An Efficient Clustering with Robust Outlier Mitigation for Wi-Fi Fingerprint based Indoor Positioning

Pampa Sadhukhan,[1,]

School of Mobile Computing & Communication, Jadavpur University, India.

Supriya Gain

TCS Research & Innovation, Kolkata, India.

Keshav Dahal,[1,]

School of Engineering & Computing, University of the West of Scotland, UK.

Samiran Chattopadhyay, Nilkantha Garain

Dept. of Information Technology, Jadavpur University, India.

Xinheng Wang[1]

Department of Mechatronics and Robotics, School of Advanced Technology, Xi′an Jiaotong-Liverpool University, China

Abstract

Wi-Fi fingerprint systems provide cost-effective and reliable solution for indoor positioning. However, such systems incur high calibration cost in the training phase and high searching overhead in the positioning phase. Moreover, huge storage requirement for the radio map of a large-scale fingerprint system is another major issue. Several solutions based on crowd-sourcing or machine learning technique have been proposed in literature to reduce the calibration overhead. On the other hand, various clustering methods have been proposed over the past decade to reduce the searching overhead. However, none of the existing systems has addressed the issue of high storage requirement for the fingerprint database constructed in the training phase. Moreover, presence of outlier in the

*Corresponding author
[1]Member IEEE
[2]Senior Member IEEE

received signal strength (*RSS*) measurements severely impacts the positioning accuracy of such systems. Thus, this paper proposes an *efficient clustering strategy* for fingerprint based positioning systems to reduce the storage overhead and searching overhead incurred by such systems and also proposes a *robust outlier mitigation technique* to improve their positioning accuracy. The performances of our proposed positioning system are evaluated and compared with five existing fingerprint techniques in both the simulation test bed as well as real indoor environment via extensive experimentation. The experimental results demonstrate that our proposed system can not only reduce the storage overhead and searching overhead but also improve the positioning accuracy compared to the other existing techniques.

Keywords: fingerprint, positioning, RSS, clustering, outlier, accuracy, storage overhead, searching overhead.

---

1. Introduction

Wireless localization or positioning, which is essential for delivering the location-aware services [1] as well as for designing the location tracking systems [2] and also some location based pervasive computing applications like smart home for elders [3], have drawn significant research interests over the past few decades. Global positioning system (GPS) can provide accurate positioning solution in outdoor environments. But it does not work well in indoor areas due to almost unavailability of GPS signal in such areas [4]. Thus, various indoor positioning systems based on wireless technologies such as infrared (IR), ultrasound, Zigbee, Bluetooth, radio frequency (RF) technologies etc., and also the magnetic field [5] have been proposed in the literature. Such positioning systems have been extensively reviewed in [6]. Among these, RF-based positioning system that relies on Wi-Fi can provide cost-effective solution because of its sufficient deployment at every public place and also its availability at almost every portable device. Various RF-based positioning techniques can be mainly divided into two groups - range-based and range-free positioning schemes [7].

The range-based positioning schemes [8-11], which employs some time delay method like time of arrival, time difference of arrival etc., cannot achieve desirable accuracy in the dense cluttered indoor areas due to existence of severe multipath effect along with scattering and attenuation in such areas. Moreover, such positioning scheme cannot determine the object′s position without knowing the positional coordinates of the base stations or access points (APs).

On the other hand, *received signal strength* (*RSS*) based fingerprint positioning techniques, which belong to range-free positioning schemes, are more robust to combat the effect of multipath fading and high non-line-of-sight (NLOS) errors in the dense cluttered indoor environments. Because, such techniques work in two phases and employ the pattern matching process to estimate the user′s position [12-14]. The first phase known as training phase constructs a *fingerprint database* or *radio map* by measuring *RSS* values from several fixed APs at a set of predetermined locations called *training locations* (TLs) and then storing each such collected *RSS* pattern along with the positional coordinates of the TLs into the database. The second phase, i.e., the positioning phase compares the currently observed *RSS* vector with all *RSS* patterns stored in the radio map and then the positional co-ordinates associated with the best matched *RSS* pattern is regarded as the estimated position of the object. However, the construction of radio map through a labor-intensive operation of RSS data collection, incurs a significant calibration overhead. Numerous solutions based on the crowdsourcing or machine learning approach to reduce the calibration overhead, have been proposed in literature over the past few years [15-21]. Another important issue to achieving a reasonable positioning accuracy by such techniques is high fluctuation or variance in the measured *RSS* values over time. This happens due to the presence of outlier signal or sparse noises into the RSS measurements and also influences of various weather parameters like temperature, humidity etc., on them. The presence of outlier signal into the RSS samples measured from various APs induces large error into the estimated position and thus, it significantly degrade the positioning accuracy of the fingerprint techniques [22]. To impede the negative effect of the time varying *RSS* values on the positioning

accuracy of fingerprint techniques, most of such existing techniques collect sufficient *RSS* samples from each AP over a certain duration of time and consider their mean value as measured *RSS* value for that AP instead of relying on a single *RSS* sample. However, the mean of a set of *RSS* samples is highly sensitive to the presence of outlier signal which takes some extreme value at either end of *RSS* data set [23]. Thus, devising a suitable technique to mitigate the effect of outlier on the positioning accuracy of fingerprint techniques is of great importance.

Apart from the above-mentioned issues, fingerprint systems need a large amount of storage space to store the measured RSS patterns along with their positional co-ordinates into the radio map. Moreover, RSS vector observed in the positioning phase is compared with all RSS patterns stored in the radio map to determine object′ position and thus, such systems incur significant searching overhead in the positioning phase. The positioning accuracy of a fingerprint localization system can be significantly enhanced by increasing the number of TLs into the localization area. But, increasing number of TLs into the radio map also increases its storage requirements as well as the searching overhead in the positioning phase. Therefore, two crucial issues with a large-scale fingerprint positioning system, e.g., a fingerprint system designed for a wireless city, are very high storage overhead as well as searching overhead. This is because such systems need to incorporate huge number of TLs into the database to achieve better positioning accuracy. A considerable amount of research have been devoted to reduce the *searching overhead* by integrating some clustering method with the fingerprint positioning [29-41]. However, none of such existing systems has paid attention to the other crucial issue of a fingerprint system, which is high storage requirement for its radio map, i.e., *storage overhead*. *Large storage requirement* also acts as a major burden to deploy such positioning systems onto the small memory portable devices and thus, it still remains as a *non-resolving issue in the literature*. The comparative performance analysis of various clustering based fingerprint systems given in [42, 46] demonstrate that our earlier work on hierarchical clustering strategy [41] can significantly reduce the search-

ing overhead as well as positioning error compared to existing three techniques. But, the issue of high storage overhead associated with such systems and also their degraded positioning accuracy due to presence of outlier in the collected RSS samples have not been addressed in our previous work. Thus, this article aims to resolve the following two highly sensitive issues of a Wi-Fi fingerprint positioning system:

- ***Degraded positioning accuracy due to presence of outlier***
- ***High storage overhead***

Therefore, we propose an outlier mitigation technique based on *Hampel filter* [24] and also an *efficient clustering strategy* (*ECS*) for fingerprint positioning system in this paper to resolve the above-mentioned issues. Our proposed outlier mitigation technique can effectively remove the outlier data from the collected RSS samples and also addresses their time-varying property (i.e., variation in measured RSS values over time) to improve the positioning accuracy of our proposed fingerprint system. On the other hand, our proposed clustering strategy for fingerprint positioning system attempts to reduce the storage requirement for the radio map as well as the searching overhead by applying fusion of similar RSS patterns belonging to the same cluster. To demonstrate the effectiveness and efficiency of our proposed system, we have evaluated and compared its performances with several existing clustering techniques proposed in [33, 34, 37, 39, 40] and also a non-clustering fingerprint technique [12]. Among the existing clustering techniques considered in this paper, affinity propagation clustering has been used in a recently proposed fingerprint system [39]. The extensive experimentation conducted in both the simulation test bed as well as real indoor environment demonstrate that our proposed system can significantly reduce the storage requirement for the fingerprint database and also the searching overhead compared to other existing techniques considered in this paper. Moreover, the experimentation in real indoor areas shows that our proposed system achieves better positioning accuracy at reduced amount of storage compared to the other techniques considered in this paper.

The remaining part of this paper is structured as follows. The background of Wi-Fi fingerprint positioning system along with a short literature review of such existing systems are presented in Section 2. In Section 3, we describe our proposed fingerprint system. Section 4 evaluates and compares the performances of our proposed system with the other existing systems considered in this paper. Finally, we conclude and also point out our future research directions in Section 5.

2. Background and Related Work

In this section, at first, the working methodology of a *Wi-Fi fingerprint positioning system* when applied to a localization area having m APs (denoted as $[AP_1, AP_2, \cdots, AP_m]$) deployed within it and n predetermined TLs (represented by $[l_1, l_2, \cdots, l_n]$), is described. The RSS vector measured at location l_i in the training phase is represented by $v_i = [v_{i1}, v_{i2}, \cdots, v_{im}]$, where v_{ij} denotes the mean value of RSS samples collected from j^{th} AP (AP_j) at l_i. The *radio map* for this localization area comprises n data records, each of which contains the positional co-ordinates of some *training location* (TL) along with RSS vector measured at that TL. According to the positioning algorithm *nearest neighbour in signal space* ($NNSS$) [12], the position of an unknown location is determined by comparing the RSS vector (r') observed at that location in the positioning phase with every RSS pattern ($v_i, 1 \leq i \leq n$) stored in the radio map and then taking the positional co-ordinates of the TL whose associated RSS pattern has the shortest euclidean distance [44] from the observed vector r', as the estimated position. The euclidean distance ($d(l_i)$) between the RSS pattern measured at l_i in the training phase (v_i) and the observed RSS vector in the positioning phase (denoted as $r' = [r'_1, r'_2, \cdots, r'_m]$ under the assumption that same set of *APs* remains active in both phases) is computed by the following equation.

$$d(l_i) = \|v_i, r'\| = \sqrt{\sum_{j=1}^{m} \left(v_{ij} - r'_j\right)^2}, \tag{1}$$

The position of unknown location (l^e) is then estimated by

$$l^e = l_j \Leftrightarrow l_j = \arg\min_{l_i, 1 \leq i \leq n} d(l_i).$$

A considerable amount of research on Wi-Fi fingerprint positioning systems have addressed the issue of high calibration effort required to construct the *radio map* in training phase or its periodic update [15-21]. Of these, several systems have adopted the crowd-sourcing approach, which involves an automatic collection of the *RSS* data along with the data of integrated inertial sensors from the smart phone and then constructing radio map based on those collected data [15-19]. However, such systems cannot achieve desirable positioning accuracy because of large deviation in the collected *RSS* data due to sufficient differences in the smart phones used for data collection as well as collection time and environment. The *gaussian process regression* models have been used in [20] to construct the full radio map for an uncalibrated localization area from a few labeled fingerprints only collected in that area. In order to avoid the periodic updating of radio map, which is needed due to variation in the environmental parameters, a deep learning based *self-calibration time-reversal fingerprint positioning technique* has been proposed in [21]. The fingerprint system proposed in [27] uses continuous and differentiable discriminant functions to represent the search space and thus, can significantly reduce the searching overhead in the positioning phase. However, the discriminant functions based positioning incurs high computational overhead at the run time and is less resilient to highly fluctuating *RSS* samples than other techniques.

On the other hand, fingerprint based positioning technique proposed in [28] uses *robust principal component analysis* in both phases to filter out the sparse noises from the collected RSS measurements and also do not consider the unstable APs in determining the set of nearest neighbor TLs during positioning estimation in order to improve the robustness and accuracy of fingerprint technique. The authors in [47] have proposed a machine learning based indoor positioning technique that at first uses the AP selection strategy to reduce the computational overhead and then applies some local feature extraction process to retrieve the important features of the observed RSS data set. Finally, the re-

constructed data set is fed into the Long Short-term Memory (LSTM) network, a type of neural network, for estimating position of the unknown location. In [48], the researchers have presented a technique of extracting the robust features of RSS data set based on *fisher score–stacked sparse autoencoder* to improve the localization performance. A hybrid localization model designed by combining the global model with sub model to reduce the coordinate localization error is also presented in [48]. An indoor positioning algorithm that combines the features of RSS and *channel state information* (*CSI*) has been proposed in [49]. The proposed technique applies the principle of coherent bandwidth to reduce the dimension of *CSI* data after filtering it in time domain and then estimates the position by using fusion of the relationship between *CSI* and *RSS* data by some confidence degree. The placement strategy of Wi-Fi access points and also the degree of beacon coverage have significant influences on the positioning accuracy of fingerprint localization system as demonstrated in our earlier work [46]. The researchers in [50] have proposed a solution to the problem of optimizing the placement strategy of APs and beacon nodes within the area of localization based on *Cramer-Rao lower bound*.

On the other hand, several clustering methods for fingerprint positioning system in order to reduce the searching overhead, have been proposed in literature over the past decade. Among these, *k-means* clustering technique [29] that partitions the whole radio map into k different subsets through a recursive method has gained popularity in fingerprint positioning due to its low computational overhead [31-33]. However, *k-means* clustering based systems cannot provide desirable positioning accuracy as the random selection of the initial cluster members or exemplars increases the possibility of false cluster selection [30]. To improve the positioning accuracy by allowing overlap among the clusters created by the *k-means* algorithm, two enhanced clustering techniques, *multi nearest-neighbour* (*MNN*) overlapping strategy and *Voronoi* (*VRN*) based overlapping strategy, have been proposed in [33]. Although both overlapping strategies achieve better positioning accuracy compared to *k-means* strategy, but they increase the searching overhead and also incur higher computational complexity

as demonstrated in [42]. The fingerprint positioning system proposed in [35] uses self-organizing map (SOM)for locating objects with room-level accuracy in the indoor environments. The SOM is a special kind of neural networks, which can map the complex high dimensional data to low dimensional one [34]. The proposed system employs three-phase methodology for position estimation in the indoor areas. The first phase attempts to remove the noises incorporated into the collected *RSS* samples. Then, self organizing maps are used in the remaining two phases for constructing the clustered radio map and also locating the objects in a certain region. The drawback of this SOM based positioning system is that it can locate the unknown object within a certain area rather than estimating its exact position.

The *affinity propagation* (*AFP*) clustering [30], which unlike the *k-means* clustering technique can yield the optimal cluster heads along with their corresponding clusters by iteratively exchanging two kinds of messages between the data points, has been effectively used by several fingerprint systems [36-39]. The major advantage of the *AFP* clustering is that it does not require the number of clusters to be formed along with the randomly selected exemplars as inputs, which are mandatory for the *k-means* clustering, to converge to the solution. However, the number of clusters to be generated by *AFP* clustering is still controlled by an user defined input called *preference*. Moreover, the negative *Euclidean* distance [44] between the *exemplar* and individual data point as a measure of similarity between them can significantly degrade the performance of such clustering when applied to a data set with complex structure [43]. Another clustering method that forms the clusters of training locations based on the virtual positions of APs for fingerprint based indoor positioning has been proposed in [40]. However, such clustering technique works perfectly in the indoor environments without linear constraints. Our proposed *hierarchical clustering strategy* (*HCS*) presented in [41], partitions the radio map into a set of non-overlapping clusters, each of which includes those training locations receiving strongest signal strength from certain number of APs determined by the level of hierarchy (e.g., only one AP in case of *1-way HCS*, two APs in

case of *2-way HCS* and so on) in a fixed order only. The order of APs providing strongest signal is used to assign an unique id to each cluster formed by the proposed *HCS*. So, the number of clusters created by our proposed scheme (*HCS*) is easily determined based on the number of APs deployed in the localization area and the level of hierarchy chosen by [41]. Moreover, our proposed *HCS* significantly reduces the searching overhead and positioning error compared to the *k-means* clustering as well as two overlapping strategies as shown in [42, 46]. Henceforth, to demonstrate the effectiveness and efficiency of our proposed fingerprint system presented in this paper, we have compared its performances with the well-known *k-means* clustering, popular *AFP* clustering, SOM based clustering and our previously proposed *HCS* among the other existing clustering based systems.

3. Proposed Fingerprint Positioning System

Our proposed Wi-Fi fingerprint positioning system which employs a robust outlier mitigation technique and also an efficient clustering method, works in two phases for position estimation as the other existing fingerprint systems. The framework of proposed positioning system is depicted in fig. 1. To remove the outlier from the RSS samples collected from various APs, we propose a robust outlier mitigation (ROM) technique for the fingerprint positioning in this paper and applies it in both phases (training phase and positioning phase) of our proposed system as shown in fig. 1. We also propose a clustering technique named as efficient clustering strategy (ECS) for fingerprint positioning herein to reduce the searching overhead as well as the size of radio map by applying fusion of similar RSS patterns. After removing outlier from the observed RSS samples, our proposed system in the positioning phase selects the appropriate cluster based on the strongest signal providing AP and then applies NNSS [12] within the selected cluster to estimate the position of unknown location as shown in fig. 1. The detailed design and working methodology of our proposed outlier mitigation technique, clustering strategy and the positioning technique adopted

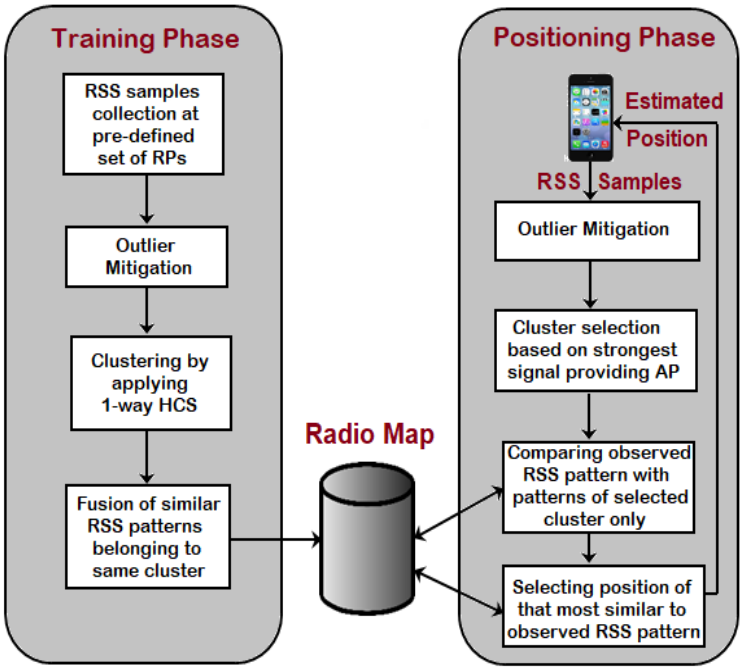


Figure 1: Framework of proposed fingerprint positioning system

by the proposed fingerprint system are provided in subsections 3.1, 3.2 and 3.3 respectively.

3.1. Proposed Outlier Mitigation Technique

The list of various abbreviations and symbols used in our proposed outlier mitigation technique is presented in table 1. The primary reason behind high variance in the values of RSS samples collected from a certain AP over time is the presence of outlier signal or sparse noises into those collected samples. Thus in this paper, we have proposed an outlier mitigation technique that can effectively remove the outlier data from the set of collected RSS samples to impede the negative effect of RSS variance problem on the positioning accuracy of fingerprint technique. To address this issue, most of the existing fingerprint systems collects sufficient number of RSS samples and computes their mean value. But the mean or average is very easily affected by the presence of out-

Table 1: List of abbreviations & symbols used in proposed outlier mitigation technique

Abbreviation/ Symbol	Meaning
APs	Access Points
HF	Hample filter
MAD	Median Absolute Deviation
RSS	Received Signal Strength
SD	Standard Deviation
TLs	Training Locations
$\overline{S}$	Mean of data set S
σ_S	Standard deviation of data set S
X_M	Median of data set X
X_{MAD}	MAD of data set X
m	Total number of APs deployed in the area of localization
n	Total number of training locationsselected in the area of localization
l_i	i^{th} training location (TL)
AP_j	j^{th} access point
V_{ij}	Set of RSS samples collected at l_i from AP_j
V_{ij}^{MAD}	MAD of RSS vector V_{ij}
V_{ij}^{o}	Resulting RSS vector after removal of outlier from V_{ij}
r_i	Outlier-free RSS pattern computed at i^{th} TL

liers within the data set. This happens because the presence of a single outlier in the data set deviates its mean from the measure of central tendency. Both the mean as well as standard deviation (SD) are highly sensitive to the presence of outlier within the data set. So, the traditional method of removing outlier which considers the data samples within the range of $mean \pm 3 \times SD$ [23] does not work for the smaller data sets.

In order to prove the above-mentioned fact, let us consider an RSS data set of 10 samples $S = [-54, -65, -63, -59, -59, -57, -61, -61, -62, -97]$. In this data set, the last sample (-97) is definitely an outlier. The mean $(\overline{S})$ and SD (σ_S) of the above data set are -63.8 and 12.07 respectively. Based on the criterion of $\overline{S} \pm \ 3 \times \sigma_S$, we find that any sample value larger than -27.58 or smaller than -100.02 can be treated as outlier. So, the above rule fails to remove the obvious outlier from the given data set. On the other hand, the median of a data set, which is computed by considering the data values located at the middle of the sorted data set, remain unaffected until more than 50% data turns to be outlier. The median of a data set X (X_M), where $X = [x_1, x_2, ..., x_n]$ and $\forall i, j, 1 \leq i < j \leq n, x_i \leq x_j$, is computed as follow.

$$X_M = \begin{cases} x_{\frac{n}{2}} + x_{\lceil \frac{n+1}{2} \rceil} & if\ n\ is\ even \\ \\ x_{\frac{n+1}{2}} & if\ n\ is\ odd \end{cases} \tag{2}$$

Median absolute deviation (MAD) [25], another estimator of data dispersion like SD, is also insensitive to the presence of outlier within the data set as median. To determine the MAD of a data set, it is at first required to compute the median of the absolute deviations of individual data element from that data set$'$ median and then multiplying it by a constant, which is set to 1.4826 based on the assumption of the data normality [26]. The MAD of the above data set X is estimated as follow [26].

$$X_{MAD} = 1.4826 \times median(|x_i - X_M|) \tag{3}$$

Since the median and MAD of a data set are highly immune to the presence of outlier, Hampel filter (*HF*), which relies on median and MAD, is more effective for removing outlier compared to other techniques [24]. *HF* considers those data elements which are within the range of median $\pm$ 3$\times$*MAD of the data set*. So, the data elements of X considered by *HF* as normal data are determined as follows.

$$X_M - 3 \times X_{MAD} < x_i < X_M + 3 \times X_{MAD} \tag{4}$$

The efficiency of *HF* in removing outlier from the smaller data set can be proved by applying it on the above-mentioned RSS data set S. As per equations 2 and 3, the median and MAD of S are -61 and 2.9652 respectively. Now, by applying equation 4, we obtain any data sample having value larger than -52.1 or smaller than -69.9 is treated as outlier. So, the data sample having value -97 is definitely removed from S by applying *HF*. Although *HF* provides more robust solution for removing outlier compared to the traditional mean and *SD* based rule, but it still has a limitation. *HF* would not work, if more than 50% samples within a data set have the same value (in which case MAD of that data set becomes zero) irrespective of whether these are outlier or normal data. Since the outlier within a data set of RSS samples collected from the Wi-Fi APs is generated by a sudden spike of noise, so it comes in very low density. Thus, our proposed outlier mitigation technique uses *HF* as an effective means for removing the outlier. To overcome the only limitation of *HF*, our proposed technique at first checks the value of MAD for a given data set. If MAD is found to have some non-zero positive value, *HF* is applied on the data set for removing outlier in case the data set is contaminated with the sparse noises. Otherwise, the data set is kept unchanged based on the assumption that some normal data populates more than 50% samples within that data set. The *median* of a data set just like the *mean* also gives the central tendency of the data set if it has some time-varying characteristic. Hence, to address the time-varying characteristics

of the RSS measurements received from a particular AP, our proposed technique computes the median of RSS data set after removing outlier from the data set and returns it as measured RSS value. The only limitation of our proposed outlier mitigation technique is that it would also fail to remove the outlier from a data set if its more than 50% samples are contaminated by some outlier. Since the probability of having more than 50% outlier data within a large data set is very low, thus our proposed outlier mitigation technique can be effectively used on a large data set rather than small data set. Our proposed outlier mitigation technique, named as *robust outlier mitigation* (ROM), is presented below.

Algorithm 1: Robust Outlier Mitigation (ROM) Scheme

Assumptions:

(i) m APs denoted as [AP_1, AP_2,$\cdots$, AP_m] are deployed in the localization area.

(ii) n training locations (*TL*) denoted as [l_1, l_2,$\cdots$, l_n] are selected in the localization area.

Input: Sets of *RSS* samples collected at each *TL* from all available APs, where k samples measured from AP_j at l_i are denoted by V_{ij}=[$v_{ij}^{(1)}$, $v_{ij}^{(2)}$, $\cdots$, $v_{ij}^{(k)}$].

Output: A set of outlier-free *RSS* patterns for all training locationsdenoted as R = [r_1, r_2,$\cdots$, r_n].

Procedure:

1. For each *TL* (l_i, where 1≤i≤n), following iterative process is executed.
   i. $V_{ij}^{MAD} \leftarrow$ MAD of set of k *RSS* samples collected from AP_j ($1 \leq j \leq m$) estimated by equation 3.
   ii. If $V_{ij}^{MAD} > 0$, apply equation 4 to obtain outlier-free *RSS* data set (V_{ij}^{o}).
   iii. Else, $V_{ij}^{o} \leftarrow V_{ij}$.
2. *RSS* pattern for i^{th} *TL* (l_i) is computed as follows.
   r_i = [$(V_{i1}^{o})_M$, $(V_{i2}^{o})_M$, $\cdots$, $(V_{im}^{o})_M$], where $1 \leq j \leq m$.

3.2. Proposed Clustering Strategy

The list of various abbreviations and symbols used in our proposed clustering strategy and positioning scheme is presented in table 2. Our proposed clustering technique works in two steps. In the first step, it partitions all the training locations into several disjoint clusters based on the working principle of *one-way hierarchical clustering strategy* ($1-way\ HCS$) proposed in our previous work [41]. The rule of $1-way\ HCS$ is that the set of training locations belonging to a cluster receive strongest signal strength from a particular AP only. So, the number of clusters created by our proposed strategy is equal to number of APs deployed in the localization area. In the second step, our proposed technique at first calculates the euclidean distances between any pair of the RSS patterns belonging to the same cluster and then adjoin those patterns whose euclidean distances are bounded by a certain value (called *threshold*) into a single RSS pattern. Even though, the RSS samples measured from a particular AP at some location fluctuates over time, the RSS patterns collected from a set of APs within some indoor area usually exhibit the *spatial correlation property*. According to *spatial correlation property*, measured RSS values taken at the nearby training locations from some AP are very similar to each other whereas those taken at the distant locations have higher degree of differences as also evidenced in [27]. The existence of this property within the RSS data sets can be theoretically proved in the following way.

Considering the effect of radio irregularity on the wireless propagation channel [45] as well as the presence of walls within the indoor areas, the value of RSS obtained at location l from k^{th} AP ($1 \leq k \leq m$) can be formulated as follows [42].

$$\begin{aligned} P_r(l, AP_k) = \ & P_t^{vsp}(AP_k) - PL^{doi}(l, AP_k) \\ & - PL^{waf}(l, AP_k) + N(0, \sigma), \end{aligned} \tag{5}$$

where $P_t^{vsp}(AP_k) \rightarrow$ power transmitted by k^{th} AP (AP_k),
$PL^{doi}(l,\ AP_k) \rightarrow$ path loss at location l from AP_k due to radio irregularity

Table 2: List of abbreviations & symbols used in proposed efficient clustering and positioning scheme

Abbreviation/ Symbol	Meaning
P_r	Amount of received power
P_t	Amount of transmit power
PL	Path loss
vsp	Variance of sending power
doi	Degree of irregularity
waf	Wall attenuation factor
$N(0, \sigma)$	Amount of Gaussian noise with mean 0 and SD σ
RM	Radio map
C_i	i^{th} cluster
$\|C_i\|$	Number of training locations within cluster C_i
δ	Value of threshold
$\|RM\|$	Total number of records within RM
w_p	pth $subsetwithinacluster$
sv_p	Representative RSS vector associated with w_p
sv_{jp}	Representative RSS vector of p^{th} subset within C_j
sx_{ji}	X co-ordinate of the median point of i^{th} subset within C_j
sy_{ji}	Y co-ordinate of the median point of i^{th} subset within C_j
k_i	Number of subsets within C_i

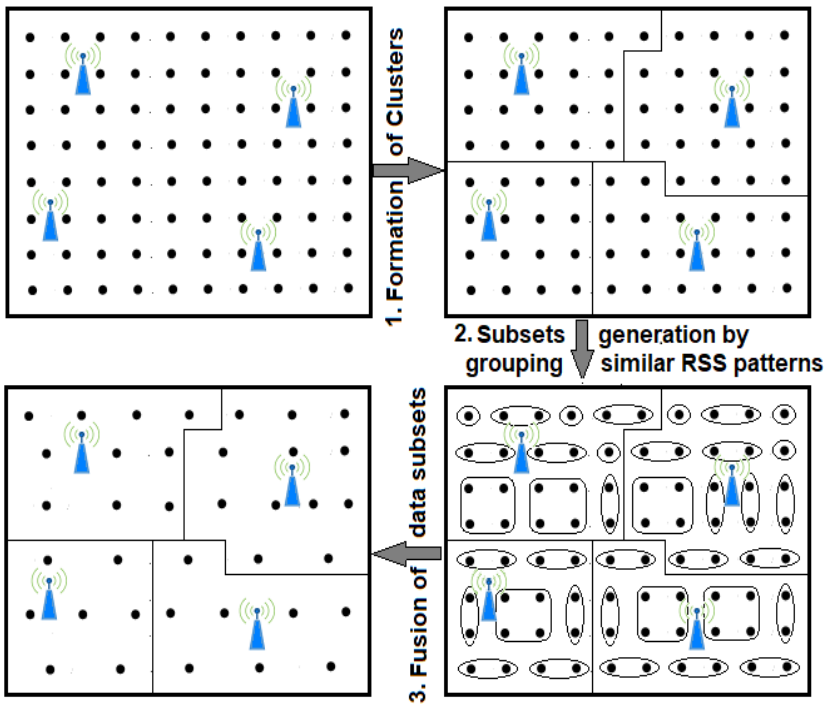


Figure 2: Working process of proposed efficient clustering strategy

properties of the wireless signal,

$PL^{waf}(l,\ AP_k) \rightarrow$ path loss at location l due to presence of walls and other obstacles between location l and AP_k,

$N(0,\ \sigma) \rightarrow$ a normally distributed random variable with mean 0 and *SD* σ to denote the noise incorporated into the *RSS* measurements. The expanded formulation of the transmitted power and various path loss components mentioned above are provided in our earlier work [42].

Now the difference in RSS measurements taken from AP_k at locations l_i and l_j (where $1 \leq i, j \leq n$ and $i \neq j$), denoted as ΔRSS_{ij}^k can be obtained as follows.

$$\begin{aligned} \Delta RSS_{ij}^k &= P_r(l_i, AP_k) - P_r(l_j, AP_k) \\ \Rightarrow \Delta RSS_{ij}^k &= (PL^{doi}(l_j, AP_k) - PL^{doi}(l_i, AP_k)) \\ &\quad + (PL^{waf}(l_j, AP_k) - PL^{waf}(l_i, AP_k)) \\ &\quad + (N(0, \sigma_i) - N(0, \sigma_j)) \end{aligned}$$

Since the number of walls and other obstacles that comes in between a certain AP and the neighboring training locations are usually almost equal, $PL^{waf}(l_i, AP_k) \approx PL^{waf}(l_j, AP_k)$, if l_i and l_j are two neighboring training locations.

On the other hand, $N(0, \sigma_i) \approx N(0, \sigma_j)$ if l_i and l_j are two neighboring training locations as the values of various weather parameters like temperature, humidity etc., at two neighboring training locations remain almost same. Based on the above reasoning, it can be inferred that the degree of differences in RSS values measured at neighboring training locations is very low compared to that taken at the far away training locations. Our proposed clustering strategy applies the *spatial correlation property* of RSS data sets belonging to the same cluster to further divide that cluster into several subsets. Each of such subsets include either only one TL or multiple training locations as shown in fig. 2. The whole working process of our proposed clustering technique when applied it to a localization area having 4 APs, are schematically depicted by fig. 2. A new parameter called *threshold* (denoted as δ) is introduced to set the upper bound value of the *euclidean distance* [44] between each pair of *specially correlated RSS patterns* that can be grouped into a subset within the same cluster. After fusion of the RSS patterns belonging to each such subset, only one radio map entry is created in order to reduce the amount of storage. Each radio map entry for a certain subset consists of the representative *RSS* vector associated with that subset and the positional co-ordinates of the median point of that subset. The representative *RSS* vector (sv) of a subset (say k^{th} subset in i^{th} cluster) and the positional co-ordinates of its median point (sx, sy) are determined by applying average fusion on the outlier-free *RSS* patterns of all training locations belonging to that subset and their positional co-ordinates respectively. Thus they are estimated as follows.

$$sv_{ik} = \frac{\sum_{j=1}^{b} r^{(j)}}{b}, \tag{6}$$

where b is the number of training locationswithin k^{th} subset in the cluster C_i and the set of outlier-free *RSS* patterns computed at those b training locationsare

denoted as $[r^{(1)}, r^{(2)}, \cdots, r^{(b)}]$.

$$sx_{ik} = \frac{\sum_{j=1}^{b} x_j}{b}, sy_{ik} = \frac{\sum_{j=1}^{b} y_j}{b}, \tag{7}$$

where $[(x_1,y_1), (x_2,y_2), \cdots, (x_b,y_b)]$ denotes the positional co-ordinates of b training locations within k^{th} subset in the cluster C_i.

After formation of the clusters by applying $1-way\ HCS$ as shown in fig. 2, a significant issue is how to select the initial TL which would begin data fusion process within each cluster. To partition each cluster into an optimum number of subsets, our proposed strategy considers the TL that obtains maximum RSS value from the strongest AP as the initial point to begin the data fusion process within a cluster. More formally, for i^{th} cluster (C_i), the TL that obtains maximum RSS from AP_i (Strongest AP for cluster C_i) is considered as initial data point to begin the fusion process. Our proposed clustering technique is named as *efficient clustering strategy* (*ECS*) because it can significantly reduce the storage requirement for radio map and also the searching overhead in the positioning phase compared to other existing clustering technique. The algorithm adopted by the proposed strategy *ECS* to construct the compressed clustered radio map is presented below.

Algorithm 2: Construction of compressed clustered radio map

Input:

i. Set of clusters $[C_1, C_2, \cdots, C_m]$.

ii. Set of positional co-ordinates of all training locations $[(x_1,y_1), (x_2,y_2), \cdots, (x_n,y_n)]$

iii. Set of outlier-free *RSS* patterns computed at n training locations $[r_1, r_2, \cdots, r_n]$

iv. Value of threshold (δ)

Output: A compressed clustered radio map

Procedure:

1. For each cluster C_i ($i = 1$ to m), following iterative process is executed.
   i. $n_i = |C_i|$ and $L_i = [l^{(1)}, l^{(2)}, \cdots, l^{(n_i)}]$, where L_i is the set of training locations belonging to C_i
   ii. Set outlier-free *RSS* pattern computed at $l^{(i)}$ to $r^{(i)}$.
   iii. Determine the training location within C_i that receives maximum RSS from AP_i, where $1 \leq m_i \leq n_i$ and set it to $l^{(m_i)}$.
   iv. k is initialized to 1 and $l^{(m_i)}$ is included in w_1, where w_1 is the first subset within cluster C_i.
   v. The outlier-free RSS pattern $r^{(m_i)}$ obtained at $l^{(m_i)}$ is set to representative *RSS* vector of subset w_1 (denoted as sv_1).
   vi. Repeat the following steps for $1 \leq j \leq n_i$ and $j \neq m_i$.
      a. Compute euclidean distances between $r^{(j)}$ and the representative *RSS* vector of existing subsets (k).
      b. If $\left\|r^{(j)}, sv_p\right\| \leq \delta$ for some $p, 1 \leq p \leq k$, then following steps are executed
         I. $l^{(j)}$ is included in w_p
         II. recompute sv_p by using equation 6.

      c. If $\left\|r^{(j)}, sv_p\right\| > \delta$ for all $p, 1 \leq p \leq k$, then a new subset is created as follows.
         I. k is incremented by 1
         II. $l^{(j)}$ is included in w_k
         III. $r^{(j)}$ is set to sv_k

   vii. Positional co-ordinates of each subset's median point ($[sx_{ip}$ $sy_{ip}]$ is estimated by equation 7.
   viii. A record containing the positional co-ordinates of median point and representative *RSS* vector of each subset ($[sx_{ip}$ sy_{ip} $sv_{ip}]$) is included in the radio map.

2. If $k_i \leftarrow$ number of subsets belonging to C_i, total number of records within the radio map is determined as follows.

$$|RM| = \sum_{i=1}^{m} k_i \tag{8}$$

By integrating our proposed outlier mitigation technique presented in subsection 3.1 with the proposed clustering strategy described above (*ECS*), we name it *efficient clustering strategy with robust outlier mitigation* (*ECS-ROM*).

3.3. Positioning phase

In this phase, our proposed outlier mitigation technique described in subsection 3.1 is at first applied on the RSS samples collected at the unknown location to generate an outlier-free RSS pattern as shown in fig. 1. Then all elementary RSS values within the generated RSS pattern are compared to determine strongest signal strength providing AP based on which the appropriate cluster is selected in this phase. In the next step, the outlier-free RSS pattern observed in this phase is compared with those data records of the radio map belonging to the selected cluster only as shown in fig. 1. Finally, the positional co-ordinates of the stored *RSS* pattern which has shortest *euclidean distance* from the observed pattern is returned as estimated position of the unknown location as depicted by fig. 1. The positioning algorithm adopted by our proposed fingerprint system is presented below.

Algorithm 3: Positioning Algorithm

Input:

i. Compressed clustered *radio map*

ii. Outlier-free *RSS* pattern observed at unknown location ($r^o = [v_1^o, v_2^o, \cdots, v_m^o]$, where m is number of *APs*)

Output: Estimated position ($[x^e \ y^e]$)

Procedure:

1. $C_j \leftarrow$ `selected cluster, if following condition satisfies` $\forall i, 1 \leq i \leq m \wedge \exists j, 1 \leq j \leq m, j \neq i, s.t., v_j^o \geq v_i^o$.
2. $RM_j \leftarrow$ `records of radio map belonging to cluster` C_j.
3. $[x^e\ y^e] = [sx_{ji}\ sy_{ji}] \Leftrightarrow sv_{ji} = \arg\min_{sv_{jp}} \|r^o, sv_{jp}\|$, `where` $i \neq p$, $1 \leq i, p \leq k$ `and` k `is number of subsets within` C_j.

The searching overhead of the proposed clustering strategy (ECS) in the positioning phase is $O(1 + \bar{k_i})$, as it comprises the summation of some fixed duration needed to select an appropriate cluster and a certain number of comparisons, which are equal to the average number of subsets within each selected cluster, required to determine the position.

4. Performance Analysis

This section evaluates and compares the performances of our proposed fingerprint system with the other existing clustering based fingerprint systems which include $k-means$ clustering [33], affinity propagation clustering ($AFPC$) [37, 39], self organizing map (SOM) based clustering [34], our previously proposed *two-way hierarchical clustering strategy* ($2-way\ HCS$) [41] and also an existing non-clustering fingerprint system *NNSS* [12] in both simulation test bed as well as real indoor environment. The performances of the above-mentioned fingerprint systems are extensively analyzed in terms of *average positioning error*, *average positioning time* and *percentage of reduction in storage*, which are defined in subsection 4.1. At first, the performance evaluations and comparisons of our proposed system with the other techniques mentioned above based on experimentation in the simulation test bed are provided in subsection 4.2. Then, the performance analysis of the same based on experimentation in some real indoor area is presented in subsection 4.3. Each experiment is conducted 50 times and their average results are provided here for performance analysis.

4.1. Performance Metrics

The following performance metrics have been considered in this paper in order to analyze the performances of our proposed system.

i. **Average Positioning Error:** The distance between an object′s true position and its estimated position determined by a positioning system, is defined as the *positioning error* acquired by that positioning system. The *average positioning error* (**APE**) acquired by a positioning system is estimated as follows.

$$APE = \frac{1}{p}\sum_{l=1}^{l=p}\sqrt{(x_l^e - x_l^t)^2 + (y_l^e - y_l^t)^2}, \tag{9}$$

where p is number of test points (TPs) whose locations are needed to be determined, $[x_l^e, y_l^e]$ denotes estimated position of l^{th} $(1 \leq l \leq p)$ TP whereas its true position is denoted by $[x_l^t, y_l^t]$.

ii. **Average Positioning Time:** The time interval between the instant estimated position is obtained and the instant the request for localization is made. The *average positioning time* (**APT**) incurred by a positioning system is computed as follows.

$$APT = \frac{1}{p}\sum_{l=1}^{l=p} pt_l, \tag{10}$$

where pt_l is estimated positioning time associated with l^{th} TP and p is number of test pointsat which positioning service is invoked.

iii. **Percentage of Reduction in Storage:** The storage requirements for the radio map yielded by a certain fingerprint positioning system reduces with the increasing value of percentage of reduction in storage (**PRS**). It is computed as follows.

$$PRS = \left(1 - \frac{|RM|}{n}\right) * 100, \tag{11}$$

where $|RM|$ denotes number of records included within the radio map created by some fingerprint system and n is number of training locations considered within the localization area.

4.2. Simulation based Performance Analysis

An area of size 100×100 $meter^2$ has been considered as the field of localization in which some horizontal and vertical walls are placed along the line

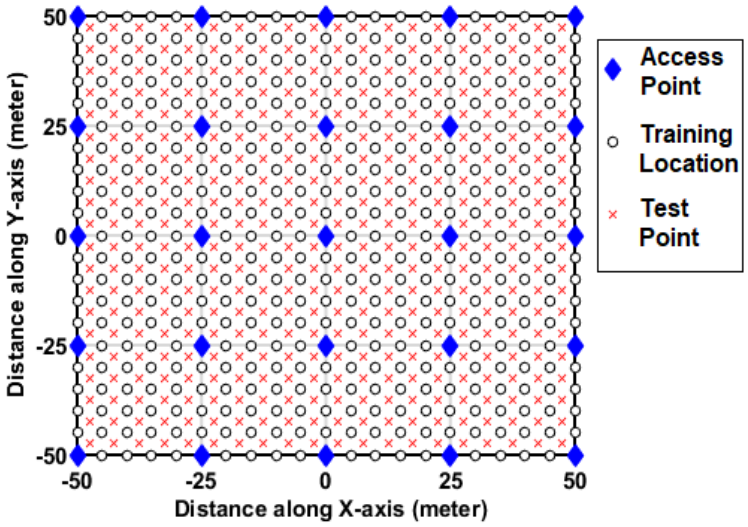
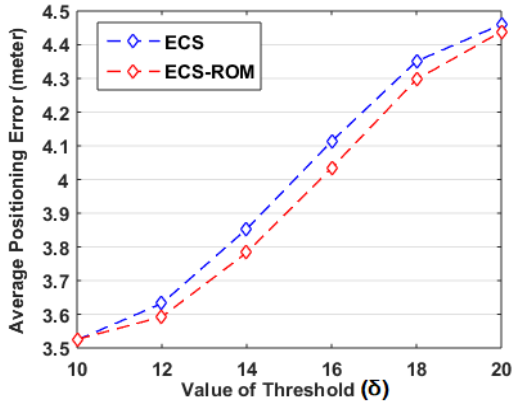


Figure 3: Simulation test bed

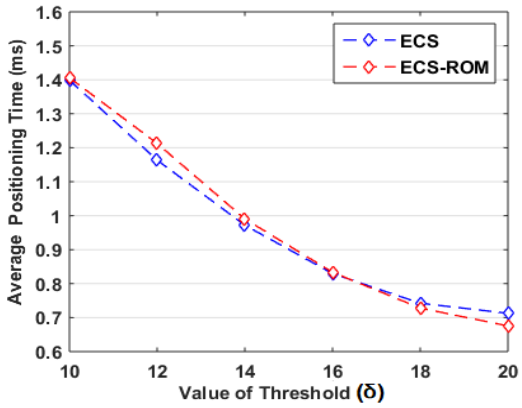
segments $([-50, i \times 25], [50, i \times 25])$ and $([i \times 25, -50], [i \times 25, 50])$, $-2 \leq i \leq 2$ as shown in fig. 3, to introduce obstacles within the localization area. The positions of the APs deployed in the simulation testbed along with selected training locations and test points are explicitly shown in fig. 3. The distance between two neighboring APs is set to 25 meter, whereas that between two consecutive training locations, which is called *training grain size* (TGS), is set to 5 meters in this simulation. The values of RSS measurements at various training locations and test points within the simulation area are determined by applying equation 5. The default values of various parameters used for modeling of RSS measurements under the presence of radio irregularities and wall attenuation effect in the indoor environment [42] are given below.

$P_t = 20$ dBm, $D_0 = 1$m, $PL(D_0) = 37.3$ dBm, $\gamma = 4$, $\sigma = 4$, $VSP = 0.2$, $DOI = 0.02$, $WAF = 3$, $N_{max} = 4$.

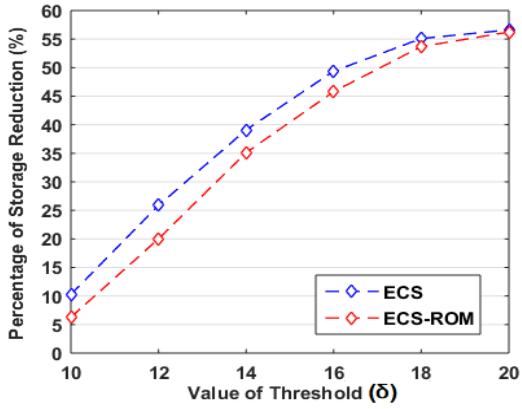
At first, the performances of our proposed clustering strategies, *ECS* and *ECS-ROM*, are compared in terms of **APE**, **APT** and *percentage of reduction in storage* under the different values of threshold (δ) to demonstrate the effectiveness of the latter (*ECS-ROM*) over the former (*ECS*) that relies on the



(a) Average positioning error vs threshold



(b) Average positioning time vs threshold

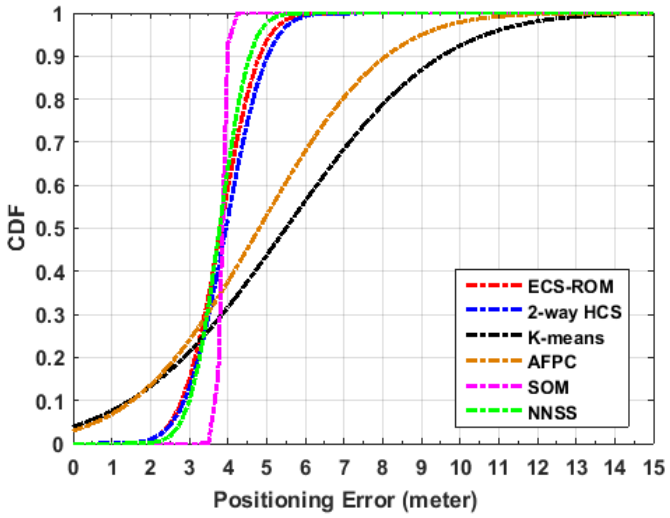


(c) Percentage of reduction in storage vs threshold

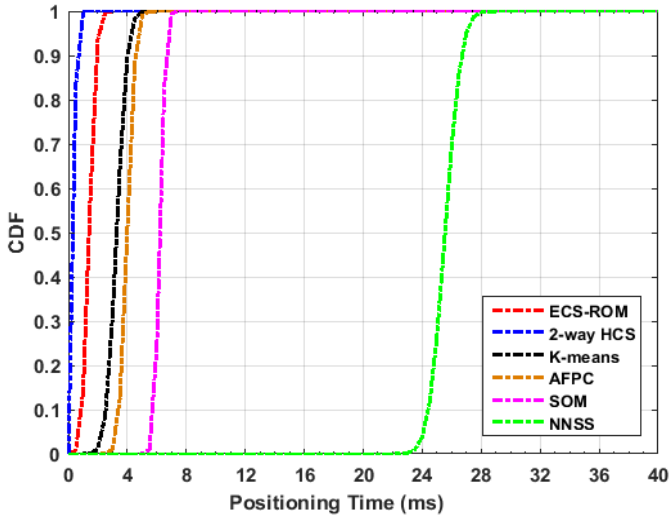
Figure 4: Impact of threshold on performances of proposed techniques *ECS* and *ECS-ROM* based on simulation results.

average value of collected RSS samples for removing the sparse noises. From this performance analysis, an optimum value for threshold is chosen so that proposed strategy, *ECS-ROM* acquires lower **APE** and **APT** under certain amount of reduced storage and then, based on the selected value of threshold (which becomes its default value), the performances of our proposed fingerprint system *ECS-ROM* are further compared with the other systems considered in this paper in terms of **APE** and **APT**. Figs. $4(a)$, $4(b)$ and $4(c)$ show the impact of threshold (δ) on the performances of our proposed techniques, *ECS* and *ECS-ROM*, in terms of **APE**, **APT** and *percentage of reduction in storage* respectively by varying its value in the range of $[10 - 20]$. It is evident from fig. $4(a)$ that **APE** acquired by the proposed strategy *ECS-ROM* is lower than that of *ECS*. Thus, our proposed outlier mitigation technique (ROM) is more effective in removing the sparse noises or outlier from the collected RSS samples compared to the traditional method that relies on their average value. Figs. $4(a)$, $4(b)$ and $4(c)$ also depict that the increasing value of threshold increases **APE** and *percentage of reduction in storage* acquired by our proposed techniques but reduces their **APT**. This happens because higher number of training locations are adjoined into each subset with the increasing value of threshold, which, in turn, decreases the average number of subsets per cluster and so increases their **APE** and percentage of reduction in storage, but reduces their **APT**. On the other hand, the *percentage of reduction in storage* for the other fingerprint systems considered in this paper remain 0 since none of these existing systems apply any RSS data fusion process. As our proposed technique *ECS-ROM* achieves optimum performances in terms of **APE**, **APT** and *percentage of reduction in storage* when the value of threshold becomes 12, henceforth the default value of threshold for proposed *ECS-ROM* is set to 12 to compare its performances with the other techniques considered in this paper in the simulation testbed.

The distributions of positioning error (**PE**) and positioning time (**PT**) incurred by our proposed fingerprint technique and other techniques considered in this paper are illustrated in figs $5(a)$ and $5(b)$ respectively. Based on the cumula-



(a) Distributions of positioning error



(b) Distributions of positioning time

Figure 5: Comparisons of positioning accuracy and positioning time of proposed *ECS-ROM* and 5 other techniques in terms of CDF based on simulation results.

Table 3: Comparisons of positioning error (PE) & positioning time (PT) acquired by various techniques based on simulation results

Positioning system	PE (meter)		PT (ms)	
	50%	95%	50%	95%
ECS-ROM	3.72	4.84	1.47	2.3
2-*way HCS*	3.94	5.5	0.41	1.0
k-means	5.66	11.75	3.25	4.5
AFPC	4.35	7.5	4.0	5.0
SOM	3.71	4.5	6.9	7.8
NNSS	3.73	4.75	25.97	27.5

tive distributions of **PE** and **PT** acquired by those techniques in the simulation test bed, 50% (average) & 95% **PE** and **PT** acquired by them have been provided in table 3. The comparative results on **PE** given in table 3 shows that our proposed system *ECS-ROM* achieves almost similar performances as that of existing non-clustering fingerprint technique *NNSS* and outperforms other existing clustering-based techniques except *SOM* based clustering in terms of distribution of the positioning error. On the other hand, comparative results on **PT** provided in table 3 depicts that the proposed *ECS-ROM* and our previously proposed $2-way\ HCS$ achieves better performances in terms of positioning time distributions compared to the other techniques considered in this paper. Figs 5(a) and 5(b) also demonstrates that non-clustering technique *NNSS* incurs very high positioning time compared to our proposed and other existing clustering techniques considered in this paper. This is so because the added time of cluster selection followed by comparison of the *RSS* pattern observed in the positioning phase with fewer data records belonging to the selected cluster only is much lower than that needed to compare the observed *RSS* pattern with all data records within the radio map as done by *NNSS*.

The performance comparison of our proposed system with the other systems considered in this paper in terms of **APE** under different amount of noises

Table 4: Comparisons of **APE** (meter) under different amount of noises.

Value of σ	*ECS-ROM*	2-*way HCS*	*k-means*	*AFPC*	*SOM*	*NNSS*
2	3.75	3.90	4.85	4.66	3.76	3.75
3	3.76	3.92	4.86	4.67	3.77	3.78
4	3.79	3.95	4.87	4.69	3.78	3.79
5	3.87	4.05	4.96	4.79	3.87	3.88
6	4.02	4.31	5.13	4.97	4.03	4.03

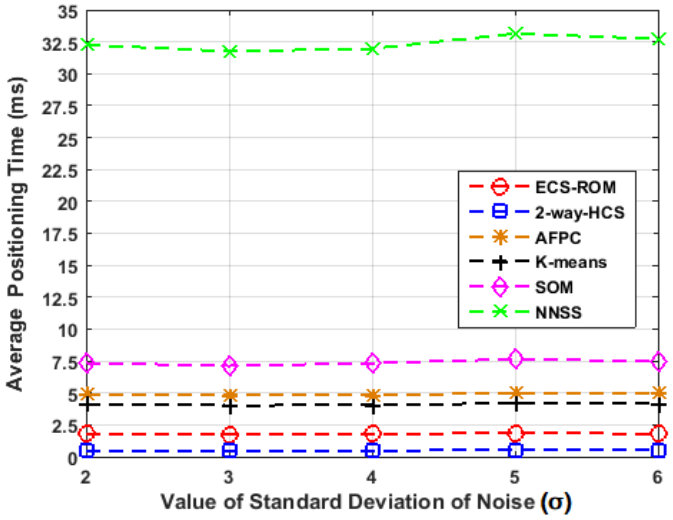


Figure 6: Comparisons of **APT** (ms) under different amount of noises

Table 5: Comparisons of **APE** (meter) w.r.t. different values of *training grain size*.

TGS (meter)	*ECS-ROM*	2-*way HCS*	*k-means*	*AFPC*	*SOM*	*NNSS*
2	2.64	2.64	2.87	3.96	2.61	2.60
4	3.26	3.37	4.62	4.18	3.25	3.22
6	4.47	4.65	5.89	5.12	4.46	4.42
8	5.72	6.04	8.29	6.42	5.74	5.71

yielded by varying the standard deviation of noise (σ) within the range $[2-6]$ is provided in table 4 and the same for **APT** is depicted by fig. 6. The comparative results given in table 4 depict that our proposed system, *ECS-ROM*, outperforms other clustering techniques considered in this paper except *SOM* based clustering and achieves almost similar performance like existing *NNSS* and *SOM* based clustering. It also shows that the presence of increasing amount of noises into the RSS measurements increases **APE** acquired by each fingerprint system considered in this paper, which is very usual phenomenon in real environment. On the other hand, fig. 6 depicts that **APT** incurred by the non-clustering fingerprint technique, *NNSS*, is much higher than that of the clustering based fingerprint techniques considered in this paper as also demonstrated in fig. 5(b). Fig. 6 also shows that **APT** incurred by all fingerprint techniques considered in this paper remain insensitive to the variance in the level of noise.

The performances of our proposed system (*ECS-ROM*) and the other systems considered in this paper in terms of **APE** and **APT** with respect to different values of training grain size (*TGS*) are presented in table 5 and 6 respectively. The experimental results given in table 5 depict that **APE** acquired by each of the fingerprint systems considered in this paper increases with the increasing value of *TGS*. This happens because the positioning error of a fingerprint system increases as the distance between two consecutive training locations, i.e., the value of *TGS* increases. With the increasing value of *TGS*, the number of training locations considered in the area of localization reduces.

Table 6: Comparisons of **APT** (ms) w.r.t. different values of *training grain size*

TGS (meter)	*ECS-ROM*	2-*way HCS*	*k-means*	*AFPC*	*SOM*	*NNSS*
2	15.52	5.64	22.85	21.52	27.38	397.8
4	4.74	1.33	8.72	10.01	16.60	88.52
6	1.42	0.38	3.97	4.43	13.57	23.04
8	1.04	0.27	3.83	3.41	9.14	14.92

Table 5 also shows that **APE** acquired by our proposed system *ECS-ROM*, existing *SOM* based clustering and *NNSS* are almost same at the different values of *TGS* and these are comparatively lower than that of the other existing clustering techniques considered in this paper. Table 6 demonstrates that **APT** incurred by our proposed techniques (*ECS-ROM* and $2-way\ HCS$) are lower than that of other existing clustering techniques considered in this paper, where as the same for non-clustering technique *NNSS* is comparatively very high at different values of *TGS*. Moreover, **APT** for *NNSS* decreases very sharply with increasing value of *TGS* while the same for each clustering based fingerprint technique considered in this paper decreases very slowly as evidenced by the data given in table 6. This happens because the positioning time for *NNSS* is directly proportional to the number of training locations within the field of localization, which sharply decreases as the value of *TGS* increases. On the other hand, the positioning time of some clustering technique is directly proportional to the average size of each cluster created by it, which slowly decreases with the increasing value of *TGS*.

The positioning accuracy of a fingerprint positioning system improves as its positioning error decreases whereas increasing positioning time indicates higher searching overhead. Thus, the above performance comparisons in terms of **APE** and **APT** between various fingerprint systems considered in this paper including our proposed one demonstrate that our proposed system *ECS-ROM* achieves better positioning accuracy compared to the other clustering based systems

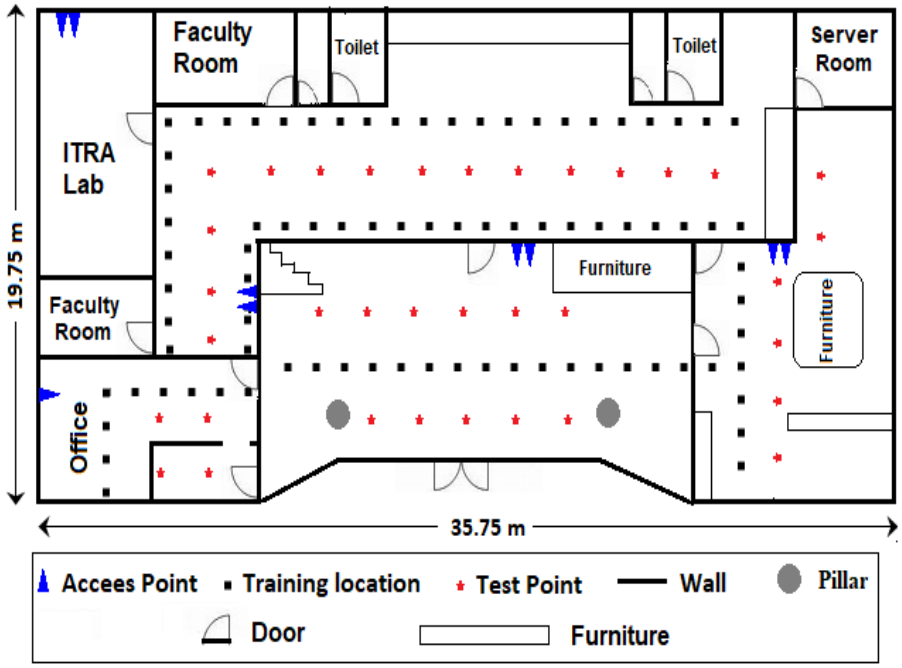


Figure 7: Experimental area inside departmental building of School of Mobile Computing & Communication.

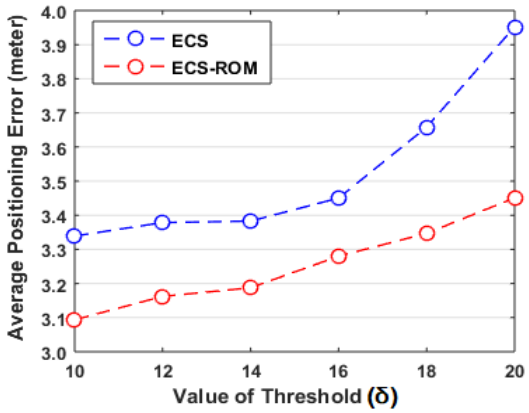
considered in this paper except *SOM* based clustering and also incurs lower searching overhead compared to them except our previously proposed $2-way$ *HCS* even by reducing about 25% storage requirement for the radio map.

4.3. Performance Analysis in Real Indoor Environment

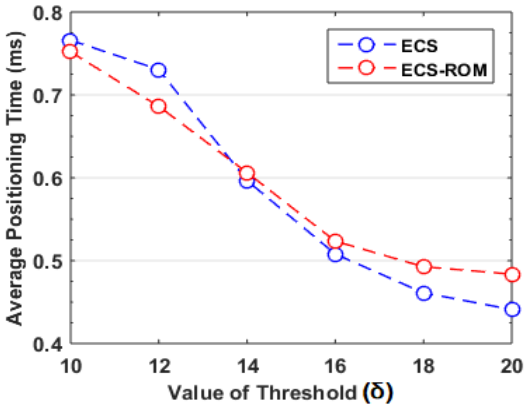
To demonstrate the effectiveness of our proposed system in the real indoor areas, experiments have been carried out extensively on the ground floor of our departmental building whose floor map along with the positions of available APs, selected training locations and test points are shown in fig. 7. We have selected 84 training locationsby maintaining the separating distance between any two consecutive training locations at 1.2 m and 35 test pointswhich are well distributed in the experimental area as shown in fig. 7. Each AP is clung to either the ceiling or side wall inside our departmental building. Physical

Figure 8: Physical placement of a certain AP on the ceiling of experimental area.

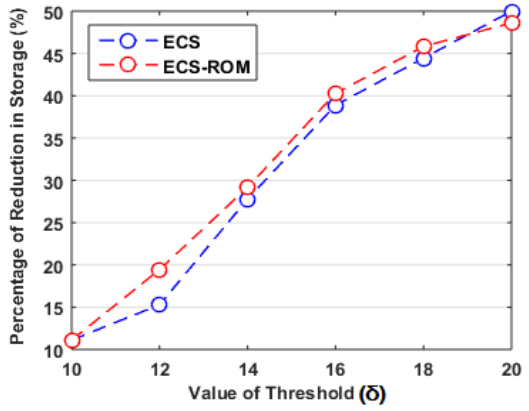
placement of a certain AP on the ground floor ceiling inside our departmental building is shown in fig. 8. Xiomi Redmi Note 4, a smart phone with $octa-core$ $aarch64$ processor and running android 7.0, has been used for collection of RSS samples in both training and positioning phase. At each TL, a total of 120 RSS samples are collected from each of the available APs (9 APs herein) by placing the above-mentioned android device at 4 different directions (east, south, west, north). The experimental site for real indoor environment includes the presence of walls and other obstacles like pillars, furniture, staircase etc., have been explicitly shown in fig. 7. Apart from the effects of above-mentioned obstacles, various other external factors have significant influences on the measured RSS values obtained from a certain AP. These include presence movements of human beings, variation in weather parameter like humidity, temperature etc. Since we have carried out the RSS data collection at the selected training locations testing points in the experimental site during office hours throughout a duration of three months, hence the effect of above said external factors on the experimen-



(a) Average positioning error vs threshold



(b) Average positioning time vs threshold



(c) Percentage of reduction in storage vs threshold

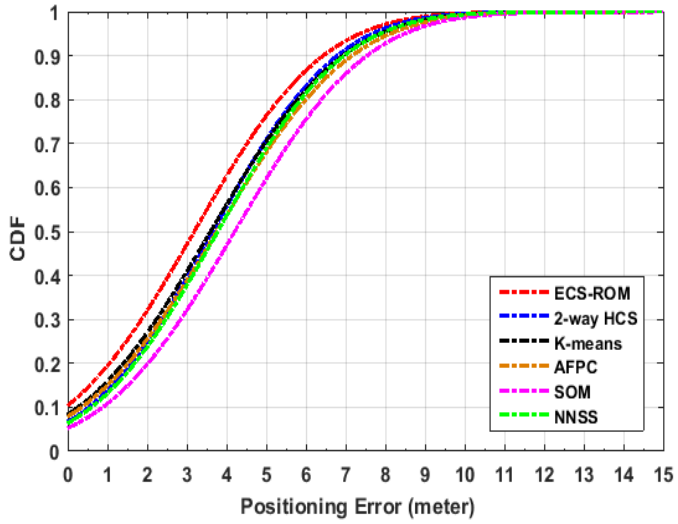
Figure 9: Impact of threshold on performances of proposed techniques *ECS* and *ECS-ROM* based on real experimental results.

Table 7: Comparisons of positioning error (PE) & positioning time (PT) acquired by various techniques in real Indoor area

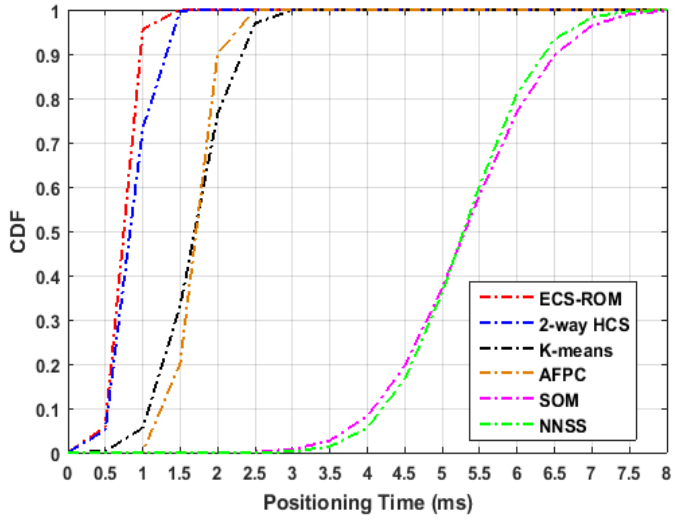
Positioning system	PE (meter)		PT (ms)	
	50%	95%	50%	95%
ECS-ROM	3.19	7.5	0.74	1.0
2-*way HCS*	3.66	7.75	0.86	1.5
k-means	3.6	8.0	1.69	2.5
AFPC	3.76	8.25	1.7	2.3
SOM	4.21	8.5	5.31	7.0
NNSS	3.78	8.0	5.29	6.8

tal results obtained in real indoor areas have already been considered. At first, the performances of our proposed clustering techniques *ECS-ROM* and *ECS* are compared in terms of **APE**, **APT** and *percentage of reduction in storage* in figs. $9(a)$, $9(b)$ and $9(c)$ respectively under the different values of threshold within the range $[10-20]$. The experimental results provided in figs. $4(a)$ and $9(a)$ show that performance accuracy improvement of the proposed technique *ECS-ROM* over *ECS* in the real environment is better than that obtained in the simulation test bed. This happens because the RSS samples collected from the APs in the real environment are contaminated with higher percentage of outliers compared to those generated in the simulation test bed. Thus, it can be inferred that our proposed outlier mitigation technique (ROM) is definitely a robust approach in removing the outliers very effectively from the RSS samples collected in the real indoor areas. The default value of threshold for our proposed system *ECS-ROM* is set to 14 at which it achieves optimum performances in terms of **APE**, **APT** and *percentage of reduction in storage* in the real environment for comparing its performances with the other existing systems considered in this paper.

The distributions of positioning error and positioning time incurred by six fingerprint techniques including our proposed one based on experimentation in real indoor areas are illustrated in figs $10(a)$ and $10(b)$ respectively. Based on

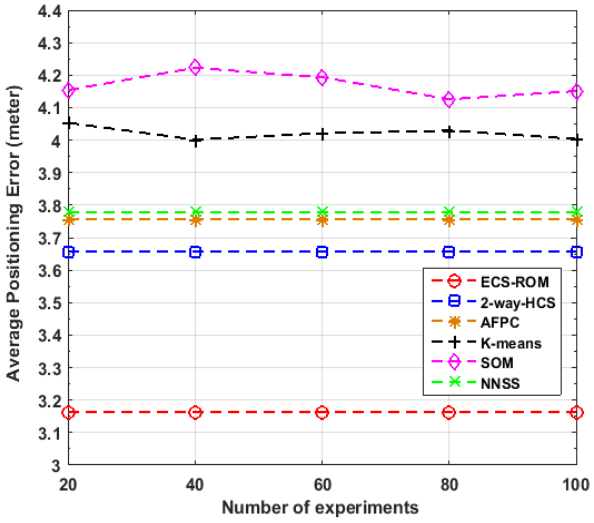


(a) Distributions of positioning error

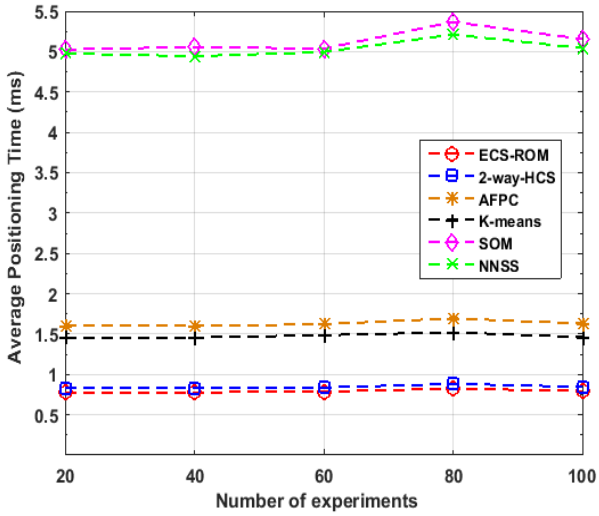


(b) Distributions of positioning time

Figure 10: Comparisons of positioning accuracy and positioning time of 6 fingerprinting techniques including proposed *ECS-ROM* in terms of CDF based on experimentation in real environment.



(a) **APE** vs no. of experiments



(b) **APE** vs no. of experiments

Figure 11: Comparisons of **APE** and **APT** w.r.t different number of experiments conducted in real environment.

the cumulative distributions of **PE** and **PT** acquired by those techniques in real indoor areas, 50% (average) & 95% **PE** and **PT** acquired by them have been provided in table 7. The comparative results given in table 7 depicts that our proposed technique *ECS-ROM* achieves better positioning accuracy in real indoor areas compared to all other existing fingerprint techniques considered in this paper. This happens because our proposed clustering adopts a robust outlier mitigation technique which is highly effective in the real indoor environments. Fig. 10(b) and table 7 also shows that the proposed *ECS-ROM* provides better performances in terms of positioning time distributions compared to other techniques considered in this paper and the performances of non-clustering technique *NNSS* compared to other clustering techniques in term of positioning time in the real indoor areas remain same as that in the simulation test bed.

On the other hand, the performance comparisons in terms of **APE** and **APT** between our proposed system *ECS-ROM* and the other existing systems considered in this paper with respect to different number of experiments (varying in the ranges of $[20 - 100]$) conducted in the real environment are depicted in figs. 11(a) and 11(b) respectively. The above graphical results depict that our proposed system *ECS-ROM* outperforms the other existing systems considered in this paper in terms of both **APE** and **APT** in the real environment. Hence, all experimental results obtained in the real indoor areas demonstrate that our proposed system *ECS-ROM* can achieve better positioning accuracy and lower searching overhead compared to all other fingerprint techniques considered in this paper even by reducing almost 30% storage requirement for the *radio map* in the real environment.

5. Conclusions and Future Research Directions

An *efficient clustering strategy* and also a *robust outlier mitigation technique* for Wi-Fi fingerprint positioning system have been proposed in this paper. The proposed clustering strategy aims to reduce the storage requirement for the radio map and also the searching overhead for fingerprint based indoor po-

sitioning system by applying our previously proposed $1-way$ ***hierarchical clustering strategy*** and then fusion of *similar RSS patterns* belonging to the same cluster. On the other hand, proposed outlier mitigation technique aims to remove the outliers as well as sparse noises from the collected **received signal strength** samples by using *Hample filter* [24] and also addresses their time-varying property to improve the positioning accuracy of the proposed system. The performances of our proposed system have been compared with four existing clustering based fingerprint techniques and one non-clustering technique in both simulation test bed as well as real indoor environment to demonstrate the efficiency and robustness of the proposed method. Simulation based experimental results given in this paper show that our proposed system achieves improved positioning accuracy compared to the existing *k-means clustering*, *affinity propagation clustering* and our previously proposed $2-way$ *hierarchical clustering strategy*, but it remains almost similar to that of existing *SOM based clustering* technique and *NNSS*. Moreover, our proposed system presented in this paper also reduces the searching overhead compared to all other existing techniques considered here except our previously proposed clustering method even by reducing about 25% storage requirement for the radio map. Simulation based experimental results also demonstrate that existing *NNSS* and *SOM based clustering* technique achieve better positioning accuracy which is almost similar to that of our proposed *ECS-ROM*, but incur higher searching overhead compared to other existing techniques considered in this paper.

On the other hand, the experimental results obtained in real indoor environment demonstrate that our proposed system outperforms five other existing techniques considered in this paper in terms of positioning accuracy and searching overhead while reduces the storage requirement for the radio map by about 30%. Among the existing techniques, *SOM based clustering* and *k-means clustering* acquire higher positioning error, whereas the former and NNSS incur higher searching overhead in the real indoor settings. Moreover, the experimentation in real indoor areas have considered the effect of the presence of walls and other obstacles and also the movement of human beings inside the experimental

site along with the effect of variation in various weather parameters on the the experimental results. Thus, the robustness of our proposed outlier mitigation technique in removing the sparse noises from the collected signal strength values as well as the efficiency of proposed system in reducing the searching overhead and storage overhead have been demonstrated via the experimental results of real indoor areas. Therefore, our proposed fingerprint based positioning system offers an effective solution to the problem of precise and real-time positioning in any complex indoor environment.

However, our proposed outlier mitigation technique has some limitation. It can be effectively used on a data set which is contaminated with less than 50% outlier data. Thus, further researches are required to overcome the above-mentioned limitation of our proposed method. To achieve this goal, we need to carry out more investigations and researches on whether stochasticity of the RSS data set can be explicitly included in the clustering techniques directly and how it can be done. Moreover, further researches are still needed to investigate the effect of redesigning our proposed clustering technique based on a different measure of dissimilarity among the data points as employed in *Gustafson Kessel fuzzy clustering* [51] instead of using traditional *Euclidean distance* which is used by most of the existing fingerprint positioning systems including our proposed one.

Acknowledgments

We appreciate the support for this research received from the European Union (EU) sponsored SmartLink project (EACEA Grant No. 2014-0858).

References

[1] P. Sadhukhan and P K Das, ″Location-aware Services in Mobile Environments,″ In proc. of Asian Mobile Computing Conference (AMOC 2006): 4th International Conference, Kolkata, India, January 4-7, 2006, pp. 152-156.

[2] M. Kadibagil and H S Guruprasad, ″Position Detection and Tracking System,″ IRACST - International Journal of Computer Science and Information Technology Security (IJCSITS), Vol. 4, No. 3, June 2014.

[3] S. Helal et al., ″Enabling location-aware pervasive computing applications for the elderly,″ In Proc. of the First IEEE international conference on pervasive computing and communications, PerCom 2003, pp. 531–536.

[4] P. Misra P. Enge, ″Special issue on global positioning system,″ Proceedings of the IEEE, Vol. 87 , Issue 1 , Jan. 1999, pp. 3–15.

[5] X. Wang et al., ″Exponentially Weighted Particle Filter for Simultaneous Localization and Mapping Based on Measurements of Magnetic Field″, IEEE Trans. Instrumentation and Measurement, Vol.66, issue 7, July 2017, pp. 1658-1667.

[6] F. Zafari, A. Gkelias and K. K. Leung, ″A Survey of Indoor Localization Systems and Technologies,″ IEEE Communications Surveys Tutorials, vol. 21, no. 3, thirdquarter 2019, pp. 2568-2599.

[7] S. He and S. -. G. Chan, ″Wi-Fi Fingerprint-Based Indoor Positioning: Recent Advances and Comparisons,″ IEEE Communications Surveys Tutorials, Firstquarter 2016, vol. 18, no. 1, pp. 466-490.

[8] P. Sadhukhan and P.K. Das, ″MGALE: A Modified Geometry-Assisted Location Estimation Algorithm Reducing Location Estimation Error in 2D Case under NLOS Environments,″ Proc. of MELT 2009, LNCS vol 5801, pp. 1-18, Springer, Heidelberg.

[9] G. Wang, H. Chen, Y. Li and N. Ansari, ″NLOS Error Mitigation for TOA-Based Localization via Convex Relaxation,″ IEEE Transactions on Wireless Communications, vol. 13, no. 8, pp. 4119-4131, Aug. 2014.

[10] B. Xu, G. Sun, R. Yu and Z. Yang, ″High-Accuracy TDOA-Based Localization without Time Synchronization,″ IEEE Trans. on Parallel and Distributed Systems, vol. 24, no. 8, pp. 1567-1576, Aug. 2013.

[11] L. Zhang, M. Chen, X. Wang and Zhi Wang, ″TOA Estimation of chirp signal in Dense Multipath Environment for Low-cost Acoustic Ranging,″ IEEE Trans. Instrumentation and Measurement, vol. 68, issue 2, Feb. 2019, pp.355-367.

[12] P. Bahl and V. N. Padmanabhan, ″RADAR: an in-building RF-based user location and tracking system,″ Proc. of 19th Annu. Joint Conf. of IEEE Comp. and Comm. Soc., 2000, vol. 2, pp. 775-784.

[13] K. Kaemarungsi and P. Krishnamurthy, ″Modeling of indoor positioning systems based on location fingerprinting,″ Proceedings of 23th Annu. Joint Conf. IEEE Comput. Commun. Soc., Hong Kong, 2004, pp. 1012–1022.

[14] V. Honkavirta, T. Perala, S. Ali-Loytty, and R. Piche, ″A comparative survey of WLAN location fingerprinting methods,″ Proceeding of 6th Workshop Positioning, Navig. Commun., Hannover, Germany, 2009, pp. 243–251.

[15] K. Chang and D. Han, ″Crowdsourcing-based radio map update automation for wi-fi positioning systems,″ Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information, Nov. 2014, pp. 24–31, https://doi.org/10.1145/2676440.2676441.

[16] C. Wu, Z. Yang and Y. Liu, ″Smartphones Based Crowdsourcing for Indoor Localization,″ IEEE Transactions on Mobile Computing, vol. 14, no. 2, Feb. 1 2015, pp. 444-457.

[17] Q. Jiang, Y. Ma, K. Liu and Z. Dou, ″A Probabilistic Radio Map Construction Scheme for Crowdsourcing-Based Fingerprinting Localization,″ IEEE Sensors Journal, vol. 16, no. 10, May 15, 2016, pp. 3764-3774.

[18] C. Wu, Z. Yang, Y. Liu, and W. Xi, ″WILL: Wireless Indoor Localization without Site Survey,″ IEEE Transactions on Parallel and Distributed Systems, Vol. 24, No. 4, April 2013, pp. 839-848.

[19] S. Jung and D. Han, ″Automated Construction and Maintenance of Wi-Fi Radio Maps for Crowdsourcing-Based Indoor Positioning Systems,″ in IEEE Access, vol. 6, pp. 1764-1777, 2018.

[20] W. Sun et al., ″Augmentation of Fingerprints for Indoor WiFi Localization Based on Gaussian Process Regression,″ IEEE Transactions on Vehicular Technology, vol. 67, no. 11, Nov. 2018, pp. 10896-10905.

[21] L. Zheng et al., ″A Deep-Learning-Based Self-Calibration Time-Reversal Fingerprinting Localization Approach on Wi-Fi Platform,″ IEEE Internet of Things Journal, vol. 7, no. 8, Aug. 2020, pp. 7072-7083.

[22] Y. Chen and J. Juang, ″Outlier-Detection-Based Indoor Localization System for Wireless Sensor Networks,″ International Journal of Navigation and Observation, Vol. 2012, Hindawi, DOI:10.1155/2012/961785.

[23] D. Cousineau and S. Chartier, ″Outliers detection and treatment: A review,″ International Journal of Psychological Research, Vol. 3(1), 2010, pp. 58–67.

[24] R. K. Pearson, ″Outliers in process modeling and identification,″ IEEE Transactions on Control Systems Technology, vol. 10, no. 1, Jan. 2002, pp. 55-63.

[25] T. Pham-Gia and T.L. Hung, ″The mean and median absolute deviations,″ Mathematical and Computer Modelling, Vol. 34, Issues 7–8, October 2001, pp. 921-936.

[26] C. Leys et al., ″Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median,″ Journal of Experimental Social Psychology, Vol. 49, 2013, pp. 764–766.

[27] S.-P. Kuo and Y.-C. Tseng, ″Discriminant minimization search for Large-Scale RF-based Localization Systems,″ IEEE Transaction on Mobile Computing, Vol. 10, No. 2 Feb. 2011, pp. 291-304.

[28] H. Li et al, ″TILoc: Improving the Robustness and Accuracy for Fingerprint-Based Indoor Localization,″ IEEE Internet of Things Journal, vol. 7, no. 4, April 2020. pp. 3053-3066.

[29] E. Gokcay and J. Principe, ″Information Theoretic Clustering,″ IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 2, Feb. 2002, pp. 158-172.

[30] B.J. Frey and D. Dueck, ″Clustering by Passing Messages Between Data Points,″ Science, vol. 315, no. 1, Feb. 2007, pp. 972-976.

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

[32] Y. Ji, S. Biaz, S. Pandey, and P. Agrawal, ″ARIADNE: A Dynamic Indoor Signal Map Construction and Localization System,″ Proc. ACM Intl Conf. Mobile Systems, Applications, and Services, 2006, pp. 151-164.

[33] S.-P. Kuo, B.-J. Wu, W.-C. Peng, and Y.-C. Tseng, ″Cluster-Enhanced Techniques for Pattern-Matching Localization Systems,″ Proceedings of IEEE Int′l Conference Mobile Ad-Hoc and Sensor Systems, 2007.

[34] Teuvo Kohonen, ″The self-organizing map,″ Neurocomputing 21, Elsevier, 1998, pp. 1-6.

[35] L. Mengual et al., ″Clustering-based location in wireless networks,″ Expert Systems with Applications 37 (2010) pp. 6165-6175.

[36] C. Feng, W. S. A. Au, S. Valaee and Z. Tan, ″Received-Signal-Strength-Based Indoor Positioning Using Compressive Sensing,″ in IEEE Transactions on Mobile Computing, vol. 11, no. 12, Dec. 2012, pp. 1983-1993.

[37] Z. Tian et al., ″Fingerprint indoor positioning algorithm based on affinity propagation clustering,″ EURASIP Journal on Wireless Communications and Networking 2013, doi: 10.1186/1687-1499-2013-272.

[38] G. Ding et al., ″Fingerprinting Localization Based on Affinity Propagation Clustering and Artificial Neural Networks,″ IEEE Wireless Communications and Networking Conference 2013 (WCNC 2013): NETWORKS, pp. 2317-2322.

[39] S. Subedi, H. Gang, N. Y. Ko, S. Hwang and J. Pyun, ″Improving Indoor Fingerprinting Positioning With Affinity Propagation Clustering and Weighted Centroid Fingerprint,″ in IEEE Access, vol. 7, pp. 31738-31750, 2019.

[40] W. Xue et al., ″APs' Virtual Positions-Based Reference Point Clustering and Physical Distance-Based Weighting for Indoor Wi-Fi Positioning,″ IEEE Internet of Things Journal, vol. 5, no. 4, Aug. 2018, pp. 3031-3042.

[41] A. Saha and P. Sadhukhan, ″A novel clustering strategy for fingerprinting-based localization system to reduce the searching time,″ Proc. of 2015 IEEE 2nd International Conference on Recent Trends in Information System (ReTIS′15), 9-11 July, 2015, Kolkata, India, pp. 538-543.

[42] P. Sadhukhan, ″Performance Analysis of Clustering-based Fingerprinting Localization Systems″, Wireless Networks, Vol. 25, Issue 5, pp. 2497–2510, July 2019, Springer US.

[43] L. Wang, H. Zhiyuan, and S. Wenjing, ″A Novel Self-Adaptive Affinity Propagation Clustering Algorithm Based on Density Peak Theory and Weighted Similarity,″ IEEE Access 7 (2019): pp. 175106-175115.

[44] R. Fabbri et al, ″2D Euclidean distance transform algorithms: A comparative survey,″ ACM Computing Surveys (CSUR), Vol. 40 Issue 1, February 2008.

[45] G. Zhou et al., ″Impact of Radio Irregularity on Wireless Sensor Networks,″ in Proceedings of the 2nd international conference on Mobile systems, applications, and services, 2004, pp. 125-138.

[46] P. Sadhukhan, K. Dahal and Z. Pervez, ″Impact of beacon coverage on clustering strategies for fingerprinting localization system,″ Proc. of 2017 International Conference on Computing, Networking and Communications (ICNC′17), Santa Clara, CA, 2017, pp. 184-188.

[47] X. Shi, J. Guo and Z. Fei, ″WLAN Fingerprint Localization with Stable Access Point Selection and Deep LSTM,″ 2020 IEEE 8th International Conference on Information, Communication and Networks (ICICN), Xi′an, China, 2020, pp. 56-62.

[48] Z. Wang et al., ″A Hybrid Wi-Fi Fingerprint-Based Localization Scheme Achieved by Combining Fisher Score and Stacked Sparse Autoencoder Algorithms,″ Mobile Information Systems 2020 (2020).

[49] S. Zhang, F. Xin-Yue and L. Xiao-Yong, ″Wireless Indoor Positioning Algorithm Based on RSS and CSI Feature Fusion,″ In International Conference in Communications, Signal Processing, and Systems, Springer, Singapore, 2019, pp. 2057-2067.

[50] Y. Tian, B. Huang, B. Jia, L. Zhao, ″Optimizing AP and Beacon Placement in WiFi and BLE hybrid localization,″ Journal of Network and Computer Applications, Volume 164, 2020.

[51] Chaomurilige, J. Yu, M.-S. Yang, ″Deterministic annealing Gustafson-Kessel fuzzy clustering algorithm,″ Information Sciences, Volume 417, 2017, pp. 435-453.