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

Riaz Uddin Mondal, Tapani Ristaniemi, Jussi Turkka

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 (AP). 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 (MDT) 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 (PE) 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], [3], [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 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 data-base 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 [41], [12]. 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 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 Section 3. Section 4 presents the experiment results and their performance comparison. Finally, Section V 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.

Table 1 Summary of two different data recording campaigns

Time of Data Collecti -on	Area of Interest (kilometers^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 (5days)	0.33	16 and 1776	21954	Samsung GT-I9305	Murata M2322007	LTE- 1800 and 800 MHz WLAN- 2.4 and 5 GHz	2 samples/sec. and 1 sample/ 5 seconds
May 2015 (6 days)	0.34	13 and 2280	87930	Samsung SM-G900F	Murata KM4220004	LTE- 1800 and 800 MHz WLAN- 2.4 and 5 GHz	2 samples/sec. and 1 sample/ 5 seconds

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 kilometres 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.

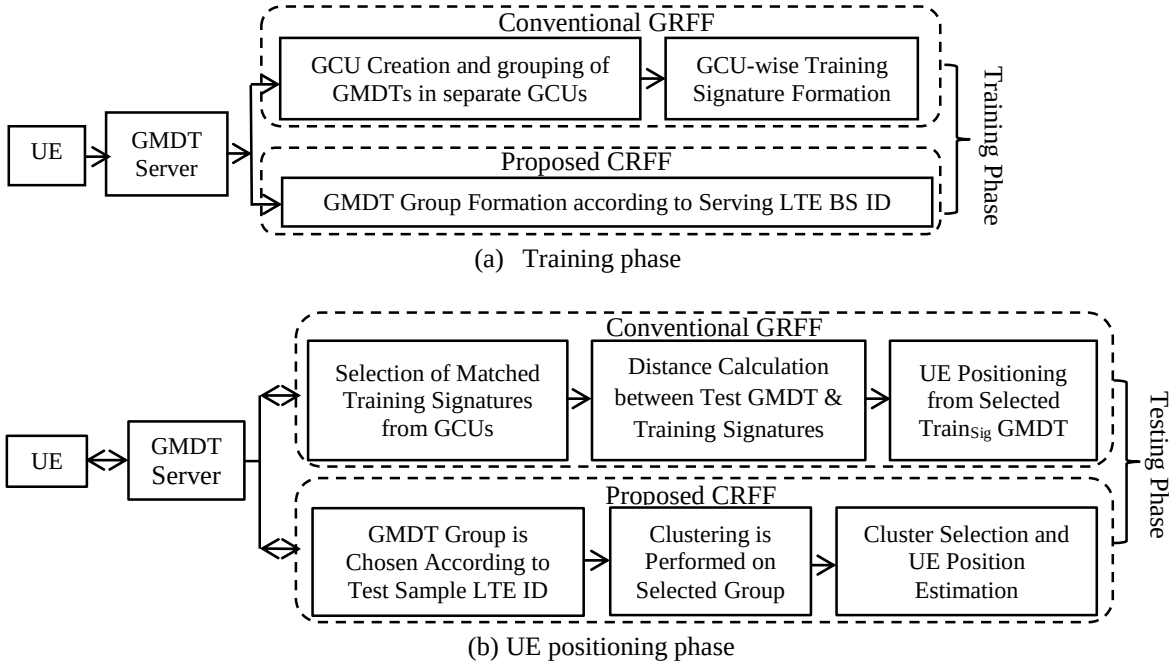


Fig. 1 Block diagram of GRFF and CRFF positioning methods

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:

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

where, LW_{ID} denotes the LTE BS IDs and WLAN AP IDs, RSS_{LW} corresponds to RSRP and RSSI values, and P_{XY} 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 10m-by-10m square grid-cell units (GCU). As shown in Fig. 1(a) 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:

$$Train^{i}_{sig} = \{TS_{ID}^{LW}, RSS_{TS}^{LW}, P_{Ref}^{XY}\} \quad (2)$$

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. 1(b) 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:

$$d(Test_{Sam}, Train_{Sig}) = \sqrt{\{(\boldsymbol{u}_{Te} - \boldsymbol{u}_{Tr})^{T} \boldsymbol{I} (\boldsymbol{u}_{Te} - \boldsymbol{u}_{Tr})\}} \quad (3)$$

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 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:

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

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.

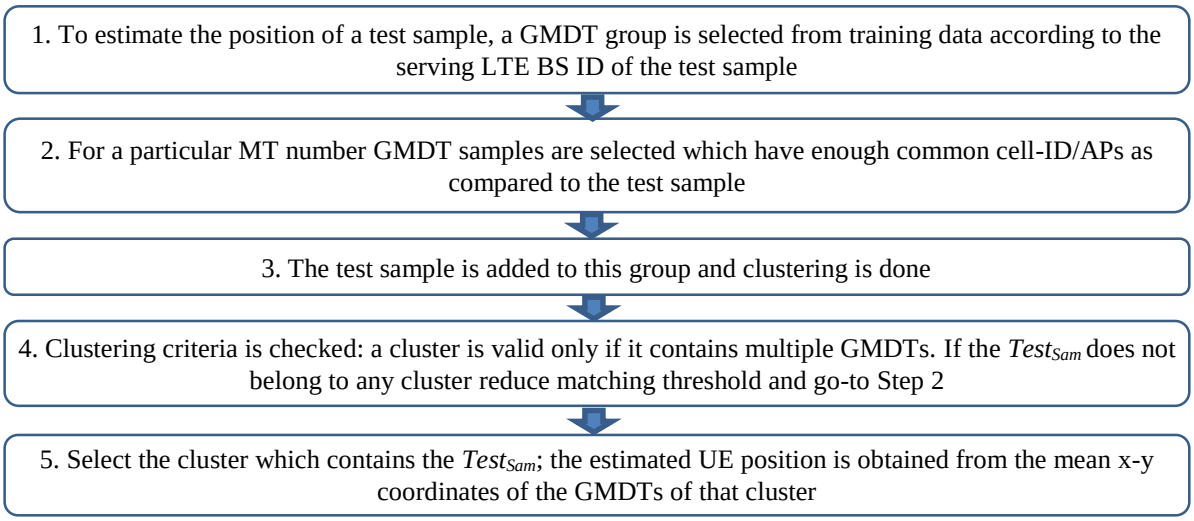


Fig. 2 Block-diagram of CRFF based UE fingerprint positioning

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,

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

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:

(i) The first centre c_1 is chosen uniformly at random from x.

(ii) A new centre c_k is chosen from x with probability $\frac{D(x_i)^2}{\sum_{i=1}^{n-1} D(x_i)^2}$

(iii) Step (ii) is repeated until all k centres are chosen.

(iv) For each c_k, data points are assigned to it which are closer to it than any other c_k.

(v) New c_k is computed from the mean of all data points that belongs to the previous c_k.

(vi) Steps (iv) and (v) 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 $GMDT_{PM}^{num} >= 20$, K is 3 if $20 > GMDT_{PM}^{num} >= 10$ and K is 2 if $10 > GMDT_{PM}^{num} >= 2$.

3.2.3 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; $D_{i,j} = (\bar{d_i}+\bar{d_j})/d_{i,j}$, where, $\bar{d_i}$ is the average distance between each point in ith cluster and centroid of the ith cluster. $\bar{d_j}$ 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.

3.2.4 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 $GMDT_{PM}^{num} >= 20$; $c = 3$ if $GMDT_{PM}^{num} < 20$ and $GMDT_{PM}^{num} >= 10$; 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]:

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

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 $(c \times n)$ containing membership function, v_i is the centre of i^{th} 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 $u_{i,k}(1 \leq k \leq n)$. The norm $\left\| D_k - v_i \right\|$ is the distance between the sample D_k and the clusters centre v_i.

3.2.5 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) to 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 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 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 23080 and 2565 respectively. Table 2 shows the 68th and 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.

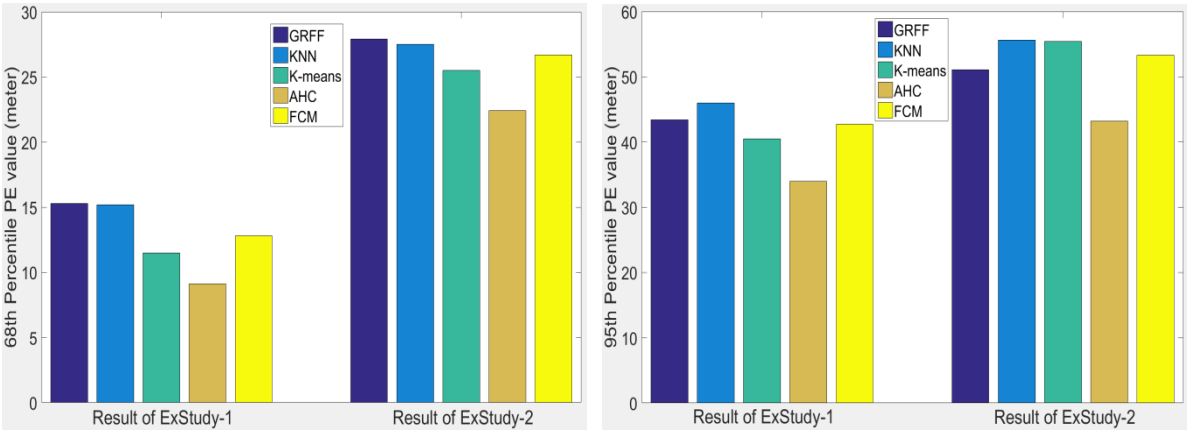
Table 3 Positioning error results of ExStudy-2 using GMDT dataset $S_{1,n}$ and $S_{2,n}$

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. (%)*
$S_{1,n}$	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

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

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

Table 3 shows results of ExStudy-2 where both $S_{1,n,}$ and $S_{2,n}$ datasets were used. These datasets contain 32791 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 three meter circular radius area to ensure the presence of sufficient number of training samples in its vicinity. It is found from Table 2, Table 3 and Table 4 that for MT-2 all the methods have analyze maximum amount of $Test_{Sam}$s.



(a) 68th percentile PE values (meters) (b) 95th percentile PE values (meters)

Fig. 3. Comparison of PE results between ExStudy-1 and ExStudy-2 for MT-2.

The bar plot of Fig. 3(a) and Fig. 3(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, 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 to 7 for data-set $S_{1,n}$ and d = 2 to 11 for data-set $S_{2,n}$ - is the data dimension of a GMDT sample; K = 2 to 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 to 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.

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) 0.6062 (for Testing)	1005.5040 (for Training) 1.0145 (for Testing)
KNN	$O(n\ d)$	0.9367	1.7078
K-means	O (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

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.

Acknowledgments

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, S. Reale, Map-Based Location and Tracking in Multipath Outdoor Mobile Networks, in *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),* Nanjing, China, pp. 369-372, 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, 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, Global 1st Party Location Network, http://www.skyhookwireless.com/about-skyhook. Accessed 24 December 2016.
8. K. Li, P. Jiang, E. L. Bodanese, 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, G. Bodog, Minimization of drive tests in 3GPP release 11, in *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 December 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)*, South Dakota, USA, pp.1-6, 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)*, Boston, USA, pp. 1-5, 2015.
15. M. Bshara, U. Orguner, F. Gustafsson, 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, in *1st IEEE International Conference on Pervasive Computing and Communication (PerCom 2003)*, Texas USA, pp. 143–150, 2003.
19. Y. Chen, Q. Yang, J. Yin, 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), Dresden, Germany, pp. 24–29, 2012.
21. P. Bahl, V. Padmanabhan, RADAR: An In-Building RF-Based User Location and Tracking System, in *IEEE INFOCOM*, Vol. 2, pp. 775–784, 2000.
22. Hossain, M. H. N. 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)*, Hong Kong, pp. 110–116, 2008.
24. 3GPP TR 37.834, Study on WLAN/3GPP radio interworking. Vol. 1.0.0, September 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, Z. Tan, Received-signalstrength-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 Data*, Springer, Berlin Heidelberg, pp.25-71, 2006.

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, 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)*, *Louisiana*, United States, pp. 1027–1035, 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*, Taipei, China, pp. 1–5, 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*, China, pp. 337–340, 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 Press, New 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*, Florida, USA, pp. 74-78, 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,* Massachusetts, USA, April 25-27, 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.

Author Photographs

Riaz Uddin Mondal

Prof. Tapani Ristaniemi

Dr. Jussi Turkka