like to thank Tuomas Hiltunen for a joint research work.

I am grateful to the reviewers of my dissertation, Professor Peter Chong and Professor Yevgeni Kucheryavy for their constructive comments and to Professor Mohammed Elmusrati for being my opponent.

I am indebted to all my Bangladeshi and other Muslim brothers who have given me hope and accompanied me in the time of need. They have eased this long journey into a memorable and enjoyable one.

Finally my warmest gratefulness goes to my mother Mrs. Nurun Nahar Mondal for all her prayers and constant encouragement for her elder son, for her patience and love. I am grateful to my late father Dr. Abdus Salam Mondal for giving me love and affection when he was beside me. I am glad that my wife Munzura Raish Ud Doula and my three daughters Ashfia Afiath, Saara Afiath and Khadija Afiath are beside me and constantly giving me all the love and peace I need to be happy in life.

Jyväskylä, Finland
26.11.2017

Riaz Uddin Mondal

LIST OF ACRONYMS

3GPP	3rd Generation Partnership Project
AECID	Adaptive Enhanced cell-ID
AGNSS	Assisted-Global Navigation Satellite System
AHC	Agglomerative Hierarchical Clustering
AOA	Angle of Arrival
AP	Access Points
BS	Base-Station
C-RFFP	Cluster-based RF Fingerprint Positioning
CDB	Correlation Data-Base
CID	Cell-ID
CIR	Channel Impulse Response
CRS	Cell-specific Reference Signal
DCM	Database Correlation Methods
E-OTD	Enhanced Observed Time Difference
ECID	Enhanced Cell-ID
eNBs	E-UTRAN Node B
FCM	Fuzzy C-Means
FCF	Frequency Coherence Function
G-RFFP	Grid-based RF Fingerprint Positioning
GA	Genetic Algorithm
GAFP	MOGA optimized RF Fingerprint Positioning
GCL	Grid-Cell Layouts
GCU	Grid-Cell Unit
GMDT	Generalized MDT
GNSS	Global Navigation Satellite System
HSC	Heterogeneous Small-Cell
KLD	Kullback-Leibler Divergence
KNN	K-Nearest Neighbours
LAI	Location Area Identifier
LEMT	Location Estimation using Model Trees
LBSs	Location-based services

LTE	Long Term Evolution
MD	Mahalanobis Distance
MDT	Minimization of Drive Tests
MOGA	Multi-Objective Genetic Algorithms
MT	Matching Threshold
NLOS	Non-Line-of-Sight
NoGCL	Non-overlapping GCL
NSGA	Non-dominated Sorting Genetic Algorithm
OFDM	Orthogonal Frequency-Division Multiplexing
OGCL	Overlapping GCL
OGL	Overlapping Grid Layout
OPEX	Operational Expenditure
OTDOA	Observed Time Difference Of Arrival
PCA	Principal Component Analysis
PCM	Pilot Correlation Method
PE	Positioning Error
PRBs	Physical Resource Blocks
QoS	Quality of Service
R_fings	Reference RF fingerprints
RAN	Radio Access Networks
RF	Radio Frequency
RFID	Radio Frequency Identification
RFFP	RF Fingerprint Positioning
RM	Regular Macro
RRM	Radio Resource Management
RSCP	Received Signal Code Power
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
RSS	Radio Signal Strength
RSSI	Received Signal Strength Indicator
RTT	Round Trip Time
SGL	Single Grid Layout
SINR	Signal-to-Interference-plus-Noise Ratio
SOM	Self-Organizing Map
SRNC	Serving Radio Network Controller

TCE	Trace Collection Entity
T_fing	Target RF fingerprint
TDOA	Time Difference of Arrival
TOA	Time of Arrival
UE	User Equipment
UTDOA	Uplink Time Difference Of Arrival
WLAN	Wireless Local Area Network
WPS	Wi-Fi Positioning System

LIST OF FIGURES

LIST OF TABLES

CONTENTS

LIST OF INCLUDED ARTICLES

PI Riaz Uddin Mondal, Jussi Turkka, Tapani Ristaniemi & Tuomas Henttonen. Positioning in Heterogeneous Small Cell Networks using MDT RF Fingerprints. *1st IEEE International Black Sea Conference on Communications and Networking*, *Batumi, Georgia,* 2013.

PII Riaz Uddin Mondal, Jussi Turkka, Tapani Ristaniemi & Tuomas Henttonen. Performance Evaluation of MDT Assisted LTE RF Fingerprinting Framework. *7th International Conference on Mobile Computing and Ubiquitous Networking (ICMU2014), Singapore,* 2014.

PIII Riaz Uddin Mondal, Jussi Turkka & Tapani Ristaniemi. An Efficient Grid-based RF Fingerprint Positioning Algorithm for User Location Estimation in Heterogeneous Small Cell Networks. *International Conference on Localization and GNSS (ICL-GNSS 2014)*, *Helsinki, Finland,* 2014.

PIV Riaz Uddin Mondal, Tapani Ristaniemi & Jussi Turkka. Genetic Algorithm Optimized Grid-based RF Fingerprint Positioning in Heterogeneous Small Cell Networks. *International Conference on Localization and GNSS (ICL-GNSS 2015), Gothenburg, Sweden*, 2015.

PV Tuomas Henttonen, Riaz Uddin Mondal, Jussi Turkka & Tapani Ristaniemi. Generic Architecture for Minimizing Drive Tests in Heterogeneous Networks. *IEEE Vehicular Technology Conference (VTC Fall), Boston, USA,* 2015.

PVI Jussi Turkka, Tuomas Henttonen, Riaz Uddin Mondal & Tapani Ristaniemi. Performance Evaluation of LTE Radio Fingerprinting using Field Measurements. *International Symposium on Wireless Communication Systems (ISWCS'15), Brussels, Belgium,* 2015.

PVII Riaz Uddin Mondal, Jussi Turkka & Tapani Ristaniemi. An Efficient Cluster-based Outdoor User positioning using LTE and WLAN Signal Strengths. *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2015), Hong Kong,* 2015.

PVIII Riaz Uddin Mondal, Jussi Turkka & Tapani Ristaniemi. Cluster-Based RF Fingerprint Positioning Using LTE and WLAN Outdoor Signals. *10th International Conference on Information, Communications and Signal Processing (ICICS 2015), Singapore,* 2015.

PIX Riaz Uddin Mondal, Tapani Ristaniemi & Jussi Turkka. Cluster-based RF Fingerprint Positioning Using LTE and WLAN Signal Strengths. *International Journal of Wireless Information Networks (IJWI), Springer,* 2017.

1 INTRODUCTION

1.1 Motivation

Locating which is a fundamental need of human beings ever since they came into existence is a process used to determine the location of one position relative to other known positions. In recent technological era it is possible to localize persons and objects in real time by exploiting wireless radio transmission characteristics. The principles of all radio frequency (RF) positioning systems rely on the properties of electromagnetic waves and simple geometric principles. The first navigational aid radio direction-finding system was patented in 1902 (Stone, 1902). After that for almost 100 years until the late 1990s, RF positioning applications were mainly used as a navigational aid to ships, aircrafts and terrestrial vehicles. One of the main motivations for civilian mobile position location is safety – a mandatory requirement for emergency call originated by dialing 112 in Europe and 911 in the U.S.A (Reed et al., 1998). With the evolution of mobile cellular telephony and increasing popularity of smart phones with positioning capability a wide range of Location-based services (LBSs) starts to play a key role in Information Society and the way people live and interact (Chen, 2012). Mobile LBSs mainly include five categories of application: 1) mapping and navigation, 2) search and information, 3) tracking, 4) social networking and 5) mobile advertising. It is expected that the global revenues coming from real-time locating systems technology will amount to more than six billion Euros in 2017 (Dardari et al., 2012). However still today there is no ubiquitous positioning solution that delivers high localization accuracy under all circumstances anytime/anywhere (indoors and outdoors).

Many research and commercial organizations tried to improve accuracy and precision using different sensing technologies including ultrasonic time-of-flight, infrared proximity, radio signal strength (RSS) and time-of-flight, optical vision and so on. First, these location systems do not offer coverage in both outdoor and indoor environments; second, these technologies require expensive infrastructure, time consuming calibration, or special tags, beacons, and sensors (Campos and Lovisolo, 2015). For outdoor mobile UE localization GNSS (e.g. GPS, GLONASS) is most popular positioning system, due to its integration into smartphones and other UEs. However it performs rather poorly in GNSS signal degraded environments such as dense urban regions or inside buildings (Ben-Moshe et al., 2011). An alternative positioning approach has been suggested by 3rd Generation Partnership Project (3GPP) in Long Term Evolution (LTE) cellular networks (3GPP, 2009a), (Knutti et al., 2015). A major driving force to estimate mobile UE location is the requirement for E-911 emergency positioning in the North American market. This requirement for network-based methods should have accuracies of 100 meters for 67% and 300 meters for 95% of cellular mobile phone calls. In 2003 the European Union also adopted a regulation for the calls originating from cell phones to the pan-European emergency number 112; initially no precision requirements were specified. Four positioning techniques have been included in 3GPP Release 11: (1) Enhanced Cell ID (ECID)- positioning accuracy depends upon cellular BS coverage size, (2) Assisted-Global Navigation Satellite System (AGNSS) - suffers from Non-Line-of-Sight (NLOS) signal propagation, (3) Observed Time Difference Of Arrival (OTDOA) uses triangulation techniques which is highly affected by multipath and NLOS propagation conditions frequently occurring in indoor and urban environments and (4) Uplink Time Difference Of Arrival (UTDOA) conceptually similar to OTDOA (Campos and Lovisolo, 2015), (3GPP, 2012a), (Vaghefi and Buehrer, 2014).

Among the non-standard positioning methods included in LTE Release 9, RF fingerprint positioning (RFFP) is the most cost-efficient solution for indoor WLAN positioning (Sun et al., 2005), (Wigren, 2007), (Shi and Wigren, 2009), (Bahl and Padmanabhan, 2000), (Chan et al., 2008), (Milioris et al., 2012) as well as for outdoor mobile cellular positioning in densely built urban environments (Zhu and Durgin, 2005), (Laitinen et al., 2001), (Liu et al., 2010). Various popular positioning applications - Google Maps, Skyhook and Find my iPhone integrate multiple technologies such as Wireless Local Area Network (WLAN), GSM/UTMS and GNSS to form a hybrid positioning system. Place Lab has shown that RF fingerprint based positioning using existing radio beacon sources from GSM BSs and WLAN APs are sufficiently pervasive and can maximize coverage in most people’s daily lives with sufficient positioning accuracies in urban, suburban and residential areas (Anthony et al., 2005).

Previous RFFP studies in the literature have mainly focused on indoor localization due to the difficulty in acquiring and analyzing large amounts of training data (Li et al., 2012). Typically such databases are created by collecting field measurements or using propagation model predictions (Zekavat and Buehrer, 2011). If propagation model predictions are used rather than real measurements, then the positioning accuracy is reduced (Aarnaes and Holm, 2004). On the other hand, if operators like to create and maintain a large accurate measurement database, then extensive and expensive periodical drive test campaigns are needed. The conventional drive test method is a time consuming manual process and the obtained data only covers certain area of the network, e.g., roads and motorways. A feature named minimization of drive tests (MDT) was adopted by 3GPP in its Release 10 & 11 specifications in order to regularly assess the quality of the cellular networks. MDT enables cellular operators to collect radio measurements from subscriber UE together with location information when the measurements are taken (Johansson et al., 2012). Thus we were motivated to study RFFP localization performances in various outdoor environments using MDT data.

1.2 Research Objectives and Scope

Conventional RFFP methods comprises of two distinct phases: 1) the training phase collects RF fingerprints and store them in RF fingerprinting database; 2) the testing or positioning phase estimates an UE position by comparing an RF fingerprint collected by the test UE with the stored reference fingerprints. An ideal positioning system should be self-learning and environmentally adaptive, capable of building up information databases that store actual observations, and employ smart data analysis mechanisms (Olsson and Mulligan, 2012). The MDT feature incorporated in 3GPP has provided the automatic collection and updating of RF fingerprints of an area of interest. So far limited number of researches haves been done on RFFP in LTE cellular networks using MDT data. The main goal of this thesis is to improve UE assisted network-based RFFP accuracy in LTE cellular environments using MDT date and address the main challenges of RFFP based UE localization: generation and maintenance of radio map; pre-processing of RF fingerprints for enhancing accuracy; selection of BSs/APs for use in UE positioning; and estimation of the distance between a new RSS observation and the fingerprints (Kushki et al., 2007). In order to generate MDT data a state-of-the-art dynamic LTE system simulator was used. Recently in 3GPP Release 12 a study has been conducted to access the potential of Radio Access Networks (RAN) level solutions for enhancing the internetworking between WLAN and LTE systems (3GPP, 2013a). As a hybrid positioning method with cellular and WLAN RSS provides excellent coverage for UE localization with adequate positioning accuracy this thesis aims to improve UE RFFP accuracy using LTE and WLAN RSS values.

The main research objectives of this study are as follows:

Utilizing LTE MDT data-

1) To build a framework for GCL-based RFFP using MDT data in LTE network deployments in rural, regular urban and heterogeneous urban environments.
2) To evaluate overlapping GCL approach for enhancing G-RFFP UE localization accuracy.
3) Use of artificial intelligence in order to optimize grid layout selection procedure and thus increasing the performance of G-RFFP.

Utilizing LTE and WLAN MDT (GMDT) data-

4) To propose a generic architecture for MDT to include and correlate WLAN RSS measurements with LTE RSS measurements and evaluate the improvement of G-RFFP performance using the hybrid RSS values.
5) To employ cluster-based RFFP approach for reducing computational complexity in training phase data pre-processing.
6) To compare UE localization performances between cluster-based and traditional RFFP approaches.
7) To evaluate the effectiveness of C-RFFP approach with variations in RF fingerprint recording device and RF fingerprint collection time between training radio map and test UE.

The research objectives were achieved in two main stages. During the first stage we proposed a framework for G-RFFP using MDT radio map that consists of only LTE RSS values. To improve the UE positioning accuracy novel methods were developed and results were compared to the traditional G-RFFP method. Artificial intelligence was used to build an efficient G-RFFP system which automates GCL selection procedure as well as improves the user localization. The RFFP methods used in the first stage depends upon two phases: (1) in training phase grid-cell unit (GCU) wise manipulation of MDT data is carried out to create training signatures; and (2) in testing phase UE positions are estimated using training signatures. During the second stage of study a generic architecture for MDT functionality was proposed to attach WLAN RSS values with LTE RSS. The performance improvement of G-RFFP with GMDT data was studied and results were compared to that of using MDT data. C-RFFP method was used to reduce training phase computational complexity and to deliver fast positioning output. Performance evaluation studies between various C-RFFP methods and G-RFFP method were conducted.

1.3 Main Contributions of the Thesis

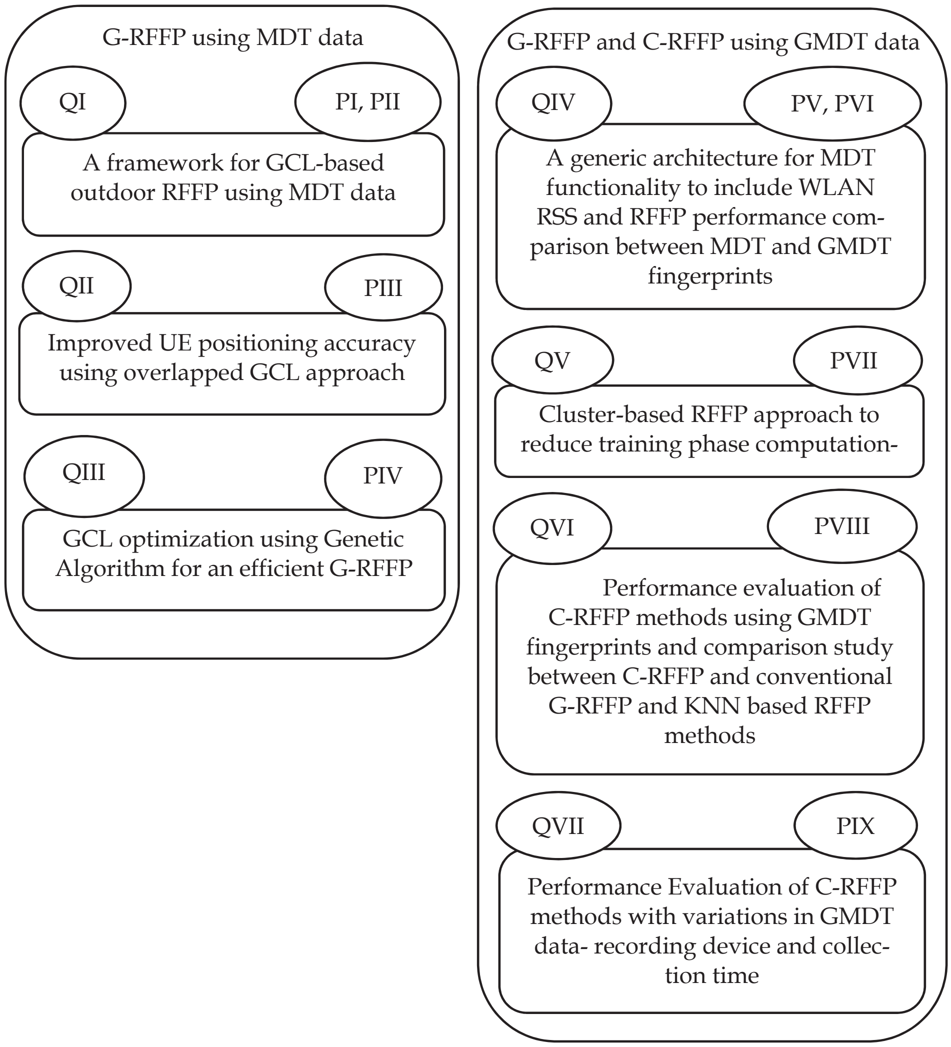


FIGURE 1 Relationship between the research questions and included articles

The relationship between the research questions and included articles are shown in Figure 1. In papers [PI] and [PII] we studied the potentials of MDT database assisted RFFP positioning in intra- and inter-frequency LTE network deployments with rural, urban and heterogeneous urban configurations. In [PI] we investigate the variations in UE positioning accuracy between Mahalanobis distance (MD) and Kullback-Leibler divergence (KLD) based RFFP methods with different GCL sizes. Here we assumed ideal cell detection criteria - every MDT sample contains equal number of neighbour BS RSS values. Then RFFP performance evaluation was carried out in interference limited LTE system taking into account the 3GPP cell detection criterion in [PII]. In order to improve positioning accuracy of conventional G-RFFP method consisting of a single grid layout (SGL) in [PIII] an overlapping grid layout (OGL) based RFFP was proposed which estimates unknown UE position through calculation of two smallest weighted KLD grids. The proposed OGL method not only improved the positioning accuracy but also analysed more test MDT samples as compared to that of traditional SGL G-RFFP method. In [PIV] a novel multi-objective Genetic Algorithm (GA) based G-RFFP method (GA-RFFP) was proposed for autonomous calibration of GCL and thereby improving test UE localization accuracy. The robustness of the proposed method was tested by employing it in two different areas of LTE heterogeneous small cell (HSC) network scenario. Simulation results showed that if sufficient amount of training data is available then GA-RFFP method can improve positioning accuracy of 56.74% over the conventional GCL-RFFP method. An enhancement to LTE MDT architecture through minor changes to the 3GPP MDT framework was proposed in [PV] which allows for the collection of location-aware radio measurements from WLAN APs. This functionality was referred to as *generalized MDT* (GMDT). To evaluate G-RFFP performance of the proposed concept real world field measurements were collected from heterogeneous LTE and WLAN networks of an urban area of Tampere, Finland. It was found that G-RFFP positioning using RSS values from 10 strongest WLAN APs along with LTE BSs could significantly reduce UE positioning error (PE) over that of using only LTE BS RSS fingerprints. To emphasize the importance of GMDT a comparison study between G-RFFP methods using LTE only and LTE-WLAN RSS data was performed in [PVI] through choosing different GCL sizes and partial fingerprint matching (PFM) criteria.

A novel cluster-based RF fingerprinting method for outdoor UE positioning was proposed in [PVII] which essentially removes the need of training data pre-processing complexity. It utilizes LTE serving BS ID to reduce computational time during UE positioning phase and delivers better results over G-RFFP approach. In [PVIII] performance comparison study between K-nearest neighbours (KNN) based RFFP, Agglomerative Hierarchical Cluster (AHC) based RFFP, Fuzzy C-Means (FCM) cluster based RFFP and G-RFFP methods were conducted using GMDT data through different PFM cases. Finally in [PIX] we have evaluated the robustness and complexity C-RFFP methods with GMDT data have time difference between training RF fingerprints and test UE measurement samples using different UE devices. Experimental results showed that even under the effect of environmental changes and device variations C-RFFP can improve positioning accuracy of outdoor UEs as compared to that of G-RFFP and KNN based positioning methods.

1.4 Author's Role in Included Articles

The author of this thesis is the main person in algorithm design, performance evaluation and writing of the articles [PIII] [PIV] [PVII] [PVIII] and [PIX]. In articles [PI] [PII] [PV] and [PVI], he contributed in data collection and analysis, system design and performance evaluation and analysis; he was responsible for writing the results and analysis parts of these articles.

1.5 Other Publications

The author has also researched other areas of nonlinear data analysis during the doctoral study and results are published in the following papers:

1) Riaz Uddin Mondal, Tapani Ristaniemi & Munzura Raish Ud Doula. Genetic Algorithm Optimized Memory Polynomial Digital Pre-distorter for RF Power Amplifiers. *International Conference on Wireless Communications & Signal Processing (WCSP), Hangzhou, China,* 2013.

2) Tuomas Henttonen, Jussi Turkka, Riaz Uddin Mondal & Tapani Ristaniemi. Performance Evaluation of LTE Radio Fingerprint Positioning with Timing Advancing. *10th International Conference on Information, Communications and Signal Processing (ICICS 2015), Singapore,* 2015.

1.6 Organization of the Thesis

This thesis is composed of following chapters:

Chapter 2: This chapter provides the relevant theoretical knowledge of conventional RF fingerprint positioning methods. The MDT functionality and the dynamic LTE system simulator are briefly explained. Then a brief description of previous RF fingerprint based methods found in literature has been presented.

Chapter 3: The research outcome of this thesis is presented in this chapter. Grid-based RFFP method and cluster-based RFFP methods evaluated in this thesis work are briefly explained and results are presented. The results of novel RFFP methods proposed in this study are compared to the conventional method.

Chapter 4: Finally concluding remarks are given and outlines the future studies relevant to the current topic.

2 RF FINGERPRINTING USING MDT DATA

2.1 Minimization of Drive Tests

Coverage estimation is the first step for cellular network deployments which consists of signal strength estimation in the area to be served. To deal this coverage prediction operators usually perform *drive tests*, which consist of geographically measuring different network metrics and indicators with motor vehicles equipped with specialized mobile radio measurement equipment and GNSS System and then analyze the collected measurements to derive optimal parameters. Once the optimal parameters have been selected, the parameters are applied to the networks and another set of driving tests is carried out to evaluate the impact of the parameter tuning. Such calibration may have to be repeated until the expected level of performance is achieved. Drive tests generate a tremendous amount of data to be processed, allowing the operators to get realistic network information close to the actual user experience which is very useful and desired information by operators (Hamalainen et al., 2012). Drive tests are quite an inefficient means to solve the coverage problems since they: (1) imply large Operational Expenditure (OPEX), (2) incur delays in detecting and predicting the coverage holes, (3) are an undesirable source of pollution, and (4) provide an incomplete picture of the "ground-truth" since they are limited to roads and other regions accessible by motor vehicles (Hamalainen et al., 2012). For this purpose, the 3GPP standardization body has been working on the minimization of the use of drive tests for LTE since Release 9 (3GPP, 2009b). In the study phase of the MDT the following use cases were considered: Coverage optimization, Mobility optimization, Capacity optimization, Parameterization for common channels, Quality of Service (QoS) verification. In Release 10, a Minimization of Drive Tests (MDT) work item (3GPP, 2012b; Hapsari et al., 2012) for Universal Terrestrial Radio Access Network (UTRAN) was also included with main focus on coverage optimization. Release 11 focuses on QoS verification, further improvements in coverage optimization, positioning enhancement and the study of other MDT use cases (Holma and Toskala, 2012).

The key idea of the MDT proposed by 3GPP is to take advantage of the geo-location measurement capabilities of the recent advanced subscribers UEs as well as of the radio measurements performed any-time and anywhere as part of the Radio Resource Management (RRM) procedures. The main characteristic of MDT is that the UEs report their *geo-located* measurements to the network upon operator request. The collected MDT measurements are at operator's direct disposition to ease any kind of (automated as well as manual) network operation, management and optimization task.

The design principle of MDT in Release 10 includes (3GPP, 2012c):

1) support for real-time and non-real-time measurement reporting;
2) acceptable end-user implications, for example, battery consumption and memory requirements should be kept reasonable;
3) correlation of the measurement results with time and location, indicating when and where the measurements were obtained, respectively.

The MDT in Release 10 is built on control plane architecture as shown in Figure 2. It consists of the OAM system, which initiates and controls MDT data collection, the RAN and the UE, where data is collected, and the trace collection entity (TCE), where MDT data is sent after collection. The core network has a general role in the signaling to control MDT and other trace features. The post processing, analysis, and visualization of collected MDT data in the TCE is not standardized. From network signaling perspective there are two methods to control MDT (3GPP, 2011a).

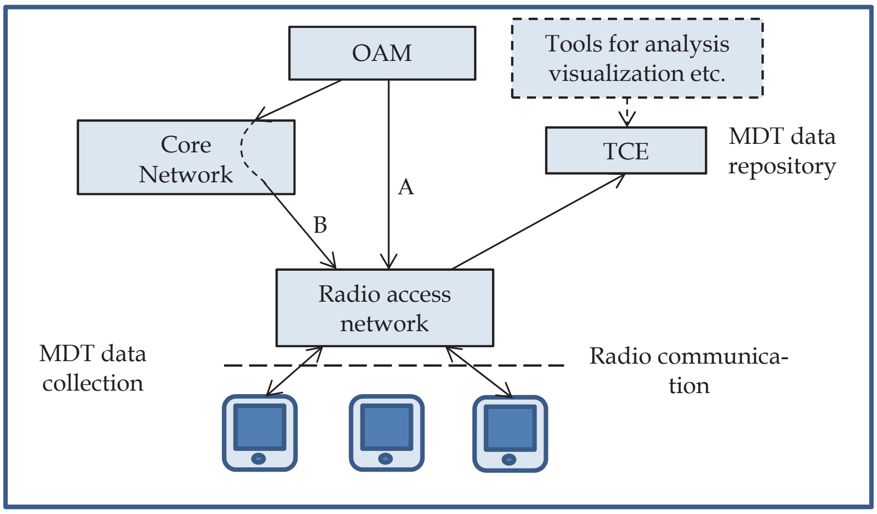


FIGURE 2 MDT architecture

The first method is called "management-based MDT," which uses control path A as shown in Figure 2. It reuses cell traffic trace feature to collect MDT measurements from randomly chosen UEs or a group of UEs that enter a certain geographical area. The second method is called "signaling-based MDT," and uses control path B. It reuses subscriber and equipment trace feature to collect from a specific subscriber or equipment (3GPP, 2011b; Hapsari et al., 2012). Regardless of which method is used to control MDT, OAM is able to obtain and associate measurements from both the UE and the RAN.

From radio configuration perspective, two types of MDT were defined (3GPP, 2011):

1) Logged MDT is the procedure by which the UE performs logging of measurement results and reporting of the logged measurement results. By the definition of logging, the UE stores and accumulates measurement results within its memory. The logged measurement results are reported afterwards when the reporting condition is met. This type of MDT is performed when the UE is in idle state (i.e., the UE has no setup connection with the RAN node).

Each log entry consists of serving cell information, neighbouring cell information, time information, and location information. The neighbouring cell and location information are included only if available. Serving cell information consists of the Evolved Cell Global Identifier and measurement results for the serving cell in terms of Reference Signal Received Power (RSRP) and Reference Signal Received Quality (RSRQ). The neighbouring cell information spans the measurement results of E-UTRA cells, UTRA cells, GERAN cells, and CDMA2000 cells, depending on the UE capabilities. The measurement results of neighbouring cells are logged in decreasing order of the ranking criterion used for cell reselection. Within each log entry, a maximum of six intra-frequency neighbouring cells and three inter-frequency neighbouring cells per frequency can be included for E-UTRAN/UTRAN/CDMA2000 cells, and three GERAN cells per frequency/set of frequencies. A time stamp is included in each log entry.

2) Immediate MDT is the procedure by which the UE performs measurements and reports the measurement results as soon as the reporting condition is met. This type of MDT is performed when the UE is in active state (i.e., the UE has setup a connection with the RAN node). The general principle of Immediate MDT follows the existing RRM measurements and reporting mechanisms defined in (3GPP, 2012d). The only enhancement required to support Release 10 Immediate MDT is that the UE can be configured to tag accurate location information, if available, onto the measurement results reported to the network. Support for Immediate MDT is mandatory for all UEs, since the procedure for Immediate MDT is based solely on existing RRM measurement procedures. The standard allows any combinations between the above MDT types.

Location information for MDT can be categorized into two different types: detailed location information and RF fingerprint. Detailed location information consists of at least latitude and longitude, whereas an RF fingerprint is a profile of measured signal strength from neighbouring cells. The RF fingerprint can be obtained by the network as part of the radio measurements reported by the UE and used to calculate the approximate location of the UE (e.g., by cell triangulation) (3GPP, 2011b). Inclusion of detailed location information as part of MDT measurement reports is very important. In contrast to the best effort location information acquisition defined in 3GPP Rel-10, Rel-11 adopts the approach that the network can request the UE to acquire location information for an MDT session that is being configured. This enhancement to improve location information acquisition for MDT is based on the assumption that the positioning method used for calculating location information is standalone GNSS/GPS. For MDT that involves UE-based measurements, both radio measurements and location information are measured at the UE. The UE tags the radio measurements with location information and sends them together in the same measurement report to the network, which then uses the UE-reported location information when performing analysis of the MDT data. However, some of the MDT measurements defined in 3GPP Rel-11 are performed by the RAN rather than by the UE.

2.2 MDT Database Creation using Dynamic System Simulator

In this work MDT data were generated using a proprietary dynamic system simulator known as *Freac*. The *Freac* simulator models an E-UTRAN system in a downlink and an uplink direction by incorporating detailed implementations of various LTE functionalities such as radio resource management, link adaptation, scheduling, radio link detection and monitoring, radio measurements for mobility management, and various models for user mobility and traffic profiles. For an accurate estimation of packet error rates using different modulation and coding schemes the simulator maps a link level signal-to-interference-plus-noise ratio (SINR) performance by following the methodology in (Brueninghaus et al., 2005). The system configuration follows mostly the 3GPP simulation assumptions in (3GPP, 2006a) and (3GPP, 2009c) in order to model radio propagation, antenna characteristics, slow fading and fast fading. First the system is initialized, in this stage a simulation world is created with pre-determined configuration and location of network elements and user positions. Here one simulation step takes 1/14000 seconds, during the time interval, spatial location of a user is recalculated and radio conditions are updated. UEs monitor the radio link quality of all E-UTRAN Node B (eNBs); the strongest eNB is the serving eNB and other eNBs interfere with the connection between the UE and the serving cell. The core part of MDT modelling is how the RSRP and RSRQ measurements are done. RSRP is defined as the linear average over the power contributions of the resource elements that carry cell-specific reference signals (CRS) within the considered measurement bandwidth (3GPP, 2012e).

In LTE, transmitted energy is divided into subcarriers in frequency domain and Orthogonal Frequency-Division Multiplexing (OFDM) symbols in time domain. One subcarrier and one OFDM symbol forms a resource element in LTE. A physical resource block contains 12 adjacent subcarriers and 14 consecutive OFDM symbols. During simulations the measurement bandwidth was configured to consist of the six centremost Physical Resource Blocks (PRBs). RSRP and RSRQ modelling begins by determining a radio link between eNB and UE. The radio link is antenna specific and is given by,

$$l_e^{iu} = \frac{pl_{iu}(d_{iu}) \cdot sf_{iu} \cdot |ff_{iu}(e_{s,sc})|^2}{g_i(\phi_{iu}, \theta_{iu})} \quad (1)$$

where, l_e^{iu} is the radio link between eNB i and UE u for one resource element $e_{s,sc}$. The total attenuation of the radio link is a product of pathloss $pl_{iu}(d_{iu})$ depending on the distance d_{iu}; lognormal slow fading factor sf_{iu} depending on UE's spatial location; average power of complex fast fading factor $|ff_{iu}(e_{s,sc})|^2$ varying for each resource element e indexed by the OFDM symbol s and subcarrier sc; and antenna gain $g_i(\varphi_{iu}, \theta_{iu})$ depending on the angular difference of the UE and the eNB antenna direction in horizontal plane φ_{iu} and vertical plane θ_{iu}. The radio link is used to determine the received power per resource element rx_e^{iu} which given by,

$$rx_e^{iu} = \frac{tx_e^i}{l_e^{iu}} \quad (2)$$

where rx_e^{iu} is the received power and tx_e^i is the transmitted power per resource element from eNB i. For calculating an instantaneous RSRP denoted by $rsrp(t)$, the received power is averaged over the resource elements in the measurement bandwidth containing Cell-Specific Reference Signal (CRS) symbols for antenna port 0. The measurement bandwidth consisted of the six centermost PRBs and the measurement period for the instantaneous RSRP is one subframe that consists of 14 OFDM symbols. The six PRBs contain 48 resource elements that transmit the CRS symbols. These resource elements are distributed in frequency and time domain. A set S={1,4,7,11} denotes the indexes of OFDM symbols that carry CRS symbols for antenna port 0 whereas a set RS_s^N denotes the indexes of subcarriers that contain the CRS symbol on the measurement band during the OFDM symbol s. The linear average of instantaneous RSRP for tth subframe is given by,

$$rsrp(t) = \frac{1}{N_s \cdot Nrs_N} \sum_{s \in S} \sum_{sc \in RS_s^N} rx_e^{iu} \quad (3)$$

where Nrs_N equals to 12 and it is the total number of subcarriers per OFDM symbol that transmit the reference symbols for antenna port 0 on the measurement bandwidth of six PRBs. The variable Ns is 4 and this corresponds to the size of S i.e., the number of OFDM symbols containing reference symbols per subframe. A detail explanation of the simulator and modelling aspects is given in Appendix A of (Turkka, 2014).

2.3 Basic Principle of RF Fingerprinting

As a fingerprint is unique to a human, the properties of a radio signal are typical for the position where it is transmitted or received. A RF fingerprint or RF signature in wireless positioning is the set of measurable signal characteristics that depend on the position of transmission or reception (Sand, 2014). Each signature is registered at a unique location - typically at points on a uniformly spaced grid throughout a given environment. The basic idea behind location fingerprinting technique is to associate physically measurable properties to discrete locations throughout a deployment area. A fingerprinting method then compares a measured fingerprint of a test UE to the fingerprint entries of that database in order to determine the database fingerprint that matches best. This best match then provides the most likely position of that UE. Thus RF fingerprinting methods are also known as 'database correlation' methods (DCMs) and 'pattern matching' methods. These properties can then act as location identifiers. The greater the spatial variability of the signatures, the greater the capacity of the system to discriminate between locations. RF fingerprinting techniques are categorized mainly by the type of properties which are collected. The three major RF properties that have been implemented to date are: RSS, the time domain Channel Impulse Response (CIR) and the Frequency Coherence Function (FCF). RSS is the most prevalent in commercially deployed wireless systems due to many factors, most notably its robustness and good penetration in NLOS conditions, its simple data structure, and the computational ease with which it can be measured (Gentile, 2013).

The process of RF fingerprinting can be divided into the following two primary phases (Vaghefi and Buehrer, 2014):

Offline Training Phase: RF fingerprints are collected through field measurements or generated using simulation models to construct a 'radio map' and are stored in a database called RF fingerprint Correlation Data-Base (CDB). Each RF fingerprint stored in the CDB is associated with a specific position.

Online Localization Phase: Upon arrival of a signal from test UE, the signal signature of interest is extracted and compared to the radio map using one or more pattern matching techniques.

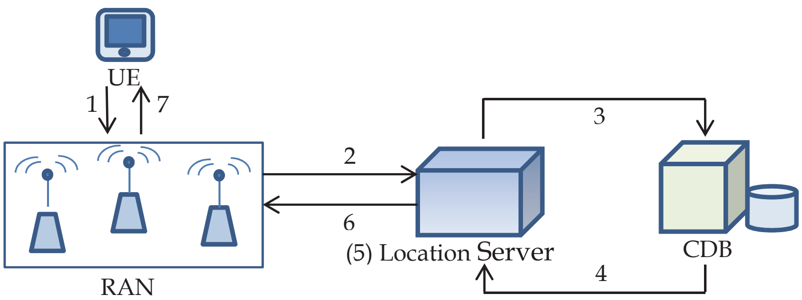


FIGURE 3 A simplified schematic diagram of RF fingerprinting

Then, the mobile position is estimated by selecting the best match or by interpolating among the best matches on the radio map.

An RF fingerprint can be classified as either a *target* or *reference* fingerprint. A target RF fingerprint (T_fing) is the RF fingerprint associated with the test UE that is to be localized. It contains signal parameters measured by the UE or by its anchor cells. The reference RF fingerprints (R_fings) are the RF fingerprints collected or generated during the training phase and stored in the CDB (Zekavat and Buehrer, 2011).

A simple RF fingerprinting operation in UE-assisted cellular network based positioning system is depicted in Figure 3 and explained as follows:

First (step 1), the UE sends a position request containing T_fing to the location server through the RAN. The RAN then (step 2) communicates with the location server, usually through a gateway. Upon receiving the position request and T_fing the location server queries the CDB (step 3) for R_fings, which is returned in sequence (step 4). The location server then compares the T_fing with returned R_fings to obtain the UE position estimate (step 5), which is sent back to RAN (step 6) and then to the UE (step 7) (Campos and Lovisolo, 2015).

RSS fingerprinting systems have been successfully deployed in dense urban and indoor environments by Skyhook Wireless and Ekahau, respectively. Skyhook Wireless's RF fingerprinting technology has attracted attention from the major players in the mobile device industry such as Apple and Google (Wortham, 2009). The technique is very practical and delivers decent accuracy (tens of meters) for mobile location applications in outdoor urban environments where Wi-Fi APs are plentiful.

RF fingerprinting techniques can be categorized into two broad categories:

Deterministic technique: Here the signal strength of an access point at a Location is represented by a scalar value and non-probabilistic approach is used to estimate the user Location (Vaghefi and Buehrer, 2014; Smailagic and Kogan, 2002).

Probabilistic technique: Information about the signal strength distributions from the access points are stored in the radio map and probabilistic techniques are used to estimate the user location (Roos et al., 2002; Youssef, 2003).

2.3.1 Cellular Mobile Network Based RFFP

In cellular mobile networks BSs have fixed geographical positions. Each BS broadcasts its CID and Location Area Identifier (LAI) to the subscriber UEs within its coverage area. The geographical coordinates of the serving BS of a UE can be used to estimate its approximate position (Trevisani and Vitaletti 2004). In (Pages-Zamora et al., 2002) Angle of Arrival (AOA) measurements of several radio links between the base stations and the UE were used to estimate client position. Here UE position is approximated according to the angle measurements and using information of BSs geographical coordinates. Authors in (Deng and Fan, 2000) have used Time of Arrival (TOA) based positioning system to convert the time difference between bursts sent by the UE into distance and trilateration is used to estimate UE position. Time Difference of Arrival (TDOA) technique was used in (Shin and Sung, 2002) and (Gustafsson and Gunnarsson, 2003) which require the simultaneous transmission of two signals having different frequencies. These signals reach the UE at different times; which is measured and converted into distance. Once three distance measurements are completed trilateration is used to estimate the UE position. Enhanced Observed Time Difference (E-OTD) technique was used in (Charitanetra and

Noppanakeepong, 2003) which require synchronization between network entities (BSs and UEs). Here each BS broadcasts messages within its coverage area and UE compares the relative times of arrival of these messages to estimate its distance from each visible base station. In AOA and other signal timing based methods UE location is calculated under the assumption of line-of-sight RF signal propagation between the BSs and the UE. In densely built urban areas severe multipath propagation characteristic of RF signal makes it difficult to detect the angle of arrival or time of arrival of the direct component, as a result the resolution of these measurements is not as good as in open propagation environments. Hence UE positioning performance of AOA and TDOA based methods is not optimal in urban environments. These methods also require hardware modifications to the network and to handsets as well. Using real life GSM measurements authors in (Laitinen et al., 2001) has shown that the positioning accuracy of RFFP method is better than that of AOA or E-OTD in urban environments and acceptable for many applications in a suburban environment also. However a disadvantage of this method is that the impulse response measurement technique used is not standardized, thus it requires software changes in the UEs. Moreover, reporting of such measurements to the location server is also not standardized, which reduces the degree of applicability of this DCM method. Hence authors in (Borkowski and Lempiaeinen, 2005) have introduced Pilot Correlation Method (PCM) for UE positioning which is entirely based on standardized measurements and procedures, therefore significantly reducing the deployment costs. Here the area of service coverage is divided into small positioning regions and for each such region the most probable situation of Common Pilot Channels visibility with corresponding results of Received Signal Code Power (RSCP) measurements is stored in Serving Radio Network Controller (SRNC). Size of the positioning regions is tuned according to the desired accuracy for planned location sensitive applications. Outdoor real data from Elisa UMTS network in urban environment in Tampere city in Finland was used for evaluating the method. The positioning accuracy obtained by the PCM method fulfilled the safety requirements for network-based solutions and due to its simple procedure of positioning the overall latency does not introduce any limits in various potential applications. In (Borkowski and Lempiainen, 2006) authors have presented an extended work where enhanced cell-ID + Round Trip Time (RTT) (ECID+RTT) based cellular positioning method was evaluated. This technique is based on multiple time measurements and requires a slight software update in the UEs. The reported accuracy of the ECID+RTT for 67% of the location estimates varies from 60m-100m in urban scenario. In (Ahonen and Laitinen, 2003) UE positioning was performed using DCM approach with simulated data generated for a UMTS microcell network layout in urban areas. Here in order to take advantage of the urban propagation environment and to avoid typical problems related to multipath channel delay profiles in locations were used for the DCM method and results were compared to two standard localization techniques OTDOA and CID. For the DCM approach only the strongest cell is measured and used in the correlation with the database. It was found from simulation results that DCM circumvents most of the urban environment related issues as compared to the OTDOA method. The accuracy achieved with DCM is adequate for most localization applications and within the range required by emergency services. In (Zimmermann et al., 2004) DCM method is utilized for network-based UE location estimation where wave propagation software generated the training database. Extended Walfisch-Ikegami Model was used to create simulated data for the urban scenario. In such an urban scenario mean location error was below 100m which outperformed simple CID based positioning. In (Lakmali and Dias, 2008) authors have proposed novel methods to improve RF fingerprint positioning using GSM signal strength values in indoor and outdoor environments. Due to multi-path propagation and small-scale fading measured RSS value at a particular location would vary time to time and also not all the detectable cells would appear in all the measurements. Hence for outdoor positioning ten measurements were used at the UE to be located. Here database search was reduced based on the serving BS of each database fingerprint and for fingerprint matching new cost functions were presented based on the Manhattan distance and the Euclidean distance. For UE positioning in the urban scenario Nearest Neighbour based location estimation has given the best result. The proposed method showed excellent performance over CID and trilateration methods for outdoor positioning and is nearly compliant with FCC-US E911 accuracy requirement. A path loss modeling based UE location estimation method was proposed in (Lin and Juang, 2005) which uses Cost231 Hata model. Here through calculating differences of downlink signal attenuations circles are drawn along which the test UE may lie and from the intersection of the curves UE position is estimated. The advantages of the proposed method include no need for a perfect path loss modeling, the reduction in shadowing impact on location, low computational complexity, and applications for existing systems without hardware development. The obtained results showed that the proposed method outperforms CID method in a real GSM cellular network environment in urban Taipei. Authors in (Varshavsky et al., 2005) have pointed out that positioning methods based on Infrared (Hopper et al., 1993), ultrasound (Priyantha et al., 2000) and Bluetooth localization systems (Aalto et al., 2004) work well indoors, but deploying these technologies to the wide area is either cost prohibitive or not technically possible. GSM-based localization systems have the potential to detect the places that people visit in their everyday lives with good localization accuracies for both indoor and outdoor environments. Devices that people do carry with them most of the time that have continuous network connectivity is the cellular mobile phones which have low power consumption, ubiquitous connectivity, established interface metaphors, wide adoption and they can offer indoor-outdoor place detection capabilities. Client-based GSM localization can provide an adequate solution both in terms of coverage and accuracy. Experimental results showed that using GSM it is feasible to achieve 2-5 meters median error and room-level indoor localization and 70-200 meters median error for outdoor UE positioning. In (Chen et al., 2006) authors have examined the feasibility of a client-side, beacon-based GSM outdoor localization system where the UE can position itself in a privacy-observant manner and can use cell towers from all surrounding cellular network operators to compute location without requiring assistance from the network operators. The approach also used existing hardware in mobile phones without requiring any additional hardware. RF fingerprints were collected having records of CIDs as well as observed signal strength in dBm for up to seven GSM cells from three different GSM network providers. External antennas for the GSM modems and GPS units were placed on the car roof to improve signal reception. The complete trace contains 6756 unique cells spanning an 18Km x 25Km region of the Seattle metropolitan area, USA collected over a period of three months and test trace was collected two weeks after the training trace was collected. Positioning accuracy of three positioning algorithms was evaluated: a simple centroid algorithm (LaMarca et al., 2005), fingerprinting (Bahl and Padmanabhan, 2000) and Monte Carlo localization with Gaussian Processes signal propagation model (Schwaighofer et al., 2003). Five factors affecting positioning accuracy were examined: location algorithm choice, scan set size, simultaneous use of cells from different providers, training and testing on different devices, and calibration data density. All three algorithms performed better in higher tower density area than lower tower density area; the fingerprinting algorithm performs much better than centroid method. Both the fingerprinting and Gaussian Processes algorithms improve in accuracy when more cells were used for positioning. Fingerprinting was the most accurate in the high-density urban area and performed well in the low-density area as well. However, fingerprinting was the most fragile of the algorithms, requiring dense training data; it was also the most fragile to the addition of new cells during the calibration drive. In (Ibrahim and Youssef, 2012) authors have proposed a probabilistic Received Signal Strength Indicator (RSSI)-based fingerprinting system- 'CellSense' for GSM mobile phones. In order to reduce complexity of constructing a probabilistic fingerprint 'CellSense' uses gridding, where the area of interest is divided into a grid and signal strength histogram is constructed for each grid cell. This increases the scalability of the technique. To further reduce the computational overhead a hybrid technique- 'CellSense-Hybrid' was proposed which combines a probabilistic estimation phase with a deterministic refinement phase; it is also robust to changes in its parameter values. Performance of 'CellSense' and 'CellSense-Hybrid' techniques were analysed under two different testbeds representing rural and urban environments using three different cellular network providers. Experimental results show that positioning accuracy of 'CellSense-Hybrid' method is better than other techniques with at least 108.57% in rural areas and at least 89.03% in urban areas with more than 5.4 times savings in running time compared with other state-of-the-art RSSI-based GSM localization techniques. A path loss (PL) model based RF fingerprinting approach was suggested in (Nurminen et al., 2012) where the number of required path loss parameters was kept small to keep down the computational complexity and the amount of information required in the positioning phase. A measurement campaign was accomplished to evaluate the performance of different algorithms in a real use case. It was found that estimating the path loss parameter uncertainties and taking them into account in the positioning phase has significant influence on the positioning performance. Here both accuracy and consistency of the proposed method were compared to the conventional methods. It was shown that a Gauss–Newton optimization algorithm can provide satisfactory accuracy and consistency with less computational cost. A probabilistic approach of UE localization in a real GSM outdoor environment was presented in (Li et al., 2013) through investigating the use of clustering and Principal Component Analysis (PCA). Here clustering was done on the RSS tuples based on the deviations of the raw RSS from the estimated path loss model for each RSS component. When compared to traditional approaches, the results indicate that applying the PCA method to transform the RSS data within each cluster and the application of kernel density estimator algorithm can improve the position estimation accuracy in complex urban environment as well as reducing the computational burden and storage. A new Adaptive Enhanced cell-ID (AECID) based fingerprint localization method was proposed in (Wigren, 2007). Here whenever a high precision A-GPS or OTDOA UE positioning measurement is received, the CIDs, the TA, the signal strengths and AOA information are measured and sent to be stored as reference points in clusters. After collecting sufficient amount of reference points in a cluster, a polygon is computed with the contraction polygon algorithm, and then it is stored in a database of fingerprinted polygons. When a positioning request is received, the list of own and neighbor CIDs; the TA are retrieved; RSS and AOA measurements are performed and quantized to create the fingerprint of the test UE. The polygon that corresponds to the fingerprint is collected from the database and reported. In (Wigren, 2011) the author tried to refine the clustering step of the AECID positioning method presented in (Wigren, 2007) through automatic cluster smoothing, outlier removal and cluster splitting algorithms for UE positioning in LTE cellular environment. The accuracy of the polygon computation scheme of the AECID algorithm was improved and the associated computational complexity was reduced. Numerical results of (Wigren, 2011) showed that the proposed algorithm significantly improves the possibility for the polygon computation algorithm to produce highly accurate polygons and thereby improving UE localization.

2.3.2 WLAN Based RFFP and Hybrid RFFP Techniques

A comparison study between RF signal propagation path loss-based RFFP (PL-RFFP) and grid-based RFFP approaches in terms of positioning accuracy for simulated outdoor cases was performed in (Laitinen et al., 2011). It was found that the G-RFFP method performed better than PL-RFFP method in terms of positioning accuracy, even if the estimates for channel characteristics in the assumed path loss model are perfect. PL-RFFP is much more sensitive to channel impairments such as fast fading and shadowing than G-RFFP approach. It was also observed that by varying the parameters in G-RFFP method (such as the grid size and the number of neighbour points used in the estimation) its positioning accuracy can be improved.

In (Kim et al., 2014) a 2-stage hybrid position estimation framework for outdoor RFFP using WLAN RSS was proposed. The system consists of three elements: grid size determination - a variable grid size is obtained depending upon the density of collected fingerprints; data filtering - the WLAN APs that have lower RSSI values compared to a threshold RSSI value are eliminated in the reference fingerprint for each grid and this threshold is coherently applied to the measured fingerprint pattern of a test UE during position estimation phase; and position estimation - first based on the reference fingerprint map a probabilistic distribution map for each AP is built. When a fingerprint pattern of a test UE is measured, all the overlapped probabilistic distribution maps of APs are retrieved and then the estimated position of the test UE is determined from the most overlapped grid. When compared to the existing RFFP approaches the proposed method achieves a better performance in both average error of estimation and deviation of errors.

In a RF fingerprint Wi-Fi positioning system (WPS) grid size determination is one of the main concerns of. The most desirable situation is the uniform assignment case where a single, effective fingerprint can be assigned to each grid. However, in most cases, fingerprint collection cannot be performed uniformly and the collected fingerprints are scattered: partially dense and partially scarce. It is hard to find any relation between grid segmentation size and position estimation quality. Hence in (Kim et al., 2014) a data clustering method was presented to perform irregular grid segmentation. Based on the statistical significance test for RF fingerprints, collected fingerprints were merged to generate grids which had an irregular form to cover the geographical area. Compared with conventional fingerprint data alignment approaches, this method achieved better performance in position estimation.

Authors in (Kim and Yeo 2014) have proposed a coherent data-filtering framework where the number of APs for each reference fingerprint is maintained, with an effective RSSI threshold level. By maintaining a common threshold for both the reference fingerprint database and test UE fingerprint an effective dimension of the fingerprints was obtained, which imparts good position estimation quality, with sufficiently fast running speed. Real life RF fingerprint data were collected from Seoul's Gangnam district, China and the proposed method was compared to the existing RFFP approaches for fingerprint filtering. It showed better performance in UE positioning accuracy.

Although GNSS offers good outdoor location accuracy of around 10m, it incurs a serious energy cost that can drain a fully charged phone battery in 8.5 hours (Gaonkar et al., 2008). Hence in (Constandache et al., 2009) authors have developed an energy-efficient localization framework called '*EnLoc*' as an alternate localization technologies, based on WiFi or GSM. The framework characterizes the optimal localization accuracy for a given energy budget, and develops prediction based heuristics for real-time use.

In a study conducted in (Roos et al., 2002) WPS positioning performance of three different methods- nearest neighbour based RFFP, Probabilistic RFFP based on kernel function and probabilistic RFFP based on histogram were compared in experimental tests. The results showed that upon a moderate amount of effort used in collecting training data produces practically applicable results with an average location estimation error below 2 meters. In the experiments performed, the two probabilistic methods produced slightly better results than the nearest neighbour method. The probabilistic methods were found to be relatively robust with respect to the number of base stations used, the amount of calibration data collected, and the length of the history used in the location estimation.

Recorded radio map can be outdated when signal-strength values change over time due to environmental dynamics and repeated data calibration is expensive. Thus a novel method called Location Estimation using Model Trees (LEMT) was proposed in (Yin et al., 2008) for estimating UE locations even when RSS samples are dynamically changing over time. This algorithm can take real-time RSS values at each time point into account and make use of the dependency between the estimated locations and reference points. This technique can effectively accommodate the variations of signal strength over different time periods without the need to repeatedly rebuild the radio maps. The effectiveness of LEMT is demonstrated using two real data sets collected from an 802.11b wireless network and a Radio Frequency Identification (RFID)-based network. Experimental results show that compared with existing adaptive techniques, LEMT is much more robust to the reduction in the number of reference points.

A low-cost system for UEs to discover and communicate their position in the physical world has long been identified as a key primitive for emerging mobile applications. Authors in (Cheng et al., 2005) conducted a comparative study between three Wi-Fi positioning algorithms - centroid based RFFP (LaMarca et al., 2005), KNN based RFFP (Bahl and Padmanabhan, 2000) and Particle filters (Hightower and Boriello, 2004) based RFFP. It was found that with dense Wi-Fi coverage, the specific algorithm used for positioning is not as important as other factors including composition of the neighbourhood, age of training data, density of training data sets and noise in the training data.

The accuracy of a Wi-Fi fingerprint localization system greatly depends on the quality of its radio map. Authors in (Kim et al., 2006) observed that the AP location estimation error could be large for war driving, seriously affecting the UE positioning accuracy. They also claimed that war walking is more effective than war driving for collecting signals because people can walk very close to buildings and walking is much slower than driving. In (Tsui et al., 2010) authors have analysed four properties (RSS values of APs, amount of fingerprints collected, GNSS labelling of calibration points and uniformity of the calibration points) of driving and walking radio maps which may affect UE positional accuracy. According to their findings to improve war driving localization following factors should be addressed: 1) choosing devices that report a smaller minimum RSS, 2) obtaining multiple samples by using multiple devices at each AP location, 3) choose uniformly distributed calibration points on the radio map and 4) offset the GNSS lag effect.

As an alternative to GPS hardware and its unavailability indoors, project Place Lab (LaMarca et al., 2005) proposed a positioning method that uses WiFi and GSM sensors for localization. It tries to achieve two goals: (1) maximizing coverage across entire metropolitan areas, and (2) providing a low barrier to entry by utilizing pre-deployed hardware. Authors create a wireless map of a region by war-driving in the area which is composed of sampled GPS locations, WiFi access points and GSM towers audible at these locations. This wireless map is then distributed to phones. When a phone travels through the mapped area, it estimates its own location by matching its list of audible WiFi APs/GSM towers to the wireless map. Place Lab experiments in downtown Seattle, USA exhibit a median positioning error of 13 to 40m with WiFi, and around 94 to 196m with GSM.

An experimental analysis for outdoor WLAN fingerprinting system was implemented in urban area of Shenzhen, China in (Liu et al., 2010). The RFFP method used KNN and utilizes mean fingerprinting (average RSS value collected from each AP) to estimate the user's location. The main advantage of this method is its simplicity to set up and reasonable level of accuracy. It was observed that fingerprinting system provides a better positioning accuracy in the dense urban areas by adopting reasonable grid spacing and the number of APs.

In (Magro and Debono, 2007) a Genetic Algorithm (GA) based location detection method was proposed which uses cellular network information such as CID and transmitted RSS as input to search the whole network coverage area for the most probable point of origin. A dynamic UMTS radio model was used to simulate real field measurements of the RSS variations. Here GA was employed in the triangulation algorithm which operates by generating an initial random population of candidate locations and repeatedly modifying this population to evolve towards better solutions. Simulation results showed that the algorithm is capable of locating UE in 72% of the cases in an urban environment within a range of 450 meters.

In order to reduce both the average location error and average time to produce a position fix, in comparison to the work presented in (Magro and Debono, 2007) authors in (Campos and Lovisolo, 2013; Campos and Lovisolo, 2010) propose an alternative for the initialization of the first generation population of GA. Instead of selecting individuals throughout the whole service area it restrains the randomly selected first population individuals to the predicted best server area of the serving BS. Test results from an 850 MHz GSM network in a metropolitan area showed that an average reduction of 91% in the time to produce a position fix could be achieved with reductions of 20% and 15% in the 50^{th} and 90^{th} percentile location errors respectively in comparison to the approach followed in (Magro and Debono, 2007).

3 DATA PROCESSING

3.1 Genetic Algorithm

Evolutionary computing was introduced in the 1960s by I. Rechenberg in the work "Evolution strategies". The most popular technique in evolutionary computation research has been the GA. In the year 1975John Holland developed this idea in his book "Adaptation in natural and artificial systems". Holland's theory has been further developed and now GAs stand up as a powerful tool for solving search and optimization problems. Search is one of the more universal problem solving methods for such problems one cannot determine a prior sequence of steps leading to a solution. Search can be performed with either *blind strategies* or *heuristic strategies* (Bolc and Cytowski, 1992). Blind search strategies do not use information about the problem domain. Heuristic search strategies use additional information to guide search move along with the best search directions. There are two important issues in search strategies: exploiting the best solution and exploring the search space (Booker, 1987). Random search is an example of a strategy which explores the search space, ignoring the exploitation of the promising regions of the search space. GA is a class of general purpose search methods combining elements of directed and stochastic search which can produce a remarkable balance between exploration and exploitation of the search space. The usual form of a GA is described by Goldberg (Goldberg, 1989). GAs are numerical optimization algorithms inspired by both natural selection and natural genetics. The algorithms are simple to understand and easy to implement. GAs have proved themselves capable of solving many complex problems where other traditional methods have experienced difficulties. An important feature of GA is that it is a *stochastic* Algorithm, and the second very important point is that GAs always consider a population of solutions. The algorithm can recombine different solutions to get better ones and so, it can use the benefits of assortment. The robustness of the algorithm should also be mentioned as something essential for the algorithm success. Robustness refers to the ability to perform consistently well on a broad range of problem types. There is no particular requirement on the problem before using GAs. All these features make GA a really powerful optimization tool and it is currently the most prominent and widely used computational models of evolution. GA, differing from conventional search techniques, start with an initial set of random solutions called *population* satisfying boundary and/or system constraints to the problem. Each individual in the population is called a *chromosome* (or *individual*), representing a solution to the problem at hand. These are usually random and spread throughout the search space. The chromosomes *evolve* through successive iterations called *generations*; during each generation, the chromosomes are *evaluated*, using some measures of *fitness*. A typical GA uses three operators: *selection, crossover* and *mutation* to direct the population over a series of generations towards convergence at the global optimum (Coley, 1999). To create the next generation, new chromosomes, called *offspring*, are formed by either merging two chromosomes from current generation using a *crossover* operator or by modifying a chromosome using a *mutation* operator. A new generation is formed by *selection*, according to the fitness values, some of the parents and offspring, and rejecting others so as to keep the *population* size constant. Fitter chromosomes have higher probabilities of being selected. At the beginning of genetic search, there is a widely random and diverse population and crossover operator tends to perform wide-spread search for exploring all solution space. As the high fitness solutions develop, the crossover operator provides exploration in the neighbourhood of each of them. After several generations, the algorithms converge to the best chromosome, which hopefully represents the optimum or suboptimal solution to the problem.

In general, a GA has five basic components, as summarized by Michalewicz (Michalewicz, 1994):

1) A genetic representation of potential solutions to the problem.

2) A way to create a population (an initial set of potential solutions).

3) An evaluation function rating solutions in terms of their fitness.

4) Genetic operators that alter the genetic composition of offspring (crossover, mutation, selection, *etc*.).

5) Parameter values that genetic algorithm uses (population size, probabilities of applying genetic operators, *etc*.).

Advantages:

GAs have proved themselves capable of solving many large complex problems where other methods have experienced difficulties (Coley, 1999). The advantages of GA mainly include the simplicity of the approach, its robust response to changing circumstances, and its flexibility (Sivanandam & Deepa, 2008):

1) A key advantage of GA is that it is conceptually simple; 2) GAs can be applied to any problems that can be formulated as function optimization problems; 3) GAs can be combined with more traditional optimization techniques; 4) The evaluation of each solution can be handled in parallel and selection only requires some serial operation; 5) Traditional methods of optimization are not robust to dynamic changes in the environment and they require a complete re-

start for providing a solution, in contrary, GA can be used to adapt solutions to changing circumstances; and 6) It also has the ability to address problems for which there is no human expertise.

Multi-objective Genetic Algorithms (MOGA):

In principle, multiple-objective (MO) optimization problems are very different from single objective optimization problems. In a single objective case, the best solution is the one which is absolutely superior to all other alternatives while in the case of MOs, there does not necessarily exist such a solution because of *conflict* among objectives. A solution may be best in one objective but worst in other objectives. Therefore, there usually exists a set of solutions for the MO case which cannot simply be compared with each other. These solutions are called *non-dominated* solutions or *Pareto optimal* solutions. The inherent characteristics of the GA demonstrate its ability to be a good solver for MO optimization problems. The basic feature of the GA is the multiple directional and global search by maintaining a population of potential solutions from generation to generation. The *population-to-population* approach is useful to explore all Pareto solutions. One special issue in MO optimization problems is the *fitness assignment mechanism,* several such mechanisms have been proposed and applied in MO optimization problems (Gen & Cheng, 2000; Deb, 2001). A popular fitness assignment method based on Pareto ranking was first suggested by Goldberg (Goldberg, 1989). In (Srinivas and Deb, 1994) authors also developed a Pareto ranking-based fitness assignment called non-dominated sorting Genetic Algorithm (NSGA) where the non-dominated solutions constituting a non-dominated front are assigned the same dummy fitness value. These solutions are shared with their dummy fitness values and ignored in the further classification process. Finally, the dummy fitness is set to a value less than the smallest shared fitness value in the current non-dominated front. Then the next front is extracted. This procedure is repeated until all individuals in the population are classified. Along with convergence to the Pareto-optimal set, it is also desired that a GA maintains a good spread of solutions in the obtained set of solutions. As an improvement of NSGA authors in (Deb et al., 2002) has developed NSGA-II which follows a fast non-dominated sorting procedure, an elitist strategy in selection, a parameter less approach in preserving population diversity and a simple yet efficient constraint-handling method.

3.2 K-Nearest Neighbours Algorithm

In pattern recognition, the KNN algorithm is a method for classifying objects based on closest training examples in the feature space. The KNN algorithm is amongst the simplest of all machine learning algorithms. It is an instance based learning method and has been used in many applications in areas such as data mining, statistical pattern recognition, image processing. Here each new instance is compared with existing ones using a distance metric and the closest existing instance is used to assign the class to the new one. This is called the nearest-neighbour classification method. If more than one nearest neighbour is used and the majority class of the closest k neighbours is assigned to the new instance then it is termed the k-nearest-neighbour method (Witten and Frank, 2005).

The KNN algorithm is comprises of the following steps (Kozma, 2008): 1) A positive integer k is specified, along with a test sample; 2) k training samples are selected from the database which are closest to the test sample; 3) The most common classification of these entries are found out and, 4) This is the classification given to the test sample.

The best choice of *k* depends upon the data; generally, larger values of *k* reduce the effect of noise on the classification, but make boundaries between classes less distinct. It is simple to implement and does not require tuning complex parameters to build a model. Since no training is involved in KNN algorithm new training examples can be added easily to the sample data-base. Although KNN algorithm is simple and effective, it is often slow. The obvious way to find which member of the training set is closest to an unknown test sample is to calculate the distance from every member of the training set and select the smallest. The time it takes to make a single prediction is proportional to the number of training samples.

3.3 Data Clustering

The goal of clustering is to assign data points with similar properties to the same groups and dissimilar data points to different groups.

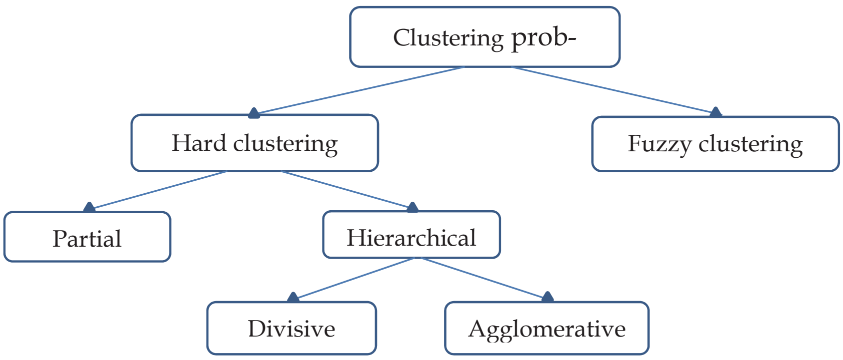


FIGURE 4 A block-diagram of clustering algorithms

As shown in Figure 4 clustering problems are mainly divided into two categories: hard clustering (or crisp clustering) and fuzzy clustering (or soft clustering). In hard clustering, data are divided into non-overlapping clusters, and in fuzzy clustering, a data sample may belong to two or more clusters with some probabilities. Hard clustering algorithms are subdivided into hierarchical algorithms and partitioning algorithms. Partitioning algorithms create a one-level non-overlapping partitioning of the data samples. Hierarchical algorithms divide data samples into a sequence of nested partitions.

3.3.1 K-means Clustering

k-Means is a partitional clustering approach and it is one of the most widely used clustering algorithms (MacQueen, 1967). In conventional *k*-means algorithm, for a given initial *k* clusters, it allocates data samples to the nearest clusters and then repeatedly changes the membership of the clusters according to an error function until the error function does not change significantly or the membership of the clusters no longer changes (Gan and Wu, 2007). If *D* is a data set with *n* samples and $C_1, C_2, \ldots, C_k$ are initial *k* disjoint clusters of *D* then the error function is defined as:

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d(\boldsymbol{x}, \mu(C_i)) \tag{4}$$

where $\mu(C_i)$ is the centroid of cluster C_i . $d(\boldsymbol{x}, \mu(C_i))$ denotes the distance between a sample $\boldsymbol{x}$ and $\mu(C_i)$. A typical distance measure can be the Euclidean distance d_{euc}. If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two data samples in *d*-dimensional space, then Euclidean distance between them is defined as:

$$d_{euc}(\boldsymbol{x}, \boldsymbol{y}) = \left[\sum_{j=1}^{d} \left(x_j - y_j\right)^2\right]^{\frac{1}{2}} \tag{5}$$

where, x_j and y_j are the values of the *j* th attribute of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively.

The *k*-means algorithm can be divided into two phases: the initialization phase and the iteration phase. In the initialization phase, the algorithm randomly assigns the samples into *k* clusters. In the iteration phase, the algorithm computes the distance between each sample and each cluster and assigns the case to the nearest cluster.

The algorithm works as follows:

1) First *k* initial cluster centroids are chosen.

2) Sample-to-cluster-centroid distances of all observations to each centroid are computed.

3) Then there are two ways to proceed: i) Batch update - Each data sample is assigned to the cluster with the closest centroid; and ii) Online update – Samples are individually assigned to a different centroid if the reassignment decreases the sum of the within-cluster, sum-of-squares sample-to-cluster-centroid distances.

4) *k* new centroid locations are computed from the average of the data samples in each cluster.

5) Steps 2 through 4 are repeated until cluster assignments do not change, or the maximum number of iterations is reached.

The k-means algorithm is easy to implement and its time complexity is $\boldsymbol{O}(n)$. A major drawback of this algorithm is that it is sensitive to the selection of the initial partition and may converge to a local minimum if the initial partition is not properly chosen (Abonyi and Feil, 2007). Unfortunately, there is no general theoretical solution to find the optimal number of clusters for any given data-set; a simple approach is to compare the results of multiple runs with different k clusters and choose the best one (Dougherty, 2013).

3.3.2 Agglomerative Hierarchical Clustering

Hierarchical clustering produces a set of nested clusters organized as a hierarchical tree, (which can be visualized as a dendrogram) that records the sequences of merges or splits. In AHC initially there are n clusters, each containing a single data sample. The number of clusters is reduced by one at each step of the algorithm by amalgamating a most similar pair of existing clusters and associating a height with the newly-formed cluster (Arabie et al., 1996). A general agglomerative algorithm for hierarchical clustering can be described as follows (Murtagh, 1983):

1) All inter-cluster dissimilarities are computed.
2) New cluster is formed from two closest clusters.
3) Dissimilarities between new cluster and other clusters are redefined.
4) Step 2 is repeated until all objects are in one cluster.

Here we need to compute the distances between old clusters and a new cluster formed by two clusters. In (Jambu, 1978) author proposed a general recurrence formula that gives the distance between a cluster C_k and a cluster C formed by the fusion of clusters C_i and C_j , i.e., $C = C_i \cup C_j$. The formula is given by:

$$D(C_k,C_i \cup C_j) = a_iD(C_k,C_i) + a_jD(C_k,C_j) + \beta D(C_i,C_j) + \gamma\,|\,D(C_k,C_i) - D(C_k,C_j)\,| + \delta_i h(C_i) + \delta_j h(C_j) + \epsilon h(C_k) \qquad (6)$$

where $D(\cdot,\cdot)$ is a distance (e.g. Euclidean distance) between two clusters and $h(C_i)$ denotes the height in the dendrogram of C_i . Different choices of the parameters a_i, a_j, β, γ, δ_i, δ_j and ϵ in (6), give different clustering strategy (algorithm). Most hierarchical clustering algorithms are variants of the single-link (Sneath and Sokal, 1973), complete-link (King, 1967), and minimum-variance (Ward, 1963) algorithms. Among these the single-link (also known as nearest neighbour method, the minimum method, and the connectedness method) and complete-link algorithms are most popular; the former uses nearest neighbour distance and the later uses farthest neighbour distance to measure the dissimilarity between two groups (Abonyi and Feil, 2007). There are some drawbacks of AHC - (a) data points that have been incorrectly grouped at an early stage cannot be reallocated and (b) different similarity measures for measuring the similarity between clusters may lead to different results (Gan and Wu, 2007).

3.3.3 Fuzzy C-Means Based Clustering

In many real life cases data clusters are not completely disjoined and data could be classified as belonging to one cluster almost as well as to another. Hence the separation of the clusters becomes a fuzzy notion and such data sets can then be more accurately handled by fuzzy clustering methods. In fuzzy clustering, each data sample can belong to more than one cluster and associated with each sample is a set of membership weights representing the strength of the association between that sample and a particular cluster (Dougherty, 2013). The fuzzy C-means (FCM) is the most widely used fuzzy clustering technique. The FCM algorithm is an extension of the *k*-means algorithm for fuzzy clustering and was first introduced in (Dunn, 1973; Bezdek, 1974).

For a data set $D = \{x_1, x_2, \ldots, x_n\}$, the algorithm is based on iterative minimization of the following objective function:

$$J(U, V) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \|x_k - v_i\|^2 \tag{7}$$

where, $V = \{v_1, \ldots\ldots, v_c\}$ are cluster centers; $U = [u_{ik}]$ is a $c \times n$ matrix, where u_{ik} is the *i*th membership value of the *k*th sample x_k and the membership values satisfy the following conditions:

$$0 \le u_{ik} \le 1 \qquad i = 1, 2, \ldots, c; \quad k = 1, 2, \ldots, n \tag{8}$$

$$\sum_{i=1}^{c} u_{ik} = 1 \qquad k = 1, 2, \ldots, n \tag{9}$$

$$0 < \sum_{k=1}^{n} u_{ik} < n \qquad i = 1, 2, \ldots, c; \tag{10}$$

$m \in (1, \propto)$ is an exponent weight factor.

The objective function is the sum of the squared Euclidean distances between each input sample and its corresponding cluster center, with the distances weighted by the fuzzy memberships.

In iterative steps of the algorithm cluster centers and membership weights are changed which are given by:

$$v_i = \frac{1}{\sum_{k=1}^{n} u_{ik}^{m}} \sum_{k=1}^{n} u_{ik}^{m} x_{ik} \qquad i = 1, 2, \ldots, c \tag{11}$$

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|x_k - v_i\|}{\|x_k - v_j\|} \right)^{\frac{2}{m-1}}} \qquad i = 1, 2, \ldots, c; \quad k = 1, 2, \ldots, n \tag{12}$$

For calculation of a cluster center all data samples are considered and the contributions of the samples are weighted by the corresponding membership values. For each sample its membership value in each cluster depends on its distance to the corresponding cluster center. The weight factor *m* controls the "fuzziness" of the resulting clusters (Bezdek, 1981). The larger the value of *m*, the smaller the influence of samples with small membership values.

The FCM clustering procedure consists of the following steps:

1) First the number of cluster centers and exponent weight *m* value are chosen. Based on an approximation cluster membership weights $U^{(0)}$ are initialized. Then Initial center centers $V^{(0)}$ are calculated and iteration counter is set to $\alpha = 1$.

2) Cluster centers are updated: given $U^{(\alpha)}$, $V^{(\alpha)}$ are calculated using equation (11).

3) Cluster membership values are updated: given $V^{(\alpha)}$, $U^{(\alpha)}$ are calculated using equation (12).

4) Iteration is stopped if $\max\left|u_{ik}^{(\alpha)} - u_{ik}^{(\alpha-1)}\right| \leq \epsilon_L$ else iteration counter is set to $\alpha = \alpha + 1$ and returned to step 2, where ϵ_L is a pre-specified number representing the smallest acceptable change in U.

3.3.4 Self-Organizing Map (SOM) Based Clustering

A SOM is a single-layer, unsupervised competitive neural network which produces a similarity graph of input data by converting the nonlinear statistical relationships between high-dimensional data into simple geometric relationships of their image points on a low-dimensional display, usually a regular two-dimensional grid of nodes (Kohonen, 2001). SOMs are particularly useful for visualization and cluster analysis as they can explore the groupings and relations within high-dimensional data.

It proceeds first by *initializing* the synaptic weights in the network by assigning them small values picked from a random-number generator. Then the following three steps are repeated until formation of the feature map has been completed:

1) *Competition* - For each input sample pattern, the neurons in the network compute their respective values of a discriminant function. The particular neuron with the largest value of discriminant function is the winner of the competition.

2) *Cooperation* - The winning neuron determines the spatial location of a topological neighbourhood of excited neurons, thereby providing the basis for cooperation among such neighbouring neurons.

3) *Synaptic Adaptation* - This last step enables the excited neurons to increase their individual values of the discriminant function in relation to the input sample pattern through suitable adjustments applied to their synaptic weights. The adjustments enhance the response of the winning neuron to the subsequent application of a similar input sample pattern.

The SOM algorithm goes through the following iterative steps (Kohonen, 1990):

1) *Initialization:* Random values for the initial weight vectors $w_j(0)$ are chosen; here $w_j(0)$ values are different for *j= 1, 2, ..., l*, where *l* is the number of neurons in the lattice. Another way of initializing the algorithm is to select the weight vectors from the available set of input vectors in a random manner; here the advantage is that the initial map will be in the range of the final map.

2) *Sampling:* A sample $\boldsymbol{x}$ is drawn from the input space with a certain probability; the vector $\boldsymbol{x}$ represents the activation pattern that is applied to the SOM lattice.

3) *Similarity matching:* Best-matching neuron, $i(\boldsymbol{x})$ is computed at time-step n by using a minimum-distance criterion-

$$i(\boldsymbol{x}) = \arg min_j \left\|\boldsymbol{x}(n) - \boldsymbol{w}_j\right\|, \qquad j = 1,2, \dots, l \tag{13}$$

4) *Updating:* The synaptic-weight vectors of all excited neurons are adjusted by using the following update formula-

$$\boldsymbol{w}_j(n+1) = \boldsymbol{w}_j(n) + \eta(n) h_{j,i(\boldsymbol{x})}(n)(\boldsymbol{x}(n) - \boldsymbol{w}_j(n) \tag{14}$$

where, $\eta(n)$ is the learning-rate parameter and $h_{j,i(\boldsymbol{x})}(n)$ is the neighbourhood function cantered around the winning neuron $i(\boldsymbol{x})$; both $\eta(n)$ and $h_{j,i(\boldsymbol{x})}(n)$ are varied dynamically during the learning phase.

5) *Continuation:* Until no noticeable changes in the feature map are observed algorithm goes back to step 2).

4 RESEARCH CONTRIBUTION

The main goal of this research is to develop and analyze a UE-assisted network-based RF fingerprint positioning framework for outdoor UE localization using MDT and GMDT data of LTE cellular networks. The utilization of MDT data has addressed two major drawbacks of RFFP based UE positioning: (1) it has minimized the huge cost and effort associated with the drive test based data collection procedure, and (2) allowed network operators to update and maintain 'radio map' fingerprint data autonomously using subscriber UEs. Here we have used two different data sets: 1) MDT data- in the beginning we have used a state-of-the-art LTE Rel'10 dynamic system simulator to generate outdoor MDT data following the 3GPP specifications defining signal bandwidth, center frequency, signal propagation, slow fading and fast fading parameterization (3GPP, 2006b). This data contains LTE serving and neighboring BS signal strength values together with UE positioning information from homogeneous and heterogeneous LTE networks. 2) GMDT data: during the second stage we have collected real life LTE and WLAN signal strength values along with corresponding GNSS positions using a popular data recording tool-Nemo handy installed in consumer mobile handsets.

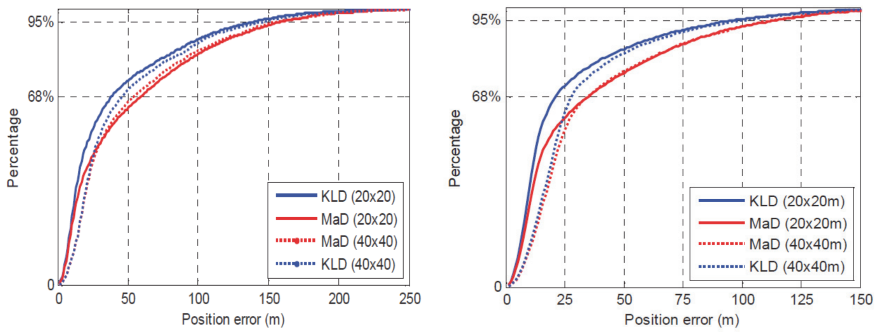
4.1 MDT Based Outdoor UE Localization Framework

The research question to be addressed here is:

Q1. To develop a framework for GCL-based outdoor RFFP using MDT data and to analyse the UE positioning performances in homogeneous and heterogeneous LTE networks in various user environments.

In [PI] we have developed a framework for G-RFFP using MDT data from LTE networks in rural, urban -regular and urban-heterogeneous configuration. For the rural and urban–regular network scenario all cellular BSs were on the same frequency layer and RF fingerprints consisted of measurements from serving BS and three neighboring BSs. While in the urban-heterogeneous LTE scenario macro eNBs and small BSs were deployed and measurements consisted of RSS values from four intra-frequency and three inter-frequency BSs. Here we were motivated to evaluate the performances of KLD and MD methods for UE positioning algorithms since in (Milioris et al., 2012) authors have used KLD for indoor WLAN positioning with good localization accuracy. We have assumed an ideal BS detection criterion in [PI] where every MDT report contains measurements from four strongest BSs.

Results: Two different rectangular GCU size were selected: 20-by-20 meters and 40-by-40 meters and positioning results of KLD and MD were obtained using 10-fold cross-validation scheme, e.g. in each evaluation 90% of the UEs were used as training database and estimation was done for the remaining 10% of UEs. In rural scenario PE is high- for both MD and KLD methods the 68% and 95% of PE are around 260 and 530 meters respectively. Figure 5 shows the cumulative distribution function positioning errors of G-RFFP method in urban-regular and heterogeneous networks. It is found that best positioning accuracy was observed in heterogeneous network using KLD with 20-by-20 meters grid-cell layout- the PE values were 21 meters and 95 meters in the 68% and 95% of PE respectively which showed an improvement of 40% and 16% respectively over that of MD positioning results. It was found that when 40-by-40 meters GCL was used 68% and 95% of PEs increase for KLD method but remains approximately same for MD method. The overall result indicates that KLD based GRFF method can provide acceptable outdoor UE positioning accuracy when compared to the North American emergency positioning requirements (3GPP, 2013b).



a) UE positioning error in urban-regular intra-frequency LTE network

b) UE positioning error in heterogeneous intra-frequency LTE network

FIGURE 5 PE CDF plot of G-RFFP in cellular LTE networks

According to 3GPP specifications a cell can only be detected and measured if Ês/Iot (energy per symbol over interference plus noise) is above -6 dB and RSRP is above -127 dBm. In [PII] RF fingerprint positioning performance is evaluated taken into account the practical cell detection criteria. UE localization performance was compared between 20-by-20 meters and 40-by-40 meters GCLs using KLD method in heterogeneous small cell network scenario (Case0) and in sparse regular macro network scenario (Case4). In both cases, inter-frequency LTE deployment consists of two layers of eNBs which operate on adjacent carriers. In the regular macro scenario, the BS sites in the higher frequency layer are deployed in a coordinated manner whereas in the heterogeneous scenario a random uncoordinated deployment of small cells is employed.

Results: Table 1 shows G-RFFP positioning error CDF percentiles of KLD method with and without considering the Ês/Iot criteria.

TABLE 1 G-RFFP PE values using KLD method

Scenario	*Ês/Iot criteria*	*For 10-by-10 m Cell*		*For 40-by-40 m Cell*	
		68% PE	*95% PE*	*68% PE*	*95% PE*
Case 4 (1750m)	n/a	127 m (-37%)	281 m (-28%)	106 m (-40%)	259 m (-28%)
	-6 dB	203 m	395 m	177 m	362 m
Case 0 (500m)	n/a	22 m (-55%)	63 m (-49%)	29 m (-42%)	65 m (-50%)
	-6 dB	49 m	124 m	50 m	131 m

The positioning error increased when Ês/Iot criterion was taken into account; in the table the minus percentages shown in first brackets indicate how much PE would decrease if ideal Ês/Iot criterion was used. It is found that for sparse Case 4 deployment, larger GCL results in better localization performance whereas in Case 0 smaller GCL provides the best performance. It can be seen form Table 1 that when sparse deployment changes to dense deployment, both the 68% and 95% of PE decreases and the obtained results fulfills easily the North American emergency positioning requirements.

4.2 Improvement of Positioning Accuracy of G-RFFP Using OGL Approach

In order to improve UE positioning accuracy of G-RFFP method we proposed a novel RF fingerprint algorithm in [PIII] based on an OGL technique and by determining the estimated UE position using the two smallest KLD weighted values among the matched training signatures. OGL method increases the amount of training signatures and reduces the distances among the surrounding grid centers and thereby improves the positioning accuracy. However it increases the position output time needed as compared to that of the traditional SGL method. In OGL approach the whole area of interest is divided into multiple overlapped square grid layouts and all the training signatures generated from these two GCLs are put together to create the whole training signature database. During the testing phase, first some training signatures are selected based upon BS id match with the test signature and then the KLD distances are calculated between the test signature and those training signatures. In order to determine the estimated position of the test signature, two smallest KLD grids are chosen and depending upon the weighted value the estimation position moves along the line joining the center locations of the corresponding grid units.

Results: Here MDT data were generated from two different network scenarios: sparse regular macro (RM) and dense HSC network. The simulator took into account the absolute and relative measurement errors and -6 dB Es/Iot cell detection criterion.

TABLE 2 PE results of SGL and OGL based G-RFFP using KLD Method

Scenario	*Training Data (%)*	*RF Fingerprint Algorithm*	***For 10-by-10 m Grid***		***For 40-by-40 m Grid***	
			68% PE (m)	***95% PE (m)***	***68% PE (m)***	***95% PE (m)***
RM (ISD 1750 M)	90%	SGL Based	29.73	165.29	43.53	196.30
		OGL Based	31.41 (+5.6%)	147.49 (-10.7%)	40.86 (-6.1%)	161.75 (-17.6%)
	10%	SGL Based	72.00	228.80	72.48	225.45
		OGL Based	63.96 (-11.1%)	206.05 (-9.9%)	65.03 (-10.2%)	203.70 (-9.6%)
HSC (ISD 500 M)	90%	SGL Based	21.12	58.08	33.73	76.43
		OGL Based	19.45 (-7.9%)	50.94 (-12.2%)	27.57 (-18.2%)	64.87 (-15.1%)
	10%	SGL Based	27.23	73.61	34.83	80.86
		OGL Based	25.14 (-7.6%)	66.47 (-9.6%)	28.28 (-18.8%)	68.71 (-15.0%)

To study how training data amount affects the performance of OGL method two different sets of training and test data sets were used. The first set contained 90% of the MDT data samples for training and the estimation was done for the remaining 10% MDT samples while the second set used 10% samples as training data and the rest were used for testing. Table 2 shows the PE results of KLD based G-RFFP method in SGL and OGL approaches. Here it was found that the number of training signatures used in OGL method is about twice of that in SGL method for every simulation cases. The percentage within the first bracket under the PE value of each of the OGL based method indicates the degradation (with a + sign) or improvement (with a - sign) in positioning as compared to the SGL method. It can be seen that in HSC for 90% training case and for 10-by-10 meter GCL size OGL method has shown 18.2% and 15.1% of improvement in 68% and 95% of PE respectively over that of the SGL method. The OGL method has constructed more training signatures with different combinations of BS IDs than the SGL method and hence it has analyzed more test samples than that of conventional SGL RFFP approach.

4.3 Improvement of Positioning Accuracy of G-RFFP Using Genetic Algorithm

The RF fingerprinting results of [PI] and [PIII] indicate that using a common GCL in different network scenarios and with varying amount of training data does not provide the best positioning accuracy in both 68% and 95% of PE values. It was also found that for different square Grid-cell layouts the 68% and 95% of PE results tend to conflict each other. Hence in [PIV] we have proposed a MOGA optimized RF fingerprint positioning (GAFP) method to make necessary adjustments to the GCL during training data processing phase. Thus using MDT data cellular operators would be able to build a large RF fingerprint training database from UE's in different network environments and update it as needed and then could utilize MOGAFP method for autonomous calibration of GCL with the updated MDT training database to deliver optimal positioning performance in both 68% and 95% of PE despite changes in the BS positions, surrounding structures or amount of MDT training samples. Since MOGA operates over a set of potential solutions it is particularly well-suited to solve multi-objective problems over other stochastic search strategies (e.g., simulated annealing, ant colony optimization or particle swarm optimization). In this work we have used the 'gamultiobj' function of Matlab R2014a which uses a controlled elitist genetic algorithm, a variant of NSGA-II (Deb, 2001).

Results: We have used two techniques to optimize the GCL- non-overlapping GCL (NoGCL) and overlapping GCL (OGCL) approaches.

TABLE 3 RFFP results of G-RFFP, GAFP-NoGCL and GAFP-OGCL methods

GAFP Type	*Positioning Error 68%-ile (m)*				*Positioning Error 95%-ile (m)*				*Analyzed Test MDT Samples (%)*			
	G-RFFP			*GAFP*	*G-RFFP*			*GAFP*	*G-RFFP*			*GAFP*
	40m	*30m*	*20m*		*40m*	*30m*	*20m*		*40m*	*30m*	*20m*	
NoGCL	48.25	47.58	44.03	24.52	127.18	152.70	135.96	55.01	63.07	62.40	64.53	65.32
OGCL	48.25	47.58	44.03	26.37	127.18	152.70	135.96	58.13	63.07	62.40	64.53	71.77

Table 3 shows results of GAFP based NoGCL, OGCL and conventional G-RFFP when applied to area '*A*'. Here 50% data was used for training; 25% data was for validation and rest of 25% of data used for testing. In both NoGCL and OGCL methods the GA population size was 100 and number of GA generations was 200. Here we can find that the with G-RFFP method the lowest PE values in 68% and 95% of PE are 44.03 and 127.18 meters whereas for the GAFP NoGCL approach the reduced PE values are 24.52 and 55.01 meters respectively. Hence GAFP was able to improve positioning accuracy of 44.31% and 56.74% in the 68% and 95% of PE as compared to the conventional grid-based RF fingerprinting. It is also found from Table 3 that positioning accuracies shown by GAFP OGCL is bit less than the NoGCL but it has analyzed more test samples as compared to NoGCL method. The robustness of the GAFP method was evaluated by using it in two different areas- '*A*' and '*B*' of 300m-by-300m. The proposed method was found to be consistent in improving UE positioning accuracy in each of these areas.

4.4 G-RFFP Performance Evaluation Using GMDT Data

Due to the recent demand of high data traffic it seems that next generation cellular networks will consist of various 3GPP and non-3GPP access techniques. 3GPP has studied the potential of RAN level solutions for enhancing the interworking between WLAN and LTE in Release 12 (3GPP, 2013). The autonomous collection of the coverage and quality information from both cellular and WLAN access networks should be supported in next generation interworking deployments for ensuring the cost efficient operation. In [PV] an enhancement to the 3GPP LTE MDT framework which we referred as GMDT has been proposed to enable cellular operators for the collection of location-aware radio measurements from WLAN access networks. The GMDT solution allows autonomous mapping of network coverage and improves RF fingerprint localization in dynamic LTE/WLAN deployment scenarios in indoor and outdoor environments. The GMDT architecture is a potential solution for two important issues: 1) by correlating WLAN radio measurements with the 3GPP radio measurements and UE location information, cellular operators can get a complete view of its heterogeneous network; 2) RF fingerprinting in LTE/WLAN deployment scenarios can be improved as compared to the case that uses only LTE RSS.

TABLE 4 G-RFFP PE using GMDT with different number of APs

Measurement Type	*Max. number of measured WLAN IDs*	*68 %-ile PE [m]*	*95 %-ile PE [m]*
WLAN + LTE (1800 MHz)	All WLAN APs	22.42	47.78
	23	22.51	48.27
	18	22.65	48.10
	13	23.14	48.21
	10	23.32	48.60
	5	26.80	57.57
	2	35.43	80.52
	1	49.41	127.67
LTE (1800 MHz)	LTE BSs only	88.29	235.45

Results: Here we have performed G-RFFP of outdoor UEs using partial fingerprint matching technique. An RF fingerprint consists of both LTE cell IDs and WLAN AP IDs - they are treated equally in the localization algorithm. GMDT contributes to an increased amount of measurement report signaling as well as to higher memory requirements for UEs.

In [PVI] we have used GMDT data that comprises of RSS field measurements from LTE network operating on 800 MHz, 1800 MHz and 2600 MHz frequency bands. Here G-RFFP was used in which per GCU a single training signature was formed from all the samples available in a particular GCU, since it is expected that the measurements recorded geographically close to each other will have similar signal characteristic and hence they can be combined to form one signature. This reduces the computational complexity by compressing the size of the search space i.e., total number of training signatures. It also reduces the amount of memory needed to store the training data. During the testing phase those training signatures that fulfill the partial matching condition are selected for distance calculation from the test signature. The best matching training signature is the one which minimize the Euclidean distance metric between the training and test signatures.

Results: Here we have gone through a 10-fold cross validation procedure - training database contains 90% chunks and remaining 10% of chunks are tested to test every data sample in turn. RF fingerprinting results were obtained using 20-by-20 meter and 40-by-40 meter of GCL sizes. Results depicts that UE positioning accuracy was worst when using only LTE 800 MHz measurements. Localization performance improved significantly when all available LTE bands were measured which increased the average dimensionality of LTE RF fingerprints. It was found from the study that both the density and the dimensionality of the fingerprints have bigger impact on UE positioning accuracy than grid density. When larger GCL was used training signature dimensionality increased and more test samples were analyzed by meeting partial fingerprinting condition. Whereas when GCL size was reduced from 40 meters to 10 meters, 68% and 95% of PE values decreased by 41% and 18% for LTE+WLAN case respectively; however percentage of discarded test samples increased from 3% to 14%.

4.5 C-RFFP Performance Evaluation Using GMDT Data

We have proposed a novel cluster-based RF fingerprinting technique in [PVII] and [PVIII] for outdoor UE positioning using GMDT data.

It has addressed the following four main challenges of RFFP (Kushki et al., 2007):

1) RF fingerprint generation and update: Here real life GMDT data were used which automates the training data collection and updating procedure. It also reduces the cost and time associated with traditional war driving data collection procedure. Since GMDT data contains LTE as well as WLAN RSS values it increases UE positioning accuracy when compared to that of using only MDT data.

2) Preprocessing of recorded training data: The C-RFFP method does not need the time consuming training signature formation as compared to the G-RFFP method. The only preprocessing needed here is to group the recorded GMDT data according to the LTE serving cell-ID. This reduces the search space for the C-RFFP method during UE localization phase and thereby enables it to estimate test UE position in short time with less computational cost.

3) Selection of BS and APs for use in positioning: Seven LTE BS and WLAN AP signals sorted in descending order of signal strength values were selected for position estimation which were found to be effective from previous RFFP research results (Youssef, 2003),(Laitinen et al., 2012). We have analyzed the UE positioning accuracy with three different GMDT data sets: the 1st set contains three LTE and four WLAN signals, 2nd set consists of two LTE and five WLAN signals and 3rd set comprises of one LTE and six WLAN signals.

4) UE position estimation method: In [PVII] effectiveness of CFFP method was compared to the conventional G-RFFP method. In [PVIII] we have evaluated performances of different cluster-based RF fingerprinting (C-RFFP) methods: K-Nearest Neighbor (KNN), Agglomerative Hierarchical Clustering (AHC) and Fuzzy C-Means (FCM) to that of a single GCL based RF fingerprint positioning method.

A brief explanation of the C-RFFP technique is given by the following block diagram:

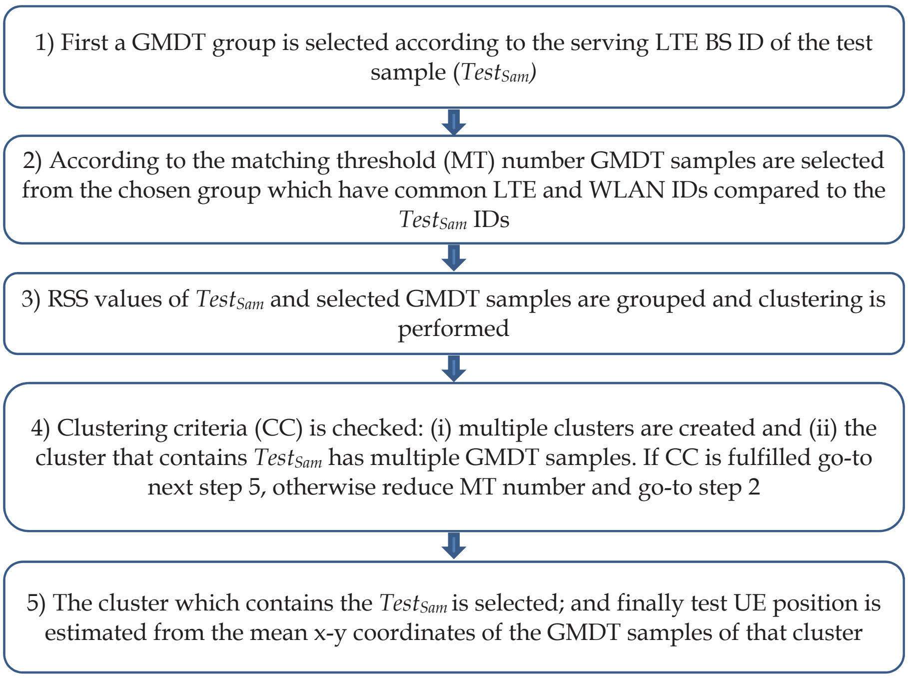


FIGURE 6 A block-diagram of C-RFFP Positioning Method

If CC is not fulfilled after reaching the minimum MT number then the $Test_{Sam}$ is considered as a non-analyzed sample.
Results: Here in the training GMDT database multiple samples which contain similar LTE BS ID and WLAN AP ID and recorded from the same x-y coordinate were merged into a single one. Training and test data-sets were created by randomly choosing chunks of 20 consecutive GMDTs samples. The G-RFFP method used a 10m-by-10m GCL which was chosen from several square grid-cell layouts according to the best delivered positioning accuracy. Ten-fold cross-validation method was used to obtain positioning results and Euclidean distance was used to measure the statistical difference between training fingerprints and test samples. Test results of G-RFFP and C-RFFP methods obtained in [PVII] with three different GMDT data-sets are shown in Table 5.

TABLE 5 G-RFFP and C-RFFP positioning results using GMDT data

LTE BS and WLAN AP	***Matchi-ng Thresho-ld***	***Using Common Test GMDTs between G-RFFP and C-RFFP***				
		G-RFFP		***C-RFFP***		***Common Test GMDT (%)***
		68% PE(m)	***95% PE(m)***	***68% PE(m)***	***95% PE(m)***	
LTE BS:3 & WLAN AP:4	80%	14.21	39.77	9.40	33.75	21.15
	60%	14.02	40.35	10.40	35.85	43.59
	40%	15.59	42.57	14.16	51.17	61.77
LTE BS:2 & WLAN AP:5	80%	11.98	36.04	7.61	26.45	21.83
	60%	11.81	36.03	7.95	27.60	46.02
	40%	12.66	38.32	9.14	33.88	62.37
LTE BS:1 & WLAN AP:6	80%	11.47	33.69	7.36	22.81	24.38
	60%	10.60	31.76	7.24	20.58	51.32
	40%	11.17	33.89	7.80	24.34	69.21

It can be seen that when 3rd GMDT data-set was used C-RFFP showed the best positioning accuracy for 80% of threshold matching: 35% improvement in PE for both the 68% and 95% of PE as compared to that of the G-RFFP. Results depicts that C-RFFP offers better positioning than G-RFFP when matching threshold is high and five or six WLAN signals are used.

RFFP results of [PVIII] are shown in Table 6 which was obtained from ten-fold cross-validation procedure for all four methods using common analyzed test GMDT samples. It can be seen that with 1st data-set PE results of G-RFFP and KNN are very similar whereas AHC and FCM have shown better positioning for MT - 6. For data set 2, KNN has reduced PE values as compared to that of GCL. AHC and FCM methods have shown further improvement in both 68th and 95th percentiles of PE values than that of G-RFFP and KNN methods. When data set 3 was used, AHC based clustering has given the best positioning accuracy as compared to the other methods; for MT-3 AHC based C-RFFP method has shown a positioning improvement of 42.3% and 39.8 % in the 68% and 95% of PE as compared to that of the G-RFFP method.

TABLE 6 Comparison of RFFP results between G-RFFP and C-RFFP methods

LTE & WLAN	*MT. No.*	*G-RFFP*		*KNN*		*AHC*		*FCM*		*Com-mo-n Test GMDT (%)*
		68% PE (m)	*95% PE (m)*	*68% PE (m)*	*95% PE (m)*	*68% PE (m)*	*95% PE (m)*	*68% PE (m)*	*95% PE (m)*	
LTE: 3 & WLAN:4	6	14.20	39.70	13.23	39.68	9.35	33.78	9.84	34.63	18.71
	5	14.06	40.22	13.30	40.43	10.50	35.73	12.01	40.47	41.02
	4	15.27	41.99	13.74	41.51	12.61	42.32	13.55	44.03	55.92
	3	15.58	42.57	14.07	42.12	14.16	51.18	14.35	47.18	61.75
LTE: 2 & WLAN:5	6	14.58	40.63	10.61	33.99	7.48	25.69	7.52	26.51	20.90
	5	13.58	40.94	11.08	37.54	7.87	26.98	8.94	33.96	45.41
	4	13.87	41.74	11.40	37.94	8.66	31.34	9.77	36.05	58.22
	3	14.22	41.86	11.73	38.68	9.13	33.87	10.25	37.68	62.33
LTE: 1 & WLAN:6	6	14.06	38.24	9.32	31.39	7.27	21.80	7.35	25.53	22.06
	5	13.14	38.00	9.73	32.78	7.18	20.16	8.08	27.71	49.56
	4	13.19	39.30	10.12	36.08	7.49	22.44	8.93	32.53	64.76
	3	13.48	40.05	10.46	36.54	7.77	24.09	9.22	33.34	69.11

These results indicate that in dense urban areas where multiple WLAN signals can be detected C-RFFP is capable of delivering better outdoor UE positioning than G-RFFP using data set two or three having seven LTE and WLAN signals strength values.

4.6 Robustness and Time Complexity Analysis of C-RFFP Method Using GMDT Data

In [PIX] we have evaluated the performances of four cluster-based RFFP methods using GMDT data when facing the following challenging issues:

1) Time difference between recorded training database samples and test samples: the characteristics of RF signal propagation and absorption by surrounding structures and human bodies causes variations in measured RSS values recorded from a fixed location; even changes in the environmental conditions, such as temperature or humidity, affect the signals to a large extent. Here we have used seven months old training data to estimate test UE positions and compare results with conventional methods.

2) Variation in recording device: under same wireless conditions measured RSS values from WLAN APs may differ significantly with the variation in

UE's hardware (Hossain et al., 2007; Kjærgaard et al., 2008). Two different devices were used to record training and test GMDT samples.

3) Variation in number of APs used for UE positioning: since considering all available APs for position estimation increases the computational complexity of the positioning algorithm we have used two data sets- first set $S_{1,n}$ comprises of serving LTE RSRP and six WLAN RSSI values and the second set $S_{2,n}$ contains serving LTE RSRP and ten WLAN RSSI values (Kushki et al., 2007; Laitinen et al., 2012).

We have also determined the position estimation time of different C-RFFP methods. Here we compare the positioning accuracies of the C-RFFP methods themselves and between conventional G-RFFP and KNN based RFFP methods.

Results: During training data collection process 'Maximum RSS' based AP selection methodology were followed where WLAN APs are sorted in descending order based on their RSS values.

TABLE 7 PE results of ExStudy-2 using GMDT dataset $S_{1,n}$ and $S_{2,n}$

D. S.	*MT*	*G-RFFP*			*KNN*			*K-means*			*AHC*			*FCM*		
		68% PE (m)	*95% PE (m)*	*Ana. Sam. (%)*	*68% PE (m)*	*95% PE (m)*	*Ana. Sam. (%)*	*68% PE (m)*	*95% PE (m)*	*Ana. Sam. (%)*	*68% PE (m)*	*95% PE (m)*	*Ana. Sam. (%)*	*68% PE (m)*	*95% PE (m)*	*Ana. Sam. (%)*
$S_{1,n}$	4	26.6	47.0	80.3	25.8	47.3	69.7	24.1	47.2	66.8	20.8	39.8	26.8	24.5	43.4	41.8
	3	27.2	49.2	96.5	27.0	53.4	94.6	25.2	51.2	93.7	21.5	41.7	60.5	25.5	50.2	78.1
	2	27.9	51.1	99.7	27.5	55.6	99.4	25.5	55.4	99.3	22.4	43.2	76.6	26.7	53.3	95.7
$S_{2,n}$	5	25.7	46.1	57.0	24.7	44.3	89.6	23.8	41.6	86.9	20.8	38.7	43.1	23.5	42.2	62.4
	4	26.7	47.6	85.5	25.8	46.8	97.5	24.5	42.8	96.6	22.0	42.0	67.0	24.3	43.0	87.4
	3	27.6	49.5	96.8	26.0	47.5	98.9	24.7	44.1	98.8	23.0	43.7	78.4	24.8	44.1	97.5
	2	28.1	50.8	99.7	26.2	49.3	99.9	24.9	46.2	99.9	23.4	45.2	82.9	25.2	46.4	99.4

Here two experimental studies (ExStudy-1 and ExStudy-2) were carried out: in ExStudy-1 both training and test samples were selected from the same time period whereas in ExStudy-2 we used 32791 training GMDTs of September 2014 and 3574 test samples ($Test_{Sam}$s) of May 2015. Each of the selected $Test_{Sam}$ is surround by more than ten training GMDTs within its three meter circular radius area to ensure the presence of sufficient number of training samples in its vicinity. Training and test data-sets comprises of randomly choosing data chunks of 20 sequentially recorded samples. Results of ExStudy-2 using G-RFFP, KNN and C-RFFP methods are shown in Table 7 and Table 8. . It was found that AHC based RFFP has outperformed other positioning methods in both 68% and 95% of PE. For MT-2 all the methods have analyze maximum amount of $Test_{Sam}$s when compared to the other MT values. For MT-2 using $S_{1,n}$

AHC based C-RFFP improved positioning accuracy by 19.71% and 15.46% in 68% and 95% of PE respectively over that of G-RFFP method. It was noticeable that when $S_{2,n}$ was used all the C-RFFP methods have analysed more test samples for different MT numbers as compared to that of $S_{1,n}$. Also the deviations of 68% and 95% of PE results between different C-RFFP methods became less obvious when $S_{2,n}$ was used.

TABLE 8 PE results of ExStudy-2 using SOM

Method	***SOM***						
Data Set	$S_{1,n}$			$S_{2,n}$			
Matching Threshold	4	3	2	5	4	3	2
68% PE (m)	22.06	23.05	25.27	24.78	23.83	24.53	24.81
95% PE (m)	34.84	39.93	45.70	42.42	41.95	44.27	45.23
Analysed Samples (%)	2.96	15.44	39.22	4.92	15.52	31.00	48.57

In this study we determined the average computation time taken by different methods. Here only G-RFFP method needs training time. It is also found that when data-set $S_{2,n}$ was used positioning time increases for all the methods as compared to that of $S_{1,n}$ due to the increase of data dimension in $S_{2,n}$. For data-set $S_{1,n}$ AHC based C-RFFP has given best positioning accuracy taking shortest time as compared to other methods.

From both the experimental studies we found that C-RFFP methods perform better than G-RFFP and KNN methods even using old training database and under recording device variation.

5 CONCLUSIONS

The advancements in smartphone technologies and popularization of it has enabled us to use various wireless positioning technologies which include GNSS, methods to exploit signal-of-opportunities, such as cellular LTE/4G, WiFi, RFID, Bluetooth, infrared (IR), magnetic field, camera, and various other hybrid solutions. Although a seamless positioning system for both outdoor and indoor environment is yet to be developed; RF fingerprint based systems have gained good popularity and commercial use in the recent years. One of the core challenges of RFFP based positioning method is the collection and updating of the training database. The recently proposed MDT functionality has opened a new opportunity for cellular operators to utilize RFFP methods for offering a cheap and ubiquitous positioning solution to their subscribers.

The present research work was conducted in two parts: In the beginning we proposed a framework for grid-based RFFP in LTE network using simulated MDT data. Methods were proposed to improve G-RFFP positioning accuracy and autonomous calibration of GCL. In the second part we have used real life GMDT data consisting of LTE and WLAN RSS values and compared the results of G-RFFP positioning using MDT and GMDT data. In here we have proposed cluster-based RFFP with vastly reduces the training phase data processing computational cost and also outperforms GRFF based UE positioning accuracy.

In publications [PI] and [PII] grid-based RF fingerprinting performance was evaluated using LTE MDT data in heterogeneous, regular urban and rural environments to study the effects of intra and inter-frequency RSS measurements in UE localization. Kullback-Leibler Divergence and Mahalanobis Distance based position estimation methods were used and two different GCLs were used over the area of interest. Simulation results indicate that GRFF method gives best performance in dense urban heterogeneous LTE network using KLD method. It was also found that when cell detection criterion was considered less number of LTE BSs were detected which consequently reduces UE positioning performance. In order to improve positioning accuracy an overlapping GCL approach was followed in [PIII] using KLD method in sparse regular macro and dense heterogeneous small cell network scenario with varying amount of training data and GCL size. This novel method used weighted geometric center from two nearest grid-cell units for estimating the test UE location. When compared to the conventional single grid-cell GRFF method the proposed method showed the best performance for 10% of training data using 40m-by-40m GCL in dense heterogeneous LTE network. In [PIV] we proposed a multi-objective Genetic Algorithm based G-RFFP method for autonomous calibration of GCL using training data to deliver better UE positioning than traditional G-RFFP method. Two different approaches were followed to optimize GCL size and robustness of the proposed method was evaluated by applying the method in two different areas of interest. The GAFP method showed 56.74% of improvement in 95% of PE value over the conventional G-RFFP method.

A generic measurement architecture referred to as GMDT for already existing MDT functionality was proposed in [PV] to correlate WLAN measurements with LTE RSS values. Here we have analyzed the RFFP performance improvements for the proposed concept with LTE and WLAN field measurements. Results showed that when 10 strongest WLAN RSS along with LTE RSS were used an improvement of 74% was achieved over that of using only LTE RSS. In [PVI] the importance of GMDT functionality was emphasized by showing the improvements of G-RFFP positioning performances over conventional approach. In this work we have used partial fingerprint matching technique for test sample positioning. Different GCLs were utilized and also two different LTE frequency bands- 800MHz (intra-frequency) and 1800MHz (intra- and inter-frequency) measurements were used. Experimental results depicts that RFFP using GMDT data and higher matching threshold gives much better positioning accuracy compared to that of LTE only measurements. Partial fingerprint matching became useful since when using 100% matching was done more than one-third test samples were discarded; and lowering the matching threshold enabled us to analyze more test samples but with a degradation of positioning accuracy.

We proposed a novel cluster-based RFFP technique for outdoor UE localization using GMDT data in [PVII] which does not go through training phase data pre-processing as compared to the traditional G-RFFP method. It utilizes AHC clustering and uses LTE cell-ID technique to reduce the search space for test GMDT positioning and also delivers better result over conventional single-grid G-RFFP method. During the evaluation of the proposed C-RFFP method we have used three different sets of GMDT data and different MT values. Through analyzing the common test sample estimation results it was found that using a data set comprising of only serving LTE BS and six WLAN AP RSS values the positioning result of C-RFFP method shows an improvement of 35% in the 68% of PE CDF values over that of the G-RFFP method. In [PVIII] four methods were used for UE positioning using three GMDT data-sets and four MT numbers. Analyzing the common analyzed test samples by the methods it was found that for a single LTE and six WLAN RSS training dataset using MT-3 AHC based C-RFFP method performs the best showing an improvement of 42.3% and 39.8% in the 68% and 95% values of PE over that of G-RFFP method.

The main goal of [PIX] was to analyze the robustness of the C-RFFP method when there are variations in the data recording device and there exists a time difference between training data collection and test data estimation. For this purpose four C-RFFP methods were used and results were compared to the G-RFFP and KNN based RFFP methods. Here two experiments were carried out: in the 1st experiment training and test data samples were selected from the same time period of recording while in the 2nd experiment training GMDT dataset was recorded in September 2014 and test data in May 2015. Each of the selected test GMDT was surrounded by ten training data samples in its three meter circular radius area in order to ensure good availability of training data samples near the test sample. Although there were degradations in positioning accuracies for every RFFP methods with variations in data-collection device and recording time but the estimated positioning results were acceptable; which shows that the C-RFFP methods are robust enough to output good UE positioning in such variations. When considering mostly analyzed test samples with variations in MT number it was found that using seven LTE and WLAN RSS and MT-2 ACH based C-RFFP provided the best UE positioning accuracy with an improvement of 15.46% and 22.30% in 95% of PE as compared to that of G-RFFP and KNN methods respectively. We have also evaluated the computational time complexity among the RFFP methods and found that AHC based RFFP also outputs UE positioning taking the shortest time - this result will vary depending mainly upon the number of training samples and data dimension.

Future work:

In our ongoing work we are trying to address the following issues:

1) In order to evaluate the performance of C-RFFP method as ubiquitous positioning technique our next goal is to use GMDT data which would be measured in outdoor as well as indoor environments.

2) There is big impact of training data density on UE positioning accuracy and percentage of analyzed test samples. Hence we can make a detail evaluation of it by analyzing every test GMDT sample and categorize them in different groups according to the number of training samples surrounding it.

3) The recording device variation should be addressed taking into account not only the mobile cellular phones but also tablets and laptops of various manufacturers.

4) In our previous work the assignment of initial number of clusters was not similar for the different C-RFFP methods, thus for a better evaluation of the cluster based RFFP methods the same initial K number could be used.

5) The amount of training samples affect the computational time needed to output a position estimation of different C-RFFP methods. In order to evaluate the real time applicability of C-RFFP methods it would be necessary to calculate the time complexity with different amount of training data.

YHTEENVETO (FINNISH SUMMARY)

Radiotaajuussormenjälkeen pohjautuvat paikannusmenetelmät

Viimeaikaiset edistysaskeleet langattoman tietoliikenteen saralla ovat avanneet kysynnän yhä tarkemmille ja luotettavimmille paikannusratkaisuille. Radiotaajuussormenjälkeen pohjautuva paikannus on yksi mielenkiintoisista ratkaisuista tällä saralla. Se käyttää olemassaolevien järjestelmien lähettämiä radiosignaaleita hyväkseen luodakseen kullekin paikalle ominaisen sormenjäljen, joka koostuu kyseisellä paikalla havaituista signaaleista ja niiden voimakkuuksista. Kyseiselle paikannustavalle yksi isoimmista haasteista on riittävän opetusinformaation kerääminen luotettavasti. Viimeaikaiset edistysaskeleet 3GPP stardardoinnissa edesauttavat tässä haasteessa, ja tämän nk. MDT-toiminnallisuuden myötä mobiilipäätelaitteet raportoivatkin sovitun käytännön mukaan havaitsemastaan radioympäristöstä. Tässä väitöskirjassa päätavoite on esittää tapoja, joilla tätä MDT toiminnallisuutta käytetään hyväksi radiotaajuuspaikannuksessa. Työssä käsitellään sekä pelkän solukkoverkkodatan hyväksikäyttöä paikannuksessa että langattomien lähiverkkojen mukanaantuomaa lisäarvoa paikannustarkkuuden parantamisessa. Työssä käytetään tekoälyyn ja koneoppismenetelmiin pohjautuvia ratkaisuja. Työn tuloksena esitetään uudenlaisia ja kustannustehokkaita tapoja paikannuksen, jonka tarkkuus yltää viranomaisten asettamiin vaatimuksiin hätätilapaikannuksen suhteen.

REFERENCES

3GPP 2006a. “Physical layer aspect for evolved Universal Terrestrial Radio Access (UTRA)”. TR 25.814, version 7.1.0.

3GPP 2006b. “Physical Layer Aspects for Evolved UTRA”. TR 36.814, version 7.1.0.

3GPP 2009a. “Evolved Universal Terrestrial Radio Access Network; Stage 2 functional specification of User Equipment (UE) positioning in E-UTRAN (Release 9)”. TS 36.305.

3GPP 2009b. “Study on minimization of drive-tests in next generation networks”. Release 9, Technical report, TR 36.805.

3GPP 2009c. “Further advancements for E-UTRA physical layer aspects”. TR 36.814, version 9.0.0.

3GPP 2011a. “Subscriber and Equipment trace; Trace Control and Configuration Management”. TS 32.422, Version 10.4.0.

3GPP 2011b. “Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Measurement Collection for Minimization of Drive Tests (MDT) Overall Description”. Stage 2, TS 37.320, v. 11.0.0.

3GPP 2012a. “Evolved Universal Terrestrial Radio Access; LTE Positioning Protocol (LPP)”. TS 36.355.

3GPP 2012b. “Universal terrestrial radio access (UTRA) and evolved universal terrestrial radio access (E-UTRA); radio measurement collection for minimization of drive tests (MDT) Stage 2 (Release 11), Technical report, TS 37.320 v11.1.0.

3GPP 2012c. “Evolved Universal Terrestrial Radio Access (E-UTRA); Study on minimization of drive-tests in next generation networks (Release 9)”. TR 36.805.

3GPP 2012d. “Radio Resource Control (RRC); Protocol Specification (Release 10)”. TS 36.331.

3GPP 2012e. “Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements”. TS 36.214, version 11.1.0.

3GPP 2013a. “Study on WLAN/3GPP radio interworking”. TR 37.834, v.1.0.0.

3GPP 2013b. “Requirements for support of radio resource management”. TS 36.133, version 11.4.0.

Aalto, L., Gothlin, N., Korhonen, J. and Ojala, T. 2004. “Bluetooth and WAP push based location-aware mobile advertising system”. In 2nd International Conference on Mobile Systems, Applications, and Services (MobiSYS ’04), ACM Press, 49–58.

Aarnaes, E. & Holm, S. 2004. “Tuning of empirical radio propagation models effect of location accuracy”. Wireless Personal Communications. Springer: 30(2-4), 267-281.

Abonyi, J. & Feil, B. 2007. “Cluster Analysis for Data Mining and System Identification”, Birkhauser Verlag AG, Berlin, Germany.

Ahonen, S. & Laitinen, H. 2003. “Database Correlation Method for UMTS Location”. In 57th IEEE Vehicular Technology Conference (VTC 2003 Spring), Jeju, South Korea.

Anthony, L., Chawathe, Y., Consolvo, S., Hightower, J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello G. & Schilit, B. 2005. “Place Lab: Device Positioning Using Radio Beacons in the Wild”. In Third International Conference on Pervasive Computing (PERVASIVE 2005). Munich, Germany, 116-133.

Arabie, P., Hubert, L. J. and Soete, G. D. 1996. “Clustering and Classification”. World Scientific Publishing Co. Pte. Ltd., River Edge, NJ, USA.

Bahl, P. & Padmanabhan, V. N. 2000. “RADAR: An In-Building RF-Based User Location and Tracking System”. In 9th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM 2000), Tel Aviv, Israel, 775-784.

Ben-Moshe, B., Elkin, E., Levi, H. & Weissman, A. 2011. “Improving Accuracy of GNSS Devices in Urban Canyons”. In Canadian Conference on Computational Geometry (CCCG'11). Toronto, Canada.

Bezdek, J. C. 1974. “Fuzzy Mathematics in Pattern Classification”. PhD thesis, Cornell University, Ithaca, NY, USA.

Bezdek, J. C. 1981. “Pattern Recognition with Fuzzy Objective Function Algorithms”. Kluwer Academic Publishers Norwell, MA, USA.

Bolc, L. & Cytowski, J. 1992. “Search Methods for Artificial Intelligence”. Academic Press, London, UK.

Booker, L. B. 1987 “Improving search in genetic algorithms”. In Davis, L. “Genetic Algorithms and Simulated Annealing”, Morgan Kaufmann Publishers Inc., Los Altos, California, USA, 61-73.

Borkowski, J. & Lempiaeinen, 2005. “Pilot correlation positioning method for urban UMTS networks”. In 11th European Wireless Conference 2005 - Next Generation Wireless and Mobile Communications and Services (European Wireless), Nicosia, Cyprus.

Borkowski, J. & Lempiainen, U. 2006. “Practical Network-Based Techniques for Mobile Positioning in UMTS”. In EURASIP Journal on Advances in Signal Processing, Hindawi Publishing Corporation, Article ID 12930, 1–15.

Brueninghaus, K., Astely, D., Salzer, T., Visuri, S., Alexiou, A., Kanger, S., & Seraji, G.-A. 2005. “Link performance models for system level simulations of broadband radio access systems”. In IEEE Personal, Indoor and Mobile Radio Communications (PIMRC 2005) conference. Berlin, Germany.

Campos, R. S. & Lovisolo, L. 2010. “Mobile Station Location using Genetic Algorithm Optimized Radio Frequency Fingerprinting”. In The 7th International Telecommunications Symposium (ITS 2010), Saarbrucken, Germany.

Campos, R. S. & Lovisolo, L. 2013. “Genetic Algorithm Optimized DCM Positioning”. In 10th Workshop on Positioning Navigation and Communication (WPNC), Dresden, Germany.

Campos, R. S. & Lovisolo, L. 2015. “RF Positioning: Fundamentals, Applications and Tools”. Boston, USA: Artech House.

Chan, E. C. L., Baciu, G. & Mak, S. C. 2008. “Wireless Tracking Analysis in Location Fingerprinting”. In IEEE International Conference on Wireless & Mobile Computing, Networking & Communication (WIMOB 2008). Avignon, France, 214-220.

Charitanetra, S. & Noppanakeepong, S. 2003. “Mobile Positioning Location Using E-OTD Method for GSM Network”. In Student Conference on Research and Development (SCORED 2003), Putrajaya, Malaysia, 319-324.

Chen, M. Y., Sohn, T., Chmelev, D., Haehnel, D., Hightower, J., Hughes, J., LaMarca, A., PotterIan, F., Smith, I., and Varshavsky, A. 2006. “Practical Metropolitan-Scale Positioning for GSM Phones”. In International Conference on Ubiquitous Computing (UbiComp 2006), Springer-Verlag Berlin Heidelberg, Germany, 225-242.

Chen, R. 2012. “Ubiquotous Positioning and Mobile Location-Based Services in Smart Phones”. Hershey PA, USA: IGI Global.

Cheng, Y., Chawathe, Y., LaMarca, A. and Krumm, J. 2005. “Accuracy Characterization for Metropolitan-scale Wi-Fi Localization”. In the 3rd International conference on Mobile systems, applications, and services (MobiSys 2005), Seattle, Washington, USA, 233-245.

Coley D. A. 1999. “An Introduction to Genetic Algorithms for Scientists and Engineers”. World Scientific Publishing Co. Pte. Ltd., London, UK.

Constandache, I., Gaonkar, S., Saylar, M., Choudhury, R. R. and Cox, L. 2009. “EnLoc Energy-Efficient Localization for Mobile Phones”. In 28th Conference on Computer Communications (IEEE INFOCOM 2009), Rio de Janeiro, Brazil, 2716-2720.

Dardari, D., Luise, M. & Falletti, E. 2012. “Satellite and terrestrial Radio Positioning Techniques: A signal processing perspective”. Waltham, MA, USA: Elsevier.

Deb, K. 2001. “Multiobjective Optimization Using Evolutionary Algorithms”. John Wiley & Sons, Ltd, Chichester, England.

Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T. 2002. “A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II”. In IEEE Transactions on Evolutionary Computation, 6 (2), 182-197.

Deng, P. & Fan, P. Z. 2000. “An AoA Assisted ToA Positioning System”. In International Conference on Communication Technology (WCC-ICCT 2000), Beijing, China.

Dougherty, G. 2013. “Pattern Recognition and Classification: An Introduction”. Springer Science+Business Media, New York, USA.

Dunn, J. C. 1973. “A fuzzy relative of the ISODATA process and its use in detecting compact well-separated cluster”. In Journal of Cybernetics, 3(3), 32-57.

Gan, G., Ma, C. and Wu, J. 2007. “Data Clustering: Theory, Algorithms, and Applications”. ASA-SIAM series on statistics and applied probability, USA.

Gaonkar, S., Li, J., Choudhury, R. R., Cox, L. and Schmidt, A. 2008. “Microblog: Sharing and querying content through mobile phones and social participation”. In the 6th international conference on Mobile systems, applications, and services (MobiSys '08), Breckenridge, CO, USA, 174-186.

Gen, M. & Cheng, R. 2000. “Genetic Algorithms and Engineering Optimization”. John Wiley & Sons, New York, USA.

Gentile, C., Alsindi, N., Raulefs, R. and Teolis, C. 2013. “Geolocation Techniques Principles and Applications”. Springer Science+Business Media, New York, USA.

Goldberg, D. E. 1989. “Genetic Algorithms in Search, Optimization and Machine Learning”, Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA.

Gustafsson, F. & Gunnarsson, F. 2003. “Positioning Using Time-Difference of Arrival Measurements”. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong, China.

Hamalainen, S., Sanneck, H., & Sartori, C. 2012. “LTE Self-Organising Networks (SON): Network Management Automation for Operational Efficiency”. John Wiley and Sons, Ltd., UK.

Hapsari, W.A., Umesh, A., Iwamura, M., Tomala, M., Gyula, B., & Sebire, B. 2012. “Minimization of drive tests solution in 3GPP”. In IEEE Communications Magazine, 50(6), 28-36.

Hightower, J. & Borriello, G. 2004. “Particle filters for location estimation in ubiquitous computing: A case study”. In International Conference on Ubiquitous Computing (UbiComp 2004), 88-106.

Holma, H. & Toskala, A. 2012. “LTE Advanced: 3GPP Solution for IMT-Advanced”. John Wiley and Sons, Ltd., West Sussex, UK.

Hopper, A., Harter, A. and Blackie, T. 1993. “The Active Badge System”. In conference on Human factors in computing systems (INTERCHI-93), Amsterdam, The Netherlands, 533-534.

Hossain, A. K. M. M., Van, H. N., Jin, Y. and Soh, W. S. 2007. “Indoor localization using multiple wireless technologies”. In IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS 2007), Pisa, Italy.

Ibrahim, M., & Youssef, M. 2012. “CellSense: An Accurate Energy-Efficient”. In IEEE Transactions on Vehicular Technology, 61(1), 286 - 296.

Jambu, M. 1978. “Classification automatique pour l’analyse de données”. Paris: Dunod., France.

Johansson, J., Hapsari, W. A., Kelley, S. & Bodog, G. 2012. “Minimization of drive tests in 3GPP release 11”. In IEEE Communications Magazine, 50(11), 36-43.

Kim, M., Feilding, J. J. and Kotz, D. 2006. “Risks of Using AP Locations Discovered through War Driving”. In the 4th International Conference on Pervasive Computing (Pervasive 2006), 67-82.

Kim, J. H., Choi, H. I., Lee, D. S. and Kang, S. Y. 2014. “A 2-stage hybrid position estimation framework in RF fingerprint WPS”. In Wireless Networks, 20(6), Springer, 1541–1556.

Kim, J. H., Min, K. S. and Yeo, W. Y. 2014. "A Design of Irregular Grid Map for Large-Scale Wi-Fi LAN Fingerprint Positioning Systems". In The Scientific World Journal, (2014), Article ID 203419, 13 pages.

Kim, J. H. & Yeo, W. Y. 2014. "A coherent data filtering method for large scale RF fingerprint Wi-Fi Positioning Systems". In EURASIP Journal on Wireless Communications and Networking, 2014:13.

King, B. 1967. "Step-wise clustering procedures". In Journal of the American Statistical Association, 62(317), 86–101.

Kjærgaard, M. B. & Munk, C. V. 2008. "Hyperbolic location fingerprinting: A calibration-free solution for handling differences in signal strength", In 6th Annual IEEE International Conference on Pervasive Computing and Communications (PerCom 2008), Hong Kong, China, 110–116.

Kohonen, T. 1990. "The self-organizing map". In Proceedings of the IEEE, 78(9), 1464-1480.

Kohonen, T. 2001. "Self-Organizing Maps". Springer-Verlag, Berlin, Germany.

Kozma, L. 2008. "k Nearest Neighbours Algorithm". Helsinki University of Technology, Helsinki.

Knutti, F., Sabathy, M., Driusso, M., Mathis,H. & Marshall, C. 2015. "Positioning Using LTE Signals". In European Navigation Conference (ENC). Bordeaux, France.

Kushki, A., Plataniotis, K. N. & Venetsanopoulos, A. N. 2007. "Kernel-Based Positioning in Wireless Local Area Networks". In IEEE Transactions on Mobile Computing. 6(6), 689-705.

Laitinen, H., Lahteenmaki, J. and Nordstrom, T. 2001. "Database Correlation Method for GSM Location". In 53rd IEEE Vehicular Technology Conference (VTC 2001 Spring), Rhodes, Greece.

Laitinen, E., Talvitie, J., Lohan, E., S. and Renfors, M. 2011. "Comparison of Positioning Accuracy of Grid and Path Loss-Based Mobile Positioning Methods Using Received Signal Strengths". In conference on Signal Processing and Applied Mathematics for Electronics and Communications (SPAMEC 2011), Cluj-Napoca, Romania.

Laitinen, E., Lohan, E. S., Talvitie, J. and Shrestha, S. 2012. "Access point significance measures in WLAN-based location". In 9th Workshop on Positioning Navigation and Communication (WPNC), Dresden, Germany.

Lakmali, B. D. S. & Dias, D. 2008. "Database Correlation for GSM Location in Outdoor & Indoor Environments". In 4th IEEE International Conference on Information and Automation for Sustainability (ICIAFS 2008), Colombo, Sri Lanka.

LaMarca, A., Chawathe, Y., Consolvo, S., Hightower J., Smith, I., Scott, J., Sohn, T., Howard, J., Hughes, J., Potter, F., Tabert, J., Powledge, P., Borriello, G., and Schilit, B. 2005. "Place Lab: Device Positioning Using Radio Beacons in the Wild". In International Conference on Pervasive Computing (Pervasive 2005), Springer-Verlag Berlin Heidelberg, Germany, 116-133.

Li, K., Jiang, P., Bodanese, E. L. & Bigham, J. 2012. "Outdoor location estimation using received signal strength feedback". In IEEE Communications Letters. 16(7), 978–981.

Li, K., Bigham, J., Tokarchuk, L. and Bodanese, E. L. 2013. "A Probabilistic Approach to Outdoor Localization Using Clustering and Principle Component Transformations". In 9th International Wireless Communications and Mobile Computing Conference (IWCMC), Sardinia, Italy.

Lin, D. B. & Juang, R. T. 2005. "Mobile Location Estimation Based on Differences of Signal Attenuations for GSM Systems". In IEEE Transactions on Vehicular Technology, 54(4), IEEE Vehicular Technology Society, 1447-1454.

Liu, X., Zhang, S., Quan, J. & Lin, X. 2010. "The Experimental Analysis of Outdoor Positioning System Based on Fingerprint Approach". In 12th IEEE International Conference on Communication Technology (ICCT), Nanjing, China, 369 - 372.

MacQueen, J. B. 1967. "Some methods for classification and analysis of multivariate observations". In the 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 281–297.

Magro, M. J. & Debono, C. J. 2007. "A Genetic Algorithm Approach to User Location Estimation in UMTS Networks". In The International Conference on Computer as a Tool (EUROCON, 2007), Warsaw, Poland.

Michalewicz, Z. 1994. "Genetic Algorithm + Data Structures = Evolution Programs". Springer-Verlag, New York, USA.

Milioris, D., Tzagkarakis, G., Papadopouli, M. & Tsakalides, P. 2014. "Low-dimensional signal-strength fingerprint-based positioning in wireless LANs". In Ad Hoc Networks. 12, Elsevier, 100-114.

Murtagh, F. 1983. "A survey of recent advances in hierarchical clustering algorithms". In The Computer Journal, 26(4), 354-359.

Nurminen, H., Talvitie, J., Ali-Löytty, S., Müller, P., Lohan, E. S., Piché, R. and Renfors, M. 2012. "Statistical path loss parameter estimation and positioning using RSS measurements". In International Conference on Indoor Positioning and Indoor Navigation (IPIN), Sydney, NSW, Australia.

Olsson, M. & Mulligan, C. 2012. "EPC and 4G Packet Networks: Driving the Mobile Broadband Revolution". Elsevier Science Publishing Co Inc., USA.

Pages-Zamora, A., Vidal, J. and Brooks, D. H. 2002. "Closed-Form Solution for Positioning Based on Angle of Arrival Measurements". In 13th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2002), Pavilhao Altantico, Lisboa, Portugal.

Priyantha, N. B., Chakraborty, A. and Balakrishnan, H. 2000. "The Cricket Location-support System". In 6th annual International conference on Mobile computing and networking (MobiCom '00), Boston, Massachusetts, USA, 32-43.

Reed, J. H., Krizman, K. J., Woerner, B. D. & Rappaport, T. S. 1998. "An overview of the challenges and progress in meeting the 911 requirement for location services". In IEEE Communication Magazine 36 (4), 30-37.

Roos, T., Myllymaki, P., Tirri, H., Misikangas, P. and Sievanen, J. 2002. "A Probabilistic Approach to WLAN User Location Estimation". In International Journal of Wireless Information Networks, 9(3), Springer, 155-164.

Sand, S., Dammann, A. & Mensing, C. 2014. "Positioning in Wireless Communications Systems". John Wiley and Sons, Ltd., West Sussex, UK.

Schwaighofer, A., Grigoras, M., Tresp, V. and Hoffmann, C. 2003. "GPPS: A Gaussian Process Positioning System for Cellular Networks". In the conference on Neural Information Processing Systems (NIPS 2003), Whistler, British Columbia, Canada.

Shin, D. H. & Sung, T. K. 2002. "Comparisons of Error Characteristics Between ToA and TDoA Positioning". In IEEE Transactions on Aerospace and Electronic Systems, 38(1), 307-311.

Shi, L. & Wigren, T. 2009. "AECID fingerprinting positioning performance". In IEEE Global Telecommunications Conference (GLOBECOM 2009), Honolulu, HI, USA, 2767-2772.

Sivanandam, S. N. & Deepa, S. N. 2008. "Introduction to Genetic Algorithms". Springer-Verlag, Heidelberg, Germany.

Smailagic, A., Kogan, D. 2002. "Location Sensing and Privacy in a Context Aware Computing Environment". In IEEE Wireless Communications, 9(5), 10-17.

Sneath, P. H. A. & Sokal, R. R. 1973. "Numerical Taxonomy". W. H. Freeman and Company, USA.

Srinivas, N. & Deb, K. 1994. "Multiobjective Function Optimization Using Nondominated Sorting Genetic Algorithms". Evolutionary Computation, 3, 221–248.

Stone, J. S. 1902. "Patent US716134-Method of determining the direction of space-telegraph signals". (US Patent Publication number US716134A).

Sun, G., Chen, J., Guo, W. & Liu, K. J. R. 2005. "Signal processing techniques in network-aided positioning - A survey of state-of-the-art positioning designs". In IEEE Signal Processing Magazine. 22 (4). 12–23.

Trevisani, E. & Vitaletti, A. 2004. "Cell-ID Location Technique Limits and Benefits: an Experimental Study". In 6th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA 2004) conference, Windermere, Cumbria, UK.

Tsui, A. W., Linn, W. and Chen, W. 2010. "Accuracy Performance Analysis between war driving and war walking". In IEEE Transactions on Mobile Computing, 9(11), 1551-1562.

Turkka, J. 2014. "Aspects of Knowledge Mining on Minimizing Drive Testsin Self-organizing Cellular Networks". PhD thesis, Tampere University of Technology, Tampere, Finland.

Vaghefi, R. M., & Buehrer, R. M. 2014. “Cooperative RF pattern matching positioning for LTE cellular systems”. In IEEE 25th Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC). Washington DC, USA.

Varshavsky, A., Chen, M. Y., Lara, E. d., Froehlich, J., Haehnel, D., Hightower, J., LaMarca, A., Potter, F., Sohn, T., Tang, K. and Smith, I. 2005. “Are GSM phones THE solution for localization?”. In 7th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA '06), Orcas Island, WA, USA.

Ward, J. H. Jr. 1963. “Hierarchical grouping to optimize an objective function”. In Journal of the American Statistical Association, 58(301), 236-244.

Wigren, T. 2007. “Adaptive Enhanced Cell-ID Fingerprinting Localization by Clustering of Precise Position Measurements”. In IEEE Transactions on Vehicular Technology, 56(5), 3199-3209.

Wigren, T., 2011. “Clustering and Polygon Merging Algorithms for Fingerprinting Positioning in LTE”. In 5th International Conference on Signal Processing and Communication Systems (ICSPCS), Honolulu, HI, USA.

Witten, I. H. & Frank, E. 2005. “Data mining : practical machine learning tools and techniques”. Morgan Kaufmann Publishers, Elsevier, San Francisco, California, USA.

Wortham, J. 2009. “Cellphone locator system needs no satellite”. New York Times, USA.

Yin, J., Yang, Q. and Ni, L. M. 2008. “Learning Adaptive Temporal Radio Maps for Signal Strength based Location Estimation”. In IEEE Transactions on Mobile Computing 7(7), 869 - 883.

Youssef, M., Agrawala, A. and Shankar, A. U. 2003. “WLAN Location Determination via Clustering and Probability Distributions”. In the First IEEE International Conference on Pervasive Computing and Communications (PerCom’03), Washington, DC, USA.

Zekavat S. A. R. & Buehrer R. M. 2011. “HANDBOOK OF POSITION LOCATION Theory, Practice, and Advances”. John Wiley and Sons, Ltd., West Sussex, UK.

Zimmermann, D., Baumann, J., Layh, A., Landstorfer, F., Hoppe, R. and Wolfle, G. 2004. “Database Correlation for Positioning of Mobile Terminals in Cellular Networks using Wave Propagation Models”. In 60th IEEE Vehicular Technology Conference (VTC 2003- Fall), Los Angeles, CA, USA.

Zhu, J., & Durgin, G. D. 2005. “Indoor/Outdoor Location of Cellular Handsets Based on Received Signal Strength”. In Electronics Letters. 41(1), IET.

ORIGINAL PAPERS

PI

POSITIONING IN HETEROGENEOUS SMALL CELL NETWORKS USING MDT RF FINGERPRINTS

by

Riaz Uddin Mondal, Jussi Turkka, Tapani Ristaniemi and Tero Henttonen 2013

1st IEEE International Black Sea Conference on Communications and Networking, Batumi, Georgia

PII

PERFORMANCE EVALUATION OF MDT ASSISTED LTE RF FINGER-PRINT FRAMEWORK

by

Riaz Uddin Mondal, Jussi Turkka, Tapani Ristaniemi and Tero Henttonen 2014

7th International Conference on Mobile Computing and Ubiquitous Networking (ICMU2014), Singapore

PIII

AN EFFICIENT GRID-BASED RF FINGERPRINT POSITIONING ALGORITHM FOR USER LOCATION ESTIMATION IN HETEROGENEOUS SMALL CELL NETWORKS

by

Riaz Uddin Mondal, Jussi Turkka and Tapani Ristaniemi 2014

International Conference on Localization and GNSS (ICL-GNSS 2014), Helsinki, Finland

An Efficient Grid-based RF Fingerprint Positioning Algorithm for User Location Estimation in Heterogeneous Small Cell Networks

Riaz Mondal[1,2], Jussi Turkka[3], Tapani Ristaniemi[1]
[1]Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland
[2]Department of Information and Communication Engineering, University of Rajshahi, Rajshahi, Bangladesh
[3]Magister Solutions Ltd, Tampere, Finland
riaz.u.mondal@student.jyu.fi, jussi.turkka@magister.fi, tapani.ristaniemi@jyu.fi

***Abstract*—This paper proposes a novel technique to enhance the performance of grid-based Radio Frequency (RF) fingerprint position estimation framework. First enhancement is an introduction of two overlapping grids of training signatures. As the second enhancement, the location of the testing signature is estimated to be a weighted geometric center of a set of nearest grid units whereas in a traditional grid-based RF fingerprinting only the center point of the nearest grid unit is used for determining the user location. By using the weighting-based location estimation, the accuracy of the location estimation can be improved. The performance evaluation of the enhanced RF fingerprinting algorithm was conducted by analyzing the positioning accuracy of the RF fingerprint signatures obtained from a dynamic system simulation in a heterogeneous LTE small cell environment. The performance evaluation indicates that if the interpolation is based on two nearest grid units, then a maximum of 18.8% improvement in positioning accuracy can be achieved over the conventional approach.**

Keywords— Grid-based RF fingerprint; Kullback-Leibler Divergence; Minimization of Drive Tests

I. INTRODUCTION

Various location-based services in wireless communication networks depend on mobile positioning. Commercial examples range from low-accuracy methods based on cell identification to high-accuracy methods combining wireless network information and satellite positioning. These methods are typically network centric, where the position is determined in the network and presented to the user via a specific service. Due to logistical reasons, the position is estimated from static snapshot measurements, possibly provided by the mobile station (MS) [1]. A major driving force to estimation mobile user location is the requirement for E-911 emergency positioning in the North American market. This requirement calls for accuracies of 50 meters (68%) and 150 meters (95%) in the terminal-dependent case and 100 meters (68%) and 300 meters (95%) in the network-dependent case [2]. Today, the Global Navigation Satellite System (GNSS) is the most effective positioning technology in the outdoor open environments [3]. However it has limitations such as poor performance in built-up areas and high power consumption. These limitations led to the development of positioning techniques based on the wireless networks. These technologies include a variety of time-of-arrival techniques (ToA and TDoA), angle-of-arrival techniques (AoA), and location fingerprinting techniques [4]. Typically, RF fingerprinting refers to a database correlation method where the position is estimated by comparing the radio measurements e.g., the RF fingerprint of the user equipment (UE) with the training fingerprints in the correlation database. The training fingerprints consist of received signal strength (RSS) radio measurements from several base stations (BS) that are used to provide a fingerprint of the radio conditions at a specific geographical location. Typically, this location is determined with an accurate positioning method, for example GNSS. Hence, fingerprinting is a positioning method which exploits the already existing infrastructures such as cellular networks [5] and WLANs [6].

One of the biggest challenges of RF fingerprinting is the burden of creating and maintaining the correlation database of the training fingerprints. Operators can maintain the correlation database by conducting extensive and expensive periodical drive test campaigns to collect the required measurements. However, the concept of Minimization of Drive Tests (MDT) provides a framework for gathering user reported location-aware radio measurements from commercial mobile phones that can be used for creating and maintaining such training databases [7], [8]. In fact, one of the benefits of MDT is that it provides an efficient way to automate the collection of training fingerprints in Universal Terrestrial Radio Access Network (UTRAN) and Evolved-UTRAN (EUTRAN) cellular systems. The MDT procedure allows operators to collect radio measurements, i.e. received signal strength and quality, with UE location information and a time stamp. It is worth noting that for immediate MDT, the network can make a request to UE to attempt to make GNSS location information available for MDT [8].

In our previous work [9], [10], we have evaluated the RF fingerprint positioning accuracy using Kullback-Leibler Divergence (KLD) and Mahalanobis distance with a single grid layout (SGL) in rural, urban and heterogeneous small cell networks. Although the positioning accuracy in dense urban scenario was good, the positioning error (PE) did not fulfill the E-911 requirements in the rural case.

In this work, we propose a novel RF fingerprint algorithm to enhance the positioning accuracies further for rural case and also in the dense urban scenario. The proposed enhancements are based on an overlapping grid layout technique to produce the training signatures and to determine the estimated position using the two smallest KLD weighted values among the matched training signatures. The effect of using the overlapping grid layout (OGL) is that it increases the training signatures as well as reduces the distances among the surrounding grid centers and thereby improving the positioning accuracy. However it increases the computational time needed to estimate the testing UE positions as compared to that of the traditional SGL method by a small amount.

This paper is organized as follows. Section 1 gives an introduction to the research problem. In Section 2, first conventional SGL RF fingerprinting and then the OGL approach using Kullback-Leibler Divergence method is described. Finally in Section 3, the performance evaluation of the enhancements is discussed in the light of extensive system simulations.

II. Grid-based RF Fingerprinting Framework

A. Single Grid Layout

In SGL approach, RF fingerprinting correlation database is compressed to present the geographical space G={g_1, g_2, …, g_N} e.g., "area of interests" as a regular tessellation of N square grid units g_i as depicted in Fig. 1. Each grid unit g_i is associated with a center point having coordinates $c_i=\{x_i,y_i\}$ and a set of RF fingerprint training signatures $S_i=\{s_{i,1}, s_{i,2}, \ldots, s_{i,k}\}$ which will depend upon different sets of same BS id MDT samples. Hence, a training signature $s_{i,j}$ is the jth signature associated with ith grid unit and it is constructed from n different MDT measurement samples. Each sample in $\mathbf{M}_{i,j}$ contains detailed location information e.g., x and y coordinate obtained from GNSS receiver, and m RSS measurement e.g., Reference Signal Received Power (RSRP) in EUTRAN, from different BSs.

$$\mathbf{M}_{i,j} = \begin{bmatrix} x_1 & y_1 & rsrp_{1,1} & \cdots & rsrp_{1,\mathrm{m}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n & y_n & rsrp_{\mathrm{n},1} & \cdots & rsrp_{\mathrm{n,m}} \end{bmatrix}, \qquad (1)$$

where, $rsrp_{1,m}$ is the 1st RSRP measurement with a BS identifier m. Moreover, Euclidean distance between the detailed locations of the n measurement samples and the center point of c_i suggests that the nearest grid unit is g_i. For ensuring an efficient and fast processing of the measurement data, the MDT training measurements are compressed e.g., for each grid g_i, the signatures in S_i stores only the means and the covariance of the received signal strengths between the detected BSs in $\mathbf{M}_{i,j}$ as discussed in [9], [10].

B. Overlapping Grid Layout

In Overlapping Grid Layout (OGL) approach, the whole area of interest is divided into square grids similar to the SGL approach but there are now multiple layouts which overlap each other. In the present work, OGL consisted of two grid

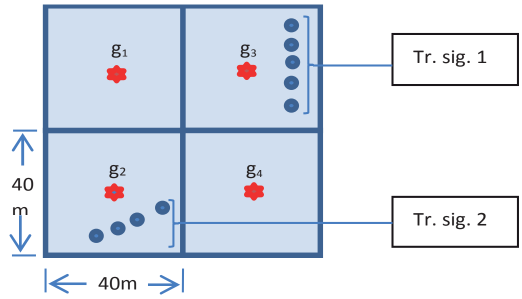


Fig. 1: Single grid layout based conventional RF fingerprint positioning

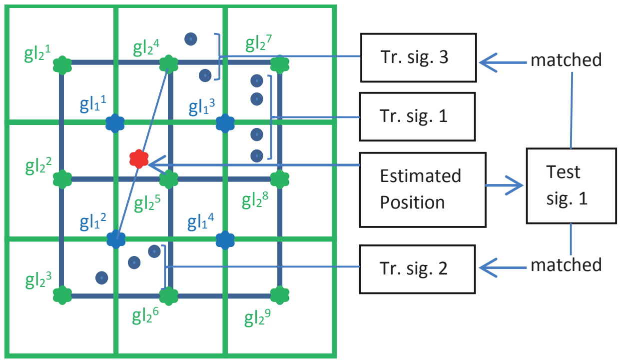


Fig. 2: Two overlapping grid-cell based RF fingerprint positioning

layouts denoted by OGL_1 and OGL_2 having same sized square grids. OGL_2 was placed in a fashion that an OGL_1 grid unit is overlapped by ¼th area of an OGL_2 grid unit. In this way OGL_1 grid unit is fully overlapped by four OGL_2 grid units as shown in Fig. 2, here the squares with blue border lines belongs to OGL_1 layout and the blue stars represents the corresponding grid centers, whereas the squares with green border lines and green stars depicts the OGL_2 layout. Thus the center points of four OGL_2 grid units fall on the four corners of the overlapped OGL_1 grid unit.

Training signatures are formed in grid-wise manner. A training signature comprises of a set MDT samples that are located within a grid unit and these MDTs were reported to have the same BS ids. The same set of training samples is used to form the training signatures for OGL_1 and then for OGL_2. Because of the offset between the layouts mostly different training signatures will be associated with the layouts. Inherently some common signatures will exist between these layouts. All the training signatures belonging to OGL_1 and OGL_2 layouts are put together to create the whole training signature database.

During the testing phase, all training signatures that match with the testing signature BS ids are searched, and then the KLD distances are calculated between the testing signature and those training signatures. In KLD method, training and testing phase signatures e.g., mean vectors $\mathbf{u}$ and covariance matrices Σ parameterize multivariate Gaussian distributions p($\mathbf{x}$|$\mathbf{u}$, Σ), and therefore, the method aims to exploit interdependencies among the received signal strengths between BSs by using covariance matrices for training and testing phase signatures as given by the following closed form KLD equation:

$$d\left(p_u \| p_{i,t}\right)=\frac{1}{2}\left(\left(\mathbf{u}_u-\hat{\mathbf{u}}_{i,t}\right)^T \hat{\boldsymbol{\Sigma}}_{i,t}^{-1}\left(\mathbf{u}_u-\hat{\mathbf{u}}_{i,t}\right)+tr\left(\boldsymbol{\Sigma}_u \hat{\boldsymbol{\Sigma}}_{i,t}^{-1}-\mathbf{I}\right)-\ln\left|\boldsymbol{\Sigma}_u \hat{\boldsymbol{\Sigma}}_{i,t}^{-1}\right|\right) \quad (2)$$

where, $\mathbf{u}_u$, and $\hat{\mathbf{u}}_{i,t}$ corresponds to the mean received signal strength values, while Σ_u and $\Sigma_{i,t}$ represents the covariance matrices of the received signal strength values of the testing and training signatures respectively. Here, *tr*() denotes the trace of matrix, Σ^{-1} denotes the inverse of covariance matrix Σ and **I** is the identity matrix. It is a non-symmetric measure of the difference between testing and training signature probability distributions p_u and $p_{i,t}$. Two grid units are selected which correspond to the two smallest KLD distance values among all the matched training signatures. Now the estimation point for the testing signature is calculated to be in between the center points of these two grids according to the KLD weight factor of the corresponding grids. For example, in Fig. 2, the testing signature 1 i.e., '*test sig.1*', has two matched training signatures: '*tr.sig. 2*' which belongs to OGL_1 and '*tr.sig 3*' of OGL_2 layout. In order to determine the estimated position for the testing signature, the KLD values of these two training signatures are used as weights and depending upon the weighted value the estimation point moves along the line joining the center points of the corresponding grid units as described in next section.

C. *Weighted Overlapping Grid Layout*

The enhanced algorithm for a grid-based RF fingerprinting uses the weighted geometric center of a set of nearest grid units to determine the location of a testing signature in a two dimensional grid. Formally, the geometric center (or centroid) is an arithmetic mean of a set of points weighted by special weight i.e., point density or mass. In the context of this paper, the KLD metric is used to find the *n* nearest grid units and to weight the arithmetic mean.

To determine the location of a testing signature s_i, a subset $G_i = \{g_1, g_2, \ldots, g_k\}$ of nearest grid units based on KLD metric is chosen depending upon the matched BS ids of the testing signature. In addition, each grid unit in G_i is assigned with a weight w_j and therefore a set of weights $W_i = \{w_1, w_2, \ldots, w_k\}$ is obtained. Since, the KLD metric is smallest for the nearest grid unit, the weighting factors are inversely proportional to KLD metric for ensuring that more weight is given to the nearest grid unit. The weighting factor w_j for *j*th grid unit g_j in subset G_i is given as,

$$w_j = \frac{\frac{1}{\varsigma_j}}{\sum_{l=1}^{k}\frac{1}{\varsigma_l}} = \frac{1}{\varsigma_j \sum_{l=1}^{k}\frac{1}{\varsigma_l}} \quad (3)$$

where, ς_j is the KLD metric between testing signature s_i and *j*th grid unit in the subset G_i of the *k* nearest grid units. Note that the weights in W_i are normalized so that they sum up to 1.

Therefore, the weighted geometric center of the testing signature is obtained from,

$$[x_i, y_i] = (\mathbf{w}_i)^{\mathrm{T}}[\mathbf{x}_i, \mathbf{y}_i], \quad (4)$$

where $\mathbf{w}_i$, $\mathbf{x}_i$, and $\mathbf{y}_i$ are column vectors containing the weights and x and y coordinates of the *k* nearest grid units in G_i. Moreover, $(.)^{\mathrm{T}}$ is a vector transpose operation. Hence, (4) gives the scalars x_i and y_i which is the location of the weighted arithmetic mean of the points in G_i and the estimated location of s_i. In this study, *k* is equal to 2, so only two nearest neighbors where considered for analyzing the location estimates for testing signatures.

III. PERFORMANCE EVALUATION

A. *Simulation Scenario*

In this study we have conducted a performance evaluation of our enhanced RF fingerprinting method having two overlapping grid layouts with the traditional single grid layout. For this purpose, we have performed two simulations in two different network scenarios, (i) sparse regular macro (RM) and (ii) heterogeneous small cell (HSC) network as illustrated in Fig. 3 and Fig. 4 respectively. Both scenarios are the same as in [10] and those were simulated by using a state-of-the-art LTE Rel'10 dynamic system simulator to model both the downlink and the uplink in an OFDM symbol resolution with several radio resource management, scheduling, mobility, handover and traffic modeling functionalities. Simulation parameters and mathematical models are based on the 3GPP specifications, (especially in the simulation assumptions in 3GPP TR 36.839) defining parameterization for used bandwidth, center frequency, path-loss, slow fading, and fast fading [11]. Moreover, UE measurements e.g., RSRP, were implemented in the simulator taking into account the technical requirements for the absolute and relative measurement errors and -6 dB Ês/Iot cell detection criterion as in [12].

The RM network consisted of 57 BSs operating on primary frequency band (CC0) and 36 sectored high power Pico BSs operating on non-overlapping secondary carrier (CC1). Inter-

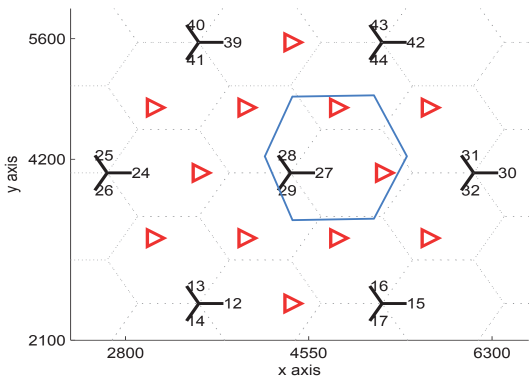


Fig. 3: Sparse regular macro scenario

site distance between the BSs on CC0 and CC1 are 1750m and 875m, respectively. Locations of three-sectored Pico sites on CC1 were shifted to be in between the macro BSs on CC0 as depicted in Fig. 3 with red triangels. In RM scenario a total of 82320 MDT samples were generated for the analysis. For this case we have picked up the MDT samples from an area enclosed by the blue hexagon as shown in Fig. 3 in order to maintain size similarity between the RM and HSC scenarios.

The HSC network consisted of 57 macro BSs having inter-site distance of 500m and operating on CC0 band. In addition,

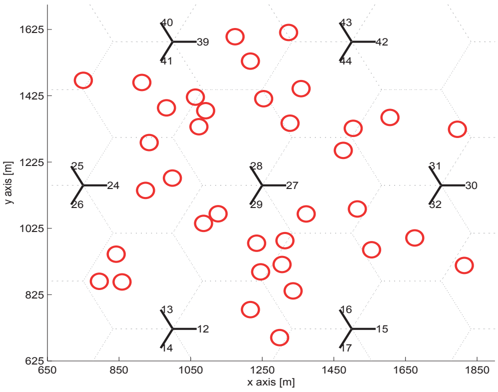


Fig. 4: Heterogeneous small cell scenario

36 small cells with omni-directional antennas were randomly deployed in the coverage area of 12 centermost macro BSs as depicted in Fig. 4 with red circles. Distance to the nearest inter-site small cell varied from 50m to 170m with average distance being 95m. In HSC scenario, UEs were moving only in the area of 12 centermost macro BSs but were able to monitor all detected cells. In RM scenario, UEs were moving in one specific area with area size similar to the simulation area in HSC scenario. In total, 1200 randomly moving outdoor vehicular UEs (30 km/h) were distributed uniformly to the simulation area. Traffic profile consisted of data generated by MDT reports which were sent once per second, however, 100% resource block loading was used for creating interference limited simulation environment which is more challenging from cell detection point of view. More details abou the used simulation parameters can be found in the Table I in [10].

B. Performance Results

In this study the positioning accuracies were analyzed using two different amounts of training data and testing data. First 90% of the MDT data samples were used in training and the estimation was done for the remaining 10% MDT samples. Then 10% of the MDT data samples were taken for training and 90% for the testing purpose. Such an approach was selected for studying how the amount of training data affects the performance of OGL method. The training and the testing database samples are selected in call-wise manner e.g., all MDT reports from single UE belongs always either on the training or the testing database. This is done to avoid too optimistic positioning results in cases where consecutive measurements from one UE are found in training and testing databases. The number of training signatures used and analyzed test signature percentages in SGL and OGL methods are given in Table I. From this table we can find that number of training signatures used in OGL is about twice the number used in SGL for all the different simulations.

TABLE I. ANALYZED RF FINGERPRINT SIGNATURES OF SGL AND OGL LAYOUTS

Grid-cell size	Scen-ario	Train-ing Data (%)	Total no. of Training Signatures (Absolute)		Analyzed Test Signatures (%)	
			SGL	OGL	SGL	OGL
10-BY-10 M	RM (ISD 1750M)	90%	16443	32758	83.19	84.86
		10%	2044	4092	62.19	69.07
	HSC (ISD 500M)	90%	48808	97687	71.66	73.39
		10%	6319	12707	47.60	55.14
40-BY-40 M	RM (ISD 1750M)	90%	7090	14222	82.64	85.72
		10%	1709	3401	64.02	73.80
	HSC (ISD 500M)	90%	22079	44219	70.50	74.62
		10%	5321	10677	47.74	61.43

TABLE II. POSITIONING ACCURACY PERFORMANCE EVALUATION

Scen-ario	Trai-ning Data (%)	RF Finger-print Algori-thm	For 10-by-10 m Grid		For 40-by-40 m Grid	
			68% PE (m)	95% PE (m)	68% PE (m)	95% PE (m)
RM (ISD 1750 M)	90%	SGL Based	29.73	165.29	43.53	196.30
		OGL Based	31.41 (+5.6%)	147.49 (-10.7%)	40.86 (-6.1%)	161.75 (-17.6%)
	10%	SGL Based	72.00	228.80	72.48	225.45
		OGL Based	63.96 (-11.1%)	206.05 (-9.9%)	65.03 (-10.2%)	203.70 (-9.6%)
HSC (ISD 500 M)	90%	SGL Based	21.12	58.08	33.73	76.43
		OGL Based	19.45 (-7.9%)	50.94 (-12.2%)	27.57 (-18.2%)	64.87 (-15.1%)
	10%	SGL Based	27.23	73.61	34.83	80.86
		OGL Based	25.14 (-7.6%)	66.47 (-9.6%)	28.28 (-18.8%)	68.71 (-15.0%)

It is noticeable from Table I that for 90% of training data the increased percentage of analyzed test signatures for OGL as compared to SGL method is about 1.7% and 3% for 10 by 10 m and 40 by 40m grid-cells respectively. Whereas for 10% of training data the improvement shown by OGL method over SGL is about 7% and 10% for 10 by 10 m and 40 by 40m grid-cells respectively in analyzing the test signatures.

The simulation results for the conventional SGL and the proposed OGL RF fingerprint positioning in RM and HSC scenarios are given in Table II. Here 68 and 95 percentiles give the positioning error (PE) in meters for SGL and OGL methods using different training data sets during simulations. The percentage within the first bracket under the PE value of each of the OGL based method indicates the degradation (with a + sign) or improvement (with a - sign) in positioning as compared to the SGL method. For example, in the HSC 10% training case with 40m by 40m grid the SGL fingerprint positioning algorithm has given 34.83 meters and 80.86 meters of PE in the 68 and 95 percentiles respectively whereas that of OGL is 28.28 meters and 68.71 meters respectively. Hence OGL method has reduced the PE by 18.8% and 15.0% in 68 and 95 percentiles respectively as compared to that of the SGL method. Also for HSC 90% training case with the same grid size OGL has shown 18.2% and 15.1% of improvement in 68 and 95 percentiles respectively over that of the SGL. Table I depicts that almost in every simulated study OGL RF Fingerprinting performs better than SGL method. Moreover, it is worth highlighting that if only 10% samples are available for constructing the training signatures, 40-by-40 meters OGL can outperform 10-by-10 meters SGL performance.

In our present study the MDT samples comprising the testing signatures tend to be distributed over several training grid units, as a result positioning accuracies were improved when an overlapping grid layout is used for the formation of training signatures and then the centroid point is determined from the two smallest weighted KLD grids. The OGL method has constructed more training signatures with different combinations of BS IDs than the SGL method. Therefore, it is more probable to find matching combination of BS IDs in OGL method. Hence the proposed algorithm not only improves the positioning accuracy but it can also analyze more testing samples.

IV. Conclusion

This paper proposed two enhancements for grid-based RF fingerprint position estimation framework. First enhancement is an introduction of overlapping grid layout to form training signatures. In the second enhancement, the location of the testing signature is estimated to be a weighted geometric center of a set of nearest grid units. This is different from the traditional grid-based RF fingerprinting where only the center point of the nearest grid unit is used for determining the user location. The performance evaluation of the enhanced RF fingerprinting algorithm was conducted by analyzing the positioning accuracy of the RF fingerprint signatures obtained from a dynamic LTE system simulation.

The proposed enhancements can increase the number of training signatures that needs to be analyzed for finding the nearest grid but in addition the positioning accuracy is increased, specifically in cases where only a limited amount of data is used for constructing the training signatures. The performance evaluation indicates that the proposed overlapping grid layout method using KLD can provide a maximum of 18.8% improvement in positioning accuracy as compared to that of the conventional single grid layout approach.

Acknowledgement

The present work was carried out within the framework of European Celtic-Plus project SHARING (Self-organized Heterogeneous Advanced Radio Networks Generation).

References

[1] F. Gustafsson and F. Gunnarsson, "Mobile Positioning Using Wireless Networks," IEEE Signal Processing Magazine, July 2005, pp. 41-53.

[2] The FCC, *Fact Sheet—FCC Wireless 911 Requirements*, Jan. 2001, FCC.

[3] E.C.L. Chan, G. Bacieu, "Wireless tracking analysis in location fingerprinting," in *Proc. IEEE International conference on Wireless and Mobile Computing*, October 2008, pp. 214–220.

[4] A. Kupper, *Location based services:* Fundamentals and Operation. John Wiley & Sons Ltd, England , August 2005.

[5] C. Takenga, K. Kyamakya, "A low-cost fingerprint positioning system in cellular networks," in *Proc. International Conference on Communications and Networking*, August 2007, pp. 915 – 920.

[6] P. Bahl, V.N Padmanabhan, "Radar: an in-building rf-based user location and tracking system," in *Proc. Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, March 2000, pp. 775 – 784.

[7] Wuri A. et al., "Minimization of Drive Tests Solution in 3GPP", *in IEEE Communications Magazine*, Vol. 50, No. 6, June 2012.

[8] Johansson J., Hapsari W.A, Kelley S. and Bodog G., "Minimization of drive tests in 3GPP release 11", *in IEEE Communications Magazine*, Vol. 50, No. 11, pp. 36-43, November 2012.

[9] Mondal R., Turkka J., Ristaniemi T. and Henttonen T., "Performance Evaluation of MDT RF Fingerprinting Framework", in *Proc. of 7th International Conference on Mobile Computing and Ubiquitous Networking (ICMU2014)*, January 2014, Singapore.

[10] Mondal R., Turkka J., Ristaniemi T. and Henttonen T, "Positioning in Heterogeneous Small Cell Networks using MDT RF Fingerprints", in *Proc. of the First IEEE International Black Sea Conference on Communications and Networking*, July 2013, Batumi, Georgia.

[11] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects", version 9.0.0, March 2009, available at http://www.3gpp.org.

[12] 3GPP TS 36.133, "Requirements for support of radio resource management", version 11.4.0, March 2013, available at http://www.3gpp.org.

PIV

GENETIC ALGORITHM OPTIMIZED GRID-BASED RF FINGERPRINT POSITIONING IN HETEROGENEOUS SMALL CELL NETWORKS

by

Riaz Uddin Mondal, Tapani Ristaniemi and Jussi Turkka 2015

International Conference on Localization and GNSS (ICL-GNSS 2015), Gothenburg, Sweden

Genetic Algorithm Optimized Grid-based RF Fingerprint Positioning in Heterogeneous Small Cell Networks

Riaz Uddin Mondal
Department of Mathematical Information Technology
University of Jyvaskyla
Jyvaskyla, Finland
riaz.u.mondal@student.jyu.fi

Tapani Ristaniemi
Department of Mathematical Information Technology
University of Jyvaskyla
Jyvaskyla, Finland
tapani.ristaniemi@jyu.fi

Jussi Turkka
Magister Solutions Ltd.
Tampere, Finland
jussi.turkka@magister.fi

Abstract— **In this paper we propose a novel optimization algorithm for grid-based RF fingerprinting to improve user equipment (UE) positioning accuracy. For this purpose we have used Multi-objective Genetic Algorithm (MOGA) which enables autonomous calibration of grid-cell layout (GCL) for better UE positioning as compared to that of the conventional fingerprinting approach. Performance evaluations were carried out using two different training data-sets consisting of Minimization of Drive Testing measurements obtained from a dynamic system simulation in a heterogeneous LTE small cell environment. The robustness of the proposed method has been tested analyzing positioning results from two different areas of interest. Optimization of GCL is performed in two ways: (1) array-wise calibration of the grid-cell units using non-overlapping GCL and (2) creating an overlapping GCL to cover of whole simulation area with different rectangular grid-cell units. Simulation results show that if sufficient amount of training data is available then the proposed method can improve positioning accuracy of 56.74% over the conventional grid-based RF fingerprinting.**

Keywords- Grid-based RF fingerprinting; Minimization of Drive Tests; Multi-objective Genetic Algorithm; Kullback-Leibler Divergence.

I. INTRODUCTION

Positioning in wireless networks is dependent on the mobility of users and the dynamic nature of both the environment and radio signals. Users expect the same level of performance whether they are indoors or outdoors in a rural or urban environment. So far no single positioning method, including GPS, works well in all environments [1]. Receivers in Global Navigation Satellite Systems (GNSS) such as GPS or GLONASS tend to output inaccurate location estimations while operating in urban regions, mostly due to the density of tall buildings, which often block a receiver's line of sight to the navigation satellites [2]. Among the non-standard positioning methods included in LTE Release 9, RF fingerprinting is the most cost-efficient solution for indoor WLAN positioning [3], [4], [5] as well as for outdoor mobile cellular positioning in densely built urban environments [6], [7]. RF fingerprinting, also known as database correlation method (DCM) finds a user's position by mapping RF measurements obtained from the UE onto an RF map, where the map is typically based on detailed RF predictions or site surveying results.

An ideal positioning system should be self-learning and environmentally adaptive, capable of building up information databases that store actual observations, and employ smart data analysis mechanisms [1]. In order to achieve such a goal, a functionality known as Minimization of Drive Tests (MDT) has been proposed in LTE Release 10 which reduces the huge cost and efforts associated with the conventional drive test measurement procedure. MDT provides a framework for gathering user reported location-aware radio measurements from commercial mobile phones that can be used for creating and maintaining RF fingerprint training databases [8], [9]. It allows operators to autonomously build and update large training database for RF fingerprinting from various locations of user experience along with available location information from UEs without extra hardware installation.

In [10] and [11] grid-based RF fingerprinting has shown good positioning performance in rural, urban and heterogeneous small cell networks scenario using MDT simulated data. An overlapping grid-cell layout (GCL) based RF fingerprinting approach was proposed in [12] which further improves the positioning accuracy. The performance results of [11] indicate that using a common GCL in different network scenarios does not provide the best positioning accuracy in both 68%-ile and 95%-ile of positioning error (PE) estimation. It was also found

The present work was carried out within the framework of European Celtic-Plus project SHARING (Self-organized Heterogoneus Advanced Radio Networks Generation).

from [12] that positioning performances varies of a GCL that is composed of same square grid-cell units when amount of training data is varied. From literature review we did not find an explicit method for optimizing GCL in grid-based RF Fingerprint positioning. In [13] GA was used to reduce the correlation space in RF fingerprinting location method to improve location accuracy and authors have claimed their method to be suitable for UE positioning in urban environments. Authors in [14] have proposed a location detection algorithm which employs cell-id positioning enhanced by triangulation. The process was accelerated through the application of a GA.

In this paper, we propose a novel method based on MOGA to develop an autonomous grid-based RF fingerprinting by optimizing the GCL in order to achieve the best possible positioning accuracy despite changes in cellular network scenarios and amounts of training data. This would render the positioning system an adaptive one which can make necessary adjustments to the grid-cell size along with the structural and environmental changes of the surrounding for an optimal RF fingerprint positioning performance.

The following section contains a brief description of the conventional RF fingerprinting using MDT measurements. In section III first a brief description of multi-objective GA is given, and then the MOGA optimized RF fingerprint positioning (GAFP) method is explained. Finally section 4 discusses the performance evaluation of GAFP in the light of extensive system simulations.

II. RF FINGERPRINTING USING MDT MEASUREMENTS

A. *Minimization of drive tests*

Conventional drive tests consume significant time and human efforts to obtain reliable data [8]. MDT is a feature introduced in 3GPP Release 10 that enables operators to utilize UE to collect radio measurements and associated location information, in order to assess network performance while reducing the large operation expenditure associated with traditional drive tests. Location information for MDT can be categorized into two different types: detailed location information and RF fingerprint. Detailed location information is typically obtained by GPS or Global Navigation Satellite System (GNSS) positioning method, but can also be obtained by other positioning methods supported by the UE and the network, e.g., Observed Time Difference of Arrival (OTDOA), Assisted -GNSS (Assisted-GPS) or Enhanced Cell ID (E-CID). Whereas in RF fingerprint type of MDT a profile of measured signal strength from neighboring cells is created. A major difference between conventional drive tests and MDT is that conventional drive tests use a controlled application with known traffic characteristics, whereas MDT uses ordinary user traffic with largely unknown characteristics [9].

Thus operators are able to build a large RF fingerprint training database from UE's in different network environments and up-date it as needed. This urges for an autonomous RF fingerprinting method which makes necessary calibration to its GCL with the updated MDT training database.

B. *Grid-cell Based RF Fingerprinting*

In a conventional grid-cell layout (CGCL) based RF fingerprinting method using Kullback-Leibler Divergence (KLD) has two main phases [3], [10], [11]:

Training Phase: First an offline processing and manipulation of MDT correlation database takes place. A layout of adjoining rectangular or square grid-cell units are formed over the whole geographical area of interest. Each of these grid units g_i is associated with a center point having coordinates $c_i=\{x_i,y_i\}$ as shown in Fig.1. Then unit-wise creation of training signatures is performed by gathering MDT measurements having signal strength values, e.g., Reference Signal Received Power (RSRP) from same set of base-stations (BSs). Hence depending upon the size of a grid-cell unit I and the availability of MDT measurement samples within that grid, it may contain multiple training signature $\boldsymbol{S}_{i,j}=\{s_{i,1}, s_{i,2}, \ldots, s_{i,j}\}$, where j is the total number of signatures. Each signature consists of multiple MDT measurements collected from that particular grid-cell unit having RSRP values from similar BS Ids along with detail location information of corresponding MDT measurement positions. For example the signature $s_{1,1}$ as shown in Fig. 1 consists of four MDT samples and can be expressed by:

$$s_{1,1} = \begin{pmatrix} c_{1,1} & ID_{1,1} & rsrp_{1,1} \\ \vdots & \vdots & \vdots \\ c_{1,4} & ID_{1,4} & rsrp_{1,4} \end{pmatrix} \tag{1}$$

where, $c_{i,j}$ is a vector of x and y coordinates for an MDT measurement, $ID_{i,j}$ stands for the cell identities which is same for all the samples that form a signature and the vector $rsrp_{i,j}$ contains RSRP values of corresponding MDT samples.

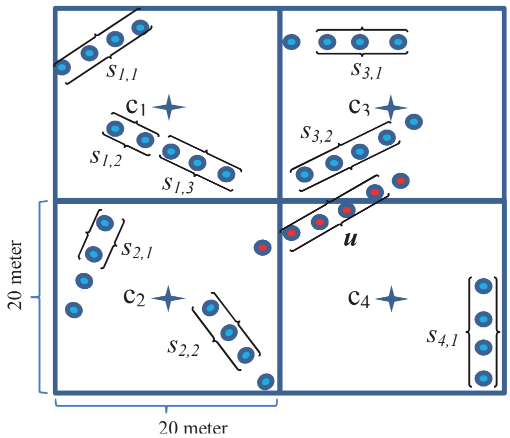


Figure 1. 20-by-20 meter grid-cell layout of a conventional RF fingerprint positioning

Fig.1 illustrates a conventional grid-based RF fingerprinting method. Here the area of interest has been divided into four square grid-cells, all of which have the same arm length of 20 meters and corresponding grid centers as depicted by c_1, c_2, c_3 and c_4. The training MDT samples are shown here by blue dots while the testing samples by red dots. It can be seen from this figure that no all the training samples which belong to a grid-cell form training signatures, but only those which matches in BS ids.

Testing Phase: During this phase testing signatures are formed from the MDT samples obtained from a particular UE. As we can see from the Fig.1 we have multiple training signatures and one testing signature depicted by $\boldsymbol{u}$ in this example. At first training signatures are collected which match with the testing signature BS IDs. Here if we assume that we have three such matched training signatures- $s_{1,3}$, $s_{2,2}$, and $s_{3,2}$, then KLD is calculated between $\boldsymbol{u}$ and matched training signature RSRP values. Now if the KLD between $\boldsymbol{u}$ and $s_{3,2}$ gives the minimum value then the center location of grid-cell 3, which is $\mathbf{c}_3$ is chosen to be the estimated position for the MDT samples of testing signature $\boldsymbol{u}$. The closed form KLD equation is given by:

$$d(p_u \| p_{i,t}) = \frac{1}{2}\left((\mathbf{u}_u - \hat{\mathbf{u}}_{i,t})^T \hat{\boldsymbol{\Sigma}}_{i,t}^{-1} (\mathbf{u}_u - \hat{\mathbf{u}}_{i,t}) + tr\left(\boldsymbol{\Sigma}_u \hat{\boldsymbol{\Sigma}}_{i,t}^{-1} - \mathbf{I}\right) - \ln\left|\boldsymbol{\Sigma}_u \hat{\boldsymbol{\Sigma}}_{i,t}^{-1}\right| \right) \quad (2)$$

where, $\mathbf{u}_u$, and $\hat{\mathbf{u}}_{i,t}$ corresponds to the mean received signal strength values, while Σ_u and $\Sigma_{i,t}$ represents the covariance matrices of the received signal strength values of the testing and training signatures respectively. Here, $tr()$ denotes the trace of matrix, Σ^{-1} denotes the inverse of covariance matrix Σ and $\mathbf{I}$ is the identity matrix. It is a non-symmetric measure of the difference between testing and training signature probability distributions p_u and $p_{i,t}$.

III. Grid-cell Layout Optimization using Genetic Algorithm

A. Motivation

From our previous work in this field it is clear that GCL optimization is necessary to make RF fingerprinting an adaptive positioning technique which would deliver its optimal performance in both 68%-ile and 95%-ile of PE despite changes in the BS positions, surrounding structures or amount of MDT training samples used [10],[11],[12]. Here our goal is to autonomously select a GCL which would deliver optimal positioning performance in both the 68%-ile and 95%-ile value of PE. From simulation results we found that using square GCs between 20m-by-20m to 40m-by-40m, for different square Grid-cell layouts the 68 and 95 percentiles of positioning error tends to be conflicting - when 68%-ile value of PE decreases then 95%-ile value increases and vice-versa. Therefore we were motivated to use a popular Multi-Objective Evolutionary Algorithm (MOEA) know as Multi-Objective GA (MOGA) to select the proper GCL for RF fingerprinting, where the objective functions are 68th and 95th percentiles of positioning error. The reason for employing MOGA is that other stochastic search strategies (e.g., simulated annealing, ant colony optimization or particle swarm optimization) do not guarantee to find the true Pareto optimal set but, instead, aim to generate a good approximation of such set in a reasonable computational time. On the other hand, MOEAs are particularly well-suited to solve multi-objective problems because they operate over a set of potential solutions. This feature allows them to generate several elements of the Pareto optimal set in a single run. Furthermore, MOEAs are less susceptible to the shape or continuity of the Pareto front than traditional mathematical programming techniques, require little domain information and are relatively easy to implement and use.

B. Genetic Algorithm Optimized RF Fingerprinting

The term genetic algorithms refer to a subset of evolutionary algorithms that model biological processes to optimize highly complex cost functions. GA is capable of yielding a robust search by implicitly sampling hyper-plane partitions of a search space. A single hyper-plane, commonly referred to as schema, is the theoretical foundation on which the algorithm was developed as first introduced by John Holland in 1975 [15]. Here a population represents a group of potential solution points and a generation represents an algorithmic iteration. A chromosome is comparable to a design point and a gene is comparable to a component of the design vector. Given a population of designs, three basic operations are applied: reproduction, crossover, and mutation. Reproduction involves selecting design vectors from the current generation to be used in the next generation and whether or not a design is selected depends on its fitness value. Fitness, which is determined by a fitness function, is an indication of how desirable a design is in terms of surviving into the next generation. The selection probability represents the chance for survival and is proportional to a design's fitness value. Once a new generation of designs is determined, crossover is conducted as a means to introduce variations into the population of designs. Crossover is the process of combining or mixing two different designs. The next operation, which also is used to introduce variations into the population, is mutation. It is a random process that entails altering part of a design's genetic string. In our simulations we have used the multi-objective GA function 'gamultiobj' of Matlab R2014a which uses a controlled elitist genetic algorithm, a variant of NSGA-II [16]. This controlled elitist GA favors individuals that can help increase the diversity of the population even if they have a lower fitness value. It is very important to maintain the diversity of population for convergence to an optimal Pareto front. This is done by controlling the elite members of the population as the algorithm progresses. Two options 'ParetoFraction' and 'DistanceFcn' are used to control the elitism. The Pareto fraction option limits the number of individuals on the Pareto front and the distance function helps to maintain diversity on a front by favoring individuals that are relatively far away on the front. Here a crowding distance for each member is calculated and it is used in the selection process in order to spread the solutions along the Pareto front.

In the present work two approaches were followed to optimize the GCL using MOGA:

(1) Non-overlapping GCL (NoGCL) approach: Here the length of a chromosome is twice the number obtained dividing the length of the area of interest by the lower bound. The first half genes of a chromosome are allocated as lengths of the grid-cell units (GCUs) sequentially in a row one after another from left to right. While the second half genes are for the heights of the corresponding grid-cell units. In this approach MOGA tries to select GCUs design which is replicated

column-wise to cover the whole area of interest. For the example shown in Fig. 2, the NoGCL approach will use a chromosome length of 60 genes for 10m-by-10m square GCL. Using MOGA operators the GAFP searches for the optimal GCL which consists of different rectangular GCUs along a row while he same GCU is used along a column.

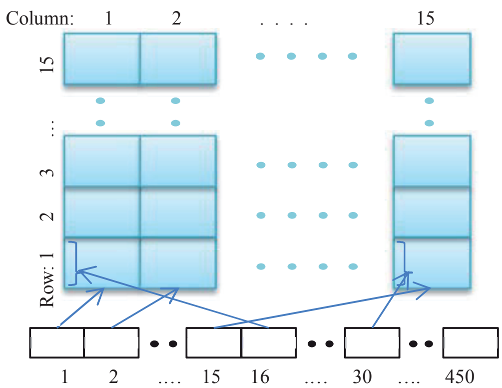


Figure 2. Allocation of a chromosome for overlapping GCL approach

(2) Overlapping GCL (OGCL) approach: In the second approach the total gene number of a chromosome equals the multiplications of the numbers obtained dividing the length and height of the area of interest by the lower bound. The chromosome allocation used in the present research is depicted in Fig. 2 where the area of interest spans over 300m-by-300m and only 20m-by-20m square GCUs constructs the GCL. So in this case there are 15 GCUs along the rows as well as 15 GCUs along the columns. As we can see from the Fig.2 a chromosome comprises of 450 genes. In each of the chromosomes 30 consecutive genes starting from the left of the chromosome string is allocated to the lowest row of GCUs as indicated by number 1 in the figure, the next 30 consecutive genes belongs to the 2nd row and so on. In each of these sets of genes the first 15 genes are allocated to the lengths of GCUs starting from left side and rest of the 15 genes are allocated to the heights of the corresponding GCUs. In this approach, we have fixed the base-position of each of the GCUs according to the lower bound of the GCU, which is 20m along the columns heights. So, only for the 20m-by-20m square GCU there will be no overlapping, otherwise for all GCL overlapping GCUs will cover the area of interest. Hence for any chromosome GCU layout begins from row 1 according to the first 30 consecutive genes and then the next 30 genes are allocated to row 2 where for each of the GCU the lower base is 20m above the base of row1. Hence there will be fixed number of overlapping GCUs along the columns, except the 20m-by20m GCL. The OGCL approach is computationally more expensive as compared to that of the NoGCL approach.

Initial population was generated using real valued chromosome genes. For the NoGCL approach the lower and upper bounds of the GCU were 10m and 30m respectively. For the OGCL approach we have selected the bounds to be 20m and 40m respectively in order to reduce the chromosome length. In both the approaches initial population contains GCL having square GCUs ranging between the lower and upper bounds covering the whole area of interest while the rest of the chromosomes comprise of random real valued genes within the lower and upper bounds of the respective approaches. It is worth mentioning that for creating the training signatures only those GUCs are selected which fall inside the area of interest, since for various chromosome structures some GCL will cover far bigger area as compared to our area of interest.

The MOGA operation is given by the following flowchart:

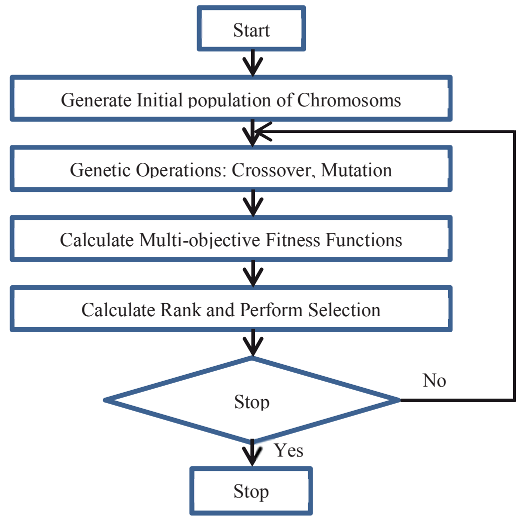


Figure 3. A flowchart of MOGA

The fitness function of the proposed GAFP method is given below:

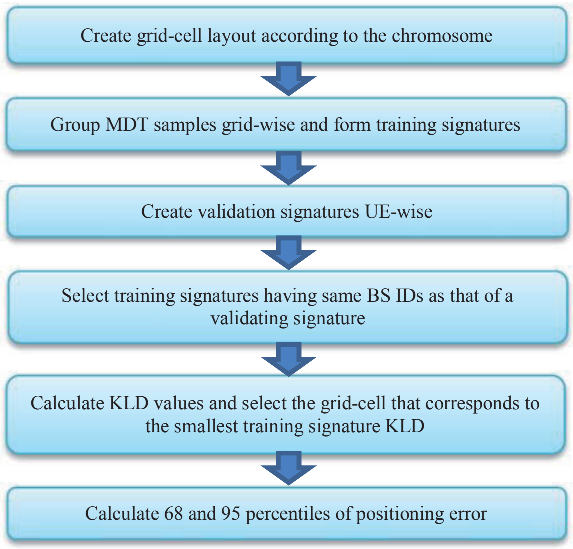


Figure 4. A block diagram representation of GAFP fitness function

We have divided the total number of MDT samples obtained from the area of interest into three sets: (i) training data, (ii) validation data and (iii) test data. As we can see from Fig.4, GAFP uses the training data to create training signatures and the validation data to calculate the fitness values of the chromosomes in different generations. Finally the test data is used to evaluate the performance of the GAFP optimized GCL.

GAFP delivers no single optimum solution but a series of equally optimal ones. In order to select a single solution we have adopted the approach of the evolution line as a posteriori preference articulation technique that uses weights to prioritize objectives after the Pareto front has been obtained [19]. The evaluation line is drawn in the objective space using a weighted function that reflects the preferences towards the objectives. The advantage of the function is that it leans the evaluation line towards the preferred objectives. Here Pareto-optimal solution is evaluated based on its point-line distance from the evaluation line. This criterion overcomes the issue of non-convex areas of the Pareto front, as each solution is evaluated based on its closeness to the evaluation line. In our simulations equal weight was given to the objective functions to select the optimal solution while evaluating the GAFP positioning accuracy.

IV. Performance Evaluation

A. Simulation Scenario

In this study we have performed test simulations in heterogeneous small cell (HSC) network scenario as illustrated in Fig. 5. It was simulated by using a state-of-the-art LTE Rel'10 dynamic system simulator to model both the downlink and the uplink in an OFDM symbol resolution with several radio resource management, scheduling, mobility, handover and traffic modeling functionalities. Simulation parameters and mathematical models are based on the 3GPP TR 36.839 specifications, defining parameterization for used bandwidth, center frequency, path-loss, slow fading, and fast fading 0. Moreover, UE measurements e.g., RSRP, were implemented in the simulator taking into account the technical requirements for the absolute and relative measurement errors and -6 dB Ês/Iot cell detection criterion as in 0.

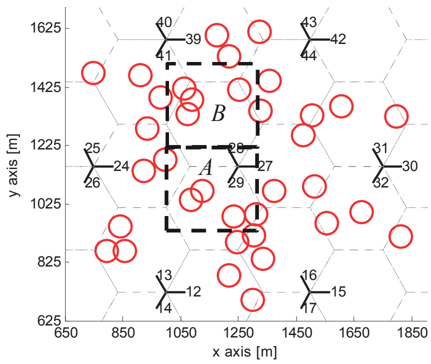


Figure 5. Heterogeneous small cell scenario

The HSC network consisted of 57 macro BSs having inter-site distance of 500m and operating on CC0 band. In addition, 36 small cells with omni-directional antennas were randomly deployed in the coverage area of 12 centermost macro BSs as depicted in Fig. 5 with red circles. Distance to the nearest inter-site small cell varied from 50m to 170m with average distance being 95m. Here, UEs were moving only in the area of 12 centermost macro BSs but were able to monitor all detected cells.

In order to check the robustness of our proposed method the total area of interest which is 300m-by-600m has been equally divided into two squares, each of 300m-by-300m. In Fig 5 these square areas are indicated by alphabetic letters- *A* and *B* with dotted black lines showing their boundaries. Within area-*A* there were 36184 MDT samples and within area-*B* the number of MDT samples are 36375. In total, 1623 randomly moving outdoor vehicular UEs (30 km/h) were distributed uniformly to the simulation area, where 100% resource block loading was used for creating interference limited simulation environment which is more challenging from cell detection point of view. Traffic profile consisted of data generated by MDT reports which were sent once per second. More details about the used simulation parameters can be found in [11].

B. MOGA Parameters

The parameters used in the MOGA are summerized in Table I.

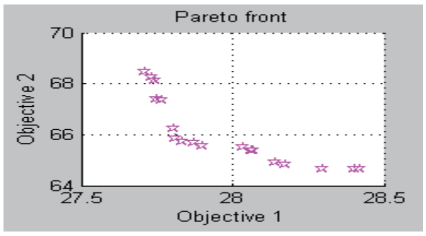
TABLE I. MOGA Parameters used in Simulations

Parameters	Type/Value
Selection type	Tournament
Crossover type and	Scattered
Crossover fraction	0.8
Mutation function	Constraint dependent
Fitness functions	68 percentile value of PE, 95 percentile value of PE
Chromosome length	60 and 450 for Non-overlapping and Overlapping approaches respectively
Population size	60 and 100
Stopping criteria	200 generations and spread of Pareto solutions less than tolerance: 0.0001
Pareto fraction	0.35

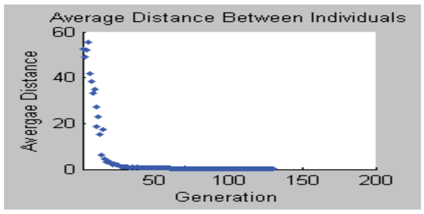
C. Simulation Results

Robustness Evaluation: For this purpose we have chosen two different areas- *A* and *B* with different training, validation and testing datasets in the simulations. In both of the simulations only 10% training data were used to form the training signatures, 30% MDT samples were used for validation purpose of the GAFP fitness evaluation and rest of 60% data for testing the optimized GCL. Here we have used the computationally less expensive NoGCL approach with a GA population size of 50; it has taken 127 and 132 GA generations for area *A* and *B* respectively. A comparison study between CGCL and NoGCL methods positioning performance is shown in Table II. Here we have results from three different CGCL grid-cell layouts: 30m-by30m, 20m-by-20m and 10m-by-10m following the conventional grid-based RF fingerprinting. We have chosen these CGCL grid-cell layouts as the grid-cell size parameter used in NoGCL simulations range between 30m to 10m. The Pareto front and average distance between chromosomes are shown in Fig. 6(a) and 6(b) respectively, obtained from NoGCL simulation in area *B*. We can see from the 68%-ile and 95%-ile results in Table II that the GAFP NoGCL approach gives less PE as compared to CGCL in both the areas *A* and *B*. In area B the NoGCL approach shows an improvement of 10.5% in the 95%-ile PE over that of the 10m-by-10m CGCL method. The analyzed test MDT sample percentage is bit higher in area *A* and bit lower in area *B* for the NoGCL than that of the CGCL method. The third row in Table II shows the performance results for the whole area of interest, i.e. combined area *A* and *B*. Here we have used the optimized GCL solutions from the NoGCL method simulated separately in areas *A* and *B*. Here also the NoGCL has better positioning accuracies in both 68%-ile and 95%-ile as compared to the CGCL. All these results show the effectiveness of the proposed GAFP method despite changes in cellular network structure.

Next we have evaluated the GAFP positioning accuracy through increasing amount of training data: 50% data for training signature creation, 25% for validation and 25% for testing. In this case area *A* was chosen for simulation and both NoGCL and OGCL approaches were used. In both methods the GA population size was 100 and number of GA generations was 200. The results are shown in Table III; here we have three different CGCL simulations: 40m-by-40m, 30m-by-30m and 20m-by-20m. For both the NoGCL and OGCL approaches the grid-cell size ranges between 40m to 20m. Here we can find that the with CGCL method the lowest PE values in 68%-ile and 95%-ile are 44.03 and 127.18 meters whereas for the GAFP NoGCL approach we have much reduced PE values of 24.52 and 55.01 meters respectively. Hence the proposed method shows an excellent improvement in positioning accuracy of 44.31% and 56.74% in the 68%-ile and 95%-ile as compared to the conventional grid-based RF fingerprinting.



(a)



(b)

Figure 6. (a) Pareto front obtained from NoGCL method
(b) Average distance between chromosomes

TABLE II. POSITIONING ACCURACY PERFORMANCE EVALUATION OF CGCL AND GAFP NOGCL METHODS WITH 10% TRAINING DATA

Area of Interest	Positioning Error 68 %-ile (m)				Positioning Error 95 %-ile (m)				Analyzed Test MDT Samples (%)			
	CGCL			*GAFP NoGCL*	*CGCL*			*GAFP NoGCL*	*CGCL*			*GAFP NoGCL*
	30 m	*20 m*	*10 m*		*30 m*	*20 m*	*10 m*		*30 m*	*20 m*	*10 m*	
AREA A	51.53	49.21	47.27	44.96	131.64	128.47	130.58	115.75	57.99	57.33	55.95	59.94
AREA B	40.12	34.93	33.02	30.73	93.29	85.21	76.32	65.85	52.84	56.16	52.60	52.82
AREA AB	45.99	40.05	42.13	38.07	115.02	109.32	110.27	98.49	55.83	54.63	57.06	56.84

TABLE III. PERFORMANCE EVALUATION OF CGCL, GAFP NOGCL AND GAFP OGCL METHODS WITH 50% TRAINING DATA

GAFP Type	Positioning Error 68 %-ile (m)				Positioning Error 95 %-ile (m)				Analyzed Test MDT Samples (%)			
	CGCL			*GAFP*	*CGCL*			*GAFP*	*CGCL*			*GAFP*
	40 m	*30 m*	*20 m*		*40 m*	*30 m*	*20 m*		*40 m*	*30 m*	*20 m*	
NoGCL	48.25	47.58	44.03	24.52	127.18	152.70	135.96	55.01	63.07	62.40	64.53	65.32
OGCL	48.25	47.58	44.03	26.37	127.18	152.70	135.96	58.13	63.07	62.40	64.53	71.77

It can be seen from Table III that positioning accuracies shown by GAFP OGCL is bit less than the NoGCL but it has analyzed 6 more percentage of test samples as compared NoGCL method. The histogram of the chosen chromosome from OGCL method in shown in Fig. 7. As we can see the widths and heights of the grid-cell units are mostly chosen to be in-between 21 to 27 meters.

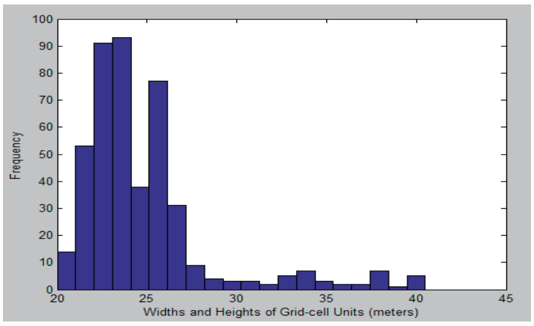


Figure 7. Histogram of the selected chromosome from GAFP OGCL method simulated in area *A*.

V. Conclusion

In this paper we propose a novel method to improve the positioning accuracy of grid-based RF fingerprinting through autonomous calibration of grid-cell layout. For this purpose multi-objective GA is used for selecting the best possible grid-cell layout over the area of interest so that RF fingerprinting gives the optimal output. Comparison study with conventional grid-based RF fingerprinting shows that GA optimized RF fingerprinting has the ability to calibrate the grid-cell layout in an optimal way for improved RF fingerprinting despite the change in the cellular network. It was found that with sufficient amount of training data the proposed method is able provide very good positioning accuracy as compared to that of the conventional RF fingerprinting.

Acknowledgement

The authors would like to thank colleagues from University of Jyvaskyla and Magister Solutions Ltd. for their constructive criticism, comments and support.

References

[1] Catherine Mulligan, "EPC and 4G Packet Networks: Driving the Mobile Broadband Revolution," ISBN: 978-0-12-394595-2, Academic Press, 2013.

[2] Ben-Moshe B., E. Elkin, H. Levi, A. Weissman, 'Improving Accuracy of GNSS Devices in Urban Canyons', *In Proc. of 23rd Canadian Conference on Computational Geometry (CCCG'11)*, August 10-12, 2011.

[3] D. Milioris et al., "Low-dimensional signal-strength fingerprint-based positioning in wireless LANs," Ad Hoc Networks, 2012, doi:10.1016/j.adhoc.2011.12.006.

[4] Chan E. C. L., G. Baciu, S.C. Mak, "Wireless Tracking Analysis in Location Fingerprinting," in *Proc. of IEEE International Conference on Wireless & Mobile Computing, Networking & Communication,* October 12-14, 2008.

[5] P. Bahl, V.N Padmanabhan, "Radar: an in-building rf-based user location and tracking system," in *Proc. of Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, March 2000, pp. 775 –784.

[6] Zhu J., G. D. Durgin, "Indoor/Outdoor Location of Cellular Handsets Based on Received Signal Strength," in *Proc. of Vehicular Technology Conference*, May 30 - 1 June, 2005.

[7] Laitinen H., J. Lahteenmaki, T. Nordstrom, "Database Correlation Method for GSM Location," in *Proc. of Vehicular Technology Conference*, May 6-9, 2001.

[8] Wuri A. et al., "Minimization of Drive Tests Solution in 3GPP," *in IEEE Communications Magazine*, Vol. 50, No. 6, June 2012.

[9] Johansson J., Hapsari W.A, Kelley S. and Bodog G., "Minimization of drive tests in 3GPP release 11," in *IEEE Communications Magazine*, Vol. 50, No. 11, pp. 36-43, November 2012.

[10] Mondal R., Turkka J., Ristaniemi T. and Henttonen T., "Performance Evaluation of MDT RF Fingerprinting Framework," in *Proc. of 7th International Conference on Mobile Computing and Ubiquitous Networking (ICMU2014)*, January 2014, Singapore.

[11] Mondal R., Turkka J., Ristaniemi T. and Henttonen T, "Positioning in Heterogeneous Small Cell Networks using MDT RF Fingerprints," in *Proc. of the First IEEE International Black Sea Conference on Communications and Networking*, July 2013, Batumi, Georgia.

[12] R. Mondal, J. Turkka and T. Ristaniemi, "An Efficient Grid-based RF Fingerprint Positioning Algorithm for User Location Estimation in Heterogeneous Small Cell Networks," in *Proc. of International Conference on Localization and GNSS (ICL-GNSS)*, Helsinki, Finland, June 24-26, 2014.

[13] Campos R. S. and L. Lovisolo, "Mobile Station Location using Genetic Algorithm Optimized Radio Frequency Fngerprinting," in *Proc. of International Telecommunications Symposium*, Manaus, Brazil, September 2010.

[14] M. J. Magro and C. J. Debono, "A Genetic Algorithm Approach to User Location Estimation in UMTS Networks," in *EUROCON - The International Conference on Computer as a Tool*, Warsaw, Poland, pp. 1136 – 1139, September 2007 .

[15] J. Holland, "Adaptation in Natural and Artificial Systems," University of Michigan Press, 1975.

[16] Deb, Kalyanmoy, "Multi-Objective Optimization using Evolutionary Algorithms," John Wiley & Sons, Ltd, Chichester, England, 2001.

[17] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects," version 9.0.0, March 2009, available at http://www.3gpp.org.

[18] 3GPP TS 36.133, "Requirements for support of radio resource management," version 11.4.0, March 2013, available at http://www.3gpp.org.

[19] Vergidis. T., Vergidis. K., Tiwari. A., "The evaluation line: A posteriori preference articulation approach," *IEEE Congress on Evolutionary Computation*, pp. 2694 - 2700, 2008.

PV

GENERIC ARCHITECTURE FOR MINIMIZING DRIVE TESTS IN HETEROGENEOUS NETWORKS

by

Tuomas Hiltunen, Riaz Uddin Mondal, Jussi Turkka and Tapani Ristaniemi 2015

IEEE Vehicular Technology Conference (VTC Fall), Boston, USA

PVI

PERFORMANCE EVALUATION OF LTE RADIO FINGERPRINTING USING FIELD MEASUREMENTS

by

Jussi Turkka, Tuomas Hiltunen, Riaz Uddin Mondal and Tapani Ristaniemi 2015

International Symposium on Wireless Communication Systems (ISWCS'15),
Brussels, Belgium

PVII

AN EFFICIENT CLUSTER-BASED OUTDOOR USER POSITIONING USING LTE AND WLAN SIGNAL STRENGTHS

by

Riaz Uddin Mondal, Jussi Turkka and Tapani Ristaniemi 2015

IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC 2015), Hong Kong

An Efficient Cluster-Based Outdoor User Positioning Using LTE and WLAN Signal Strengths

Riaz Uddin Mondal[1], Jussi Turkka[2], Tapani Ristaniemi[1]
[1]Department of Mathematical Information Technology, University of Jyvaskyla, Jyvaskyla, Finland
[2]Magister Solutions Ltd, Tampere, Finland
riaz.u.mondal@student.jyu.fi, jussi.turkka@magister.fi, tapani.ristaniemi@jyu.fi

Abstract— **In this paper we propose a novel cluster-based RF fingerprinting method for outdoor user-equipment (UE) positioning using both LTE and WLAN signals. It uses a simple cost effective agglomerative hierarchical clustering with Davies-Bouldin criterion to select the optimal cluster number. The positioning method does not require training signature formation prior to UE position estimation phase. It is capable of reducing the search space for clustering operation by using LTE cell-ID searching criteria. This enables the method to estimate UE positioning in short time with less computational expense. To validate the cluster-based positioning real-time field measurements were collected using readily available cellular mobile handset equipped with Nemo Handy software. Output results of the proposed method were compared with a single grid-cell layout based RF fingerprinting method. Simulation results show that if a single LTE and six WLAN signal strengths are used then the proposed method can improve positioning accuracy of 35% over the grid-based RF fingerprinting.**

Keywords-component; LTE cell-ID; Grid-based RF fingerprinting; Hierarchical Clustering; Minimization of Drive Tests.

I. INTRODUCTION

Location is a vital component in consumer services like social media, search, advertising and navigation. For authorities, mobile location is mandatory for emergency-call location, and can also be used for road-traffic management and machine-to-machine purposes. GPS-based consumer navigation devices have reached mass market status, benefits are undeniable: it is ubiquitous and always available, high-accuracy positioning [1]. However, GPS has two major drawbacks: the signals broadcast by the satellites are too weak to be received indoors in places such as shopping malls, and in dense urban environments not enough satellites are visible to obtain positioning fixes in a reasonable time. The popularity of IEEE 802.11 infrastructures, their low deployment cost, and the advantages of using them for both communication and positioning, make them an attractive choice. Therefore, authors in [2] have proposed a portable positioning system, that utilizes both GPS and Wi-Fi-based pattern matching methods to estimate the position [2][3]. To improve this combined GPS and Wi-Fi-based Pattern Matching Method, in [4] authors proposed to assign weights to different weather conditions, determined the position of the mobile terminal by the Euclidean distance, and adjusted the weights according to the environment. Wi-Fi positioning system based on fingerprinting was evaluated in the Sydney CBD area where Wi-Fi APs are densely deployed and test results show that it works well for outdoor localization with errors in the tens of meters [4]. Also in [5] authors have carried out experimental analysis for outdoor fingerprinting system, implemented over the WLAN and demonstrated that it is feasible to perform outdoor positioning with reasonable accuracy using 802.11-based positioning. A three-phase methodology (measurement, calibration and estimation) for locating mobile stations (MS) in an indoor environment using wireless technology was proposed in [6] where combination of fingerprint and cluster based positioning system was developed to overcome the problem of the relative effect of doors and walls on signal strength and the system is independent of the hardware technology manufacturer. A new algorithm was proposed in [7] for enhancing the performance of adaptive enhanced cell-ID (AECID) fingerprint positioning in LTE, where clustering was employed to increase the accuracy of the polygon computation scheme of the AECID algorithm. The basic positioning method in most cellular communication systems is the cell-identity (cell-ID) method which has the advantage of short response time and thus it fulfills the time to first fix (TTFF) requirement for E-911 emergency positioning in the North American market which is specified to be below 30s. This method is applicable in all situations where there is cellular coverage.

One major requirement of RF fingerprint based positioning is to create and maintain the big correlation database in order to update the training fingerprints with surrounding structural and environmental changes. Operators usually conduct extensive and expensive periodical drive test campaigns to fulfill this requirement. The operational expenditure (OPEX) associated with traditional drive tests can be removed by a feature introduced in 3GPP Release 10, known as Minimization of Drive Tests (MDT) which enables operators to utilize users' equipment to collect radio measurements and associated location information [8]. MDT provides a framework for gathering user reported location-aware radio

The present work was carried out within the framework of European Celtic-Plus project SHARING (Self-organized Heterogeneous Advanced Radio Networks Generation).

measurements from commercial mobile phones that can be used for creating and maintaining such training database. This procedure allows operators to collect radio measurements, i.e. received signal strength and quality, with UE location information and a time stamp [9].

In our previous work grid-cell based RF fingerprinting (GRFFP) has shown good positioning accuracy in dense urban scenario using MDT samples obtained from a dynamic LTE system simulator [10][11]. To improve the user equipment (UE) positioning accuracy using grid-based RF fingerprinting, weighted Kullback-Leibler Divergence based overlapping grid-cell layout method was proposed in [12]. However GRFFP delivers good positioning when two requirements are fulfilled: (i) training signatures need to be updated in regular interval of time, (ii) an optimal grid-cell layout needs to be chosen for different cellular network scenario and for the amount of available training MDT samples.

In this study we propose a simple cluster-based RF fingerprinting (CRFFP) method which does not go through any training phase to estimate UE position. It uses MDT samples comprising of both LTE and Wi-Fi signals; we refer to this as *generalized MDT* (GMDT). The CRFFP takes advantage of the LTE serving cell-ID based searching technique to deliver UE positioning in short time. Here we also analyze the UE positioning accuracy with three different combinations of LTE and Wi-Fi signals and results were compared with the traditional GRFFP method.

The following section contains a brief description of the GMDT field measurements used in this study and then the conventional GRFFP method is explained. In Section III, first description of the proposed CRFFP is given, then test results obtained with GRFFP and CRFFP positioning methods are presented, and finally concluding remarks are given.

II. Grid-based RF Fingerprinting

A. Generalized MDT

According to 3GPP specifications MDT enables the operation, administration, and maintenance (OAM) system to collect radio measurements from the UE, together with location information if available when the measurements are taken [8].

Here we propose a GMDT that is capable of collecting Wi-Fi signal strengths along with the LTE and UE location information. There are two main reasons behind this: (i) RF fingerprinting gives very good positioning accuracy using Wi-Fi signal strengths in outdoors (especially in dense urban areas), (ii) to decrease the search space in CRFFP positioning which also shortens the operation time. In order to create the GMDT database we have used Samsung Galaxy S3 (LTE capable) which was installed with a handheld drive test software application- Nemo Handy. This application is very suitable for performing measurements both outdoors and indoor spaces while the device being simultaneously used as a regular mobile phone [13]. LTE reference signal received power (RSRP) and WLAN received signal strength indicator (RSSI) measurements were recorded from a residential urban area in Tampere, Finland during September 2014 as shown in Figure 1. Two measurement campaigns were done for 800 and 1800 MHz LTE bands, in the 1800 MHz case, inter-frequency measurements were also reported according to the measurement configuration provided by the network. Hence in this study we have used GMDT samples from LTE 1800 MHz measurements.

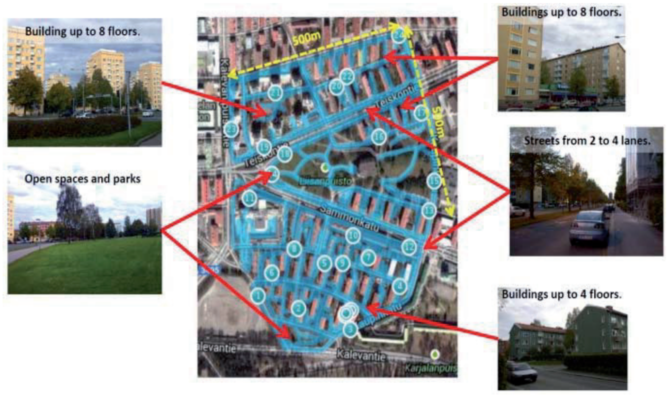


Figure 1: GMDT field measurement area in Tampere, Finland

More than 150 kilo-metres of measurements were collected by feet, bicycle and car covering approximately an area of 0.33 square kilo-metres. In all measurements, the route was repeated at least twice to ensure that enough measurement samples are collected for each grid unit.

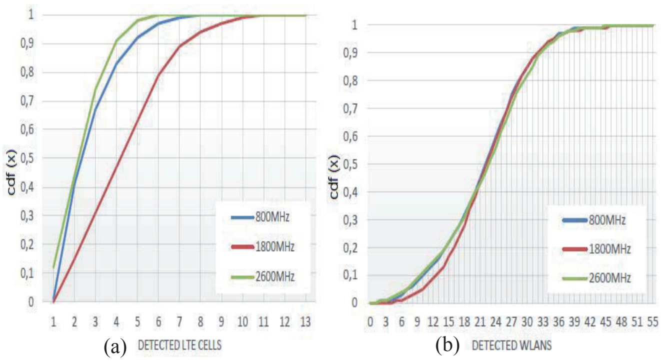


Figure 2: (a) Number of detected LTE cells and (b) WLAN APs per measurement sample.

As we can find from Figure 2 that every GMDT sample contains at least 1 serving LTE base-station (BS) signal and 98% of the samples comprises of more than 5 WLAN access points (AP). A study conducted in [14], measured the significance of Wi-Fi APs for UE position estimation where good results were obtained by limiting the Wi-Fi AP number to seven for all analyzed samples. In [15] the selection of APs was done based on the largest signal strength values recorded at each location. Hence we were motivated to use seven signals in total including both LTE BSs and WLAN APs. So, every GMDT contains the serving LTE BS ID of the recording mobile handset. Both LTE and WLAN signals were sorted in descending order of signal strength values. We have also used three different sets of GMDT samples by choosing different combinations of LTE and WLAN signals from the total database.

A set of GMDT measurement can be defined by:

$$M_j = \{s_{j,1}, s_{j,2}, \dots, s_{j,N}\} \quad (1)$$

where, j=1, 2 and 3 referring different GMDT set, N is the total number of measurement samples of a particular set. The nth GMDT sample of a set can be presented by a row vector:

$$s_{j,n} = \{LW_{ID}, RSS_{LW}, P_{XY}\} \quad (2)$$

where, LW_{ID} denotes the LTE BS IDs and WLAN AP IDs, RSS_{LW} stands for the corresponding RSRP and RSSI values, and P_{XY} contains the x-y coordinates of the UE obtained from GNSS position information.

B. A Simple Grid-based RF Fingerprinting Method

Here we have used a single grid-cell layout based RF fingerprinting method by segmenting the whole geographical area of interest with square grid-cell units (GCU). We have used Euclidean distance to measure the statistical difference between training fingerprints and test samples since it has previously been used for outdoor RF fingerprinting in order to obtain good UE positioning accuracy [16].

Training Phase of GRFFP: In conventional GRFFP method multiple training signatures are formed within a single GCU [10][11]. To reduce the searching time to find the best match training signature for a test sample and also to reduce the related computational cost, a single training signature ($Train_{Sig}$) is created from all the training GMDT samples (GMDTs) that belong to a one particular GCU. The $Train_{Sig}$ formed from all the GMDT samples of ith GCU ($GMDTs_i^{All}$) is define as follows:

$$Train^{i}_{Sig} = \{TS^{LW}_{ID}, RSS^{LW}_{TS}, P^{XY}_{Ref}\} \quad (3)$$

where, TS^{LW}_{ID} contains all unique LTE BS IDs and WLAN AP IDs obtained from $GMDTs_i^{All}$, RSS^{LW}_{TS} is a vector of the corresponding LTE RSRP and WLAN RSSI values, and P^{XY}_{Ref} is the reference x-y coordinate calculated from the mean values of x and y coordinates of $GMDTs_i^{All}$.

An example training signature creation and test phase matching of GRFFP method is illustrated in Figure 3. Here for simplicity only two GCUs are shown, the blue dots inside a GCU represent its GMDT samples and the small red triangle depicts the corresponding reference position. We can find from Figure 3 that GCU 2 has two samples: $GMDT_2^1$ and $GMDT_2^2$ represented by two row vectors. Here black squares containing LW_{ID} (L1 indicates LTE BS ID number 1, W1 is for WLAN AP ID-1), the corresponding RSS_{LW} values (S_L^1 indicates RSRP of LTE BS ID-1, S_W^1 is RSSI of WLAN AP ID-1) are within the green squares and P_{XY} are inside the blue squares. For GCU 2 a single training signature- $Train^2_{Sig}$ has been created from $GMDT_2^1$ and $GMDT_2^2$. It has three parts: (i) TS^{LW}_{ID} contains all unique LTE and WLAN IDs. (ii) RSS^{LW}_{TS} comprises of mean RSRP and RSSI values for common LTE and WLAN IDs, otherwise the RSRP values are copied from either $GMDT_2^1$ or $GMDT_2^2$. And (iii) P^{XY}_{Ref} gives the reference x-y coordinates calculated from the mean x and y coordinates of $GMDT_2^1$ and $GMDT_2^2$.

Test Phase of GRFFP: To test a GMDT we first compare its LTE and WLAN IDs with all the training signatures available and select those signatures which meet the least matching threshold. For example in Figure 3, $Train^2_{Sig}$ contains four matching IDs: L1, L4, W2 and W5 which are common to the Test sample ($Test_{Sam}$), hence we have 57% of ID match between $Test_{Sam}$ and $Train^2_{Sig}$. Now if the minimum matching threshold is set to 50% then $Train^2_{Sig}$ is selected for distance measurement. As shown by red dotted arrows in Figure 3, only common RSRP and RSSI values are used to calculate the Euclidean distance between $Test_{Sam}$ and $Train^2_{Sig}$. A simplified Mahalanobis distance equation is used for distance calculation where the inverse covariance matrix is replace by an identity matrix:

$$d(TestSam, TrainSig) = \sqrt{(\boldsymbol{u_{Te}} - \boldsymbol{u_{Tr}})^T \boldsymbol{I} (\boldsymbol{u_{Te}} - \boldsymbol{u_{Tr}})} \quad (4)$$

where, $\boldsymbol{u_{Te}}$ and $\boldsymbol{u_{Tr}}$ denotes the RSRP and RSSI values of the $Test_{Sam}$ and a selected $Train_{Sig}$ respectively and $\boldsymbol{I}$ is the identity matrix. After separate calculation of all the distances between a $Test_{Sam}$ and the selected training signatures; the $Train_{Sig}$ corresponding to the smallest Euclidean distance is chosen for positioning purpose. The estimated position of that $Test_{Sam}$ is given by P^{XY}_{Ref} of the chosen $Train_{Sig}$.

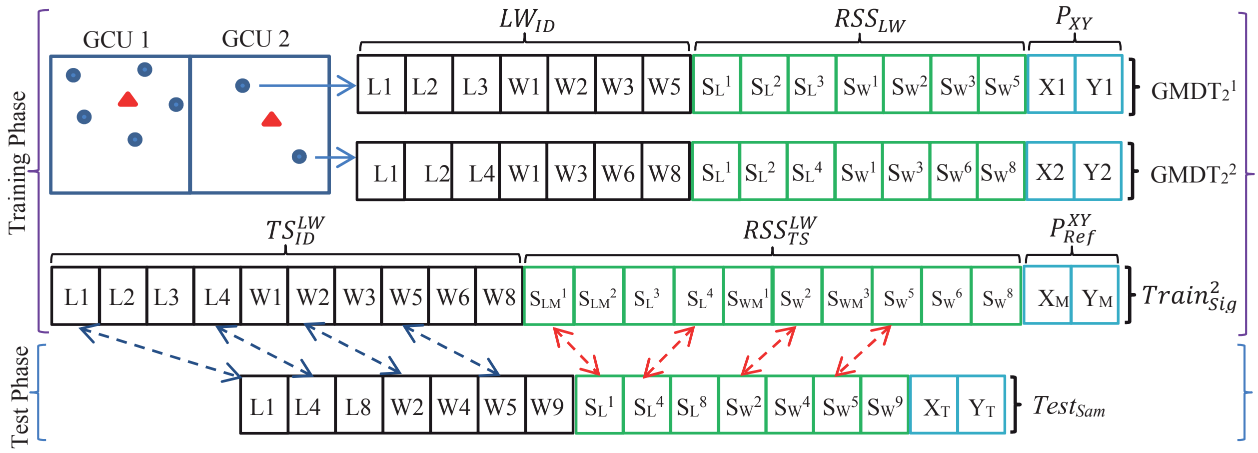


Figure 3: Training and test phases of grid-cell based RF fingerprinting

III. CLUSTER-BASED RF FINGERPRINTING

A. An Efficient Cluster-based Positioning Algorithm

At first the GMDT samples of the total data-base are sorted into different GMDT groups according to the serving LTE BS ID. For testing a sample the group that matches the serving LTE BS ID of that $Test_{Sam}$ is selected. From this selected group GMDT samples are selected which fulfill the least matching threshold- the matching is similar to the one described in section II (B), the only difference is that here matching is between $Test_{Sam}$ and a GMDT sample of the selected group. Now for the clustering purpose, the RSS_{TS}^{LW} values of the $Test_{Sam}$ and the selected GMDTs are put together in the same pool. We have used a simple agglomerative hierarchical clustering with Davies-Bouldin criterion to select the optimal cluster number [17]. This criterion is based on a ratio of within-cluster and between-cluster distances. The Davies-Bouldin index (DB) is defined by the follow equation:

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}\{D_{i,j}\} \qquad (5)$$

where, $D_{i,j}$ is the within-to-between cluster distance ratio for the ith and jth clusters. $D_{i,j}$ is given by,

$$D_{i,j} = \frac{\left(d_i^- + d_j^-\right)}{d_{i,j}} \qquad (6)$$

where, d_i^- is the average distance between each point in the ith cluster and the centroid of the ith cluster. d_j^- is the average distance between each point in the ith cluster and the centroid of the jth cluster. $d_{i,j}$ is the Euclidean distance between the centroids of the ith and jth clusters. The optimal cluster number is obtained between 1 to 6 clusters using the smallest Davies-Bouldin index value. After multiple clusters are formed, clustering criteria (CC) is checked: the cluster which contains the $Test_{Sam}$ must have two GMDTs. If CC is met, then the cluster that contains the $Test_{Sam}$ is selected and $Test_{Sam}$ UE position is calculated from the mean x-y coordinates of all GMDTs of that cluster. If CC is not fulfilled the matching threshold is reduced and clustering is performed again in order to analyze the $Test_{Sam}$. Thus CRFF method does not need any prior training before the test phase; it utilizes the cell-ID advantages in reducing the search space thereby reduces position estimation time. Hence it offers a computationally less expensive RF fingerprinting method which can be implemented in real-time using GMDT. CRFFP positioning method is described in Figure 4.

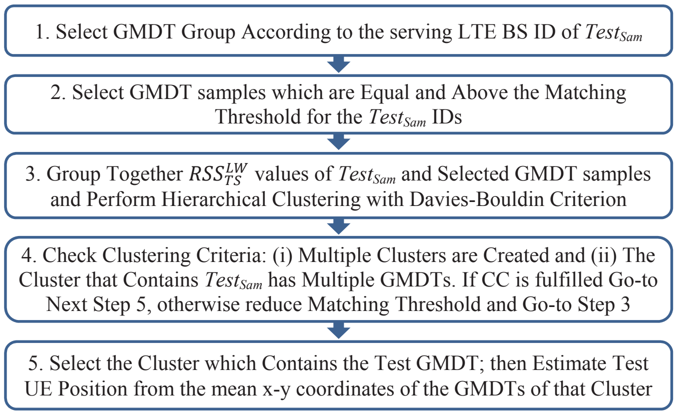


Figure 4: Block-diagram of the CRFFP Positioning Method

B. Experimental Results: Outdoor UE positioning

In the total GMDT data-set we have merged multiple samples into a single one which contain similar LTE BS ID and WLAN AP ID and were recorded from the same x-y coordinate. In order to avoid over-optimal results consecutive GMDTs have been grouped into chunks of 20 samples in sequence. Training and test data-sets were created by randomly choosing such data chunks.

TABLE I: RESULTS OF GRFFP AND CRFFP METHODS USING LTE AND WLAN SIGNALS

LTE and WLAN Combination	Matc-hing Thres -hold	Using Total Test GMDT Samples						Using Common Test GMDTs between GRFFP and CRFFP				
		GRFFP			CRFFP			GRFFP		CRFFP		Comm. Test GMDT (%)
		68% PE (m)	95% PE (m)	Test GMDT (%)	68% PE (m)	95% PE (m)	Test GMDT (%)	68% PE (m)	95% PE (m)	68% PE (m)	95% PE (m)	
LTE BS:3 & WLAN AP:4	**80%**	16.91	45.41	85.86	9.40	33.74	21.17	14.21	39.77	9.40	33.75	21.15
	60%	17.65	47.78	97.68	10.39	35.84	43.60	14.02	40.35	10.40	35.85	43.59
	40%	17.88	48.99	99.96	14.16	51.17	61.77	15.59	42.57	14.16	51.17	61.77
LTE BS:2 & WLAN AP:5	**80%**	15.16	42.69	85.36	7.58	26.39	21.94	11.98	36.04	7.61	26.45	21.83
	60%	16.04	44.76	97.51	7.94	27.59	46.04	11.81	36.03	7.95	27.60	46.02
	40%	16.33	45.92	100	9.14	33.88	62.37	12.66	38.32	9.14	33.88	62.37
LTE BS:1 & WLAN AP:6	**80%**	14.29	40.31	85.43	7.33	22.80	24.59	11.47	33.69	7.36	22.81	24.38
	60%	15.09	42.54	97.02	7.24	20.56	51.35	10.60	31.76	7.24	20.58	51.32
	40%	15.35	43.42	99.97	7.80	24.34	69.21	11.17	33.89	7.80	24.34	69.21

In the simulations we have used 23080 training GMDTs and 2565 samples were tested. The GRFFP method uses a 10m-by-10m grid-cell layout which was chosen from several square grid-cell layouts according to the delivered positioning accuracy. Tenfold cross-validation method was used to obtain positioning results for both GRFFP and CRFFP methods with three different GMDT data-sets as shown in Table I. For the 1st GMDT data-set each sample is constructed with maximum 3 LTE BS signals (LTE_{BS}) and 4 WLAN AP signals ($WLAN_{AP}$); in 2nd GMDT data-set there are 2 LTE_{BS} and 5 $WLAN_{AP}$; and the 3rd set comprises of only 1 LTE_{BS} and 6 $WLAN_{AP}$. In Table I the second column indicates the different matching threshold used in both of the methods. After that UE positioning error (PE) results (68%-ile and 95 %-ile values) along with the analyzed test sample percentages are given when both methods use all the test samples. From these results we can find that when 80% matching threshold was used the 68 %-ile and 95 %-ile PEs of CRFFP are lower than that of respective GRFFP PE values, but the GRFFP method has analyzed more test samples than CRFFP. Hence on right side of Table I we have PE results considering only those test samples which were analyzed in both methods. This help to compare the proposed method with the GRFFP in the best possible way. It is found from the simulation results with common test results that when 1st data-set was used, CRFFP has given lower PE in both 68%-ile and 95 %-ile values for 80% and 60% of matching threshold. However when the threshold is lowered to 40%, GRFFP has given better positioning in 95 %-ile than that of CRFFP. With the 2nd data-set CRFFP outperforms GRFFP in both percentile values and also for all three matching thresholds used. The best positioning accuracy given by CRFFP is with the 3rd data-set and for 80 % of threshold: it has shown 35% improvement in positioning accuracy for both the 68%-ile and 95%-iles as compared to that of the GRFFP. It is clear from the results that CRFFP offers better positioning than GRFFP when matching threshold is high and 5 or 6 WLAN signals are used. Hence in dense urban areas where multiple WLAN signals can be detected CRFFP is capable of providing good outdoor UE positioning with GMDT samples.

IV. Conclusion

In this paper we propose a novel cluster-based RF fingerprinting method for outdoor UE positioning which uses LTE and WLAN signals. It provides better positioning accuracy as compared to that of grid-based RF fingerprinting. The benifit of the cluster-based approach is that it uses simple clustering method and no prior training phase for estimating test UE positioning. During cluster operation it reduces the searching space by utilizing LTE cell-ID; thus delivers output result in short time. The proposed method is capable of providing good positioning by using only serving LTE BS signal and six WLAN AP signals. Hence the present research outcome suggests that the next MDT functionality should include WLAN signals into consideration; which would benefit cellular operators to develop cost-effective solutions for developing real-time positioning systems.

Acknowledgement

The authors would like to thank colleagues from University of Jyvaskyla and Magister Solutions Ltd. for their constructive criticism, comments and support.

References

[1] T. S. Rappaport, J. H. Reed, and D. Woerner, "Position location using wireless communications on highways of the future," *IEEE Communications Magazine*, vol. 34, no. 10, pp. 33–41, 1996.

[2] Son H. C., J. G. Lee and G. I. Jee, "Mobile station location using hybrid GPS and a wireless network," The 57th IEEE Semiannual Vehicular Technology Conference 2003, VTC 2003-Spring, vol. 4, pp. 2716 – 2720, April 2003.

[3] Sheng-Cheng Yeh, Wu-Hsiao Hsu, Ching-Hui Chen and Rung-Huei Liang," A Context-aware Blogging and Learning Service in Mobile IPv6 Networks", IEEE Wireless Communications and Networking Conference, March 2007.

[4] Li B., I. J. Quader, A. G. Dempster, "On outdoor positioning with Wi-Fi," *Journal of Global Positioning Systems*, Vol. 7, No. 1, pp. 18-26, 2008.

[5] Liu X., S. Zhang, J. Quan, X. Lin, "The Experimental Analysis of Outdoor Positioning System Based on Fingerprint Approach," 12th IEEE International Conference on Communication Technology (ICCT), 2010.

[6] Mengual L., O. Marbán, S. Eibe, "Clustering-based location in wireless networks," *Expert Systems with Applications,* 37, pp. 6165–6175, 2010.

[7] Wigren, T., "Clustering and Polygon Merging Algorithms for Fingerprinting Positioning in LTE," 5th International Conference on Signal Processing and Communication Systems (ICSPCS), December 2011.

[8] Johansson J., Hapsari W.A, Kelley S. and Bodog G., "Minimization of drive tests in 3GPP release 11", *in IEEE Communications Magazine*, Vol. 50, No. 11, pp. 36-43, November 2012.

[9] Wuri A. et al., "Minimization of Drive Tests Solution in 3GPP", *in IEEE Communications Magazine*, Vol. 50, No. 6, June 2012.

[10] Mondal R., Turkka J., Ristaniemi T. and Henttonen T., "Performance Evaluation of MDT RF Fingerprinting Framework", in *Proc. of 7th International Conference on Mobile Computing and Ubiquitous Networking (ICMU2014)*, Singapore, January 2014.

[11] Mondal R., Turkka J., Ristaniemi T. and Henttonen T, "Positioning in Heterogeneous Small Cell Networks using MDT RF Fingerprints", in *Proc. of the First IEEE International Black Sea Conference on Communications and Networking*, Batumi, Georgia, July 2013.

[12] Mondal, R. U., J. Turkka, T. Ristaniemi, "An Efficient Grid-based RF Fingerprint Positioning Algorithm for User Location Estimation in Heterogeneous Small Cell Networks", International Conference on Localization and GNSS (ICL-GNSS), Helsinki, Finland, June 2014.

[13]Web-link:http://www.anite.com/businesses/network-testing/products/nemo-handy-world%E2%80%99s-most-widely-used-handheld-drive-test-tool#.VT_jNSGeDRY

[14] E. Laitinen, E. S. Lohan, J. Talvitie and S. Shrestha, "Access point significance measures in WLAN-based location", In proc of. Workshop on Positioning, Navigation and Communication, March 2012.

[15] M. Youssef, A. Agrawala, and A. U. Shankar, "WLAN location determination via clustering and probability distributions," in *Pervasive Computing and Communications*, pp. 143–150, 2003.

[16] Kim Jae-H., K. SikMin,2 and W. Y. Yeo, "A Design of Irregular Grid Map for Large-Scale Wi-Fi LAN Fingerprint Positioning Systems," *The Scientific World Journal*, Vol 2014, Article ID 203419, 2014.

[17] Davies, D. L., and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-1, No. 2, pp. 224–227, 1979.

PVIII

CLUSTER-BASED RF FINGERPRINT POSITIONING USING LTE AND WLAN OUTDOOR SIGNALS

by

Riaz Uddin Mondal, Jussi Turkka and Tapani Ristaniemi 2015

10th International Conference on Information, Communications and Signal Processing (ICICS 2015), Singapore

Cluster-Based RF Fingerprint Positioning Using LTE and WLAN Outdoor Signals

Riaz Uddin Mondal, Tapani Ristaniemi
Department of Mathematical Information Technology,
University of Jyvaskyla
Jyvaskyla, Finland
riaz.u.mondal@student.jyu.fi, tapani.ristaniemi@jyu.fi

Jussi Turkka
Magister Solutions Ltd
Tampere, Finland
jussi.turkka@magister.fi

Abstract— **In this paper we evaluate user-equipment (UE) positioning performance of three cluster-based RF fingerprinting methods using LTE and WLAN signals. Real-life LTE and WLAN data were collected for the evaluation purpose using consumer cellular-mobile handset utilizing 'Nemo Handy' drive test software tool. Test results of cluster-based methods were compared to the conventional grid-based RF fingerprinting. The cluster-based methods do not require grid-cell layout and training signature formation as compared to the grid-based method. They utilize LTE cell-ID searching technique to reduce the search space for clustering operation. Thus UE position estimation is done in short time with less computational cost. Among the cluster-based methods Agglomerative Hierarchical Cluster based RF fingerprinting provided best positioning accuracy using a single LTE and six WLAN signal strengths. This method showed an improvement of 42.3 % and 39.8 % in the 68th percentile and 95th percentile of positioning error (PE) over the grid-based RF fingerprinting.**

Keywords- LTE cell-ID; Grid-based RF fingerprinting; K-Nearest Neighbor; Hierarchical Clustering; Fuzzy C-means; Minimization of Drive Tests

I. INTRODUCTION

Over the next decade the integration of location services into our day-to-day life will increase significantly as technologies mature and accuracy improves. Currently, as an accurate and reliable outdoor localization system Global Navigation Satellite System (GNSS) has revolutionized navigation-based applications running on automotive GNSS-enabled devices and smart phones. However, GNSS relies on special hardware support, has high complexity, high battery consumption and the access to GPS signals is limited in some environments, such as urban areas with many high buildings, mountainous terrain and indoor areas [1]. Received Signal Strength (RSS) based fingerprinting localization has been the most widely used technique for user positioning during the last few decades [2-3]. Researchers are studying how to conduct radio signal positioning through signals from existing wireless infrastructure, such as cellular networks [2], WiMaX [3] and WiFi [4-5] networks. The rapid expansion of Wi-Fi access points (AP) across the urban/indoor environments made it possible for researchers to envision alternatives to TOA-based systems. One success story for deployment in the urban environment is Skyhook Wireless [6]. Skyhook realized the potential of exploiting Wi-Fi signals emitted from residential homes and offices that are continuously in use. They have improved localization by building databases of Wi-Fi signatures tied to locations that could be integrated to aid in the localization process. Wi-Fi based fingerprint positioning system was evaluated in the Sydney CBD area and test results show that it works well for outdoor localization with errors in the tens of meters [7]. In [8] authors have carried out outdoor fingerprinting over WLAN and achieved good accuracy using 802.11-based positioning.

In this study we have evaluated cluster-based RF finger-printing approaches which have taken into account four key challenges of fingerprint positioning [5]:

1) RF fingerprint generation: the factors affecting fingerprint generation are the placement and number of survey points and time samples. Most approaches have selected such parameters experimentally. In order to avoid such difficulties we have used real life Minimization of Drive Tests (MDT) date - a feature introduced in 3GPP Release 10 which enables operators to utilize users' equipment to collect radio measurements and associated location information [9].

2) Preprocessing of recorded training data for reducing computational complexity: in [4] authors have proposed an offline clustering of locations aiming to reduce the search space to a single cluster. Chen et al. in [10] consider the similarity of signal values, as well as the covering APs, to generate a set of clusters using K-means to improve the power efficiency of mobile devices. Both of the above clustering techniques are carried out offline based on the training data. This hampers the operation of the system over time since WLAN infrastructures are highly dynamic and APs can be easily moved or discarded, in contrast to their base-station counterparts in cellular systems, which generally remain intact for long periods of time [5]. Therefore, we have used a combination of LTE and WLAN signal strengths, *generalized MDT* (GMDT) which gives us the opportunity to use LTE serving cell-ID based searching technique to deliver user-equipment (UE) positioning in short time with less computational cost.

3) Selection of APs for use in positioning: in a typical dense urban WLAN environment the number of available APs is much higher than three and using all available APs increases the computational complexity of the positioning algorithm. In this research we have chosen seven LTE and WLAN signals for position estimation which has been found to be effective from previous Wi-Fi positioning results [11].

4) User equipment (UE) position estimate based on a new RSS observation: in the simplest case, the Euclidean distance is used to find the distance between the new RSS observation and the center of the training RSS vectors at each survey point or grid cell units [12][13]. However, choosing an optimal grid-cell layout requires computational power and training time [14]. Hence, we have selected cluster-based RF fingerprinting (CRFFP) methods: K-Nearest Neighbor (KNN), Agglomerative Hierarchical Clustering (AHC) and Fuzzy C-Means (FCM) which do not create training signatures through grid-cell layout during the training phase. To verify the effectiveness of CRFFP methods UE positioning results were compared to that of the conventional grid-cell based RF fingerprint positioning (GRFFP).

The rest of the paper is organized as follows: Section II contains a brief description of the recorded GMDT field measurements and then the conventional GRFFP method is described. In section III we explain three different CRFFP methods. Experimental test results of GRFFP and CRFFP methods are shown in section IV. Finally we draw some concluding remarks in section V.

II. Grid-based RF Fingerprinting using GMDT

A. Generalized MDT Measurements

Drive tests are the main source for collecting measurement data from cellular networks which is costly and time consuming. The problem that drive tests need human effort to collect measurement data and that only spot measurements can be performed, has led to automated solutions which include the UEs from the end user. The feature for this evolution in the 3GPP standard is named MDT [14]. Here we were motivated to use GMDT data which is an enhancement to the LTE Minimization of Drive Tests architecture allowing the collection of location-aware radio measurements from WLAN access networks as well. Grid-based RF fingerprinting test results show that GMDT data containing only the single strongest WLAN measurement in addition to the LTE RF fingerprint can improve the 67th percentile location accuracy from 88.2 m to 49.4 m [15]. The GMDT database were created with the help of a popular drive test software application known as Nemo Handy installed in Samsung Galaxy S3 (LTE capable) [16]. This handheld drive test tool is very suitable for performing measurements both outdoors and in crowded indoor spaces while being simultaneously used as a regular mobile phone. In our research we have recorded reference signal received power (RSRP) values of LTE serving and neighboring base station (BS) signals and received signal strength indicator (RSSI) values of WLAN APs. About 150 kilo-metres of measurements were collected by feet, bicycle and car covering approximately an area of 0.33 square kilo-metres of a residential urban area in Tampere, Finland during September 2014 as shown in Fig. 1.

Fig. 1: GMDT field measurement area in Tampere, Finland

The GMDT samples used in this study were from LTE 1800 MHz measurements, in which 800 MHz inter-frequency measurements were also reported according to the measurement configuration provided by the network. Every route was repeated at least twice to ensure that enough measurement samples were collected for each grid unit. From the measurements we have found that all the GMDT samples contain at least one serving LTE BS RSRP and 98% of the samples comprises of more than five WLAN RSSI values. Authors in [15] have selected WLAN APs based on the largest signal strength values recorded at each location. Hence we have chosen seven signal strength values in total including both LTE RSRPs and WLAN RSSIs. Both RSRP and RSSI values were sorted in descending order of signal strength values. We were interested to see how different combinations of LTE and WLAN signals affect the UE positioning performance using the same fingerprinting method. Thus three different sets of GMDT samples were created by choosing different combinations of LTE and WLAN signals from the total database. A GMDT measurement set is defined by:

$$M_j = \{s_{j,1}, s_{j,2}, \ldots, s_{j,N}\} \tag{1}$$

where, j=1, 2 and 3 refers to the different GMDT sets, N is the total number of measurement samples of any particular set. The nth GMDT sample of a set is given by a row vector:

$$S_{j,n} = \{LW_{ID}, RSS_{LW}, P_{XY}\} \tag{2}$$

where, LW_{ID} denotes the LTE BS IDs and WLAN AP IDs, RSS_{LW} comprises of the corresponding RSRP and RSSI values, and P_{XY} contains the x-y coordinates of the UE obtained from GNSS information.

B. Grid-cell based RF Fingerprint Positioning

A conventional single grid-cell layout based RF fingerprinting method was used, which segmented the whole geographical area of interest into 10m-by-10m square grid-cell units (GCU). Euclidean distance was used to measure the statistical difference between training fingerprints and test samples, as previous WLAN-based UE positioning research suggests it to be effective [17].

Training Phase: To reduce the searching time of the best match training signature for a test sample and also to reduce the related computational cost, a single training signature ($Train_{Sig}$) is created from all the training GMDT samples that belong to a particular GCU. The $Train_{Sig}$ formed from all the GMDT samples of ith GCU ($GMDTs_i^{All}$) is defined as follows:

$$Train^i_{sig} = \{TS_{ID}^{LW}, RSS_{TS}^{LW}, P_{Ref}^{XY}\} \quad (3)$$

where, TS_{ID}^{LW} contains all unique LTE BS IDs and WLAN AP IDs obtained from $GMDTs_i^{All}$, RSS_{TS}^{LW} is a vector of the corresponding mean LTE RSRP and WLAN RSSI values, and P_{Ref}^{XY} is the reference x-y coordinate calculated from the mean values of x and y coordinates of $GMDTs_i^{All}$.

Test Phase: To test a GMDT sample we first compare its LTE and WLAN IDs with all the training signatures available and select those signatures which meet a least matching threshold. The minimum matching threshold (MT) was set to three, so in this case all the training signatures that contain at least three or higher number of LTE and WLAN IDs similar to that of test sample will be chosen. The maximum MT number was set to six. A simplified Mahalanobis distance equation is used for distance calculation where the inverse covariance matrix is replaced by an identity matrix:

$$d(Test_{Sam}, Train_{Sig}) = \sqrt{\{ (\mathbf{u_{Te}} - \mathbf{u_{Tr}})^T \mathbf{I} (\mathbf{u_{Te}} - \mathbf{u_{Tr}}) \}} \quad (4)$$

where, $\mathbf{u_{Te}}$ and $\mathbf{u_{Tr}}$ denotes the RSRP and RSSI values of the $Test_{Sam}$ and a selected $Train_{Sig}$ respectively and $\mathbf{I}$ is the identity matrix. After separate calculation of all the distances between a $Test_{Sam}$ and the selected training signatures; the $Train_{Sig}$ corresponding to the smallest Euclidean distance is chosen for test UE positioning. The estimated position of that $Test_{Sam}$ is given by P_{Ref}^{XY} of the chosen $Train_{Sig}$.

III. CLUSTER-BASED RF FINGERPRINT POSITIONING

A. *K-nearest Neighbors Cluster-Based Positioning*

KNN is one of the basic algorithms used for UE positioning using RF fingerprint [18]. In this work we have chosen K to be 5 which has given good positioning result in WLAN positioning performed in [19]. Here the only processing required during the data collection phase is to group the GMDT samples according to the LTE serving BS ID. During the positioning phase the first task is to choose the group of GMDT samples according to the LTE serving BS ID of the test GMDT sample. Then for selecting training GMDT samples ($Train_{Sam}$) we start with the highest MT number: 7 and select n $Train_{Sam}$ which match with the $Test_{Sam}$ IDs. If we do not get any $Train_{Sam}$ corresponding to the chosen MT then MT is lowered to the next integer number and select n $Train_{Sam}$ that matches with the $Test_{Sam}$ IDs. This process continues until we get multiple matched $Train_{Sam}$ or the lowest MT is reached. Now Euclidean distance is used to choose five closest GMDTs with the KNN algorithm:

$$D_{Train_{Sam}, Test_{Sam}} = \sqrt{\{\sum_{j=1}^{n} (Train_{RSS} - Test_{RSS}) \}} \quad (5)$$

where, $Train_{RSS}$ and $Test_{RSS}$ are vectors of LTE RSRP and WLAN RSSI values of $Train_{Sam}$ and $Test_{Sam}$ respectively.

The test UE position is estimated from the mean x-y coordinate value of the five selected $Train_{Sam}$.

B. *Agglomerative Hirarchical Cluster-based Positioning*

The AHC clustering method uses Davies-Bouldin criterion to select the optimal cluster number [20]. This criterion is based on a ratio of within-cluster and between-cluster distances. The Davies-Bouldin index (DB) is defined by the follow equation:

$$DB = (1/k)\{\sum_{i=1}^{k} \max_{j\neq i} (D_{i,j}) \} \quad (6)$$

where, k is the number of clusters, $D_{i,j}$ is the within-to-between cluster distance ratio for the ith and jth clusters. $D_{i,j}$ is given by: $D_{i,j} = (\bar{d_i} + \bar{d_j})/d_{i,j}$ (7)

where, $\bar{d_i}$ is the average distance between each point in the ith cluster and the centroid of the ith cluster. $\bar{d_j}$ is the average distance between each point in the ith cluster and the centroid of the jth cluster. $d_{i,j}$ is the Euclidean distance between the centroids of the ith and jth clusters. During evaluation optimal cluster number is set between 1 to 6 using the smallest Davies-Bouldin index value. When multiple clusters are formed, clustering criteria (CC) is followed: the cluster which contains the $Test_{Sam}$ must contain at least two GMDTs. AHC-based positioning method is described in Fig. 2.

1. Select GMDT Group According to the serving LTE BS ID of $Test_{Sam}$

2. Select GMDT samples which are Equal and Above the Matching Threshold for the $Test_{Sam}$ IDs

3. Group Together RSS_{TS}^{LW} values of $Test_{Sam}$ and Selected GMDT samples and Perform AHC Clustering with Davies-Bouldin Criterion

4. Check Cluster Criteria: (i) Multiple Clusters are Created and (ii) The Cluster that Contains $Test_{Sam}$ has Multiple GMDTs. If CC is fulfilled Go-to Next Step 5, otherwise reduce Matching Threshold and Go-to Step 3

5. Select the Cluster which Contains the Test GMDT; then Estimate Test UE Position from mean x-y coordinates of the GMDTs of that Cluster

Fig. 2: Block-diagram of the AHC-based Positioning Method

C. *Fuzzy C-Means Cluster-Based Positioning*

FCM has effectively been used in WLAN indoor localization [21]. Here we have used it for outdoor positioning using GMDT data. Its implementation steps are similar to that of the AHC-based fingerprint positioning. In this method in step 3 as shown in Fig. 2, we have added another criterion that if the number of selected GMDT samples is more than six than initial number of clusters assigned to FCM method is 6 otherwise 2. FCM starts with an initial guess for the cluster centers, which are intended to mark the mean location of each cluster and it also assigns every data point a membership grade for each cluster. By iteratively updating the cluster centers and the membership grades for each data point, it moves the cluster centers to the right location. This iteration is based on minimizing the objective function for the partition of the selected GMDT data-set [22]:

$$J_m(u,v) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{i,k}^{m} \| D_k - v_i \|^2 \quad (8)$$

where, J_m is the objective function, n is the number of samples in the data set, c is the number of clusters ($1 \le c \le n$), $u_{i,k}$ is the element of partition matrix U of size (c x n) containing the membership function, v_i is the center of the i^{th}

cluster, and m is a weighting factor that controls fuzziness of the membership function. The matrix U is constrained to contain elements in the range of [0, 1] such that $\sum_{i=1}^{c} u_{ik} = 1$ for each $u_{i,k}(1 \leq k \leq n)$. The norm $\|D_k - v_i\|$ is the distance between the sample D_k and the clusters center v_i .

IV. EXPERIMENTAL RESULTS: OUTDOOR UE POSITIONING

Before the positioning phase we have processed the whole GMDT data-set by merging multiple samples into a single one which were recorded from the same x-y coordinate and also contain similar LTE BS and WLAN AP IDs. In order to avoid over-optimal results consecutive GMDTs have been grouped into chunks of 20 samples in sequence. Training and test data-sets were created by randomly choosing such data chunks. Test results were derived from three different GMDT sets: The 1st set contains 3 LTE and 4 WLAN signals, 2nd set consists of 2 LTE and 5 WLAN signals and 3rd set is for 1 LTE and 6 WLAN signals. The number of training and test GMDTs were 23080 and 2565 respectively. Results shown in Table I were obtained from 10 fold cross-validations for testing all GMDTs. The 1st and 2nd columns of Table I corresponds to the different LTE-WLAN sets and the matching threshold numbers and after that we can find the 68th and 95th percentile values of UE positioning error for each of the methods. The analyzed test GMDT percentage is attached to positioning error (PE) values for each of the methods. Table I shows that GCL method is capable of analyzing almost 100% of test GMDT and the PE values for any given data set remain the same for different MT values. The KNN and FCM perform better for MT-6 and similar results were obtained for MT-5 case as compared to that of the GCL method, both KNN and FCM have analyzed less percentages of test samples than that of GCL. Thus for a better comparison between the methods we have prepared Table II where the PE results of all four methods are calculated for common analyzed test GMDT samples as given in the last column of Table II.

TABLE I. RESULTS OF GRFFP AND CRFFP METHODS USING ALL GMDT TEST DATA

LTE & WLAN	MT. No.	GRFFP			KNN			AHC			FCM		
		68% PE (m)	*95% PE (m)*	*Test GMDT (%)*	*68% PE (m)*	*95% PE (m)*	*Test GMDT (%)*	*68% PE (m)*	*95% PE (m)*	*Test GMDT (%)*	*68% PE (m)*	*95% PE (m)*	*Test GMDT (%)*
LTE:3 & WLAN:4	**6**	17.89	49.17	99.52	12.95	40.34	63.12	9.34	33.66	19.05	12.35	38.16	30.02
	5	17.88	49.06	99.89	15.76	47.55	89.27	10.49	35.64	41.13	16.56	51.45	68.49
	4	17.89	49.05	99.99	16.86	50.29	98.02	12.60	42.30	55.96	18.51	56.83	91.21
	3	17.89	49.05	99.99	17.15	51.56	99.87	14.16	51.18	61.75	19.28	59.58	98.93
LTE:2 & WLAN:5	**6**	16.29	45.82	99.72	11.00	36.83	66.38	7.45	25.63	21.57	8.89	33.58	29.89
	5	16.30	45.85	99.86	14.88	44.75	90.92	7.87	26.88	45.78	13.48	43.46	69.89
	4	16.31	45.89	99.90	15.92	46.75	98.60	8.64	31.31	58.32	15.56	46.57	93.13
	3	16.33	45.92	99.99	16.15	47.20	99.87	9.13	33.87	62.35	16.11	48.25	99.09
LTE:1 & WLAN:6	**6**	15.28	43.32	99.57	9.75	33.25	64.03	7.27	21.38	24.16	7.80	28.64	29.42
	5	15.31	43.34	99.69	13.48	41.87	89.47	7.18	19.91	51.02	11.01	35.98	67.21
	4	15.31	43.34	99.83	15.02	45.56	98.61	7.49	22.41	65.09	13.37	42.91	92.60
	3	15.32	43.36	99.87	15.22	46.07	99.75	7.77	24.09	69.12	14.01	44.57	98.77

TABLE II. RESULTS OF GRFFP AND CRFFP METHODS USING COMMON GMDT TEST DATA

LTE & WLAN	MT. No.	GRFFP		KNN		AHC		FCM		*Common Test GMDT (%)*
		68% PE (m)	*95% PE (m)*	*68% PE (m)*	*95% PE (m)*	*68% PE (m)*	*95% PE (m)*	*68% PE (m)*	*95% PE (m)*	
LTE:3 & WLAN:4	**6**	14.20	39.70	13.23	39.68	9.35	33.78	9.84	34.63	18.71
	5	14.06	40.22	13.30	40.43	10.50	35.73	12.01	40.47	41.02
	4	15.27	41.99	13.74	41.51	12.61	42.32	13.55	44.03	55.92
	3	15.58	42.57	14.07	42.12	14.16	51.18	14.35	47.18	61.75
LTE:2 & WLAN:5	**6**	14.58	40.63	10.61	33.99	7.48	25.69	7.52	26.51	20.90
	5	13.58	40.94	11.08	37.54	7.87	26.98	8.94	33.96	45.41
	4	13.87	41.74	11.40	37.94	8.66	31.34	9.77	36.05	58.22
	3	14.22	41.86	11.73	38.68	9.13	33.87	10.25	37.68	62.33
LTE:1 & WLAN:6	**6**	14.06	38.24	9.32	31.39	7.27	21.80	7.35	25.53	22.06
	5	13.14	38.00	9.73	32.78	7.18	20.16	8.08	27.71	49.56
	4	13.19	39.30	10.12	36.08	7.49	22.44	8.93	32.53	64.76
	3	13.48	40.05	10.46	36.54	7.77	24.09	9.22	33.34	69.11

Here it is found that with the 1st GMDT set PE results of GCL and KNN are very similar. AHC and FCM have shown better positioning results for MT number of 5 and 6, but for MT-3 and MT-4 PEs are higher than that of GCL and KNN. For data set 2, KNN has reduced the PEs a bit in both the PE-percentile values as compared to that of GCL; whereas FCM has shown further improvement in 68th percentile and 95th percentile of PE values than that of KNN method. With the same data set AHC has outperformed FCM based positioning performance. Also with data set 1, AHC based clustering has given the best positioning accuracy as compared the other methods. For MT-3 in data set 1 common analyzed test GMDT percentage was the highest 69.11 and in this case AHC method has shown an improvement in positioning accuracy of 42.3% and 39.8 % in the 68th percentile and 95th percentile values of PE as compared to that of the GCL method. These results indicate that in dense urban areas where multiple WLAN signals can be detected CRFFP is capable of delivering better outdoor UE positioning than GRFFP using data set 2 or 3 having seven LTE and WLAN signals strength values.

V. CONCLUSION

Here we proposed cluster-based RF fingerprinting methods for outdoor UE positioning using seven LTE and WLAN signals. Only seven LTE RSRP and WLAN RSSI signal strength values were used to perform a comparative study between the GRFFP and CRFFP methods. The cluster-based positioning reduces the search space by taking advantage of LTE cell-ID based searching technique. As a result CRFFP methods give faster UE positioning output with less computational cost. They do not need training signature formation during offline phase as compared to the conventional GRFFP method. Among the three CRFFP methods used in this research AHC based RF fingerprinting provided the best UE positioning accuracy using only serving LTE RSRP and six WLAN RSSI values. Hence as a cost-effective real time RF fingerprinting method, AHC based positioning would certainly be a good choice for any cellular operator.

ACKNOWLEDGMENT

The authors would like to thank colleagues from University of Jyvaskyla and Magister Solutions Ltd. for their constructive criticism, comments and support.

REFERENCES

[1] Li K., J. Bigham, L.Tokarchuk and E. L. Bodanese, “A probabilistic approach to outdoor localization using clustering and principal component transformations”, 9th International Wireless Communications and Mobile Computing Conference, 1-5 July 2013.

[2] Mazuelas S., F. Lago, J. Blas, A. Bahillo, P. Fernandez, R. Lorenzo and E. Abril, “Prior NLOS measurement correction for positioning in cellular wireless networks”, IEEE Trans. Veh. Technol., vol. 58, no. 5, pp. 2585-2591, Jun. 2009.

[3] Bshara M., U. Orguner, F. Gustafsson and L. Van Biesen, “Fingerprinting localization in wireless networks based on received-signal-strength measurements: A case study on WiMAX networks”, IEEE Trans. Veh. Technol., vol. 59, no. 1, pp. 283-294, Jan. 2010.

[4] M. Youssef, A. Agrawala and A.U. Shankar, “WLAN Location Determination via Clustering and Probability Distributions,” 1st IEEE Int. Conf. on Pervasive Computing and Commun., pp. 143-150, Texas, USA, 2003.

[5] A. Kushki, K. N. Plataniotis and A. N. Venetsanopoulos, “Kernel-Based Positioning in Wireless Local Area Networks”, IEEE Trans. on Mobile Computing, vol. 6, no. 6, pp. 689-705, June 2007.

[6] http://www.skyhookwireless.com/about-skyhook

[7] Li B., I. J. Quader, A. G. Dempster, “On outdoor positioning with Wi-Fi,” *Journal of Global Positioning Systems*, Vol. 7, No. 1, pp. 18-26, 2008.

[8] Liu X., S. Zhang, J. Quan, X. Lin, “The Experimental Analysis of Outdoor Positioning System Based on Fingerprint Approach,” 12th IEEE International Conference on Communication Technology, pp. 369 - 372, China, Nov. 2010.

[9] Johansson J., Hapsari W.A, Kelley S. and Bodog G., “Minimization of drive tests in 3GPP release 11”, *in IEEE Communications Magazine*, Vol. 50, No. 11, pp. 36-43, November 2012.

[10] Y. Chen, Q. Yang, J. Yin, and X. Chai, “Power-Efficient Access-Point Selection for Indoor Location Estimation,” IEEE Trans. Knowledge and Data Eng., vol. 18, no. 7, pp. 877-888, July 2006.

[11] E. Laitinen, E. S. Lohan, J. Talvitie and S. Shrestha, “Access point significance measures in WLAN-based location”, In proc of. Workshop on Positioning, Navigation and Communication, Germany, pp. 24-29, March 2012.

[12] K. Kaemarungsi and P. Krishnamurthy, “Modeling of Indoor Positioning Systems Based on Location Fingerprinting,” Proc. INFOCOM, vol. 2, pp. 1012-1022, March 2004.

[13] Mondal, R. U., J. Turkka, T. Ristaniemi, “An Efficient Grid-based RF Fingerprint Positioning Algorithm for User Location Estimation in Heterogeneous Small Cell Networks”, International Conference on Localization and GNSS (ICL-GNSS), Helsinki, Finland, June 2014.

[14] Johansson J., Hapsari W.A, Kelley S. and Bodog G., “Minimization of drive tests in 3GPP release 11”, *in IEEE Communications Magazine*, Vol. 50, No. 11, pp. 36-43, November 2012.

[15] Hiltunen T., R. U. Mondal, J. Turkka, T. Ristaniemi, “ Generic Architecture for Minimizing Drive Tests in Heterogeneous Networks”, IEEE 82nd Vehicular TechnologyConference (VTC-fall), USA, 6–9 September 2015.

[16] Weblink:http://www.anite.com/businesses/network-testing/products/nemo-handy-world’s-most-widely-used-handheld-drive-test-tool#.Vc8_nPmqpBd

[17] C. Feng, W. S. A. Au, S. Valaee, and Z. Tan, “Received-signalstrength-based indoor positioning using compressive sensing”, *IEEE Transactions onMobile Computing*, vol. 11,no. 12,pp. 1983–1993, 2012.

[18] Bahl, P., and V.N. Padmanabhan,: “RADAR: An in-building RFbased user location and tracking system”, Proc. of IEEE Infocom 2000, Vol. 2, pp. 775–784, Tel Aviv, Israel, March 2000.

[19] Li B., Y. Wang, H.K. Lee, A.G. Dempster, and C. Rizos, “Method for yielding a database of location fingerprints in WLAN**”,** *IEE Proceedings-Communications, Volume 152, Issue 5*, pp. 580-586, October 2005.

[20] Davies, D. L., and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-1, No. 2, pp. 224–227, 1979.

[21] Zhou H. , N. N. Van, “Indoor Fingerprint Localization Based on Fuzzy C-means Clustering”, 6th International Conference on Measuring Technology and Mechatronics Automation, China, pp. 337 - 340, 10-11 Jan. 2014.

[22] Bezdec, J.C., “Pattern Recognition with Fuzzy Objective Function Algorithms”, Plenum Press, New York, 1981.

PIX

CLUSTER-BASED RF FINGERPRINT POSITIONING USING LTE AND WLAN SIGNAL STRENGTHS

by

Riaz Uddin Mondal, Tapani Ristaniemi and Jussi Turkka 2017

International Journal of Wireless Information Networks (IJWI), Springer

Int J Wireless Inf Networks (2017) 24:413–423
DOI 10.1007/s10776-017-0369-9

Cluster-Based RF Fingerprint Positioning Using LTE and WLAN Signal Strengths

Riaz Uddin Mondal[1] · **Tapani Ristaniemi**[1] · **Jussi Turkka**[2]

Received: 27 December 2016 / Accepted: 8 August 2017 / Published online: 16 August 2017

Abstract Wireless Local Area Network (WLAN) positioning has become a popular localization system due to its low-cost installation and widespread availability of WLAN access points. Traditional grid-based radio frequency (RF) fingerprinting (GRFF) suffers from two drawbacks. First it requires costly and non-efficient data collection and updating procedure; secondly the method goes through time-consuming data pre-processing before it outputs user position. This paper proposes Cluster-based RF Fingerprinting (CRFF) to overcome these limitations by using modified Minimization of Drive Tests data which can be autonomously collected by cellular operators from their subscribers. The effect of environmental changes and device variation on positioning accuracy has been carried out. Experimental results show that even under these variations CRFF can improve positioning accuracy by 15.46 and 22.30% in 95 percentile of positioning error as compared to that of GRFF and K-nearest neighbour methods respectively.

Keywords RF fingerprint positioning · K-nearest neighbors · K-means clustering · Hierarchical clustering · Fuzzy C-means clustering

1 Introduction

Location systems have long been identified as an important component of a wide set of applications such as for E-911 emergency positioning, personal navigation and Location-Based Services in outdoor environments. The role of a positioning system is to estimate and report geographical location information pertaining to the user for the purposes of management, enhancement, and personalization of services. At present Global Navigation Satellite System (GNSS) is the most popular positioning system for mobile devices in outdoor environments. However, GNSS geolocation performs poorly in dense urban areas and inside buildings, where satellites are not visible by mobile user equipment (UE) [1]. With the rapid increase in Wireless Local Area Network (WLAN) access points (AP) in metropolitan areas and due to their ubiquitous coverage in large environments, outdoor location systems based on WLAN have gained recent attention in research and commercial applications [2–4]. WLAN positioning works better than GNSS in dense metropolitan areas, both outdoors and indoors owing to its greater received signal strength and lower attenuation [3]. WLAN received signal strength (RSS) measurements can be obtained relatively effortlessly and inexpensively without the need for additional hardware [5]. Moreover, RSS-based positioning is non-invasive, as all sensing tasks can be carried out on the mobile UE, eliminating the necessity for central processing [6]. Skyhook [7] has used Wi-Fi signals emitted from residential homes and offices to build a cost-effective location system on a global scale. Several existing WLAN methods have aimed to use theoretical path loss (PL) models whose parameters are estimated based on training data [8]. Given an RSS measurement and PL model, the distances from the UE to at least three APs are determined, and trilateration is used to obtain the UE position. The limitations of such an approach are the dependence on prior topological information and assumption of isotropic RSS contours [9]. Alternatively, the RSS-position relationship has been characterized implicitly using a training-based

✉ Riaz Uddin Mondal
riaz.u.mondal@student.jyu.fi

1 University of Jyväskylä, Jyväskylä, Finland

2 European Communications Engineering Ltd., Espoo, Finland

method known as location fingerprinting. Positioning results from urban and sub-urban areas with WCDMA and GSM networks in [10] shows that radio-frequency (RF) fingerprinting is a better method than PL model based localization. An RF fingerprint-based positioning system has two phases. First, offline training phase: RSS and corresponding location data are collected to create a 'radio map' with sufficient representation of spatiotemporal RSS properties of the area. Second, online location determination phase: the system uses the signal strength samples received from a test UE to 'search' the radio map to estimate the user location.

In order to enhance WLAN RSS based indoor positioning pedestrian dead reckoning (PDR) is often used. PDR uses an inertial measurement unit (IMU) which has three-axis accelerometers and gyroscopes to detect a user direction changes between footsteps. The user heading change is computed by projecting the gyroscope measurements to the horizontal plane. Authors [42] have proposed a novel linear model for PDR and compared it to conventional nonlinear models. For this purpose they have used Kalman filter (KF), the extended Kalman filter (EKF), and the unscented Kalman filter (UKF). The evaluation shows that despite being simpler than the traditional methods, it performs especially well in situations where the initial heading and position are not known.

In this work, cluster-based RF fingerprinting (CRFF) method is used with data similar to Minimization of Drive Tests (MDT) data [11]. CRFF method divides a group of a MDT data-set into a certain number of subsets or clusters, so that the members in the same cluster are similar in terms of their RSS values. The proposed CRFF confronts the following main challenges of RF fingerprint based UE positioning:

1.1 RF Fingerprint Collection and Updating

The conventional way of creating fingerprint training database is to periodically conduct extensive drive test campaigns which are time-consuming and unpractical for building a metropolitan-scale radio map of the locating system [12, 41]. A major drawback of this method is to update the training radio map when new APs are deployed and existing APs are decommissioned. The accuracy of any location estimation system is highly dependent on the density of the set of collected fingerprints which is difficult to achieve through conventional drive test methods [13]. To solve this issue we have used generalized MDT (GMDT) data that allows UEs to collect location-aware radio measurements from LTE BSs as well as WLAN access networks [14]. GMDT allows cellular operators to collect and update big RF fingerprint data-base autonomously using subscribers UE without any additional hardware instalment. This is the most cost effective solution to build and maintain fine-grained radio map to increase the accuracy of UE localization.

1.2 Pre-processing of Training Data

In most cellular-communication systems the basic positioning method is based upon cell-identity (cell-ID) which reports the identity of the cell to which the terminal is connected to [15]. It has sort response time but the accuracy is low [16]. Author in [17] has proposed an adaptive enhanced cell-ID localization method which uses an offline cluster based fingerprinting to enhance the positioning performance. To reduce computational complexity and search space in WLAN positioning authors in [18] and [19] have conducted offline clustering of locations based on the training data. However the operation of these systems are hampered over time since WLAN infrastructures are highly dynamic and APs can be easily moved or discarded, in contrast to the BS counterparts in cellular systems, which generally remain intact for long periods of time. Our proposed CRFF method utilizes GMDT data to output result in sort time and does not go through time consuming training data processing phase.

1.3 AP Selection for UE Positioning

In a typical urban environment, the number of detected WLAN APs is greater than usually necessary for UE position estimation. RSS is dependent on the relative distance of the UE and each AP. It is affected by the topology of the surrounding environment in terms of obstacles causing non line-of-sight RF signal propagation; thus subsets of available APs may report correlated readings. Hence considering all available APs for position estimation increases the computational complexity of the positioning algorithm [6]. To simplify the training data collection process we have adopted the 'Maximum RSS' (MRSS) based selection methodology where APs are sorted in descending order based on their maximum RSS value and a certain part is chosen to create the training database [20].

1.4 Position Estimation Using New RSS Observation and Radio Map

This essentially involves a distance calculation between the RSS observation of a test UE and the training records; Euclidean distance has been used in this study [21]. UE location estimation using RSS measurements is a difficult task due to the noisy characteristics of signal propagation and absorption by surrounding structures and human bodies. Even changes in the environmental conditions, such as temperature or humidity, affect the signals to a large

Table 1 Summary of two different data recording campaigns

Time of data collection	Area of interest (km^2)	No. of BSs and APs	No. of GMDT samples	Mobile device	Wi-Fi module	LTE and WLAN signal frequency	Sampling frequency of LTE and WLAN
Sept. 2014 (5 days)	0.33	16 and 1776	21,954	Samsung GT-I9305	Murata M2322007	LTE-1800 and 800 MHz WLAN-2.4 and 5 GHz	2 samples/sec. and 1 sample/5 s
May 2015 (6 days)	0.34	13 and 2280	87,930	Samsung SM-G900F	Murata KM4220004	LTE-1800 and 800 MHz 4 and 5 GHz	2 samples/sec. and 1 sample/5 s

extent. As a consequence, the signal strength recorded from an AP at a fixed location varies with [19]. Moreover RSS values measured from WLAN APs may differ significantly with the UE's hardware even under the same wireless conditions [22, 23]. In order to study the effect time and device variation on UE positioning we have collected GMDT data using different devices in two different times of a year.

The main goal of this research is to use four popular clustering algorithms namely: k-means, Hierarchical Clustering, Fuzzy C-Means Clustering and Self-Organizing Map based clustering in conjunction to our proposed CRFF method and also to compare these CRFF methods with GRFF and KNN in terms of positioning accuracy and computational time complexity. Thereby we can evaluate which clustering algorithm performs the best using the proposed CRFF technique. The rest of the paper is organized as follow. Section 2 describes the GMDT data collection and pre-processing steps. The conventional grid-based RF fingerprinting (GRFF) method, K-nearest neighbours (KNN) based positioning and CRFF methods are explained in Sect. 3. Section 4 presents the experiment results and their performance comparison. Finally, Section 5 concludes the paper and gives some future directions to this effort.

2 Offline Data Collections and Pre-processing

2.1 GMDT Data Measurement

The 3rd Generation Partnership Project (3GPP) has been studying solutions for enhancing the interworking between WLAN and LTE in Release 12 and 13 [24]. Authors in [14] have proposed an enhancement to the LTE MDT referred to as GMDT with minor changes to the 3GPP MDT framework which enables WLAN APs to be added to the MDT report containing LTE network measurements as well as the UE location information.

To build the GMDT data-base commercially available mobile phones installed with drive test software known as 'Nemo Handy' was used [25]. This enabled us to measure reference signal received power (RSRP) values of Long Term Evolution (LTE) serving and detected Base Stations (BS) and received signal strength indicator (RSSI) values of WLAN APs with corresponding GNSS locations of the UEs. Both LTE and WLAN signal strengths were recorded in dBm and GNSS latitude and longitude values were converted to Universal Transverse Mercator (UTM) coordinate system values. About 150 km of measurements were recorded by feet, bicycle and car from a residential urban area in Tampere, Finland. In order to collect enough measurement samples from the area of interest every route was repeated at least twice during the data recording period. Table 1 summarizes the parameters of two data collection campaigns.

2.2 GMDT Data Pre-processing

Our proposed positioning system is network-based system where a positioning server (GMDT server) is used to store and update the 'radio map' through merging multiple GMDT samples recorded from the same x–y coordinate comprising of similar LTE BS and WLAN AP IDs to form a single fingerprint of mean RSS values of the constituent GMDTs. Since the strongest APs provide good probability of coverage over time [18]; we have chosen a subset of APs with the highest observation RSS values. In indoor WLAN positioning seven WLAN RSSI values were used by authors in [20] to obtain acceptable positioning accuracies. Authors in [14] have noticed that increasing WLAN APs after ten provides little to no gain in UE positioning performance. Hence in this study we have compare the UE positioning performances of two different sets of RSS values $S_{j,n}$ where, $j = 1$ and 2 refers to different GMDT data-sets and n is the total number of GMDT samples. The first set $S_{1,n}$ comprises of serving LTE RSRP and six WLAN RSSI values while the second set $S_{2,n}$ contains serving LTE RSRP and ten WLAN RSSI values. We can represent a GMDT sample of a set by a row vector:

$$S_{j,n} = \{LW_{ID}, RSS_{LW}, P_{XY}\} \tag{1}$$

where, LW_{ID} denotes the LTE BS IDs and WLAN AP IDs, RSS_{LW} corresponds to RSRP and RSSI values, and P_{XY}

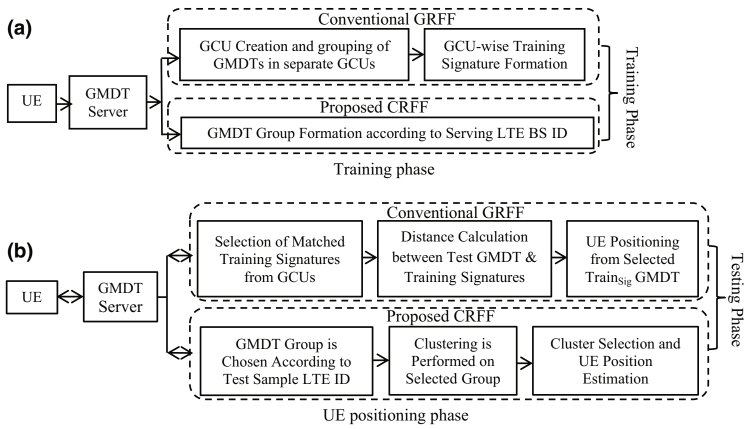


Fig. 1 Block diagram of GRFF and CRFF positioning methods. **a** Training phase, **b** UE positioning phase

contains x–y coordinates of the UEs obtained from GNSS positioning information.

Training phase of GRFF method: We have used a conventional single grid-cell layout based fingerprinting. The whole geographical area of interest is segmented into 10 m-by-10 m square grid-cell units (GCU). As shown in Fig. 1a the GMDT samples of a given data-set $S_{j,n}$ are grouped in different GCUs. For any particular GCU a single training signature $Train_{Sig}$ is formed from all its samples. This shortens the searching time during the UE position estimation phase and reduces the computational cost. The $Train_{Sig}$ formed from all the GMDT samples of ith GCU can be defined by:

$$Train_{sig}^{i} = \left\{TS_{ID}^{LW},\, RSS_{TS}^{LW},\, P_{\mathrm{Ref}}^{XY}\right\} \tag{2}$$

where, TS_{ID}^{LW} contains all unique LTE BS IDs and WLAN AP IDs obtained from samples of the GCU, RSS_{TS}^{LW} is a vector of the corresponding mean LTE RSRP and WLAN RSSI values, and P_{Ref}^{XY} is the reference x–y coordinate calculated from the mean values of x and y coordinates of the samples.

Training phase of CRFF method: The GMDT samples of a given data set $S_{j,n}$ are grouped according to unique LTE serving BS IDs. Hence literally it does not require any data-processing during the training phase.

3 Position Estimation Phase

The test UE first sends a positioning request to the GMDT server along with the recorded cell-IDs and associated RSS values. After matching and data processing GMDT server sends the position estimation information to the test UE.

3.1 Test Phase of GRFF Method

As shown in Fig. 1b the LW_{ID} of test GMDT sample ($Test_{Sam}$) is compared to TS_{ID}^{LW} of all the training signatures of the data server to select those signatures which meet a minimum matching threshold (MT) value. In our study this minimum MT number for both GMDT sets were set to two. Therefore for MT-2 all the training signatures that contain at least two or higher number of LW_{ID} as compared to the test GMDT are selected: a partial ID match procedure. The maximum MT numbers for $S_{1,n}$ and $S_{2,n}$ were four and five respectively. Euclidean distance was used to measure the statistical difference between a test sample and selected training signatures which was found to be effective in WLAN-based indoor UE positioning [26]. Here we have used a simplified Mahalanobis distance (MD) equation where the inverse covariance matrix is replaced by an identity matrix:

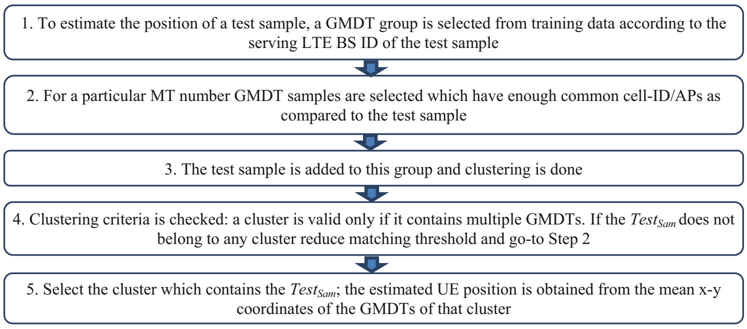
$$d\left(Test_{Sam,}Train_{Sig}\right) = \sqrt{\{(\boldsymbol{u_{Te}} - \boldsymbol{u_{Tr}})^{T}\boldsymbol{I}(\boldsymbol{u_{Te}} - \boldsymbol{u_{Tr}})\}} \tag{3}$$

where, $\boldsymbol{u_{Te}}$ and $\boldsymbol{u_{Tr}}$ denotes the RSRP and RSSI values of the $Test_{Sam}$ and a $Train_{Sig}$ respectively and $\boldsymbol{I}$ is the identity matrix. Separate calculations are done to measure all the distances between a $Test_{Sam}$ and training signatures. The $Train_{Sig}$ that corresponds to the smallest Euclidean distance is chosen for UE positioning. The estimated position of the $Test_{Sam}$ is obtained from P_{Ref}^{XY} of the chosen $Train_{Sig}$.

3.2 Test Phase of KNN Based Positioning

The most well-known pattern matching algorithm is K nearest neighbour (KNN) [5]. In order to satisfy the acceptable localization accuracy with low computation effort KNN has been used for WLAN UE positioning by

Fig. 2 Block-diagram of CRFF based UE fingerprint positioning



several researchers [3, 21, 27, 28]. Here first we select the training GMDT group ($Train_{Grp}$) according to the LTE serving BS ID of the $Test_{Sam}$. Then multiple GMDT samples are selected from $Train_{Grp}$ according to the partial ID matching. The partial matching begins with the highest MT number and until multiple partially matched training samples ($GMDT_{PM}$) are obtained MT number is sequentially lowered towards the minimum. Now according to the lowest Euclidean distance a maximum of five closest GMDTs are chosen using the following KNN equation:

$$d(GMDT_{PM}, Test_{Sam}) = \sqrt{\{\sum\nolimits_{j=1}^{n} (GMDT_{RSS} - Test_{RSS})\}^2} \quad (4)$$

where, $GMDT_{RSS}$ and $Test_{RSS}$ are vectors of LTE RSRP and WLAN RSSI values of $GMDT_{PM}$ and $Test_{Sam}$ respectively. The estimated position of a test UE is calculated from mean x–y coordinates of the selected $GMDT_{PM}$ samples.

3.3 Test Phase of CRFF Methods

The main steps of the proposed CRFF method is depicted in Fig. 2.

3.3.1 K-means Cluster Based Positioning

The k-means method is a widely used clustering technique in scientific and industrial applications [29]. Although it offers no accuracy guarantee, its simplicity and speed are very appealing in practical RF fingerprint positioning. It has been successfully used in indoor mobile localization and also in outdoor positioning as an energy efficient RF fingerprinting method [30, 31]. Here k-means^{++} algorithm was used which is faster to implement and also improves the performance of Lloyd's algorithm [32]. The methods begins with a set of x_i data points where $i = 1,2,\ldots,n$ and a pre-defined maximum cluster number K. The task is to choose K centres c_k so as to minimize the following distance function,

$$d(x, c) = \sum\nolimits_{i=1}^{n} |x_i - c_k|. \quad (5)$$

Here each centroid is the component-wise median of the sample points in that cluster. Assuming $D(x_i)$ denotes the shortest distance from a data point to the already chosen cluster centre k-means^{++} algorithm performs the following steps:

1. The first centre c_1 is chosen uniformly at random from x.
2. A new centre c_k is chosen from x with probability $\frac{D(x_i)^2}{\sum_{i=1}^{n-1} D(x_i)^2}$.
3. Step (2) is repeated until all k centres are chosen.
4. For each c_k, data points are assigned to it which are closer to it than any other c_k.
5. New c_k is computed from the mean of all data points that belongs to the previous c_k.
6. Steps (4) and (5) are repeated until c no longer changes.

Depending upon number of $GMDT_{PM}$ samples ($GMDT_{PM}^{num}$) different K values were assigned for k-means^{++} algorithm so that clustering takes place even with less $GMDT_{PM}^{num}$. K is set to 6 if $GMDT_{PM}^{num} \geq 20$, K is 3 if $20 > GMDT_{PM}^{num} \geq 10$ and K is 2 if $10 > GMDT_{PM}^{num} \geq 2$.

3.3.2 Agglomerative Hierarchical Cluster Based Positioning

Hierarchical clustering is a technique that constructs a tree-like nested structure of clusters. In agglomerative hierarchical clustering (AHC), one starts by considering each data point as a single cluster and follows by merging two neighbouring clusters at each step of the process [33]. In this study we have used weighted-linkage based AHC clustering since it has shown good positioning performance in GSM outdoor UE localization [34]. The neighbouring clusters are chosen based on a *linkage* criterion where weighted average distance determines the distance between two clusters. In order to select the optimal cluster number in AHC method we have used Davies-Bouldin criterion [35]. This criterion is based on a ratio of within-cluster and between-cluster distances. Minimum Davies–Bouldin index *(DB)* indicates the potential number of clusters in the data:

$$DB(K) = (1/K)\{\sum_{i=1}^{k} max_{j\neq i}(D_{i,j})\} \quad (6)$$

where, K is the initial maximum number of clusters, $D_{i,j}$ is the within-to-between cluster distance ratio for the ith and jth clusters. $D_{i,j}$ is given by; $D_{i,j} = (d_i^{-} + d_j^{-})/d_{i,j}$, where,$d_i^{-}$ is the average distance between each point in ith cluster and centroid of the ith cluster d_j^{-} is the average distance between each point in jth cluster and centroid of the jth cluster $d_{i,\ j}$ *is* the Euclidean distance between centroids of the ith and jth clusters. Here we have selected $K = 6$ if $GMDT_{PM}^{num} > 10$ and $K = 2$ when $GMDT_{PM}^{num} < 10$, so that clustering still takes place when there is lees number of $GMDT_{PM}^{num}$ samples.

3.3.3 Fuzzy C-Means Cluster Based Positioning

Fuzzy C-means (FCM) is a data clustering technique—a dataset is partitioned into multiple clusters with every data-point in the dataset belonging to every cluster to a certain degree. Authors in [36] and [37] have used FCM in WLAN indoor localization to obtain good positioning accuracy and also to reduce the computation time as compared to a conventional GRFF method. We have assigned different initial cluster size c depending on number of $GMDT_{PM}$ samples: $c = 6$ if $GMDT_{PM}^{num} \geq 20$; $c = 3$ if $GMDT_{PM}^{num} < 20$ and $GMDT_{PM}^{num} \geq 10$; and $c = 2$ if $GMDT_{PM}^{num} < 10$ and $GMDT_{PM}^{num} > 2$. FCM starts with an initial guess for the cluster centres, which are intended to mark the mean location of each cluster and it also assigns every data point a membership grade for each cluster. By iteratively updating the cluster centres and the membership grades for each data point, it moves the cluster centres to the right location. This iteration is based on minimizing the objective function for subdividing the selected GMDT data-set [38]:

$$J_m(u, v) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{i,k}^{m}||D_k - v_i||^2 \quad (7)$$

where, n is the number of samples in the data set, c is the number of clusters $(1 \leq c \leq n)$,$u_{i,k}$ is the element of partition matrix U of size $(c\ x\ n)$ containing membership function, v_i is the centre of ith cluster, and m is a weighting factor that controls fuzziness of membership function. The matrix U is constrained to contain elements in the range of [0, 1] such that $\sum_{i=1}^{c} u_{ik} = 1$ for each $u_{ik}(1 \leq k \leq n)$. The norm $||D_k - v_i||$ is the distance between the sample D_k and the clusters centre v_i.

3.3.4 Self-Organizing Map Based Positioning

SOM was introduced as an unsupervised competitive learning algorithm of the artificial neural networks by Finnish Professor Teuvo Kohonen in the early 1980s, SOM is also called the Kohonen map. A Self Organizing Map (SOM) is a single layer neural network, where neurons are set along an n-dimensional grid. Each neuron has as many components as the input patterns. Training a SOM requires a number of steps to be performed in a sequential way. For an input sample the SOM training phase consists of three steps: (1) to evaluate the distance between input sample and each neuron of the SOM; (2) to select the neuron (node) with the smallest distance from the sample; and (3) to correct the position of each node according to the results of step 2), in order to preserve the network topology. Steps 1–3) can be repeated more than once for each input sample until stopping criteria is reached. The SOM technique is simple yet effective in capturing the properties of the input space and it can be used for clustering input data.

In [43] and [44] authors have used SOM to compute virtual coordinates that are effective for location-aided routing in Wireless Sensor Networks (WSN). In [44] synchronous readings collected by all the sensor nodes were used to build the training set for the SOM. After training the model, the localization task was performed using new sensor readings to sort nodes on the basis of their proximity to a virtual grid of nodes. In [45] authors have used SOM to develop an indoor locating and tracking system using Wi-Fi RSS values. They have achieved good positioning accuracy by using SOM technique. In this study we have employed SOM as another CRFF method for outdoor user localization using GMDT data.

4 Experimental Results and Discussion

To evaluate the robustness of the positioning methods with changes in recording device and surrounding environment two experimental studies (ExStudy-1 and ExStudy-2) were carried out. In ExStudy-1 both training and test samples were selected from the same time period—September 2014. Here training and test data-sets comprises of randomly choosing data chunks of 20 sequentially recorded samples.

Table 2 shows the UE positioning results of ExStudy-1 obtained from 10 fold cross-validations. In this study only GMDT data-set $S_{1,n}$ was used. In each of experimental studies the number of training and test GMDTs were 23,080 and 2565 respectively. Table 2 shows the 68th and

Table 2 Positioning error results of ExStudy-1 using GMDT dataset $S_{1,n}$

MT	GRFF			KNN			K-means			AHC			FCM		
	68% PE (m)	95% PE (m)	Ana. Sam. (%)	68% PE (m)	95% PE (m)	Ana. Sam. (%)	68% PE (m)	95% PE (m)	Ana. Sam. (%)	68% PE (m)	95% PE (m)	Ana. Sam. (%)	68% PE (m)	95% PE (m)	Ana. Sam. (%)
4	15.3	43.1	99.3	15.0	45.5	98.6	10.0	36.3	84.6	8.2	31.2	74.9	11.0	38.8	84.6
3	15.3	43.3	99.8	15.2	46.0	99.7	11.5	40.4	94.9	9.0	33.8	80.8	12.8	42.6	96.2
2	15.3	43.4	99.9	15.2	46.0	99.8	11.5	40.5	95.0	9.1	34.0	81.0	12.8	42.7	96.3

Table 3 Positioning error results of ExStudy-2 using GMDT dataset $S_{1,n}$ and $S_{2,n}$

D. S.	MT	GRFF			KNN			K-means			AHC			FCM		
		68% PE (m)	95% PE (m)	Ana. Sam. (%)	68% PE (m)	95% PE (m)	Ana. Sam. (%)	68% PE (m)	95% PE (m)	Ana. Sam. (%)	68% PE (m)	95% PE (m)	Ana. Sam. (%)	68% PE (m)	95% PE (m)	Ana. Sam. (%)
$\boldsymbol{S_{1,n}}$	4	26.6	47.0	80.3	25.8	47.3	69.7	24.1	47.2	66.8	20.8	39.8	26.8	24.5	43.4	41.8
	3	27.2	49.2	96.5	27.0	53.4	94.6	25.2	51.2	93.7	21.5	41.7	60.5	25.5	50.2	78.1
	2	27.9	51.1	99.7	27.5	55.6	99.4	25.5	55.4	99.3	22.4	43.2	76.6	26.7	53.3	95.7
$\boldsymbol{S_{2,n}}$	5	25.7	46.1	57.0	24.7	44.3	89.6	23.8	41.6	86.9	20.8	38.7	43.1	23.5	42.2	62.4
	4	26.7	47.6	85.5	25.8	46.8	97.5	24.5	42.8	96.6	22.0	42.0	67.0	24.3	43.0	87.4
	3	27.6	49.5	96.8	26.0	47.5	98.9	24.7	44.1	98.8	23.0	43.7	78.4	24.8	44.1	97.5
	2	28.1	50.8	99.7	26.2	49.3	99.9	24.9	46.2	99.9	23.4	45.2	82.9	25.2	46.4	99.4

Table 4 Positioning error results of ExStudy-2 using SOM with GMDT dataset $S_{1,n}$ and $S_{2,n}$

Method	SOM						
Data set	$\mathbf{S_{1,n}}$			$\mathbf{S_{2,n}}$			
Matching threshold	4	3	2	5	4	3	2
68% PE (m)	22.06	23.05	25.27	24.78	23.83	24.53	24.81
95% PE (m)	34.84	39.93	45.70	42.42	41.95	44.27	45.23
Analysed Samples (%)	2.96	15.44	39.22	4.92	15.52	31.00	48.57

95th percentile cumulative distribution function (CDF) values of positioning error (PE) for each of the positioning methods along with the percentage of analysed $Test_{Sam}$s corresponding to different MT values.

Table 3 shows results of ExStudy-2 where both $S_{1,n,}$ and $S_{2,n}$ datasets were used. These datasets contain 32,791 training GMDTs of September 2014 and 3574 $Test_{Sam}$s of May 2015. Here each of the selected $Test_{Sam}$ is surround by more than ten training GMDTs within its 3 m circular radius area to ensure the presence of sufficient number of training samples in its vicinity. It is found from Tables 2, 3 and 4 that for MT-2 all the methods have analyze maximum amount of $Test_{Sam}$s.

The bar plot of Fig. 3a, b shows 68th and 95th percentile PE values respectively corresponding to MT-2 of both studies using dataset $S_{1,n}$. In every study AHC based RFFP has outperformed other positioning methods in both 68%-ile and 95%-ile of PE. For MT-2 in ExStudy-1 AHC has shown an improvement of 40.52% and 21.66% in 68%-ile and 95%-ile of PE respectively as compared to that of the GRFF method. For the same MT value and using $S_{1,n}$ in ExStudy-2 AHC improves positioning accuracy by 19.71% and 15.46% in 68%-ile and 95%-ile of PE respectively over that of GRFF method. In ExStudy-2 AHC outperforms KNN by 18.54% and 22.30% in 68%-ile and 95%-ile of PE respectively. However in both of the studies AHC has analyzed lower percentages of $Test_{Sam}$s. From Table 3 it was found that when $S_{2,n}$ is used in ExStudy-2 positioning performences of K-means and FCM does not differ significantly from that of the AHC method for MT values of 2, 3 and 4. It is also noticeable that corresponding to each of these MT values K-means and FCM have analyzed more $Test_{Sam}$s than AHC based positioning.

In Table 4 gives the PEs of SOM based RFFP for ExStudy-2 using GMDT dataset $S_{1,n}$ and $S_{2,n}$. It has given better positioning accuracies when compared to GRFF,

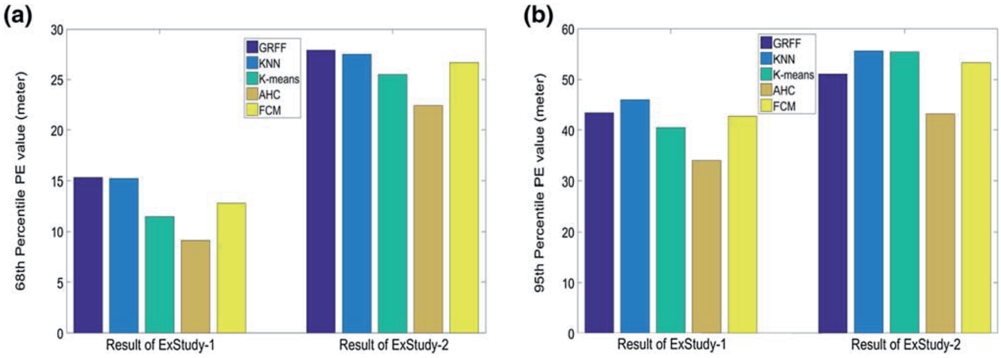


Fig. 3 Comparison of PE results between ExStudy-1 and ExStudy-2 for MT-2. **a** 68th percentile PE values (meters), **b** 95th percentile PE values (meters)

Table 5 Execution time analysis of different methods in ExStudy-2

Methods	Time complexity	Average elapsed time (seconds) for $S_{1,n}$	Average elapsed time (seconds) for $S_{2,n}$
GRFF	*Depends on n, N_{GCU}, and d*	591.9551 (for training)	1005.5040 (for training)
		0.6062 (for testing)	1.0145 (for testing)
KNN	$O(n\ d)$	0.9367	1.7078
K-means	$O(\mathrm{n}\ K\ d\ T)$	1.1492	1.9058
AHC	$O(n^2\ logn)$	0.0788	0.1302
FCM	*Near* $O(n)$	1.1003	1.7887
SOM	$O(d + K_n)$	5.94	12.24

KNN, K-means and FCM based RFFP but with significant reduction of analyzed $Test_{Sam}$s. For MT-2 its 68%-ile and 95%-ile results closely resemble that of AHC results. For higher MT values the analyzed percentages of $Test_{Sam}$s are even less.

The average computation time taken by the GRFF and cluster based methods are shown in Table 5; where $n = 3574$ is the total number of GMDT data samples; $N_{GCU} = 5478$ is the total number of GCUs in GRFF method, $d = 2$–7 for data-set $S_{1,n}$ and $d = 2$–11 for data-set $S_{2,n}$—is the data dimension of a GMDT sample; $K = 2$–6 is the number of initial clusters; $K_n = 100$ is the number of neurons in SOM and $T = 1$ to 6 for data-set $S_{1,n}$ and $T = 1$–10 for data-set $S_{2,n}$—is the number of iterations taken by an algorithm to converge. The computation time of all the positioning methods other than GRFF depend upon the T. We can find from Table 5 that only the GRFF needs training time—which is very long compared to the testing time of any method. It is also found that UE position estimation time increases for all the methods when data-set $S_{2,n}$ was used as compared to that of $S_{1,n}$—due to the increase in data dimension.

AHC has taken the least amount of time for UE positioning in both of the experimental Studies. But due to its high computational complexity, which is at least $O(N^2)$ it may not be a suitable method for a large-scale data-set. Since K, d, and T are usually much less than N, the time complexity of K-means method is approximately linear; hence this algorithm scales well to large-scale data-sets [39, 40]. SOM based RFFP has taken much longer time to output position estimation as compared to rest of the methods. It is worth mentioning that depending upon the choice of the initial cluster size K both the performances and execution time of the methods might differ. Hence as a future work we intend to compare positioning accuracies of the methods with variations in K numbers. Also it worth comparing the results with less number of training samples in the vicinity of a test sample.

5 Conclusion

The conventional grid-based RF fingerprinting positioning heavily depends on training phase data-processing and also the output result varies upon the chosen grid-cell size. In this study we have used GMDT data for outdoor UE positioning in urban area using cluster-based fingerprint positioning that does not go through a training phase data processing. Proposed CRFF method can provide improved positioning accuracy with less computational cost over traditional GRFF and KNN methods. CRFF continues to perform better than GRFF and KNN even when facing recording device variation and environmental changes. For lower MT value SOM performs similar to AHC method but it fails to analyze considerable amount of test samples and also it takes the longest execution time for positioning. With data-set having eleven RSS K-means and FCM based CRFF improves positioning accuracies and analyzes 99% test data. From this study it is found that using GMDT data consisting of seven RSS values AHC based CRFF has given best positioning accuracy taking shortest time as compared to other methods. Hence using GMDT data cellular operators can utilize AHC based RF fingerprinting to provide fast and acceptable results for outdoor UE positioning.

Acknowledgements The authors would like to thank colleagues from University of Jyvaskyla and European Communications Engineering, Finland for their constructive criticism, comments and support.

References

1. E. Kaplan, C. Hegarty, *Understanding GPS: Principles and Applications.* Artech House, Inc., 2005.
2. M. Anisetti, C. A. Ardagna, V. Bellandi, E. Damiani and S. Reale, Map-based location and tracking in multipath outdoor mobile networks, *IEEE Transactions on Wireless Communications*, Vol. 10, No. 3, pp. 814–824, 2011.
3. J. H. Kim, K. S. Min, W. Y. Yeo, A design of irregular grid map for large-scale Wi-Fi LAN fingerprint positioning systems, *The Scientific World Journal,* Vol. 2014, ID 203419, 2014.
4. X. Liu, S. Zhang, J. Quan, X. Lin, The experimental analysis of outdoor positioning system based on fingerprint approach. In *12th IEEE International Conference on Communication Technology (ICCT),* pages 369–372, Nanjing, China, 2010.
5. M. Yousief, *Horus: A WLAN-Based Indoor Location Determination system*, in PhD thesis, University of Maryland, 2004.
6. A. Kushki, K. N. Plataniotis and A. N. Venetsanopoulos, Kernel-based positioning in wireless local area networks, *IEEE Transactions on Mobile Computing*, Vol. 6, No. 6, pp. 689–705, 2007.
7. Skyhook. In *Global 1st Party Location Network*. http://www.skyhookwireless.com/about-skyhook. Accessed 24 Dec 2016.
8. K. Li, P. Jiang, E. L. Bodanese and J. Bigham, Outdoor location estimation using received signal strength feedback, *IEEE Communications Letters*, Vol. 16, No. 7, pp. 978–981, 2012.
9. R. Singh, L. Macchi, C. S. Regazzoni, K. N. Plataniotis. A statistical modelling based location determination method using fusion technique in WLAN. In *International Workshop on Wireless Ad-Hoc Networks*, London, UK, 2005.
10. J. Talvitie. *Algorithms and Methods for Received Signal Strength Based Wireless Localization*, in PhD thesis, Tampere University of Technology, 2016.
11. J. Johansson, W. A. Hapsari, S. Kelley and G. Bodog, Minimization of drive tests in 3GPP release 11, *IEEE Communications Magazine*, Vol. 50, No. 11, pp. 36–43, 2012.
12. 3GPP TR 36.805. Study on minimization of drive-tests in next generation networks. Accessed Dec 2009.
13. M. H. A. Meniem, A. M. Hamad, E. Shaaban. Relative RSS-based GSM localization technique. In *IEEE International Conference on Electro/Information Technology (EIT)*, pages 1–6, South Dakota, USA, 2013.
14. T. Hiltunen, R. U. Mondal, J. Turkka, T. Ristaniemi. Generic architecture for minimizing drive tests in heterogeneous networks. In *IEEE 82nd Vehicular Technology Conference (VTC Fall),* pages 1–5, Boston, USA, 2015.
15. M. Bshara, U. Orguner, F. Gustafsson and L. V. Biesen, Fingerprinting localization in wireless networks based on received signal-strength measurements: a case study on WiMAX networks, *IEEE Transactions on Vehicular Technology*, Vol. 59, No. 1, pp. 283–294, 2010.
16. H. Liu, Y. Zhang, X. Su, X. Li, N. Xu, Mobile localization based on received signal strength and Pearson's correlation coefficient, *International Journal of Distributed Sensor Networks,* Vol. 2015, ID. 157046, 2015.
17. T. Wigren, Adaptive enhanced cell-ID fingerprinting localization by clustering of precise position measurements, *IEEE Transactions on Vehicular Technology*, Vol. 56, No. 5, pp. 3199–3209, 2007.
18. M. Youssef, A. Agrawala, A. U. Shankar. WLAN location determination via clustering and probability distributions. *1st IEEE International Conference on Pervasive Computing and Communication (PerCom 2003),* pages 143–150, Texas USA, 2003.
19. Y. Chen, Q. Yang, J. Yin and X. Chai, Power-efficient access-point selection for indoor location estimation, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, No. 7, pp. 877–888, 2006.
20. E. Laitinen, E. S. Lohan, J. Talvitie, S. Shrestha. Access point significance measures in WLAN-based location. In *9th Workshop on Positioning Navigation and Communication (WPNC)*, pages 24–29, Dresden, Germany 2012.
21. P. Bahl and V. Padmanabhan, RADAR: An in-building RF-based user location and tracking System, *IEEE INFOCOM*, Vol. 2, pp. 775–784, 2000.
22. M. H. N. Hossain, Van, Y. Jin, W. S. Soh. Indoor localization using multiple wireless technologies. In *IEEE MASS*, Pisa, Italy, 2007.
23. M. B. Kjærgaard, C. V. Munk. Hyperbolic location fingerprinting: A calibration-free solution for handling differences in signal strength. In *6th Annual IEEE International Conference on Pervasive Computing and Communications (PerCom 2008)*, pages 110–116, Hong Kong, 2008.
24. 3GPP TR 37.834, Study on WLAN/3GPP radio interworking. Vol. 1.0.0, 2013.
25. Nemo Handy: handheld drive test software, http://www.anite.com/businesses/network-testing/products/nemo-handy-world's-most-widely-used-handheld-drive-test-tool#.Vc8_nPmqpBd. Accessed June 2016.
26. C. Feng, W. S. A. Au, S. Valaee and Z. Tan, Received-signal-strength-based indoor positioning using compressive sensing, *IEEE Transactions on Mobile Computing*, Vol. 11, No. 12, pp. 1983–1993, 2012.
27. I. J. Quader, B. Li, W. Peng, A. G. Dempster. Use of fingerprinting in Wi-Fi based outdoor positioning. In *International Global Navigation Satellite Systems Society IGNSS Symposium*, The University of New South Wales, Sydney, Australia, 2007.

28. F. Yu, M. Jiang, J. Liang, X. Qin, M. Hu, T. Peng, X. Hu, 5G WiFi signal-based indoor localization system using cluster k-nearest neighbor algorithm, *International Journal of Distributed Sensor Networks,* Vol. 2014, ID 247525, 2014.
29. P. Berkhin, *Survey of Clustering Data Mining Techniques*, Grouping Multidimensional DataSpringer, Berlin Heidelberg, 2006. pp. 25–71.
30. A. Razavi, M. Valkama, E. S. Lohan. *K*-Means fingerprint clustering for low-complexity floor estimation in indoor mobile localization. In *IEEE GLOBCOM Workshop on Localization and Tracking: Indoors, Outdoors and Emerging Networks*, Washington DC, USA, 2015.
31. A. Arya, P. Godlewski, M. Campedel and G. Che´ne´, Radio database compression for accurate energy-efficient localization in fingerprinting systems, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 6, pp. 1368–1379, 2013.
32. A. David, S. Vassilvitskii. K-means ++: The advantages of careful seeding. In *18th Annual ACM-SIAM Symposium on Discrete Algorithms* (*SODA 2007)*, pages 1027–1035, *Louisiana,* United States, 2007.
33. A. C. Rencher, *Methods of Multivariate Analysis*. Wiley, Inc., 2002.
34. A. Arya, P. Godlewski, P. Melle. A hierarchical clustering technique for radio map compression in location fngerprinting systems. In *International Conference on Vehicular Technology*, pages 1–5, Taipei, China, 2010.
35. D. L. Davies, D. W. Bouldin, A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-1, No. 2, pp. 224–227, 1979.
36. H. Zhou, N. N. Van. Indoor fingerprint localization based on fuzzy C-means clustering. In *6th International Conference on Measuring Technology and Mechatronics Automation*, pages 337–340, China, 2014.
37. D. J. Suroso, P. Cherntanomwong, P. Sooraksa, J. Takada. Location fingerprint technique using Fuzzy C-Means clustering algorithm for indoor localization. In *IEEE TENCON,* Indonesia, 2011.
38. J. C. Bezdec, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum PressNew York, 1981.
39. R. Xu, D. C. Wunsch II, *Clustering*, John Wiley and Sons, Inc., 2009.
40. O. A. Abbas, Comparisons between data clustering algorithms, *The International Arab Journal of Information Technology*, Vol. 5, No. 3, pp. 320–325, 2008.
41. X. Liu, S. Zhang, H. Lu, X. Lin. Method for efficiently constructing and updating radio map of fingerprint positioning. In *IEEE GLOBCOM 2010 Workshop on Heterogeneous, Multi-hop Wireless and Mobile Networks,* pages 74–78, Florida, USA, 2010.
42. M. Raitoharju, H. Nurminen, R. Piché, Kalman filter with a linear state model for PDR + WLAN positioning and its application to assisting a particle filter, *EURASIP Journal on Advances in Signal Processing*, 2015.
43. E. Ertin and K. Priddy. Self-localization of wireless sensor networks using self-organizing maps. In *Proceedings of SPIE*, 2005.
44. G. Giorgetti, S. K. S. Gupta, G. Manes. Wireless localization using self-organizing maps. In *Proceedings of IPSN'07,* pages 25–27, Massachusetts, USA, April 2007.
45. T. Mantoro, M. A. Ayu, A. Nuraini, S. M. Amin. Self-organizing map approach for determining mobile user location using IEEE 802.11 signals. In *Proceeding of International Symposium on Information Technology (ITSim),* Kuala Lumpur, Malaysia, 2010.

Riaz Uddin Mondal has received his M.Sc. and B.Sc. degrees from the Department of Applied Physics and Electronics, University of Rajshahi, Rajshahi, Bangladesh in 2003 and 2001 respectively. On November 2004 he joined the Department of Information and Communication Technology, University of Rajshahi, as a Lecturer. From November 2007 to April 2010 he served the University of Rajshahi in the same Department as an Assistant Professor. At present he is pursuing his doctoral degree from the Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland. His research area spans in the field of RF Fingerprint Positioning, Power Amplifier Linearization, Wireless Communications, and Artificial Intelligence. He has authored or co-authored more than 15 Journal and conference publications.

Tapani Ristaniemi (SM'11) received the M.Sc. degree in mathematics, the Ph.Lic. degree in applied mathematics, and the Ph.D. degree in wireless communications from the University of Jyväskylä, Jyväskylä, Finland, in 1995, 1997, and 2000, respectively. In 2001, he was a Professor with the Department of Mathematical Information Technology, University of Jyväskylä. In 2004, he was with the Department of Communications Engineering, Tampere University of Technology, Tampere, Finland, where he was appointed as a Professor of Wireless Communications. In 2006, he moved back to the University of Jyväskylä to take up his appointment as a Professor of Computer Science. In 2013, he was a Visiting Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He is currently an Adjunct Professor with the Tampere University of Technology. He has authored or co-authored over 200 publications in journals, conference proceedings, and invited sessions. He served as a Guest Editor of the IEEE WIRELESS COMMUNICATIONS in 2011. He is an Editorial Board Member of Wireless Networks and the International Journal of Communication Systems. His research interests include brain and communication signal processing and wireless communication systems. Besides academic activities, Prof. Ristaniemi is also active in the industry. In 2005, he co-founded a start-up, Magister Solutions, Ltd., Finland, specializing in wireless systems (Research and Development) for telecom and space industries in Europe. He serves as a consultant and the Chairman of the Board.

Jussi Turkka received the M.Sc. and B.Sc.(Tech.) degrees from the Tampere University of Technology, Finland, in 2008 and 2014, respectively. From 2008 to 2016, he was with the Magister Solutions Ltd working as a senior research consultant. Since 2016 he has been with European Communications Engineering Ltd., as a Principal Engineer. He has authored over 15 scientific publications, contributing to 3GPP LTE specifications and filed several patents. His areas of expertise are in the field of self-organizing cellular networks, wireless communications, machine learning and knowledge mining.