doi:10.3772/j.issn.1006-6748.2020.04.006

An effective indoor radio map construction scheme based on crowdsourced samples①

Guo Ruolin (郭若琳), Qin Danyang②, Zhao Min, Xu Guangchao
(Provincial Key Laboratory of Electronic Engineering, Heilongjiang University, Harbin 150080, P. R. China)

Abstract

The crowdsourcing-based WLAN indoor localization system has been widely promoted for the effective reduction of the workload from the offline phase data collection while constructing radio maps. Aiming at the problem of the inaccurate location annotation of the crowdsourced samples, the existing invalid access points (APs) in collected samples, and the uneven sample distribution, as well as the diverse terminal devices, which will result in the construction of the wrong radio map, an effective WLAN indoor radio map construction scheme (WRMCS) is proposed based on crowdsourced samples. The WRMCS consists of 4 main modules: outlier detection, key AP selection, fingerprint interpolation, and terminal device calibration. Moreover, an online localization algorithm is put forward to estimate the position of the online test fingerprint. The simulation results show that the proposed scheme can achieve higher localization accuracy than the peer schemes, and possesses good effectiveness and robustness at the same time.

Key words: localization fingerprint, crowdsourced samples, radio map, fingerprint interpolation

0 Introduction

With the rapid spread of smart mobile terminals and the large-scale deployment of the Internet, location based service (LBS) has been widely used in various fields of daily life. In some special cases like emergency rescue, public safety and smart city construction, it is necessary to obtain the exact location of people or objects in the indoor environment. Since wireless local area networks (WLANs) have been deployed on a large scale in public places such as schools, hospitals, shopping malls, etc., it is possible to estimate the location of users by relying solely on software development without using any additional hardware facilities, which makes the indoor localization using WLAN systems the current mainstream and the most promising method in the future.

At present, the most widely used Wi-Fi-based indoor localization algorithm is a location fingerprint localization algorithm based on crowdsourced samples, where the localization fingerprint[1] sets up a mapping from the location in the physical environment to a single or multi-dimensional fingerprint of some kind so as to ensure one location for each unique fingerprint. The fingerprint in the Wi-Fi location fingerprint location algorithm refers to the received signal strength (RSS) of the access point (AP)[2]. Crowdsourcing technology[3] can reduce or even eliminate the huge workload of site surveying, which will hand over the construction of radio map to a large number of users, and integrate a small amount of RSS data collected by each user to obtain the radio map data for a large area. Such technology allows the users participating in the update of radio map fingerprint data while enjoying the location service.

Although the use of crowdsourced samples for indoor radio map construction can reduce the cost of updating the fingerprint database effectively in the offline phase, some new challenges arise. First of all, in the offline phase of the traditional WLAN fingerprint location system, a series of reference points are set in the to-be-positioned area to collect the RSS values of the surrounding detectable APs. The position coordinates of the reference points and the corresponding RSS are collected and stored together in the fingerprint database. The crowdsourced samples, however, may not

① Support by the National High Technology Research and Development Program of China (No. 2012AA120802), National Natural Science Foundation of China (No. 61771186), Postdoctoral Research Project of Heilongjiang Province (No. LBH-Q15121) and Undergraduate University Project of Young Scientist Creative Talent of Heilongjiang Province (No. UNPYSCT-2017125).

② To whom correspondence should be addressed. E-mail: qindanyang@hlju.edu.cn

Received on Oct. 23, 2019

be acquired at the specified reference point, which may cause incorrect location annotation of the sample as well as constructing the wrong radio map. Secondly, only a few of the Wi-Fi APs detected by mobile terminals can contribute to localization. If all the APs are taken to construct the radio map, it will not only take up too much storage space, but cause heavy computational overhead. Moreover, the samples collected at the same location by the same mobile terminal may also contain different detectable source APs to cause inconsistent dimensions so as to affect the normalization of radio map. Thirdly, the lack of mandatory requirement for the crowdsourcing process may result in uneven distribution of the samples collected throughout the indoor environment. Finally, the different mobile terminals used to collect fingerprint data may cause serious device diversity problems, no matter in online phase or offline phase.

Aiming at the problems discussed above, a new WLAN indoor radio map construction scheme based on crowdsourced samples (WRMCS) is proposed in this paper. The density-based algorithm in unsupervised machine learning framework is established, and a source selection algorithm is set up based on AP acceptance rate. Moreover, the fingerprint interpolation algorithm is introduced based on surface fitting techniques and the inter-device calibration algorithms are optimized based on receiver pattern analysis, which can help achieve low-cost and high-precision radio map construction.

The subsequent sections are arranged as follows: Section 1 will analyze the problems encountered in the construction of indoor radio map and typical solutions comprehensively. Section 2 will establish the source selection algorithm based on AP acceptance rate and the fingerprint interpolation algorithm based on surface fitting technology to construct the offline radio map. An improved nearest neighbor (NN) online localization algorithm will be proposed in Section 3 associated with the constructed offline radio map to estimate the position of the online test fingerprint. Simulating results will be analyzed in Section 4 and the conclusion will be provided in the end.

1 Related work

The problems of inaccurate sample location annotation, collected samples containing invalid access points (APs), uneven sample distribution, and diversity of terminal devices have been studied since such factors always have serious effect on the indoor positioning performance and practical application. Ref. [4] proposed a bottom-up hierarchical clustering algorithm to distinguish correctly labeled samples from all samples so as to avoid the wrong sample annotation. Such hierarchical clustering algorithm, however, requires a random selection of initial samples, which may introduce samples with erroneous markings. The most famous research on the selection of key APs is the information-based InfoGain algorithm[5], which used the information gain to measure the average RSS to decide the discriminability of different APs. Ref. [6] studied the correlation between different APs to improve the InfoGain algorithm. Considering the inherent defects in indoor environment, an enhanced machine learning indoor localization scheme was proposed in Ref. [7] combined with AP selection and signal strength reconstruction effectively, which can help enhance the robustness in noisy environments. A nonlinear auto-encoder was proposed in Ref. [8] to reduce the dimensionality of the radio map. Machine learning-based methods generally exhibit better performance than information-based methods, but the heavy computing load brought in by machine learning methods during offline and online processing cannot be ignored.

The terminal difference is another factor to affect the positioning results. Ref. [9] achieved normalization received signal strength indicator (RSSI) distribution of various types of terminal devices using the kernel density estimation to solve the problem of device diversity. However, the RSSI probability density distribution estimation algorithm adopted the absolute received intensity value of the RSSI, the instability of which cannot be avoided due to the occlusion of obstacles and the changes of the indoor environment. Hossain et al.[10] concluded that the signal strength difference (SSD) had stronger stability than the RSS value by studying the stability of the SSD between different APs. However, SSD only considers the proportional term of the linear transformation equations. It is necessary to combine the research results of the offset term in the RSS difference fingerprint method to obtain more complete results.

This paper will propose a new radio map constructing method with crowdsourced samples to solve the above 4 problems effectively by outlier detection, key AP selection, fingerprint interpolation and terminal device calibration, so as to realize the construction of radio map with low cost and high precision.

2 Construction of offline radio maps

2.1 System model

The target environment will be divided into differ-

ent sub-regions according to the functional layout of the indoor environment and wall partitions, such as classrooms, corridors, etc. Each sub-region is then further divided into non-overlapping grid cells of the same size. Users participating in crowdsourcing will collect samples at each grid center, and each sample has a location annotation in the grid. Finally, these samples are represented in the form of data cubes, that is, each grid has a data cube to form a grid fingerprint, thereby constructing an offline radio map M for each sub-region. Table 1 in the appendix gives the main symbols of this article.

Let S and $\boldsymbol{F}$ represent the sample set of a grid and its fingerprint, respectively. The sample set S is represented by a data cube, where each vertical slice of the data cube represents a sample acquired by a different device D, and each row vector in the slice represents a sample consisting of collected RSS values from different APs. $\boldsymbol{F}$ is a grid fingerprint vector, and each element in the vector is an RSS value from a specific AP. All grids are divided into 2 categories: a sufficient grid and a deficient grid. The former one is defined as the grid with at least one device containing enough samples; and the latter means that there are not enough samples being collected in such grid even for one device type.

The proposed offline system consists of 4 modules: (1) outlier detection (OD), (2) key AP selection (KAS), (3) fingerprint interpolation (FI), and (4) terminal device calibration (TDC), with the help of which the proposed scheme WRMCS can solve all the 4 core problems. Each grid fingerprint is constructed by the corresponding process on the original data cube. The system architecture of the proposed scheme WRMCS is shown in Fig. 1 with the specific processing as follows. Firstly, the outliers are detected and deleted from the crowdsourcing sample S. Only one subset of all APs being detected is selected to constitute the device-specific fingerprint f. For the deficient grid, a fingerprint will be interpolated to make the spatial distribution of the sample uniform. After that, the fingerprints from different devices will be calibrated and fused into a single, device-independent grid fingerprint. Finally, an improved online location algorithm is established to estimate the location of online test fingerprints. Fig. 2 is the data processing flow of the proposed scheme. The blank grid in data grid and sub-region sections represents the defect grid. The data cube can be processed as follows. After the outlier detection

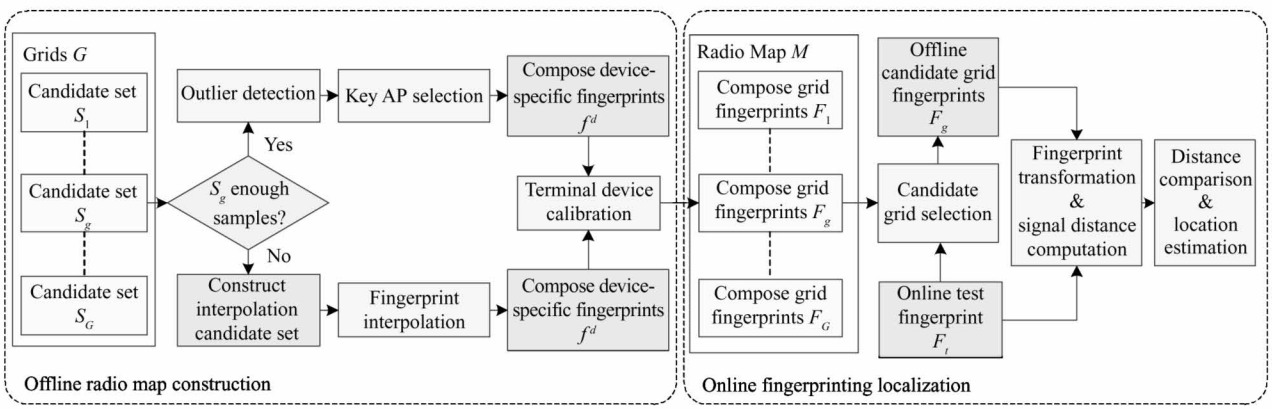


Fig. 1 The proposed system diagram

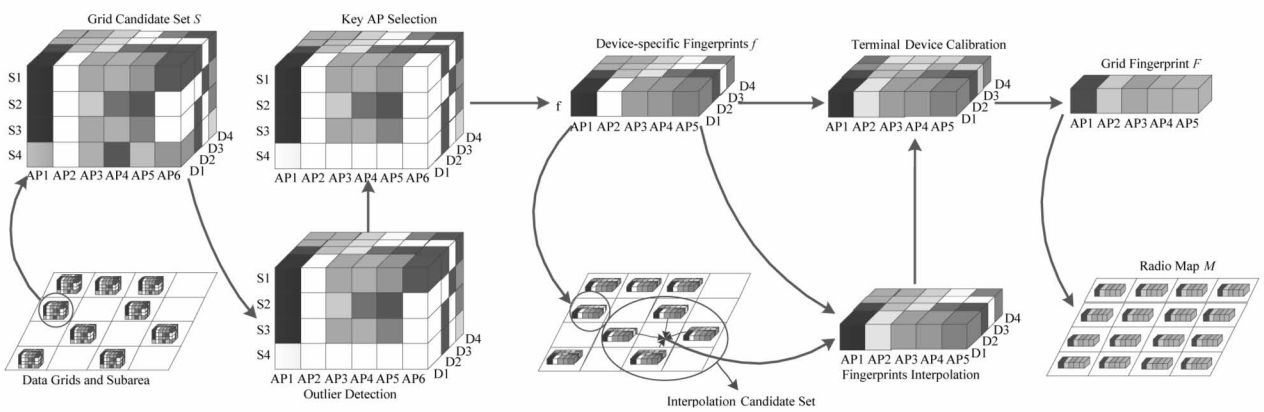


Fig. 2 Data processing flow of the proposed system

is completed, the sample $\boldsymbol{s}_4$ collected by device D_1 will be deleted, and AP_6 detected by device D_1 will also be deleted after every device selects the key AP. A device-specific fingerprint f should be built up for each device. For the deficient grid, the device-specific fingerprint will first be constructed by the fingerprint interpolation module with the terminal device calibration being performed. Then the grid fingerprint F is composed for each grid and a radio map M is constructed for a sub-area. The 4 modules of the offline system are introduced below.

2.2 Outlier detection

Crowdsourcing samples need some locational annotations to build the training fingerprint. Some samples, however, may not be acquired at the specified reference point to make the annotations inaccurate, which may cause wrong constructions of the training fingerprint and the radio map. Specifically, the AP set being detected usually consists of the APs appearing once or many times in their samples when constructing a grid fingerprint. In addition, in the subsequent construction steps, the average calculation for RSS is usually performed for each AP being detected in the sample. Therefore, the sample with errors will cause at least 2 problems: the changes of the AP set being detected in the grid and the deviation of the average RSS from the true value.

To solve such problems, the clustering method in unsupervised machine learning algorithms is adopted for outlier detection. Specifically, the density-based algorithm[11] is selected with samples being clustered according to the similarity-based local density. Here the sample being measured within its annotated grid is called a normal sample; otherwise, the sample is an outlier. Theoretically, the similarity between normal samples would be higher than that between normal samples and outliers, and also higher than that between outliers.

Such process will be performed on each slice in every data cube by the proposed scheme since the outlier detection is specific to the device. Let $S_D = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_N\}$ denote the set of samples collected by a particular device D, each of which is a vector consisting of RSS values, and N is the total number of samples. Let A_i denote the set of APs detected in sample $\boldsymbol{s}_i$. $M = |\cup A_i|$ denotes the total number of all APs being detected in S_D. The fact that not all devices can detect all APs makes the values of M and N be different for different devices and bandwidths. An $N \times M$ RSS matrix $\boldsymbol{R}^{(0)}$ is constructed for each device, where the element r_{nm} represents the RSS value received from the mth AP by the nth sample. The signal distance between the 2 samples is computed by Eq. (1).

$$d_{nn'} = \sqrt{\sum_{m \in A_{nn'}^{\text{int}}} (r_{nm} - r_{n'm})^2} \tag{1}$$

where $A_{nn'}^{\text{int}} = A_n \cap A_{n'}$ is the set of APs detected in both samples. The smaller the signal distance $d_{nn'}$ is, the more similar the 2 samples will be.

Algorithm 1 Cluster-based iterative outlier detection

Require: The set of samples S_D
Ensure: The normal set S_d^n
Compute $d_{nn'}$ between all sample pairs in S_D
Sort the $N \times N$ distances $d_{nn'}$ into D
Compute d_c as the β percentile distance of D
Compute B_n and ρ_n for each $\boldsymbol{s}_n \in S_D$
Set $S_d^n = S_D$, $S_d^o = \varnothing$
Cluster each $\boldsymbol{s}_n \in S_D$ into $S_d^n(S_d^o)$ based on $\rho_n \geqslant \rho_T (\rho_n < \rho_T)$
while $S_d^o = \varnothing$ **do**
    Pop a sample $\boldsymbol{s}^*$ from S_d^o
    for each sample $\boldsymbol{s}_i \in S_d^n$ **do**
        if $\boldsymbol{s}^* \in B_i$ **then** $B_i = B_i / \{\boldsymbol{s}^*\}$
        if $|B_i| \leqslant \rho_T$ **then** $S_d^n = S_d^n / \{\boldsymbol{s}_i\}$, $S_d^o = S_d^o / \{\boldsymbol{s}_i\}$
    end for
end while
Return S_d^n

Two thresholds are further defined[12] as the density threshold ρ_T and the cutoff distance d_c. If there is $d_{nn'} < d_c$, the sample $\boldsymbol{s}_n$ is called the neighbor of $\boldsymbol{s}_{n'}$. Let $B_n = \{n' \mid d_{nn'} < d_c\}$ denote the neighbor set of the nth sample $\boldsymbol{s}_n$. The local density ρ_n of the sample is defined as the number of its neighbors as $\rho_n = |B_n|$. According to the clustering algorithm of Ref. [13], the sample $\boldsymbol{s}_n$ will be considered as an abnormal value if the number of neighbors of the sample $\boldsymbol{s}_n$ is less than the density threshold, namely $\rho_n < \rho_T$; otherwise, it is considered to a normal sample.

The proposed WRMCS scheme will iterate over the outlier detection, with the pseudo code given as in Algorithm 1. Firstly, all samples are divided into the normal value set S_d^n and the abnormal value set S_d^o according to the number of sample neighbors. When there is an outlier $\boldsymbol{s}^*$ in S_d^o, $\boldsymbol{s}^*$ will be popped up and the neighbor of each normal sample $\boldsymbol{s}_i \in S_d^n$ will be updated. After that, the local density of sample $\boldsymbol{s}_i$ may be lower than the density threshold, to make $\boldsymbol{s}_i$ being detected as the outlier and listed in S_d^o. The iteration will terminate until $S_d^o = \varnothing$.

During the outlier detection, the sample $\boldsymbol{s}_i$ being

detected as the outlier may make its neighbor sample $\boldsymbol{s}_j$ with a small distance be an outlier as well. In fact, if the local density satisfies $\rho_j > \rho_T$ at the first iteration, the sample $\boldsymbol{s}_j$ may be a normal value at this time, but the abnormal value $\boldsymbol{s}_i$ is already included in the calculation of ρ_j (the number of neighbors of $\boldsymbol{s}_j$). After deleting the outliers, such problem can be solved by the proposed scheme through recalculating the local density of the samples for each iteration.

After the outlier detection, S_d^n can still be used to represent the normal sample set, and a new RSS matrix $\boldsymbol{R}^{(1)}$ with $N^{(1)} \times M^{(1)}$ can be constructed for each device in the grid.

2.3 Key AP selection

In fact, there are many APs that can be detected in S_d^n, but only a few of them will contribute to indoor localization. On the one hand, the samples even collected at the same location may contain different detectable sources APs due to the changes in radio propagation or in the collector's direction, which may cause the inconsistent sample dimensions. On the other hand, an AP may exist in the undeleted outlier $\boldsymbol{s}_i$ in S_d^n, which may introduce an unexpected AP so as to compose a wrong grid fingerprint.

To solve this problem, a source selection algorithm based on the AP acceptance rate is proposed to select a proper subset of APs for effective localization to help construct device-specific fingerprints for each device. Let N_m denote the number of non-empty elements in the mth column in $\boldsymbol{R}^{(1)}$. The acceptance rate of the mth AP is defined as $P_m = N_m / N^{(1)}$ and the acceptance rate threshold is P_T. If there is $P_m < P_T$, the mth column of the matrix $\boldsymbol{R}^{(1)}$ will be deleted, that is, the RSS value of the mth AP detected in each sample will be deleted.

After selecting the key AP, a new $N^{(2)} \times M^{(2)}$ RSS matrix $\boldsymbol{R}^{(2)}$ can be constructed for each device in the grid, with the element r_{nm} in the matrix representing the RSS value obtained from the mth AP in the nth sample. Next, a common RSS averaging method is used to construct the device-specific fingerprint f for each grid. The average RSS of the mth AP can be obtained by

$$r_m = \frac{1}{|r_{\cdot m}|} \sum_{n=1}^{N^{(2)}} r_{nm} \tag{2}$$

where $r_{\cdot m}$ is the mth column vector of $\boldsymbol{R}^{(2)}$, so as to construct the device specific fingerprint as $f = (r_1, \cdots, r_{M^{(2)}})$.

2.4 Fingerprint interpolation

The fact that crowdsourcing is adopted to collect the samples randomly may result in uneven distribution in the entire indoor environment. Some grids may have little or no crowdsourcing samples at all. For example, there may be fewer crowdsourced samples at the edge of the classroom than those at the central area of the classroom. Aiming at such problem, the proposed WRMCS scheme established a fingerprint interpolation module for the deficient grid lack of samples. A fingerprint will be selected by such module from a neighboring sufficient grid in the same subarea to be inserted as a device-specific fingerprint.

A fingerprint interpolation algorithm is proposed based on surface fitting technology. An interpolated fingerprint candidate set will be constructed before the fingerprint interpolation process with the pseudo code of the proposed algorithm given in Algorithm 2, where G is the set of all the grids in one subarea, and G_S is the set of interpolated fingerprint candidates.

Algorithm 2 Perform fingerprint interpolation on the deficient grid

Set **finished** = FALSE, $G_S = \varnothing$
while not finished and $G/G_S \neq \varnothing$ **do**
  Update G_S = includeOneSurroundingGrid (G_S, g)
  Update the set of devices D
  for each $d \in D$ **do**
    Compute Q^d = countSupportFingerprint (D, d)
    if $Q^d > \gamma$ **then finished** = TRUE, **break**
  end for
end while
for each $d \in D$ with $Q^d > \gamma$ **do**
  Compute $A_{\text{int}}^d = \cap_{g \in G_S} A_g^d$
  for each $m \in A_{\text{int}}^d$ **do**
    Construct $\phi_m(\cdot)$ according to Eqs(1) and (2)
    Compute $\hat{r}_m^d = \phi_m(\cdot)$
  end for
  Compose an interpolated fingerprint $f_g^d = (\hat{r}_m^d)_{m \in A_{\text{int}}^d}$
end for

The process of constructing an interpolated fingerprint candidate set is as follows. A neighboring grid of the deficient grid is added into G_S. The function includeOneSurroundingGrid (G_S, g) is set up to include the surrounding grid g in G/G_S into G_S, while grid g is the neighboring grid of a deficient grid or any grid in G_S. Let D_g and $D = \cup_{g \in G_S} D_g$ denote the device set in the grid and that in G_S, respectively; f_g^d denote the device-specific fingerprint of device d in grid g. The function countSupportFingerprint (D, d) is established to record the number of fingerprints of device d in dif-

ferent grids of the interpolated fingerprint candidate set Q^d. An interpolation support threshold γ is also defined with a typical value as a small integer. If the G/G_S is not an empty set (that is, there are other grids in the subarea except the grid set G_S) or none of device d in all sets D contains more than γ fingerprints, G_S will be built continually by including its other surrounding grids. Otherwise, the building process of G_S will terminate.

After the fingerprint interpolation candidate set is constructed, the fingerprint interpolation is performed as follows for each device d in D with the condition of $Q^d > \gamma$. Let A_g^d denote the set of APs detected by device d in grid g; $A_{\text{int}}^d = \cap_{g \in G_S} A_g^d$ is the set of APs to perform fingerprint interpolation for the deficient network. For each AP m in A_{int}^d, a function $\phi_m(x_g, y_g)$ will be constructed to minimize the sum of squared error of the following error based on the least squares principle, as shown in Eq. (3).

$$\min\theta = \sum_{g \in G_S,\ m \in A^d} (\phi_m(x_g, y_g) - r_{gm}^d)^2 \tag{3}$$

where, (x_g, y_g) is the center coordinate of the grid g, and r_{gm}^d is the RSS value of the mth AP detected by device d in grid g. The surface fitting function ϕ_m is formed using a binary polynomial function:

$$\phi_m(x, y) = \sum_{c=1}^{p} \sum_{d=1}^{q} \omega_{cd} x^{c-1} y^{d-1} \tag{4}$$

where, ω_{cd} is a polynomial coefficient. The specific position coordinates input will achieve a deterministic RSS results by ϕ_m.

The interpolation of A_{int}^d into the deficient grid and the construction of ϕ_m for each AP in A_{int}^d will make device d insert the RSS value $\hat{r}_m^d$ of the mth AP into the deficient grid as the output of ϕ_m. Such output $\hat{r}_m^d$ being added into each AP in A_{int}^d can make up the interpolated device-specific fingerprint for the deficient grid.

2.5 Terminal device calibration

Different types of mobile phones participating in crowdsourced fingerprint collection have different antennas and receiver gains, which may cause at least 2 problems: (1) Sample measurements from the same source may be different even at the same place; (2) Much storage space and computing time may be taken to create multiple grid fingerprints for one specific device.

Therefore, a new inter-device calibration algorithm is proposed based on receiver pattern analysis to calibrate specific fingerprints for different devices and combine the fingerprints collected by multiple devices into a single device-independent grid fingerprint. Fig. 3 gives the comparisons of the RSS values at the same location by using 4 different devices, which shows that there are similar differences between different APs.

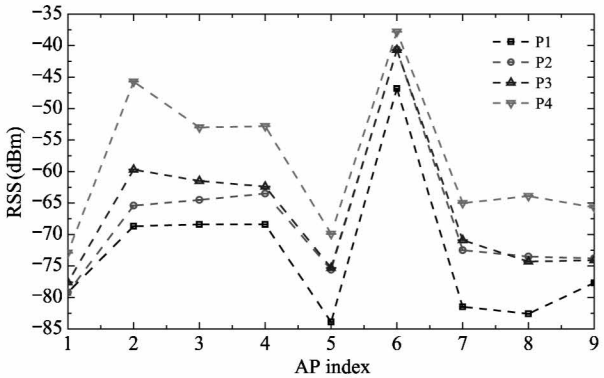


Fig. 3 Data processing flow of the proposed system

Let $F = \{f^1, \cdots, f^d, \cdots, f^{N_d}\}$ denote the set of device-specific fingerprints, and A^d is set to represent the set of APs collected by device d; $A_{\text{uni}} = \cup_{d=1}^{N_d} A^d$ and $A_{\text{int}} = \cap_{d=1}^{N_d} A^d$ represent the union and intersection of APs detected by all device types N_d, respectively; $\overline{A^d} = A_{\text{uni}} - A^d$ represents the set of AP being not detected by device d. Let $M^{(3)} = |A_{\text{uni}}|$ denote the total number of APs from all devices. Each device's specific fingerprint $f^d = (r_1^d, \cdots, r_m^d, \cdots, r_{M^{(3)}}^d)$ should have $M^{(3)}$ RSS values. However, in actual situations, the fact that not every device can detect the RSS value of all APs may cause some r_m^d being lost.

The missing values will be calibrated as follows. Let $\bar{r}_m = 1/N_d \sum_{d=1}^{N_d} r_m^d$, $m \in A_{\text{int}}$, denote the mean RSS value of the mth AP detected by all devices. A calibration factor Δ^d can be defined for each device d as Eq. (5).

$$\Delta^d = \frac{1}{|A_{\text{int}}|} \sum_{m \in A_{\text{int}}} (r_m^d - \bar{r}_m),\ d = 1, \cdots, N_d \tag{5}$$

The inter-device fingerprint calibration is performed for each AP m in set $A_{\text{uni}} - A_{\text{int}}$. Let D_m denote the device set with corresponding r_m^d missed, which lies in the complement $\overline{D}_m$. The value of $\tilde{r}_m^d$ (calibration value) of all $d \in D_m$ can be obtained according to the linear Eq. (6).

$$\tilde{r}_m^d - \frac{1}{N_d}\left(\sum_{d \in \overline{D}_m} r_m^d + \sum_{d \in D_m} \tilde{r}_m^d\right) = \Delta^d,\ d \in D_m \tag{6}$$

Therefore, it is always possible to calculate a unique solution $\tilde{r}_m^d$ to fill the missing value r_m^d for each device $d \in D_m$.

After the missing RSS values are populated, an $N_d \times M^{(3)}$ RSS matrix $\boldsymbol{R}^{(3)}$ can be constructed for each grid with the element r_m^d representing the original/cali-

brated RSS value from the mth AP of device d. Finally, a column-by-column averaging calculation is performed to obtain an independent device grid fingerprint $F_g = (r_g^1, \cdots, r_g^m, \cdots, r_g^{M(3)})$.

3 Online localization algorithm

Based on the constructed offline radio map, an improved nearest neighbor (NN) online localization algorithm is proposed by selecting the candidate grid through the number of mutual sources and determining the target grid based on the comparison of distances between transformed fingerprints.

Let $F_g = (r_g^1, \cdots, r_g^m)$ denote the offline fingerprint of grid g; $F_t = (r_t^1, \cdots, r_t^m)$ is the online test fingerprint. Let A_g and A_t represent the set of detectable APs of fingerprints F_g and F_t, respectively, and the AP intersection of the 2 fingerprints will be $A_{\text{int}} = A_g \cap A_t$. The candidate grid should be selected for distance calculation firstly. All available grids will be sorted in descending order according to the number of mutually detected APs $K_{gt} = |A_{\text{int}}|$, with the top ρth grids being selected as the candidate grids.

The online fingerprint conversion is then performed. F_g is adopted again to represent the candidate grid fingerprint. Let $F_g^{\text{int}} = (r_g^m)_{m \in A_{\text{int}}}$ and $F_t^{\text{int}} = (r_t^m)_{m \in A_{\text{int}}}$ denote the grids and test fingerprints of the RSS values consisting only of those mutually detectable APs in intersection A_{int}, respectively. The averages of F_g^{int} and F_t^{int} can be obtained by

$$\bar{r}_g = \frac{1}{|A_{\text{int}}|} \sum_{m \in A_{\text{int}}} r_g^m \tag{7}$$

$$\bar{r}_t = \frac{1}{|A_{\text{int}}|} \sum_{m \in A_{\text{int}}} r_t^m \tag{8}$$

The transformed fingerprint used for distance calculation is defined as

$$\tilde{F}_g = F_g^{\text{int}} - \bar{r}_g = (r_g^m - \bar{r}_g)_{m \in A_{\text{int}}} \tag{9}$$

$$\tilde{F}_t = F_t^{\text{int}} - \bar{r}_t = (r_t^m - \bar{r}_t)_{m \in A_{\text{int}}} \tag{10}$$

The converted fingerprints $\tilde{F}_g$ and $\tilde{F}_t$ are independent to the device receiver gain, which can solve the device diversity problem. The signal distance can be calculated using the average Euclidean distance between 2 fingerprints as

$$D(\tilde{F}_g, \tilde{F}_t) = \frac{1}{|A_{\text{int}}|} \sqrt{\sum_{m \in A_{\text{int}}} ((r_g^m - \bar{r}_g) - (r_t^m - \bar{r}_t))^2} \tag{11}$$

The target grid will be determined by the grid with the minimum signal distance, and the corresponding grid center will be selected as the estimated position of the test fingerprint.

The candidate grid selection has the following advantages: (1) only a part of the grids are selected to perform grid fingerprint conversion so as to reduce the online calculation time greatly; (2) candidate grids with more mutually detectable APs are selected, to make more APs be shared between online and offline fingerprints, which means the radio environment in grid g is more like the radio environment for testing fingerprints, so as to increase the locating accuracy.

4 Experimental and simulation results analysis

4.1 Experiment setup

On-site measurements are performed in the Electronic Engineering College Lab Building. A total of 1 460 grids are established with the size of each grid being 0.5 m × 0.5 m. The RSS measurements are taken from the existing WLAN APs using 5 different smartphones labeled as P1, P2, P3, P4 and P5, respectively. Ten samples are acquired in each grid by P1, P2, P3 and P4 to produce a total of 58 400 training samples. Each experimental device can collect 650 online test fingerprints evenly distributed in the environment, so a total of 3 250 test fingerprints can be obtained.

The grid lying less than 1m away from the center of the given grid will be considered as a normal grid; otherwise, the grid will be considered as an outlier. Four device-specific online test data sets are set up as Test1, Test2, Test3, Test4, Test5, and an independent device hybrid online test data set is established as well.

4.2 Experimental results

4.2.1 Performance analysis of radio maps based on survey samples

The proposed scheme is applied to a field survey to construct an offline radio map, being referred to as RM-SS, where each training sample is obtained within a particular sufficient grid. In this experiment, only the device calibration module is adopted to verify the ability of the proposed scheme to handle device diversity issues.

Radio maps are constructed for each device based on the samples from smartphone surveys, being labeled as RM-P1, RM-P2, RM-P3 and RM-P4, respectively. In addition, a device fusion algorithm is developed to obtain the RSS averages of survey samples from all 4 devices, and a radio map called RM-DF is constructed correspondingly. Fig. 4 is the comparison of the aver-

age localization errors (ALE) of the radio maps constructed using these 4 device-specific test data sets.

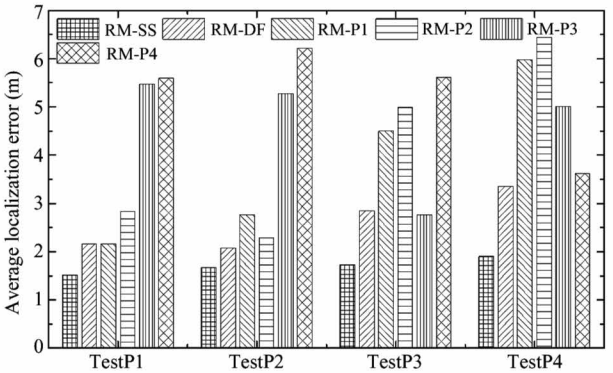


Fig. 4 Comparison of average positioning errors of radio maps

It can be clearly seen from Fig. 4 that RM-SS can achieve the best localization performance. In particular, it can decrease ALE by 30.09%, 19.32%, 39.65% and 43.28%, respectively, compared to the radio map RM-DF. The results verify the effectiveness of the device calibration algorithm of the proposed scheme.

4.2.2 Performance analysis of the proposed RM-CS

The comparison is performed between the outlier detection algorithm with iterative process and that with once detection . The classification results are shown in Fig. 5 when the ratio of the deficient grid is set to 46.9%.

It can be seen intuitively from Fig. 5 that the localization accuracy and the ability to detect outliers are worsened with the outliers increasing in both conditions. This is because the local density of outliers is proportional to the numbers within a given range, which may cause more outliers being determined as normal samples according to the density-based algorithm. In addition, the simulating results also indicate that the iterative process can achieve a lower recall rate to classify the normal values and outliers more accurately.

The localization performance of RM-CS constructed by crowdsourced samples from different situations is simulated. The experiments will focus on 2 factors as the ratio of outliers in a sufficient grid and the ratio of deficient grids in all grids. For this purpose, different numbers of outliers are added to different sufficient grids to create multiple deficient grids in different degrees based on actual environmental conditions, which are mainly divided into the following cases as E1 (46.9%), E2a (34.4%), E2b (34.4%), E3 (24%) and E4 (22.9%). The numbers in parentheses represent the proportion of the deficient grids.

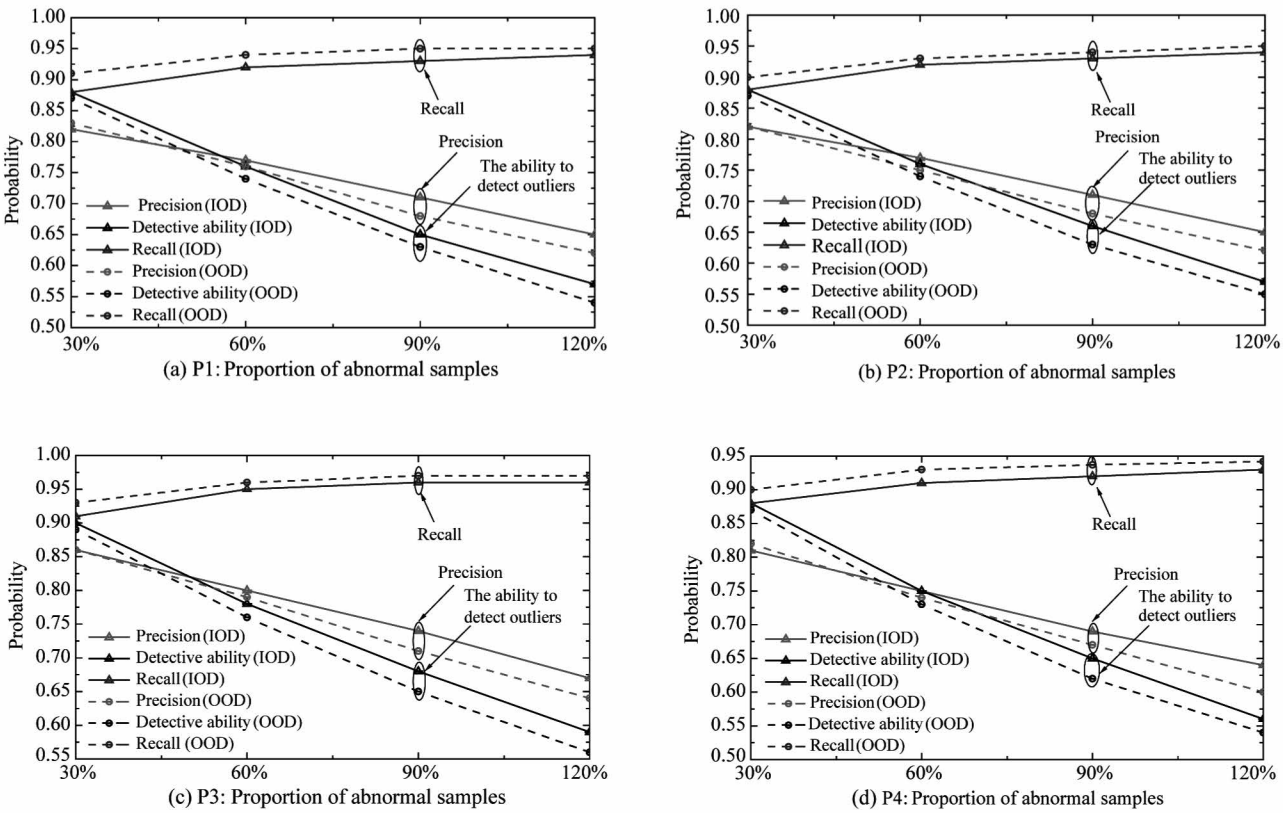
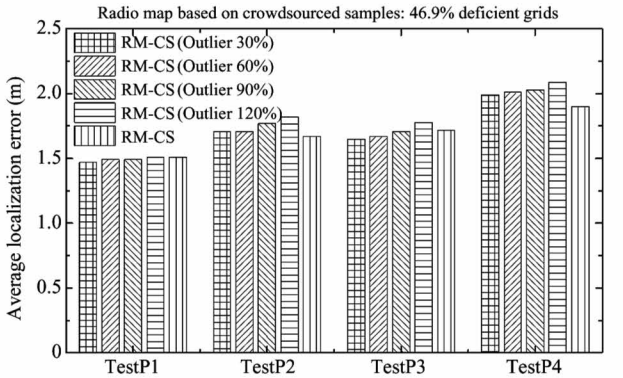
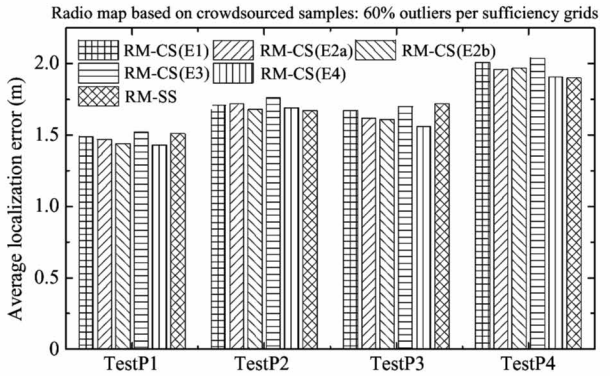


Fig. 5 Performance comparison of outlier detection using different devices with the ratio of the deficient grid being set as 46.9%

Fig. 6(a) shows the localization performance of the RM-CS containing different numbers of outliers with the ratio of the deficient grid being set as 46.9%. Fig. 6(b) plots the localization performance of the RM-CS with different proportions of deficient grid when the outlier ratio in a sufficient grid is 60%. First of all, it can be observed that as the ratio of outliers in a sufficient grid and the ratio of deficient grid in all grids increase, the localization performance of the RM-CS decreases. This is because the localization is also affected by those outliers that have not been deleted, so that detecting and deleting all outliers from the sample set becomes more difficult as the outlier ratio increases. When observing Fig. 6(b), it is worth noting that the localization performance of E4 is slightly better than other cases because it contains the least deficient grid that needs to perform the fingerprint interpolation module. On the other hand, in these practical scene, it can be seen from the experimental results of the 4 device-specific test data sets that the degradation of the localization performance is not obvious, which can verify the robustness of the proposed scheme.



(a) The deficient grid ratio is 46.9%, including different numbers of outliers



(b) The deficient grid ratio is 60%, including different numbers of outliers

Fig. 6 RM-CS localization performance

Finally, the RM-CS localization performance is investigated with some of the proposed modules disabled to produce the common cases for the conditions with the outlier ratio as 60% and the deficient grid ratio as 46.9%. Fig. 7 reveals the average localization errors of RM-CS by 4 device-specific online testing data sets. It indicates that when some modules of the proposed scheme are disabled, the ALE of the RM-CS may be increased by up to 37.91%.

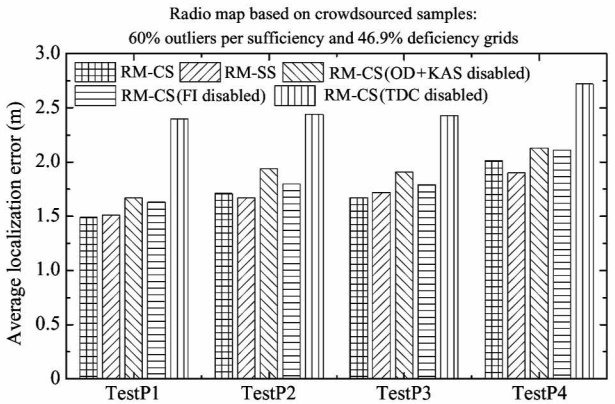


Fig. 7 Average localization error of RM-CS in 4 device-specific online test data sets

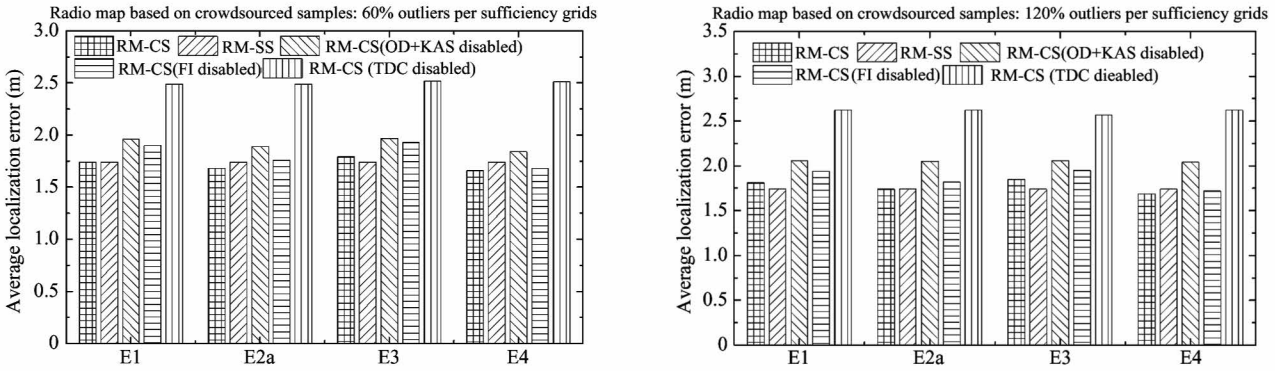
In addition, the positioning performance of RM-CS is simulated in the equipment hybrid test data set with some modules of the proposed scheme being deactivated. The average localization error of the radio map containing 60% and 120% outliers in each of the sufficient grids, respectively, are shown in Fig. 8(a) and Fig. 8(b). The proposed scheme with crowdsourced samples can achieve the similar localization performance compared to the scheme using survey samples from 2 test data sets (that is, using RM-CS and RM-SS for localization). The results not only verify the validity of the proposed modules, but also verify the effectiveness of the overall scheme of radio map construction.

4.2.3 Performance analysis using new equipment data sets

In fact, online test fingerprints may come from new devices which are not applied to offline radio map construction. A test fingerprint from another smartphone P5 is adopted to check the applicability of the radio map. In the experiments, the localization performance of RM-CS constructed by crowdsourcing samples collected by P5 is compared in the case of E1, E2a, E3 and E4 with some of the proposed modules disabled.

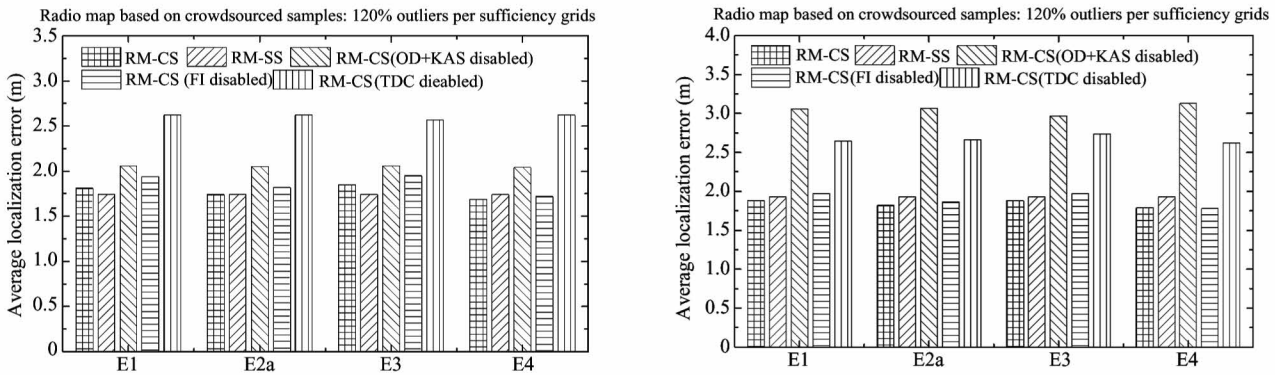
Fig. 9(a) and Fig. 9(b) plot the average localization error of a radio map containing 60% and 120% outliers in each of the sufficient grids. It shows that

when some modules of the proposed scheme are disabled, the ALE of the radio map RM-CS constructed with the new device will increase, and the maximum possible increase is 42.81%. The results verify the applicability of the proposed scheme.



(a) 60% abnormal sample in each sufficient grid (b) 120% abnormal sample in each sufficient grid

Fig. 8 Average localization error for radio maps containing 60% and 120% anomalous samples in each sufficient grid



(a) 60% abnormal sample in each sufficient grid (b) 120% abnormal sample in each sufficient grid

Fig. 9 Locating performance of RM-CS constructed using crowdsourced samples collected by P5 when deactivating some of the proposed modules

5 Conclusion

Aiming at the problems of inaccurate location annotation of samples, invalid access points (AP), uneven distribution of samples and diversity of terminal devices, a new WLAN radio map constructing scheme (WRMCS) is proposed based on the crowdsourced samples. This scheme can not only detect and delete those samples with wrong annotations, but can select valid APs to form device-specific fingerprints, which will be merged into a single device-independent grid fingerprint. The solution can also perform fingerprint interpolation on the defect grid, which improves the positioning performance. Simulating results show that the proposed scheme can achieve lower average localization error than other schemes and can be applied for a variety of terminal equipment, so as to verify the effectiveness and applicability of the scheme. Future research will focus on how to combine the data collected by sensors such as magnetometers and barometers of mobile terminals with traditional WLAN localization systems to improve the localization accuracy of the entire system.

References

[1] Zhou B, Li Q, Mao Q, et al. A robust crowdsourcing-based indoor localization system[J]. *Sensors*, 2017, 17(4):864

[2] Ashrafi S, Feng C, Roy S. Compute-and-forward for rndom-access: the case of multiple access points[J]. *IEEE Transactions on Communications*, 2018, 99:1-1

[3] Stol K J, Caglayan B, Fitzgerald B. Competition-based crowdsourcing software development: a multi-method study from a customer perspective[J]. *IEEE Transactions on Software Engineering*, 2017, 45(3):237-260

[4] Zhang Q, Shi C, Niu Z, et al. HCBC: a hierarchical case-based classifier integrated with conceptual clustering [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2019, 31(1): 152-165

[5] Chen Y, Qiang Y, Jie Y, et al. Power-efficient access-point selection for indoor location estimation[J]. *IEEE*

Transactions on Knowledge and Data Engineering, 2006, 18(7):877-888

[6] Jiang X, Liu J, Chen Y, et al. Feature adaptive online sequential extreme learning machine for lifelong indoor localization [J]. *Neural Computing and Applications*, 2016, 27(1):215-225

[7] Chou H J, Chang R Y. Joint mode selection and interference management in device-to-device communications underlaid MIMO cellular networks [J]. *IEEE Transactions on Wireless Communications*, 2017, 16(2):1120-1134

[8] Liao Y, Wang Y, Liu Y. Graph regularized auto-encoders for image representation [J]. *IEEE Transactions on Image Processing*, 2016, 26(6):2839-2852

[9] Li Z, Luo L, Sheng G, et al. UHF partial discharge localization method in substation based on dimension-reduced RSSI fingerprint [J]. *IET Generation, Transmission and Distribution*, 2018, 12(2):398-405

[10] Hossain A K M M, Jin Y, Soh W S, et al. SSD: a robust RF location fingerprint addressing mobile devices' heterogeneity [J]. *IEEE Transactions on Mobile Computing*, 2013, 12(1):65-77

[11] Rahman M A, Ang L M, Seng K P. Unique neighborhood set parameter independent density-based clustering with outlier detection [J]. *IEEE Access*, 2018, 6: 44707-44717

[12] Mehmood R, Shaikh M U, Bie R, et al. IoT-enabled Web warehouse architecture: a secure approach [J]. *Personal and Ubiquitous Computing*, 2015, 19(7): 1157-1167

[13] Hung W L, Yang J H, Song I W, et al. A modified self-updating clustering algorithm for application to dengue gene expression data [J]. *Communications in Statistics Simulation and Computation*, 2019, 12:1-18

Guo Ruolin, born in 1996. She received her B. Sc. degree in electronic information engineering from Heilongjiang University in 2018. Currently, she studys at the Department of Communication Engineering of Heilongjiang University, P. R. China. Her researches include indoor vision positioning and crowd source sensing.

Appendix:

Table 1 Symbol notations

Symbol	Definition
M	The indoor offline radio map
S	The sample set of a grid
$S_D = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_N\}$	The set of samples collected by a particular device D, where $\boldsymbol{s}_i$ is a vector of RSS values and N is the total number of samples
S_d^n, S_d^o	A collection of normal and abnormal samples detected by device d
N, $N^{(1)}$, $N^{(2)}$	The total number of samples available to a device, the number of samples after an outlier detection for a device and the number of samples after selecting a critical AP
N_d	The number of available devices
M, $M^{(1)}$, $M^{(2)}$, $M^{(3)}$	The total number of all hearable APs for one device, the number of APs after an outlier detection for a device, and the number of APs after a critical AP is selected and the total number of all detectable APs after fingerprint interpolation for all devices
F, F_g, F_t	The grid fingerprint of a grid, the grid fingerprint of grid g and the online test fingerprint
$f = (r_1, \cdots, r_{M(2)})$	Select the device-specific fingerprint that is constructed after the key AP
$F = \{f^1, \cdots, f^d, \cdots, f^{N_d}\}$	The set of device specific fingerprints
F_g^{int}, F_t^{int}	The grids and test fingerprints representing RSS values consisting only of mutually detectable APs in the intersection A_{int}, respectively
$\bar{F}_g$, $\bar{F}_t$	The transformed fingerprints of F_g and F_t, respectively
A_i, A_g, A_t	The set of APs detected in samples s_i, F_g, and F_t, respectively
A_{uni}, A_{int}	The union set and intersection set of hearable APs, respectively
r_{nm}	The nth sample receives the RSS value from the mth AP for one device
r_m	The average RSS value of the mth AP in all samples collected by one device
r_{gm}^d	The RSS value of the mth AP detected by the device d in the grid g
$\hat{r}_m^d$	The interpolated RSS value of the mth AP collected by device d
r_m^d	The raw/calibrated RSS values from the mth AP of device d

Table 1 continued

$\bar{r}_m$	The average RSS value from all devices that can hear the mth AP
$\tilde{r}_m^d$	The calibration RSS value of the mth AP for the device d
r_g^m, r_t^m	The RSS values of the mth AP in F_g and F_t, respectively
$\bar{r}_g$, $\bar{r}_t$	The average RSS value of F_g^{int} and F_t^{int}, respectively
$d_{nn'}$, d_c	The signal distance and cutoff distance of samples n and n', respectively
ρ_T, ρ_n	The density threshold and the local density of the nth sample, respectively
B_n	The neighbor set of the nth sample s_n
N_m	The number of non-empty elements in the mth column in matrix $R^{(1)}$
P_T, P_m	The acceptance ratio threshold and the acceptance rate of the mth AP, respectively
G, G_S	The set of all grids and an interpolated fingerprint candidate set in a subarea
D, D_g	The set of all devices in G_S and device sets in one grid $g \in G_S$
Q^d	The number of fingerprints of device d in different grids of the interpolated fingerprint candidate set
γ	The interpolation support threshold
ω_{cd}	The polynomial coefficient
D_m, $\bar{D}_m$	The corresponding r_m^d -deleted and existing device set
Δ^d	The calibration factor for device d
$D(\tilde{F}_g, \tilde{F}_t)$	The average Euclidean distance between $\tilde{F}_g$ and $\tilde{F}_t$