Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5530396, 15 pages
https://doi.org/10.1155/2021/5530396

Research Article

Linear Regression Algorithm against Device Diversity for the WLAN Indoor Localization System

Liye Zhang, Xiaoliang Meng, and Chao Fang

School of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong 255000, China

Correspondence should be addressed to Xiaoliang Meng; xiaoliang@sdut.edu.cn

Received 3 February 2021; Revised 24 February 2021; Accepted 19 March 2021; Published 8 April 2021

Academic Editor: Xin Liu

Recent years have witnessed a growing interest in using WLAN fingerprint-based methods for the indoor localization system because of their cost-effectiveness and availability compared to other localization systems. In this system, the received signal strength (RSS) values are measured as the fingerprint from the access points (AP) at each reference point (RP) in the offline phase. However, signal strength variations across diverse devices become a major problem in this system, especially in the crowdsourcing-based localization system. In this paper, the device diversity problem and the adverse effects caused by this problem are analyzed firstly. Then, the intrinsic relationship between different RSS values collected by different devices is mined by the linear regression (LR) algorithm. Based on the analysis, the LR algorithm is proposed to create a unique radio map in the offline phase and precisely estimate the user's location in the online phase. After applying the LR algorithm in the crowdsourcing systems, the device diversity problem is solved effectively. Finally, we verify the LR algorithm using the theoretical study of the probability of error detection. Experimental results in a typical office building show that the proposed method results in a higher reliability and localization accuracy.

1. Introduction

In recent years, people's daily life is becoming more and more convenient owing to the development of "5G", smart cities, and Internet of Things [1, 2], and the network connectivity and spectrum efficiency of Internet of Things are greatly improved by "5G" [3]. As a major part of smart cities, location-based services (LBS) are attracting an enormous amount of attention [4, 5]. Furthermore, the explosive growth of intelligent mobile devices has promoted the research of LBS technology [6]. The typical LBSs mainly involve human navigation in unfamiliar environment, robot path planning and guidance, health care inside modern hospitals, location-based enhanced sensing, entity and storage tracking, and management. We can see from all these services that the reliable, accurate, and real-time localization technologies are required to locate the users' position at the beginning of LBS [7].

In outdoor environments, the Global Navigation Satellite System (GNSS) is the most prominent positioning technology and provides precise positioning. It has made a great success and brought great convenience to people's lives [8]. However, due to line-of-sight (LOS) limitations, satellite signal receiving devices cannot be used indoors and in areas where satellites are blocked by tall buildings. Various solutions have been proposed as alternatives to GPS for indoor environments [9].

Since the RADAR system is proposed in [10], the RSS fingerprint-based WLAN localization method has attracted the attention of many researchers due to its cost-effectiveness and availability [11, 12]. After years of research and evolution, thousands of access points (APs) have already been deployed in indoor environments, such as campuses, hospitals, airports, and shopping malls, which provide great opportunities for the development of indoor location estimation services. The RSS fingerprint-based WLAN localization system can reuse the existing WLAN infrastructures in indoor environments, which significantly reduces the cost.

Typically, The RSS fingerprint-based WLAN indoor localization system contains two phases: the offline phase and the online phase [10]. In the offline phase, a number of reference points (RPs) are set in the indoor area, and

researchers collect RSS values from the existing access points (APs) at all RP throughout. A radio map is then constructed using the collected RSS values and its corresponding geographic coordinates. In the online phase, users' locations can be estimated by comparing the RSS value at their current location and those in the radio map [13, 14].

A conclusion can be drawn from the structure of the fingerprinting localization systems that RSS values are the foundation of realizing positioning. In most of the existing experimental systems, researchers build the radio map using a mobile device in the offline phase, and the user's location is computed using the same device in the online phase. However, in the actual localization system, this assumption is generally invalid, and the mobile devices used by different users are different in both the offline phase and the online phase. In the offline phase, aiming to reduce the labor and time costs of radio map construction, the crowdsourcing method has been proposed in the indoor localization domain, which brings a variety of distinct mobile devices [12, 14, 15]. In the online phase, the mobile devices which are used to localize the users may be different from the mobile devices used in the establishment of the radio map. The device diversity problem has been deeply studied in the last decade. The research results show that the main cause of such problem is the difference in hardware performance, which makes the differences between the RSS values collected by different devices may exceed 25 dB [16–18]. As a result, because of the adverse impact of the device diversity problem, the positioning accuracy of the crowdsourcing system is greatly reduced.

Although the establishment of a radio map for each device can obtain the highest positioning accuracy, this method is not practicable obviously as a result of numerous devices. In [19, 20], the linear regression algorithm (LR) is proposed to eliminate the device diversity for the crowdsourcing WLAN indoor localization system. However, the descriptions and simulations of the LR algorithm in [19, 20] are very simple. Therefore, in this paper, we discuss the proposed method in detail and get the complete simulation results. More importantly, the formula of error detection probability of the proposed LR algorithm is derived. The problem of device diversity and the adverse effects caused by this problem are analyzed at the beginning. Then, the LR method is applied to deal with the discussed problem. The device diversity will be diminished greatly with respect to the proposed algorithm. The advantages of this method include the following points: the system has a low computational complexity, does not need any training period, and can be finished automatically without user's intervention.

The main contributions of this paper are as follows.

(1) The linear relationship between the RSS data collected by different devices is proved. We obtain the RSS points by comparing the RSS vectors collected by different devices, and the slope and the intercept of the straight line determined by any two points are calculated. Since all the slopes and intercepts are equal, all the RSS points are on the same line. Therefore, the relationship of RSS data collected by different devices is linear

(2) The fast least trimmed squares (FAST-LTS) algorithm is proposed to eliminate the device diversity problem. Since the relationship between RSS values collected by different devices is linear, using the linear regression algorithm, all the RSS values can be mapped into the same signal space. Because the outliers appear in the collected RSS values frequently and seriously affect the performance of the linear least squares (LLS) algorithm, the FAST-LTS algorithm is used in this paper. Simulation results verify the effectiveness of the proposed algorithm, and all the RSS data are mapped into the same signal space

(3) We derived the probability of error detection of all fingerprints in the radio map. By deducing the formula, we can obtain that the probability of error detection is greater when the two fingerprints are closer. Hence, these fingerprints in the set of candidate nearest neighbor fingerprints contribute most of the error and need to be dealt with carefully

The rest of the paper is organized as follows. The related works are discussed in Section 2. In Section 3, we state that the problem statement on indoor localization and the linear relationship between the RSS data collected by different devices is proved. The linear regression method is proposed to solve the device diversity problem in Section 4. Section 5 analyzes the probability of error detection in the indoor localization system. The simulation and experimental results are presented in Section 6. Finally, Section 7 concludes the paper.

2. Background and Related Works

To handle the device diversity problem, the establishment of a radio map for each device can obtain the highest positioning accuracy. However, this method is not practicable obviously as a result of numerous devices [21]. Therefore, various solutions have been proposed as the alternations.

The device diversity problem was first discussed in [21]. Haeberlen et al. collected RSS values using different mobile devices at the same time and location in the test area, and they repeated this process at different locations. Then, the linear relationship between different RSS values collected by different mobile devices was inferred and used to eliminate the differences caused by different mobile devices. However, when an unknown device was used to find the current location, how to calculate the linear regression coefficients of the device had not been solved. In addition, although the authors in [21] had found the solution to overcome the device diversity problem, the solution was not applied to any localization systems.

Since collecting labeled RSS values for each mobile devices is a labor-intensive and time-consuming process, a semisupervised method is proposed in [17] to solve the device diversity problem using a small number of labeled RSS values. To solve the device diversity problem, multiple devices are treated as multiple learning tasks in this paper.

A latent feature space and a regression function are learned at the beginning, and then the signal spaces of all

devices are mapped to the latent feature space by the regression function. Accordingly, the differences between different devices have been significantly reduced, and the positioning accuracy has been greatly improved.

In [18], Tsui proposed an unsupervised learning algorithm to solve the problem in the WiFi localization system. In this paper, the Pearson product-moment correlation coefficient is used to label the RSS readings roughly collected by an unknown device at the beginning in the online phase. Then, different algorithms, such as regression algorithm and expectation maximization algorithm, are applied to train the transformation function. In [22], another solution using unsupervised learning algorithm is proposed to overcome the device diversity problem. In this paper, the probabilistic model is built to calculate the RSS values, and kernel estimation with a wide kernel width is used to reduce the difference in probability estimates. Although these methods reported some gain in localization accuracy, the unsupervised learning algorithm could not take the desired effect when the set APs detected by different devices are different.

Kjærgaard et al. utilized hyperbolic location fingerprinting (HLF) to solve the device diversity problem [23, 24]. In the training phase, the radio map is built by the signal strength ratio between two APs instead of an absolute RSS value from a single AP. Since the signal strength ratio is more stable than the absolute RSS value, the localization accuracy is significantly improved. However, the ratio term of the linear transformation function is not the only factor to be considered. If this offset component is significant in the linear relation or the set of the APs detected by different devices is different, this method is expected to fail.

In [25–27], three different of signal strength methods are proposed to reduce the impact of device diversity. In [25], Dong et al. used the different between all possible AP pairs, called DIFF, to build the radio map. In this radio map, the cost of the DIFF method is $\mathcal{O}(n^2)$ and may increase dramatically when the number of the AP increases. In [26], the signal strength difference (SSD) method approach subtracts the RSS value of an anchor AP from the other RSS values in the fingerprint. Therefore, each fingerprint contains only $n-1$ RSS differences, and the dimension of SSD method is $\mathcal{O}(n)$. As a result, the DIFF method achieves higher localization accuracy than the SSD method. Laoudias proposed the mean differential fingerprint (MDF) method in [27] which uses the mean RSS value to calculate the RSS differences to create the RSS fingerprint. The MDF method maintains the advantages of DIFF and SSD, which can achieve the high positioning accuracy as DIFF, while keeping the computational overhead similar to SSD.

In [28, 29], convolutional neural networks (CNN) are used to eliminate the device diversity problem. Cai et al. [28] proposed a device-free indoor localization system based on channel state information (CSI) in IEEE 802.11n through CNN, and the space diversity, time diversity, and frequency diversity of CSI are combined to design the more abundant localization features. In [29], the database is constructed using the magnetic pattern (MP) in the offline phase, and the location is calculated using the CNN algorithm in the online phase to eliminate the device diversity problem. Although this is a magnetic positioning system, it can give us a lot of inspiration.

In [30–33], the LR method is used to mine the internal relationship between data and eliminate the differences between data through linear regression. In [30], an automatic device transparent RSS-based indoor localization system has been proposed, and the linear least squares (LLS) algorithm is applied in this system. Combining the offset component and ratio term makes the LLS algorithm a complete algorithm. Moreover, the algorithm complexity of LLS is much lower than the other algorithms discussed above. Li et al. [31] presented a prototype model of a multiple-surveyor-multiple-client system in the crowdsourcing localization system. The linear regression model is applied to calibrate across participating training devices, and a geometric distribution is used to obtain a conditional likelihood that the client observes invisible access points in the training phase. Ye et al. [32] proposed a device calibration algorithm to fuse samples from different devices to obtain grid fingerprints and a two-step online positioning algorithm to localize user's position. In [33], the FAST-LTS algorithm is proposed to deal with large data sets. Due to the FAST-LTS algorithm, the LTS estimator becomes available as a tool for analyzing large data sets and to detect outliers or deviating substructures. This algorithm provides us with a very good idea to deal with the problem of device diversity. Therefore, we propose a FAST-LTS algorithm to eliminate the device diversity problem in [19, 20], which achieves good results. In this paper, we further improve this algorithm.

3. Problem Formulation

For the RSS fingerprint-based localization method, the localization accuracy greatly depends on the mapping relation between the fingerprint and its corresponding coordinates stored in the radio map. In the offline phase, the localization area is divided into a discrete grid with n RPs and m APs, which are deployed in this area. We collect RSS values from the APs at each RP, and a fingerprint radio map is constructed that holds the RSS for m APs and n RPs. The system diagram of the RSS fingerprint-based WLAN indoor localization system is shown in Figure 1. In traditional WLAN positioning methods, hundreds of fingerprints are collected over time for each RP. After RSS preprocessing, a $1 \times m$ vector $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{im})$ is generated at each RP, where x_{ik} is the received signal strength measured by the training device at the i-th RP from the k-th AP. The i-th fingerprint $(\mathbf{x}_i, \mathbf{c}_i)$ is the combination of the RSS measurement $\mathbf{x}_i$ and the coordinate $\mathbf{c}_i = (c_{i1}, c_{i2})$ of RP S_i. All the fingerprints are tabulated into the radio map that can be represented in Figure 1.

In the online phase, the user collects the RSS value $\mathbf{y} = (y_{j1}, y_{j2}, \cdots, y_{jm})$ at an unknown position S_j by a mobile device, and then the user's position could be estimated by comparing $\mathbf{y}$ with the radio map. Usually, if $\mathbf{y}$ is similar to the fingerprint $\mathbf{x}_i$ in the radio map, we reason that user's location S_j must be close to S_i.

In the actual localization system, the mobile devices used by different users are distinct from each other. Since the

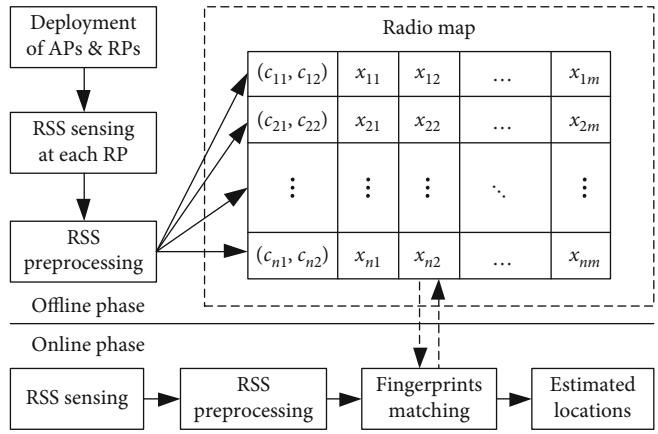

Figure 1: Typical WLAN indoor localization system.

WLAN signal receivers with different performance are equipped with the different mobile devices, so the different mobile devices may have different signal sensing capacities and yield different RSS values. To illustrate, we used five different mobile devices to collect RSS values from a single AP at a particular location and plotted the histogram in Figure 2. Due to the large fluctuation of the RSS value in the indoor environment, if only 2-3 RSS is collected on one RP, the accurate RSS distribution cannot be obtained. Therefore, we collected 100 RSS values in each RP. As shown in Figure 2, the RSS values collected by different devices can be quite different even at the same location. This directly results in erroneous location estimations if we use one device's data for training and another device's data for locating.

We define $\mathcal{X}$ and $\mathcal{Y}$ as the signal space for the radio map $\mathbf{X}$ built by the training device and the online RSS values collected by the localization device, respectively. Let the fingerprint $\mathbf{x}^*$ in the radio map $\mathbf{X}$ is the nearest neighbor online RSS value $\mathbf{y}$. Because of the different signal receiving capability of the different mobile devices, the RSS values collected at the close physical locations are obvious different. Hence, one of the key challenges arises: how to process these RSS values collected by different devices to make the $\mathbf{x}^*$ in closer to $\mathbf{y}$. Mathematically,

$$\mathcal{Y} \approx \mathcal{F}(\mathcal{X}). \tag{1}$$

By learning $\mathcal{F}$, the radio map $\mathbf{X}$ build by the training device could be used to localize any other devices.

Next, we will explore the relationship between different RSS values collected by different mobile devices. The signal processing diagram of the mobile device is shown in Figure 3.

We suppose that the transmit power of the AP is P_t, and the receiving power of the mobile device is

$$P = P_t + P_L + G_t + G_r + \alpha, \tag{2}$$

where P is the receiving power of the antenna, P_L is the path loss at distance d form AP, G_t is the transmit antenna gain, G_r is the receiving antenna gain, and α is the power amplifier magnification.

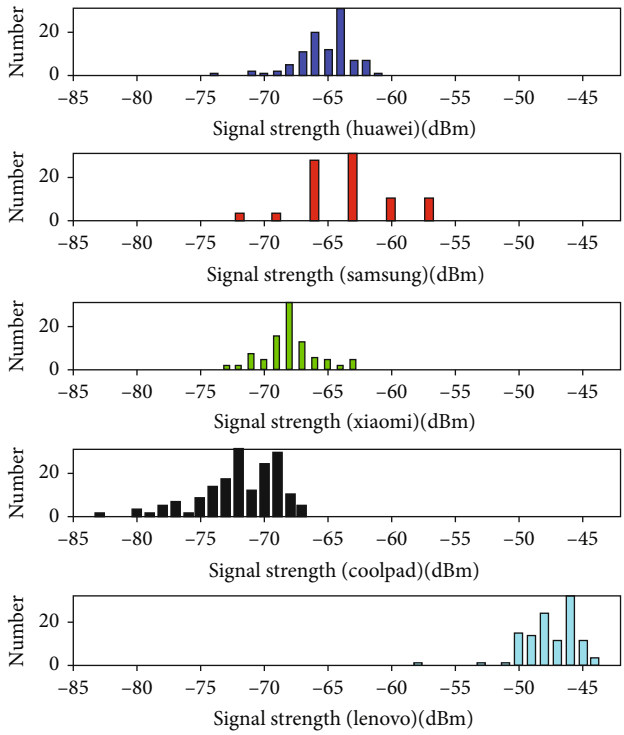

Figure 2: RSS values collected by five different devices at the same time and location from the same APs..

Assume that $\mathbf{x}^A = [P_1^A, P_2^A, \cdots, P_m^A]$ and $\mathbf{x}^B = [P_1^B, P_2^B, \cdots, P_m^B]$ represent the RSS vectors collected by two distinct devices A and B at the same location and the same time. In order to obtain the mapping function in Eq. (1), we compare $\mathbf{x}^A$ and $\mathbf{x}^B$ and get m points whose coordinates are (P_i^A, P_i^B), $i = 1, 2, \cdots, m$, and the RSS value collected by devices A and B from the i-th AP are

$$P_i^A = P_{ti} + P_{Li} + G_{ti} + G_r^A + \alpha^A, \tag{3}$$

$$P_i^B = P_{ti} + P_{Li} + G_{ti} + G_r^B + \alpha^B, \tag{4}$$

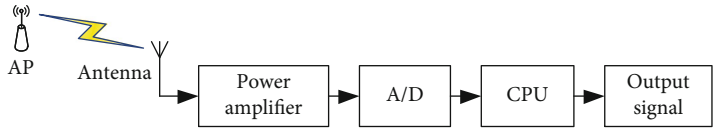

Figure 3: Signal processing diagram of the mobile device.

where the subscript i represents the i-th AP, and the superscripts A and B represent the device A and the device B, respectively.

Figure 4 shows three points determined by matrix $\mathbf{x}^A$ and $\mathbf{x}^B$, any two points can determine a straight line, and then we get 3 lines in this figure. Suppose the slopes of the 3 lines are a_{12}, a_{13}, and a_{23}, and the intercept of the 3 lines are b_{12}, b_{13}, b_{23}, if $a_{12} = a_{13} = a_{23}$ and $b_{12} = b_{13} = b_{23}$, all the 3 points are in the same line; that is, there is a linear relationship between $\mathbf{x}^A$ and $\mathbf{x}^B$. Therefore, if the slopes and intercepts of the lines determined by any two of the n data points are equal, it can be determined that all the n data points are on the same line, which means the relationship between $\mathbf{x}^A$ and $\mathbf{x}^B$ is linear. In this paper, using this method, the relationship between $\mathbf{x}^A$ and $\mathbf{x}^B$ can be proved.

The linear equation determined by two points (P_i^A, P_i^B) and (P_j^A, P_j^B) is

$$P^B = a_{ij}P^A + b_{ij}. \tag{5}$$

The slope and intercept of the equation are

$$a_{ij} = \frac{P_i^B - P_j^B}{P_i^A - P_j^A}, \tag{6}$$

$$b_{ij} = \frac{P_i^A P_j^B - P_j^A P_i^B}{P_i^A - P_j^A}. \tag{7}$$

Substituting the values of Eq. (3) and Eq. (4) into Eq. (6) and Eq. (7), then we can get

$$\mathrm{a}_{ij} = 1, \tag{8}$$

$$b_{ij} = G_r^B + \alpha^B - G_r^A - \alpha^A. \tag{9}$$

As can be seen from Eq. (8) and Eq. (9), in the ideal case, the slope and intercept of the line determined by two points (P_i^A, P_i^B) and (P_j^A, P_j^B) are only related to the antenna gains and amplification factor of the device A and device B. Therefore, the m points determined by the vectors $\mathbf{x}^A$ and $\mathbf{x}^B$ are on the same line.

$$\mathbf{x}^B = a\mathbf{x}^A + b, \tag{10}$$

where the linear equation coefficients are $a = a_{ij}$ and $b = b_{ij}$.

Assume that the RSS vectors $\mathbf{x}^A$ and $\mathbf{x}^B$ are collected at the same location and at different times. In an ideal case, the transmitting power of the AP, the working status of the mobile terminal, and the path loss remain constant; therefore, the linear relationship between $\mathbf{x}^A$ and $\mathbf{x}^B$ is same as Eq. (10).

In practice, the instability of the AP transmit power and the complexity of the indoor electromagnetic environment make the RSS values collected by the mobile device unstable at different times.

$$P_i^A = P_{ti}^A + P_{Li}^A + G_{ti} + G_r^A + \alpha^A + \xi_m^A, \tag{11}$$

$$P_i^B = P_{ti}^B + P_{Li}^B + G_{ti} + G_r^B + \alpha^B + \xi_m^B, \tag{12}$$

$$P_j^A = P_{tj}^A + P_{Lj}^A + G_{tj} + G_r^A + \alpha^A + \xi_m^A, \tag{13}$$

$$P_j^B = P_{tj}^B + P_{Lj}^B + G_{tj} + G_r^B + \alpha^B + \xi_m^B, \tag{14}$$

where $P_{ti}^B = P_{ti}^A + \delta_{Pti}$, $P_{Li}^B = P_{Li}^A + \delta_{PLi}$, $P_{tj}^B = P_{tj}^A + \delta_{Ptj}$, and $P_{Lj}^B = P_{Lj}^A + \delta_{PLj}$.

Substituting the values of Eq. (11), Eq. (12), Eq. (13), and Eq. (14) into Eq. (6),

$$\begin{aligned}
a_{ij} = \frac{P_i^B - P_j^B}{P_i^A - P_j^A} &= \left(P_{ti}^B + P_{Li}^B + G_{ti} + G_r^B + \alpha^B + \xi_m^B - P_{tj}^B - P_{Lj}^B \right.\\
&\quad \left. - G_{tj} - G_r^B - \alpha^B - \xi_m^B\right) / \left(P_{ti}^A + P_{Li}^A + G_{ti} + G_r^A + \alpha^A + \xi_m^A \right.\\
&\quad \left. - P_{tj}^A - P_{Lj}^A - G_{tj} - G_r^A - \alpha^A - \xi_m^A\right)\\
&= \frac{P_{ti}^B + P_{Li}^B + G_{ti} - P_{tj}^B - P_{Lj}^B - G_{tj}}{P_{ti}^A + P_{Li}^A + G_{ti} - P_{tj}^A - P_{Lj}^A - G_{tj}}\\
&= \left(P_{ti}^A + \delta_{Pti} + P_{Li}^A + \delta_{PLi} + G_{ti} - P_{tj}^A - \delta_{Ptj} - P_{Lj}^A - \delta_{PLj} - G_{tj}\right)\\
&\quad / \left(P_{ti}^A + P_{Li}^A + G_{ti} - P_{tj}^A - P_{Lj}^A - G_{tj}\right)\\
&= \left(P_{ti}^A + P_{Li}^A + G_{ti} - P_{tj}^A - P_{Lj}^A - G_{tj} + \delta_{Pti} + \delta_{PLi} - \delta_{Ptj} - \delta_{PLj}\right)\\
&\quad / \left(P_{ti}^A + P_{Li}^A + G_{ti} - P_{tj}^A - P_{Lj}^A - G_{tj}\right).
\end{aligned} \tag{15}$$

When the indoor environment approaches an ideal environment, which means the difference of AP transmitting power at different times $\delta_{Pt} \to 0$, and the difference of path loss at different times $\delta_{PL} \to 0$, then

$$\lim_{\substack{\delta_{Pt}\to 0 \\ \delta_{PL}\to 0}} a_{ij} = 1. \tag{16}$$

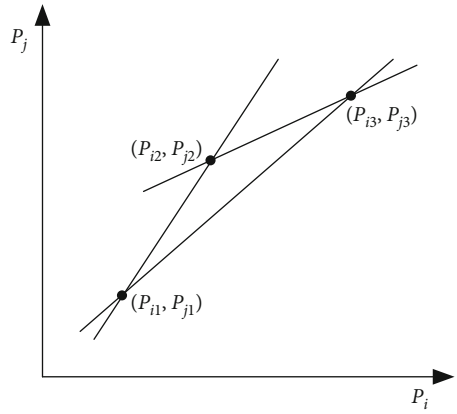

Figure 4: Schematic diagram of relation between data points.

Similarly, we can get

$$\lim_{\substack{\delta_{Pt}\to 0\\ \delta_{PL}\to 0}} b_{ij} = G_r^B + \alpha^B - G_r^A - \alpha^A. \tag{17}$$

As a result, the linear relationship between $\mathbf{x}^A$ and $\mathbf{x}^B$ can be proved by Eq. (16) and Eq. (17).

In an actual indoor environment, the working status of the AP is unstable, and the electromagnetic environment in the room is very complicated. Therefore, the difference δ_{Pt} of the transmission power of the AP and the difference δ_{PL} of the path loss are not equal to 0, and as the result, the linear equation coefficients are

$$a_{ij} = 1 + \delta_a, \tag{18}$$

$$b_{ij} = G_r^B + \alpha^B - G_r^A - \alpha^A + \delta_b, \tag{19}$$

where $\delta_a = \delta_{ij}/P_i^A - P_j^A$ is the error of slope, and $\delta_b = P_i^A\delta_{APj} - P_j^A\delta_{APi}/P_i^A - P_j^A$ is the error of intercept, $\delta_{ij} = \delta_{APi} - \delta_{APj}$, $\delta_{APi} = \delta_{Pti} + \delta_{PLi}$, and $\delta_{APj} = \delta_{Ptj} + \delta_{PLj}$.

Figure 5 shows the slope and intercept of Eq. (18) and Eq. (19). Considering the noise, we can see from Eq. (18), Eq. (19), and Figure 5 that the slope and intercept of the straight line are centered on the ideal slope and intercept and fluctuate within an error range. Therefore, the relationship between RSS data vectors $\mathbf{x}^A$ and $\mathbf{x}^B$ is approximately linear.

In Figure 6, the comparison results of RSS values collected by five distinct devices are plotted. In this figure, each point represents the RSS values collected by two distinct devices at the same RP from the same AP. For example, the top left subplot in Figure 6 represents the RSS values collected by Lenovo laptop and Huawei mobile device. Figure 6 verifies the results of Eq. (18), Eq. (19), and Figure 5, and the linear correlation can be drawn by the RSS values collected by Lenovo laptop and other devices.

As a result, we apply the linear regression method (LR) as the mapping function in this paper. The LR model is defined by specifying how the signal space of localization device $\mathcal{X}$ is mapped into $\mathcal{Y}$ in the signal space of the training device. Based on the LR method, a unique radio map in the offline phase could be built and improve the localization accuracy in the RSS-based crowdsourcing localization system.

4. Linear Regression Algorithm against Device Diversity for the Crowdsourcing Fingerprint Indoor Localization System

In this section, a preprocessing procedure is used to stabilize the acquisition of RSS values at the beginning. Then, the RSS values collected by an unknown device are labeled automatically with a rough location estimation using a correlation ratio computed from the Pearson product-moment correlation coefficient. Finally, the linear regression algorithm (LR) is proposed as the mapping function to solve the device diversity problem, and the fast least trimmed squares (FAST-LTS) is applied for the LR method to provide a more robust performance.

4.1. Preprocessing of RSS Values. The first step in our work is to mitigate the RSS fluctuations caused by the complexity of the indoor environment. Typically, when building the radio map, we measure a large number of RSS values at each RP to eliminate the noise. Let $\mathbf{RSS}_{li} = \{rss_1, rss_2, \cdots, rss_p\}$ be the set of RSS values collected at location l from the i-th AP. When the RSS values are collected to build the radio map or estimate the current location, the length of all **RSS** must be the same. However, for some reasons, some APs cannot work properly. To ensure that the RSS measurements are the same length, we use a value of −110 dBm to fill the missing RSS value, and we denote it as an outlier. These outliers could affect the linear regression process and produce erroneous location estimations, as shown in Figure 7.

Figure 7(a) plots the linear regression function of the RSS values for two different devices, where the traditional average is used for estimation. The uncertainty in the RSS samples can be seen clearly in this figure. To achieve higher positioning accuracy, the original RSS measurements should be preprocessed prior to the localization process. Average, mode, and median are the common data preprocessing methods in mathematics. Because the average takes all the RSS values into consideration and uses the data more efficiently, so we average the RSS values to overcome the RSS fluctuations. However, the average is susceptible to the outliers of -110 dBm. The existence of outliers could seriously affect the accuracy of the average and produce erroneous location estimations. Hence, the truncated average is used in our work to stabilize the collected RSS samples:

$$x_{li} = \frac{\sum_{j=1}^{p} rss_j \mathbf{I}\left(rss_j \neq -110\text{dBm}\right)}{\sum_{j=1}^{p} \mathbf{I}\left(rss_j \neq -110\text{dBm}\right)}, \tag{20}$$

where $\mathbf{I}(\cdot)$ is an indicator function.

There is also a special case when calculating the truncated average. Consider a situation that only one sample reports valid reading $rss_i = \alpha$ dBm when all the other RSS values are -110 dBm, the result of Eq. (20) is $x_{li} = \alpha$ dBm. Therefore, when applying Eq. (20), we first calculate the ratio t of the

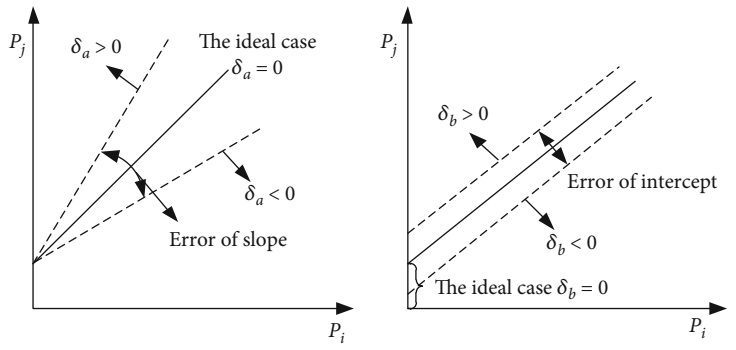

Figure 5: The slope and intercept between RSS values collected by different devices. (a) Slope diagram. (b) Intercept diagram.

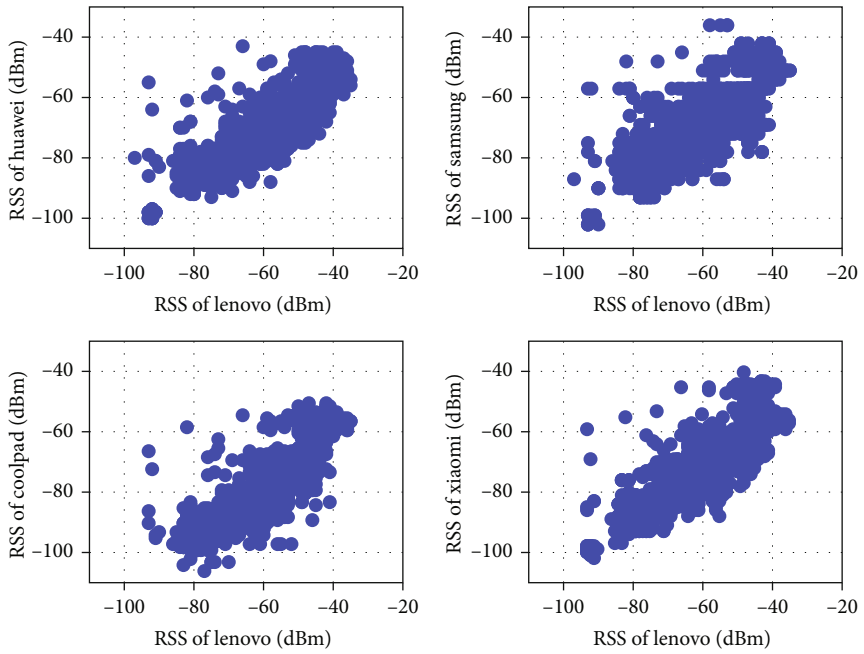

Figure 6: Linear correlation between RSS values for different devices.

normal RSS value in the collected vector and set a threshold t_{th}, and then we can get the truncated average

$$x_{li} = \begin{cases} x_{li}, & \text{if } t > t_{th} \\ -110\text{dBm}, & \text{if } t \le t_{th} \end{cases}. \tag{21}$$

4.2. Rough Location Estimation. After completing the preprocessing of RSS data, we use the linear regression method as the mapping function.

$$\mathbf{y} = a_i \mathbf{x}_i + b_i \mathbf{1}, \tag{22}$$

where $\mathbf{x}_i$ is the i-th fingerprint in the radio map collected by the training device, $\mathbf{y}$ is the RSS values measured in the online phase by localization device, and (a_i, b_i) are the coefficients in the mapping function.

Based on the mapping function in Eq.(22), the RSS values collected by different devices can be transformed to the same signal space. Accordingly, the device diversity problem can be solved. However, the RSS values collected in the online phase are unlabeled and cannot be processed using the linear regression algorithm. Therefore, the correlation ratio computed from the Pearson product-moment correlation coefficient is proposed to roughly label the RSS values collected in the online phase as [18].

$$r(\mathbf{y}, \mathbf{x}_i) = \frac{\sum_{k=1}^{m} (y_k - \bar{\mathbf{y}})(x_{ik} - \bar{\mathbf{x}}_i)}{\sqrt{\sum_{k=1}^{m} (y_k - \bar{\mathbf{y}})^2 \sum_{k=1}^{m} (x_{ik} - \bar{\mathbf{x}}_i)^2}}, \tag{23}$$

where m is the number of APs, y_k and x_{ik} are the RSS values measured from the k-th AP, $\bar{\mathbf{y}} = 1/m \sum_{k=1}^{m} y_k$ is the average of the RSS values from the tracking device, and $\bar{\mathbf{x}}_i = 1/m \sum_{k=1}^{m} x_{ik}$ represents the mean of the RSS values measured by the training device in the i-th fingerprint. The range of the Pearson correlation ratio is $[-1, 1]$, where 1 indicates the highest

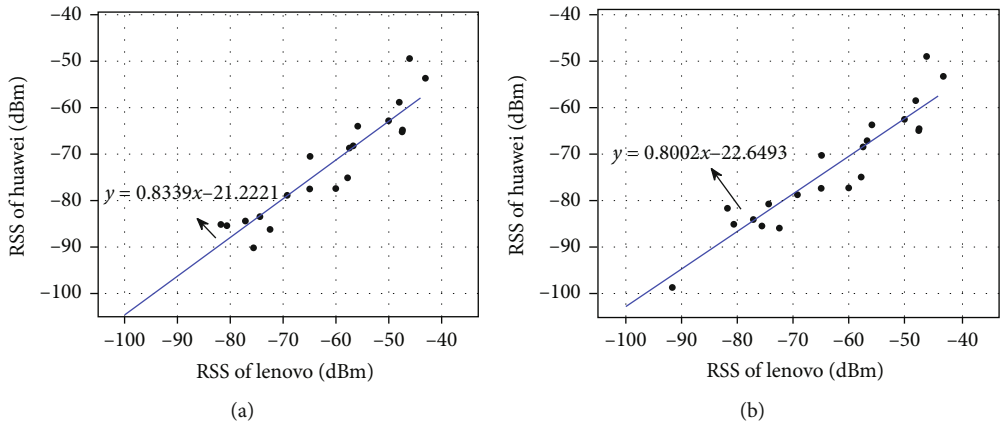

Figure 7: Preprocessing of RSS values using trimmed average. (a) Traditional average. (b) Trimmed average.

positive linear correlation between RSS values and the fingerprints with a negative r should be excluded.

When we get the online RSS values $\mathbf{y}$, the Pearson correlation ratio r between $\mathbf{y}$ and all fingerprints $\mathbf{x}_i$ in $\mathbf{X}$ can be computed using Eq. (23). By setting a correlation coefficient threshold r_{th}, we can obtain the set of nearest neighbor fingerprints in the radio map $\mathbf{X}$ for $\mathbf{y}$.

$$\mathbb{A} = \{\mathbf{x}_i \in \mathbf{X} \mid r(\mathbf{y}, \mathbf{x}_i) > r_{th}, 0 \le r_{th} \le 1\}. \tag{24}$$

The fingerprints in $\mathbb{A}$ have the strongest linear correlation with the online RSS values, the online RSS values can be labeled roughly, and the linear regression mapping function can be obtained more accurately.

4.3. Linear Regression Algorithm against Device Diversity Problem. In Eq. (22), the important parameters, a_i and b_i, need to be computed first. Because the linear least squares (LLS) algorithm is more sensitive to the outliers, we use the fast least trimmed squares (FAST-LTS) algorithm to compute the parameters in Eq. (22).

Assume that the amount of the nearest neighbors in $\mathbb{A}$ is c , the FAST-LTS solution for linear regression with intercept is given by

$$\min_{a_i, b_i} \sum_{i=1}^{h} d(i)^2, \tag{25}$$

where $h = \text{int}\,[(c+2)/2]$, $d(i) = \|\mathbf{y} - (a_i\mathbf{x}_i + b_i\mathbf{1})\|$, and $\|\bullet\|$ is norm 2 of a vector, $d(i)^2$ are the ordered squared residuals: $d(1)^2 \le d(2)^2 \le \cdots \le d(i)^2 \le \cdots \le d(c)^2$.

Given the h-subset H_{old} of all nearest neighbors, the C − step is used to compute a_i and b_i as follows [33]:

(1) Compute $\mathbf{a}_{\text{old}}$ and $\mathbf{b}_{\text{old}} :=$ least squares regression estimator based on H_{old}

(2) Compute the residuals $d_{\text{old}}(i)$ for $i = 1, \cdots, c$

(3) Sort the absolute values of these residuals, $|d_{\text{old}}(1)| \le |d_{\text{old}}(2)| \le \cdots \le |d_{\text{old}}(c)|$

(4) Arrange the absolute values of the residuals in ascending order, let H_{new} be a subset consisting of the nearest neighbors corresponding to the first h absolute values of the residuals in the sequence

(5) Compute $\mathbf{a}_{\text{new}}$ and $\mathbf{b}_{\text{new}} :=$ least squares regression estimator based on H_{new}

Repeating C − step with numerous H_{old}, a lot of regression coefficients will be gotten. The approximate solution is the coefficient corresponding to the least $\sum_{i=1}^{h} d(i)^2$. Using the parameters a_i and b_i, $\mathbf{x}$ can be transformed to the signal space of the online data $\mathbf{y}$:

$$\mathbf{x}_i' = a_i\mathbf{x}_i + b_i\mathbf{1}, \tag{26}$$

where $\mathbf{x}_i' \in \mathcal{Y}$. Since both $\mathbf{x}_i'$ and $\mathbf{y}$ belong to the same signal space, the KNN algorithm based on the RSS Euclidean distance can be used to estimate the user's location.

5. Analysis of Probability of Error Detection of the Proposed Algorithm

In this section, we analyzed the probability of error detection in the crowdsourcing indoor localization system. In the offline training phase, the radio map $\mathbf{X}$ consists of n fingerprints that are built by the training device D^T. Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are two fingerprints in the radio map $\mathbf{X}$, $\mathbf{y}$ is the RSS value collected by the localization device in the online phase. Assume that the fingerprint $\mathbf{x}_1$ is the nearest neighbor to the online RSS value $\mathbf{y}$. Based on the linear regression model, the relation between the RSS vector $\mathbf{y}$ and the fingerprint $\mathbf{x}_1$ can be expressed as

$$\mathbf{y} = a_0\mathbf{x}_1 + b_0\mathbf{1} + \boldsymbol{\varepsilon}, \tag{27}$$

where a_0 and b_0 are the real linear regression coefficient between two RSS vectors, and $\boldsymbol{\varepsilon}$ is a $1 \times m$ noise vector. We suppose that $\boldsymbol{\varepsilon}$ is the Gaussian distribution of $N(0, \sigma_\varepsilon^2)$, and the variance σ_ε^2 is unknown.

Using the proposed LR method, the online RSS vector $\mathbf{y}$ and the radio map $\mathbf{X}$ can be transferred to the same signal space

$$\hat{\mathbf{x}}_1 = a_1\mathbf{x}_1 + b_1\mathbf{1}, \tag{28}$$

$$\hat{\mathbf{x}}_2 = a_2\mathbf{x}_2 + b_2\mathbf{1}. \tag{29}$$

In this paper, the KNN algorithm is used to estimate the location of the online RSS vector $\mathbf{y}$. First of all, the Euclidean distance between the online point $\mathbf{y}$ and the fingerprints in the radio map $\mathbf{X}$ should be calculated by

$$d_1 = \|\mathbf{y} - \mathbf{x}\wedge_1\|^2, \tag{30}$$

$$d_2 = \|\mathbf{y} - \mathbf{x}\wedge_2\|^2. \tag{31}$$

In the KNN algorithm, we choose the fingerprints with the smallest Euclidean distance as the nearest neighbor of $\mathbf{y}$. Assume that $k = 1$, if $d_1 > d_2$, then $\mathbf{x}_1$ is the nearest neighbor of $\mathbf{y}$; on the contrary, $\mathbf{x}_2$ is the nearest neighbor of $\mathbf{y}$. We have assumed that $\mathbf{x}_1$ in the radio map $\mathbf{X}$ is the nearest neighbor of $\mathbf{y}$, and an error occurs when the calculation results show that $d_1 < d_2$. Therefore, we can get the probability of localization error from Eq. (32) as given as

$$P_e = P\{\mathrm{NN}(\mathbf{y}) = \mathbf{x}_2 \mid \mathrm{NN}(\mathbf{y}) = \mathbf{x}_1\} = P\left\{\left\|\mathbf{y} - \mathbf{x}\wedge_2{}^2\right\| < \|\mathbf{y} - \mathbf{x}\wedge_1\|^2\right\}, \tag{32}$$

where $\mathrm{NN}(\cdot)$ is the nearest neighbor of the online data $\mathbf{y}$.

If the nearest neighbor of the online data $\mathbf{y}$ estimated by the localization system is correct, then, $\mathrm{NN}(\mathbf{y}) = \mathbf{x}_1$, the linear regression coefficient could be calculated by Eq. (25),

$$\begin{aligned}(a_1, b_1) &= \arg\ \min_{a,b}\|\mathbf{y} - (a\mathbf{x}_1 + b\mathbf{1})\|^2 \\ &= \arg\ \min_{a,b}\|a_0\mathbf{x}_1 + b_0\mathbf{1} + \boldsymbol{\varepsilon} - (a\mathbf{x}_1 + b\mathbf{1})\|^2 \\ &= \arg\ \min_{a,b}\|(a_0 - a)\mathbf{x}_1 + (b_0 - b)\mathbf{1} + \boldsymbol{\varepsilon}\|^2.\end{aligned} \tag{33}$$

Obviously, in Eq. (33), when $a = a_0$ and $b = b_0$, we get the minimizer of Eq. (30), and the linear regression coefficient between $\mathbf{y}$ and $\mathbf{x}_1$ is

$$\begin{aligned}a_1 &= a_0, \\ b_1 &= b_0.\end{aligned} \tag{34}$$

Substituting the values of Eq. (34) into Eq. (33), we can get

$$d_1 = \|\mathbf{y} - \mathbf{x}\wedge_1\|^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \sum_{i=1}^{m}\varepsilon_i^2 \sim \chi_m^2, \tag{35}$$

where χ_m^2 is the chi-square distribution with degrees of freedom of m.

When a localization error occurs, then $\mathrm{NN}(\mathbf{y}) = \mathbf{x}_2$, and the linear regression coefficient between $\mathbf{y}$ and $\mathbf{x}_2$ becomes

$$\begin{aligned}(a_2, b_2) &= \arg\ \min_{a,b}\|\mathbf{y} - (a\mathbf{x}_2 + b\mathbf{1})\|^2 \\ &= \arg\ \min_{a,b}\|a_0\mathbf{x}_1 + b_0\mathbf{1} + \boldsymbol{\varepsilon} - (a\mathbf{x}_2 + b\mathbf{1})\|^2 \\ &= \arg\ \min_{a,b}\|(a_0\mathbf{x}_1 - a\mathbf{x}_2) + (b_0 - b)\mathbf{1} + \boldsymbol{\varepsilon}\|^2.\end{aligned} \tag{36}$$

By making use of triangular inequality,

$$\|(a_0\mathbf{x}_1 - a\mathbf{x}_2) + (b_0 - b)\mathbf{1} + \boldsymbol{\varepsilon}\| \le \|a_0\mathbf{x}_1 - a\mathbf{x}_2\| + \|(b_0 - b)\mathbf{1}\| + \|\boldsymbol{\varepsilon}\|. \tag{37}$$

From Eq. (37), $b_2 = b_0$ is the solution when we get the minimizer of Eq. (36). Therefore, Eq. (36) can be equivalent to

$$\begin{aligned}a_2 &= \arg\ \min_{a,b}\|a_0\mathbf{x}_1 - a\mathbf{x}_2\|^2 \\ &= \arg\ \min_{a,b}(a_0\mathbf{x}_1 - a\mathbf{x}_2)^T(a_0\mathbf{x}_1 - a\mathbf{x}_2) \\ &= \arg\ \min_{a,b}\left(a_0^2\mathbf{x}_1^T\mathbf{x}_1 - aa_0\mathbf{x}_1^T\mathbf{x}_2 - aa_0\mathbf{x}_2^T\mathbf{x}_1 + a^2\mathbf{x}_2^T\mathbf{x}_2\right).\end{aligned} \tag{38}$$

Because $\mathbf{x}_1$ and $\mathbf{x}_2$ are $m \times 1$ vector, then $\mathbf{x}_1^T\mathbf{x}_2 = \mathbf{x}_2^T\mathbf{x}_1$; therefore, Eq. (38) can be written as

$$a_2 = \arg\ \min_{a,b}\left[a_0^2\|\mathbf{x}_1\|^2 - 2aa_0\mathbf{x}_1^T\mathbf{x}_2 + a^2\|\mathbf{x}_2\|^2\right]. \tag{39}$$

In Eq. (39), a_0, $\mathbf{x}_1$, and $\mathbf{x}_2$ are already known. Let $A = a_0^2\|\mathbf{x}_1\|^2$, $B = a_0\mathbf{x}_1^T\mathbf{x}_2$, and $C = \|\mathbf{x}_2\|^2$, and then Eq. (39) can be expressed as

$$a_2 = \arg\ \min_{a,b}\left(A - 2Ba + Ca^2\right). \tag{40}$$

From Eq. (40), we can get the solution of a_2 by making the derivative of the quadratic equation equal to 0. Let $\Delta = A - 2Ba + Ca^2$ and $\partial\Delta/\partial a = 0$ yields

$$\frac{\partial\Delta}{\partial a} = 2Ca - 2B = 0. \tag{41}$$

Then, we can get

$$a_2 = \frac{B}{C} = \frac{a_0\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_2\|^2}. \tag{42}$$

At last, the linear regression coefficient between $\mathbf{y}$ and $\mathbf{x}_2$ can be calculated by

$$\begin{aligned}a_2 &= \frac{B}{C} = \frac{a_0\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_2\|^2}, \\ b_2 &= b_0.\end{aligned} \tag{43}$$

Substituting the linear regression coefficients into Eq. (31), then we can get

$$d_2 = \|\mathbf{y} - \hat{\mathbf{x}}_2\| = \left\| a_0\mathbf{x}_1 + b_0\mathbf{1} + \boldsymbol{\varepsilon} - \frac{a_0\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_2\|^2}\mathbf{x}_2 - b_0\mathbf{1} \right\|^2 = \|a_0\mathbf{x}_1 - a_2\mathbf{x}_2 + \boldsymbol{\varepsilon}\|^2. \quad (44)$$

Using Eq. (35) and (44), the probability of localization error in Eq. (32) can be computed by

$$\begin{aligned} P_e &= P\{\|\mathbf{y} - \mathbf{x}\wedge_2\|^2 < \|\mathbf{y} - \mathbf{x}\wedge_1\|^2\} = P\{\|a_0\mathbf{x}_1 - a_2\mathbf{x}_2 + \boldsymbol{\varepsilon}\|^2 < \|\boldsymbol{\varepsilon}\|^2\} \\ &= P\{a_0^2\mathbf{x}_1^T\mathbf{x}_1 - a_0a_2\mathbf{x}_1^T\mathbf{x}_2 + a_0\mathbf{x}_1^T\boldsymbol{\varepsilon} + a_2^2\mathbf{x}_2^T\mathbf{x}_2 - a_0a_2\mathbf{x}_2^T\mathbf{x}_1 \\ &\quad - a_2\mathbf{x}_2^T\boldsymbol{\varepsilon} + a_0\boldsymbol{\varepsilon}^T\mathbf{x}_1 - a_2\boldsymbol{\varepsilon}^T\mathbf{x}_2 + \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} < \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}\} \\ &= P\{a_0^2\mathbf{x}_1^T\mathbf{x}_1 - 2a_0a_2\mathbf{x}_1^T\mathbf{x}_2 + 2a_0\mathbf{x}_1^T\boldsymbol{\varepsilon} - 2a_2\mathbf{x}_2^T\boldsymbol{\varepsilon} + a_2^2\mathbf{x}_2^T\mathbf{x}_2 < 0\} \\ &= P\{2(a_0\mathbf{x}_1^T - a_2\mathbf{x}_2^T)\boldsymbol{\varepsilon} < 2a_0a_2\mathbf{x}_1^T\mathbf{x}_2 - a_0^2\mathbf{x}_1^T\mathbf{x}_1 - a_2^2\mathbf{x}_2^T\mathbf{x}_2\} \\ &= P\{2(a_0\mathbf{x}_1^T - a_2\mathbf{x}_2^T)\boldsymbol{\varepsilon} < 2a_0a_2\mathbf{x}_1^T\mathbf{x}_2 - a_0^2\|\mathbf{x}_1\|^2 - a_2^2\|\mathbf{x}_2\|^2\}. \end{aligned} \quad (45)$$

For the right side of the inequality in Eq. (45), we set

$$\begin{aligned} \xi &= 2a_0a_2\mathbf{x}_1^T\mathbf{x}_2 - a_0^2\|\mathbf{x}_1\|^2 - a_2^2\|\mathbf{x}_2\|^2 \\ &= 2a_0\frac{a_0\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_2\|^2}\mathbf{x}_1^T\mathbf{x}_2 - a_0^2\|\mathbf{x}_1\|^2 - \left(\frac{a_0\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_2\|^2}\right)^2\|\mathbf{x}_2\|^2 \\ &= a_0^2\frac{(\mathbf{x}_1^T\mathbf{x}_2)^2}{\|\mathbf{x}_2\|^2} - a_0^2\|\mathbf{x}_1\|^2 = a_0^2\frac{(\mathbf{x}_1^T\mathbf{x}_2)^2 - \|\mathbf{x}_1\|^2\|\mathbf{x}_2\|^2}{\|\mathbf{x}_2\|^2}. \end{aligned} \quad (46)$$

Using Cauchy-Schwartz inequality,

$$(\mathbf{x}_1^T\mathbf{x}_2)^2 \le \|\mathbf{x}_1\|^2\|\mathbf{x}_2\|^2. \quad (47)$$

Therefore, $\xi \le 0$ is in Eq. (46).

For the left side of the inequality in Eq. (45), we set $\eta = \mathbf{s}^T\boldsymbol{\varepsilon}$. In this equation,

$$\mathbf{s}^T = 2(a_0\mathbf{x}_1^T - a_2\mathbf{x}_2^T) = 2\left(a_0\mathbf{x}_1^T - \frac{a_0\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_2\|^2}\mathbf{x}_2^T\right) = 2a_0\mathbf{x}_1^T\left(\mathbf{I} - \frac{\mathbf{x}_2\mathbf{x}_2^T}{\|\mathbf{x}_2\|^2}\right), \quad (48)$$

where $\mathbf{I}$ is the identity matrix.

Since $\boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2)$, then

$$\boldsymbol{\eta} \sim N(0, \sigma^2), \quad (49)$$

where

$$\begin{aligned} \sigma^2 &= \sigma_\varepsilon^2\mathbf{s}^T\mathbf{s} = \sigma_\varepsilon^2\left[2a_0\mathbf{x}_1^T\left(\mathbf{I} - \frac{\mathbf{x}_2\mathbf{x}_2^T}{\|\mathbf{x}_2\|^2}\right)\right]\left[2a_0\mathbf{x}_1^T\left(\mathbf{I} - \frac{\mathbf{x}_2\mathbf{x}_2^T}{\|\mathbf{x}_2\|^2}\right)\right]^T \\ &= 4a_0^2\sigma_\varepsilon^2\mathbf{x}_1^T\left(\mathbf{I} - \frac{\mathbf{x}_2\mathbf{x}_2^T}{\|\mathbf{x}_2\|^2}\right)\left(\mathbf{I} - \frac{\mathbf{x}_2\mathbf{x}_2^T}{\|\mathbf{x}_2\|^2}\right)^T\mathbf{x}_1 \\ &= 4a_0^2\sigma_\varepsilon^2\mathbf{x}_1^T\left(\mathbf{I} - \frac{2\mathbf{x}_2\mathbf{x}_2^T}{\|\mathbf{x}_2\|^2} + \frac{\mathbf{x}_2\mathbf{x}_2^T\mathbf{x}_2\mathbf{x}_2^T}{\|\mathbf{x}_2\|^4}\right)\mathbf{x}_1 \\ &= 4a_0^2\sigma_\varepsilon^2\left(\mathbf{x}_1^T\mathbf{x}_1 - \frac{2\mathbf{x}_1^T\mathbf{x}_2\mathbf{x}_2^T\mathbf{x}_1}{\|\mathbf{x}_2\|^2} + \frac{\mathbf{x}_1^T\mathbf{x}_2\mathbf{x}_2^T\mathbf{x}_2\mathbf{x}_2^T\mathbf{x}_1}{\|\mathbf{x}_2\|^4}\right) \\ &= 4a_0^2\sigma_\varepsilon^2\left(\|\mathbf{x}_1\|^2 - \frac{2(\mathbf{x}_1^T\mathbf{x}_2)^2}{\|\mathbf{x}_2\|^2} + \frac{\mathbf{x}_1^T\mathbf{x}_2\|\mathbf{x}_2\|^2\mathbf{x}_2^T\mathbf{x}_1}{\|\mathbf{x}_2\|^4}\right) \\ &= 4a_0^2\sigma_\varepsilon^2\left(\|\mathbf{x}_1\|^2 - \frac{2(\mathbf{x}_1^T\mathbf{x}_2)^2}{\|\mathbf{x}_2\|^2} + \frac{(\mathbf{x}_1^T\mathbf{x}_2)^2}{\|\mathbf{x}_2\|^2}\right) \\ &= 4a_0^2\sigma_\varepsilon^2\frac{\|\mathbf{x}_1\|^2\|\mathbf{x}_2\|^2 - (\mathbf{x}_1^T\mathbf{x}_2)^2}{\|\mathbf{x}_2\|^2}. \end{aligned} \quad (50)$$

Using Eq. (46) and Eq. (48), the probability of localization error in Eq. (45) can be rewritten as

$$\begin{aligned} P_e &= P\{\eta < \xi\} = \int_{-\infty}^{\xi}\frac{1}{\sqrt{2\pi}\sigma}e^{-\left(\frac{\eta^2}{2\sigma^2}\right)}d\eta = \Phi\left(\frac{\xi}{\sigma}\right) \\ &= \Phi\left(\frac{a_0^2\left((\mathbf{x}_1^T\mathbf{x}_2)^2 - \|\mathbf{x}_1\|^2\|\mathbf{x}_2\|^2/\|\mathbf{x}_2\|^2\right)}{\sqrt{4a_0^2\sigma_\varepsilon^2\left(\|\mathbf{x}_1\|^2\|\mathbf{x}_2\|^2 - (\mathbf{x}_1^T\mathbf{x}_2)^2/\|\mathbf{x}_2\|^2\right)}}\right) \\ &= \Phi\left(-\frac{a_0}{2\sigma_\varepsilon}\sqrt{\frac{\|\mathbf{x}_1\|^2\|\mathbf{x}_2\|^2 - (\mathbf{x}_1^T\mathbf{x}_2)^2}{\|\mathbf{x}_2\|^2}}\right), \end{aligned} \quad (51)$$

where $\Phi(\cdot)$ is the standard normal distribution function.

Assume that the actual nearest neighbor of the online data $\mathbf{y}$ in the radio map $\mathbf{X}$ is $\mathbf{x}_2$, when the localization system wrongly detect the fingerprint $\mathbf{x}_1$ is the nearest neighbor, the probability of error detection is shown in Eq. (51). Generally, there are n fingerprints in the radio map $\mathbf{X}$. Assume that the fingerprint $\mathbf{x}_i$ in the radio map $\mathbf{X}$ is the nearest neighbor of the online data $\mathbf{y}$. Then, an error occurs when any other fingerprint $\mathbf{x}_j$ in the radio map $\mathbf{X}$ is chosen as the nearest neighbor of $\mathbf{y}$. Therefore, the probability of error detection is

$$P_e(i,j) = \Phi\left(-\frac{a_i}{2\sigma_\varepsilon}\sqrt{\frac{\|\mathbf{x}_i\|^2\|\mathbf{x}_j\|^2 - (\mathbf{x}_i^T\mathbf{x}_j)^2}{\|\mathbf{x}_j\|^2}}\right), i, j = 1, 2, \cdots, n, i \ne j. \quad (52)$$

Each of the standard normal distribution function Φ represents the probability of wrongly detecting the fingerprint $\mathbf{x}_j$ as the nearest neighbor of the online data $\mathbf{y}$ when the actual nearest neighbor is the fingerprint $\mathbf{x}_i$. In the radio map $\mathbf{X}$, if

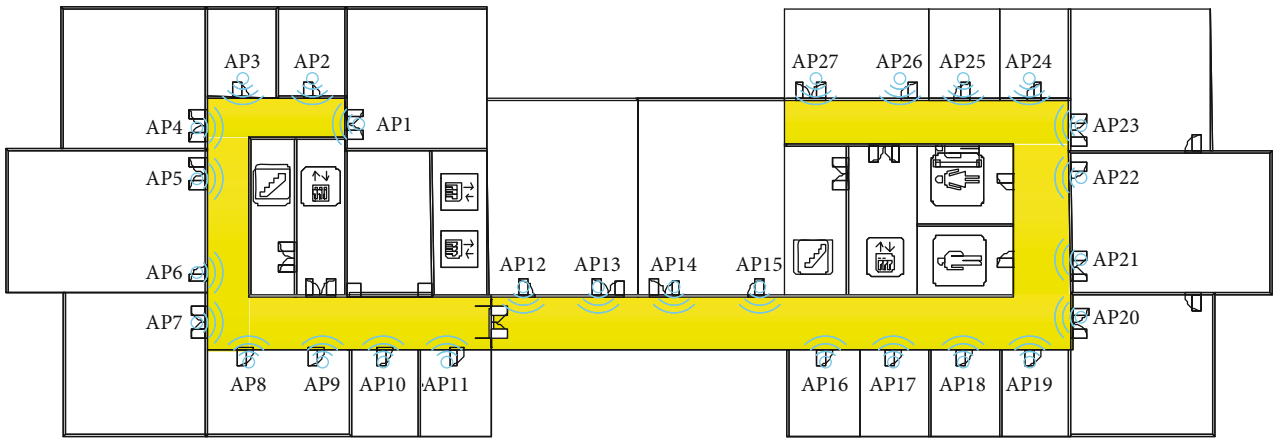

Figure 8: Floor plan for indoor localization, where the area colored in yellow is used for testing.

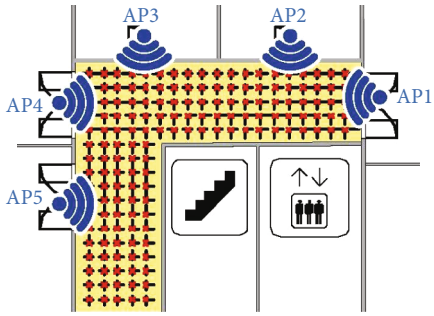

Figure 9: Illustration of the reference point distribution in the interesting area for localization.

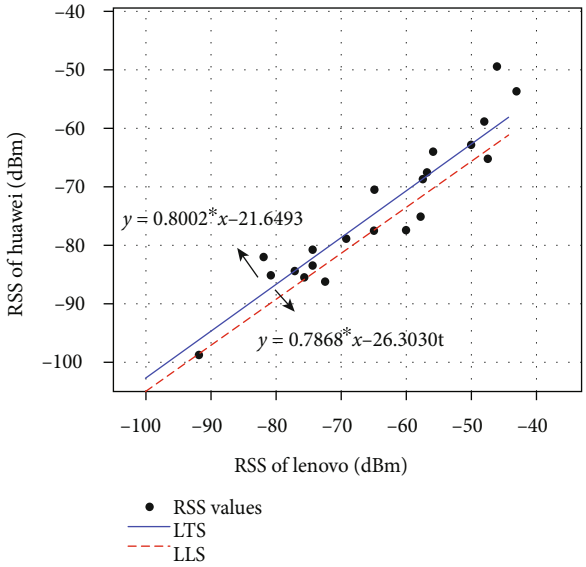

Figure 10: Linear correlation between Huawei and Lenovo.

the fingerprint is closer to the online data $\mathbf{y}$, it is determined as the nearest neighbor by the localization system with higher probability; so, when an error occurs, the probability of error detection is larger. This is because the fingerprints in the nearest neighbor set $\mathbb{A}$ calculated by Eq. (23) have a high probability to be the nearest neighbor; so, these fingerprints in $\mathbb{A}$ contribute most of the probability of error detection in Eq. (52) and should be dealt with carefully. Therefore, it is possible to obtain a more accurate nearest neighbor set by using the preprocess of RSS values, so as to realize the linear regression of RSS values more precisely.

6. Experimental Results and Analysis

The effectiveness of the proposed LR method is studied and analyzed through experiments and simulations in this section. Figure 8 shows the indoor localization experiment system. The localization area is the corridor with 49.4 m in length and 14.1 m in width, which is illustrated with yellow color. In the offline phase, we deployed 27 access points (Linksys WRT54G) with IEEE 802.11b/g mode. As we can see from Figure 9, in the indoor positioning system, the larger the interval between two adjacent reference points in radio map, the lower the final positioning accuracy, but the interval cannot be too small; therefore, the corridor is divided into several grids of 0.5 m × 0.5 m, which means the interval between any two adjacent reference points is 0.5 m, and 823 reference points are set in the corridor. At each RP, 100 RSS values are collected from all APs for each orientation by using a Lenovo laptop, so that 400 samples of RSS reading per RP are collected as the raw RSS readings. Then, the radio map is built based on the RSS values and their corresponding coordinates. In the online phase, four distinct test devices, namely, Huawei, Samsung, Xiaomi, and Coolpad are used to estimate the user's location and are verified with the proposed LR method. At the same time, the Lenovo laptop is used to obtain the optimal localization result for comparison. In our experiment, we take the RSS values in one direction to test the proposed LR method.

To verify the LR method, several RSS values are collected by the test devices and are used to find the candidate fingerprints in the radio map at the beginning. Based on the candidate fingerprints and the measured data, the linear regression coefficients are calculated. Then, the signal space of the radio map and the online RSS values can be mapped to the same signal space, and we can obtain the accurate localization result.

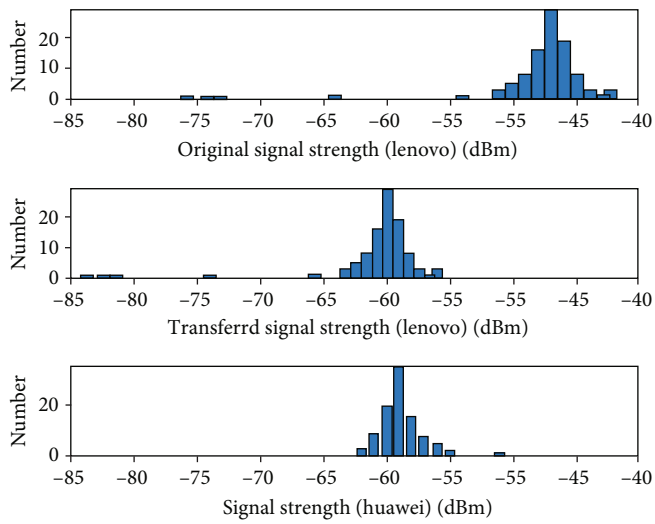

Figure 11: Comparison of signal distributions.

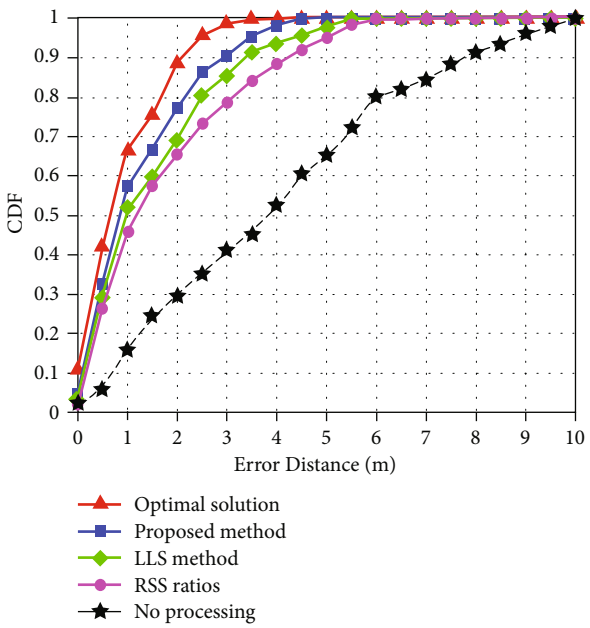

Figure 12: Comparison of the proposed algorithm with other algorithms.

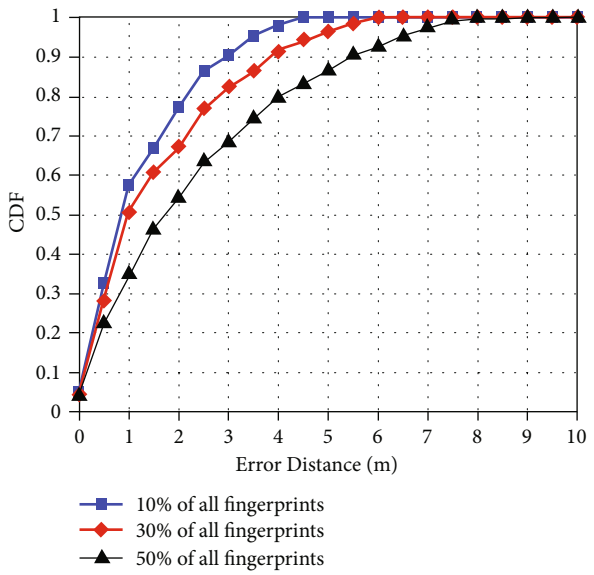

Figure 13: Localization performance for different numbers of candidate nearest neighbors.

Next, we take (Lenovo, Huawei) pair as an example to illustrate the effectiveness of the LR algorithm. As a comparison, the linear regression coefficients are calculated by the LTS algorithm and LLS algorithm, and the linear regression functions are shown in Figure 10. In Figure 10, compare with the result of the LTS algorithm, the linear regression function calculated by the LLS algorithm is closer to the outliers, which result in a large error. The LLS algorithm is more susceptible to the outliers, and this is due to the fact that the LLS algorithm deals with all the measured RSS values equally without any special treatment of the outliers. After getting the linear regression functions, the signal space of the radio map can be mapped to the signal space of the online RSS data.

In order to demonstrate the ability of linear regression algorithm more intuitively, the comparison of RSS values before and after using LTS algorithm is illustrated in Figure 11. In Figure 11, the distributions of Huawei device and Lenovo device are -62 dBm to -55 dBm and -51 dBm to -41 dBm. Obviously, the minimum and maximum RSS difference between Huawei device and Lenovo device is 4 dBm and 21 dBm, respectively. If the radio map is built by the Lenovo device and the user's location is estimated by Huawei device, the localization accuracy is considerably low. Using the LTS algorithm, the RSS values collected by Lenovo device are transformed, and the signal distributions of Huawei device and Lenovo device are basically the same. As a result, the localization accuracy can be improved significantly.

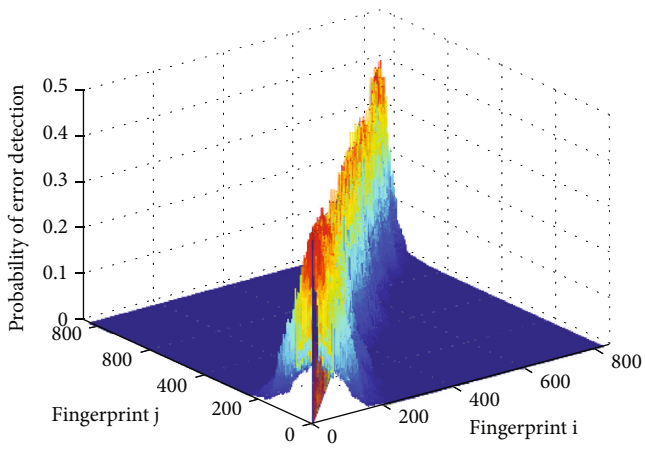

Figure 14: Probability of error detection.

After applying the LTS algorithm, the RSS values in the radio map are transformed, and the KNN algorithm ($K = 3$) is used to estimate the user's current location. As a comparison, the LLS algorithm and the RSS ratios are also used to solve the device diversity problem. The error between the estimate and the truth locations is expressed by Euclidean distances. The CDF curves of the localization error of all algorithms are displayed in Figure 12. When the devices used to build the radio map and estimate the user's current location are the same, we can obtain the optimal solution, as the red line shown in Figure 12. As we can see from the rest caves in Figure 12, if the devices used in the offline phase and the online phase are different, the localization accuracy of the crowdsourcing localization system is greatly reduced. In this paper, we do our best to eliminate the device diversity problem, so that the localization accuracy can be as close as possible to the optimal result. As we can see from Figure 12, the localization accuracy has been improved by applying different algorithms. It is clear that the proposed LTS algorithm outperforms the other methods, and the localization accuracy is closest to the optimal solution. Notably, the maximum localization error has been reduced from 10 m to 4.5 m, and the average error is reduced from 3.72 m to 2.31 m.

In Eq. (24), the correlation coefficient threshold r_{th} is set to choose the nearest neighbor fingerprints in the radio map for the online RSS data. However, if all the correlation ratios calculated by Eq. (23) are less than r_{th}, then $\mathbb{A} = \varnothing$. To be guaranteed $\mathbb{A} \neq \varnothing$, we choose 10% of the fingerprints with the highest r to form the candidate nearest neighbor set. The CDF curves of the localization error using different size of $\mathbb{A}$ are plotted in Figure 13. As shown in Figure 13, since more fingerprints with low probability to be the nearest neighbors are included in $\mathbb{A}$, the localization accuracy decreases as the candidate set size increases.

The probability of error detection in Eq. (52) for all fingerprints in the radio map is shown in Figure 14. In the simulation, since the variance of the noise has no effect on the trend of the probability distribution, we set $\sigma_\varepsilon = 10$. In addition, we assume that all the RSS values collected in the online phase have an ideal nearest neighbors in the radio map; that is $\mathrm{NN}(\mathbf{y}) = \mathbf{x}_i$, so we set $a_i = 1$ for all $i = 1, 2, \cdots, n$. Because $P_e(i, j)$ has no physical meaning when $i = j$, we make $P_e(i, i) = 0$ when drawing Figure 14. From Figure 14, it can be concluded that the probability of error detection of fingerprints closer to the nearest neighbor is higher than others, which means that these fingerprints are more likely to be detected as the nearest neighbors. Thus, the fingerprints in the nearest neighbor set in Eq. (24) contribute the most errors in Eq. (52) and should be chosen more carefully.

7. Conclusions

In this paper, the linear regression (LR) method is proposed to overcome the device diversity problem for the RSS fingerprint-based WLAN indoor localization system using crowdsourced data. The intuition behind this technique is that the RSS values between different devices have a linear relationship. The Pearson correlation coefficient is used to label the RSS values with rough location estimation at the beginning, and the regression coefficients are calculated by the LTS algorithm. Based on the LR algorithm, the RSS values collected by distinct devices can be shifted into the same signal space, and the device diversity problem can be solved. We did a theoretical study of the probability of error detection, and the proposed algorithm is validated through it. Furthermore, we tested the proposed method in a typical office environment, and the experimental results demonstrate that the proposed method leads to significant improvements in localization accuracy.

Data Availability

The radio map data used to support the findings of this study were supplied by Liye Zhang under license and so cannot be made freely available. Requests for access to these data should be made to Liye Zhang (zhangliye@sdut.edu.cn).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Liye Zhang provided the conception. Liye Zhang and Xiaoliang Meng made the analysis and experiment. Liye Zhang, Xiaoliang Meng, and Chao Fang reviewed and edited this paper.

Acknowledgments

This paper is supported by the Shandong Provincial Natural Science Foundation, China (grant number ZR2019BF022) and National Natural Science Foundation of China (grant number 62001272). Reference [20] shows the previous research result, which has been published in IEEE GLOBECOM conference. The previous work is supported by the Associate Professor Lin Ma, Professor Yubin Xu, and Professor Cheng Li. I would like to express my heartfelt thanks to the three professors. Since Associate Professor Lin Ma, Professor Yubin Xu, and Professor Cheng Li have little contribution to this paper, they have agreed not to appear in the author list of this paper.

References

[1] X. Liu, X. B. Zhai, W. Lu, and C. Wu, "QoS-guarantee resource zllocation for multibeam satellite industrial internet of things with NOMA," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 3, pp. 2052–2061, 2021.

[2] X. Liu and X. Zhang, "Rate and energy efficiency improvements for 5G-based IoT with simultaneous transfer," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 5971–5980, 2019.

[3] X. Liu and X. Zhang, "NOMA-based resource allocation for cluster-based cognitive industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5379–5388, 2020.

[4] X. Guo, N. Ansari, L. Li, and L. Duan, "A hybrid positioning system for location-based services: design and implementation," *IEEE Communications Magazine*, vol. 58, no. 5, pp. 90–96, 2020.

[5] D. Zou, W. Meng, S. Han, K. He, and Z. Zhang, "Toward ubiquitous LBS: multi-radio localization and seamless positioning," *IEEE Wireless Communications*, vol. 23, no. 6, pp. 107–113, 2016.

[6] Y. Wang, X. Jia, M. Zhou, L. Xie, and Z. Tian, "A novel F-RCNN based hand gesture detection approach for FMCW sys-tems," *Wireless Networks*, pp. 1–14, 2019.

[7] J. Yu, Z. Na, X. Liu, and Z. Deng, "WiFi/PDR-integrated indoor localization using unconstrained smartphones," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, no. 1, 2019.

[8] M. Li, L. Qu, Q. Zhao, J. Guo, X. Su, and X. Li, "Precise point positioning with the BeiDou navigation satellite system," *Sensors*, vol. 14, no. 1, pp. 927–943, 2014.

[9] F. Gu, X. Hu, M. Ramezani et al., "Indoor localization improved by spatial context—a survey," *ACM Computing Surveys*, vol. 52, no. 3, pp. 1–35, 2019.

[10] P. Bahl and V. Padmanabhan, "Radar: an in-building RF-based user location and tracking system," in *INFOCOM 2000 Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 2, pp. 775–784, Tel Aviv, Israel, 2000.

[11] C. Feng, W. S. A. Au, S. Valaee, and Z. Tan, "Received-signal-strength-based indoor positioning using compressive sensing," *IEEE Transactions on Mobile Computing*, vol. 11, no. 12, pp. 1983–1993, 2012.

[12] B. Wang, Q. Chen, L. T. Yang, and H.-C. Chao, "Indoor smartphone localization via fingerprint crowdsourcing: challenges and approaches," *IEEE Wireless Communications*, vol. 23, no. 3, pp. 82–89, 2016.

[13] C. Yang and H. R. Shao, "WiFi-based indoor positioning," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 150–157, 2015.

[14] L. Zhang, S. Valaee, L. Zhang, Y. Xu, and L. Ma, "Signal propagation-based outlier reduction technique (SPORT) for crowdsourcing in indoor localization using fingerprints," in *Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on, IEEE*, pp. 2008–2013, Hong Kong, China, 2015.

[15] B. Lashkari, J. Rezazadeh, R. Farahbakhsh, and K. Sandrasegaran, "Crowdsourcing and sensing for indoor localization in IoT: a review," *IEEE Sensors Journal*, vol. 19, no. 7, pp. 2408–2434, 2019.

[16] K. Kaemarungsi, "Distribution of wlan received signal strength indication for indoor location determination," in *Wireless Pervasive Computing, 2006 1st International Symposium*, pp. 2952–2957, Phuket, Thailand, 2006.

[17] V. W. Zheng, S. J. Pan, Q. Yang, and J. J. Pan, "Transferring multi-device localization models using latent multi-task learning," in *Proceedings of the 23rd national conference on Artificial intelligence*, pp. 1427–1432, Chicago, IL, USA, 2008.

[18] A. W. Tsui, Y. H. Chuang, and H. H. Chu, "Unsupervised learning for solving RSS hardware variance problem in WiFi localization," *Mobile Networks and Applications*, vol. 14, no. 5, pp. 677–691, 2009.

[19] L. Zhang, S. Valaee, Y. Xu, L. Ma, and F. Vedadi, "Graph-based semi-supervised learning for indoor localization using crowdsourced data," *Applied Sciences*, vol. 7, no. 5, p. 467, 2017.

[20] L. Zhang, L. Ma, Y. Xu, and C. Li, "Linear regression algorithm against device diversity for indoor WLAN localization system," in *2017 IEEE Global Communications Conference*, pp. 1–6, Singapore, 2017.

[21] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach, and L. E. Kavraki, "Practical robust localization over large-scale 802.11 wireless networks," in *Proceedings of the 10th annual international conference on Mobile computing and networking*, pp. 70–84, Philadelphia, PA, USA, September 2004.

[22] J. G. Park, D. Curtis, S. Teller, and J. Ledlie, "Implications of device diversity for organic localization," in *INFOCOM, 2011 Proceedings IEEE*, pp. 3182–3190, Shanghai, China, 2011.

[23] M. B. Kjærgaard and C. Munk, "Hyperbolic location fingerprinting: a calibration-free solution for handling differences in signal strength," in *Pervasive Computing and Communications, 2008 Sixth Annual IEEE International Conference*, pp. 110–116, Hong Kong, China, March 2008.

[24] M. B. Kjærgaard, "Indoor location fingerprinting with heterogeneous clients," *Pervasive and Mobile Computing*, vol. 7, no. 1, pp. 31–43, 2011.

[25] F. Dong, Y. Chen, J. Liu, Q. Ning, and S. Piao, "A calibration-free localization solution for handling signal strength variance," in *International Conference on Mobile Entity Localization and Tracking in Gps-Less Environments (MELT), Springer-Verlag*, pp. 79–90, Orlando, USA, 2009.

[26] A. Mahtab Hossain, Y. Jin, W.-S. Soh, and H. N. Van, "SSD: a robust RF location fingerprint addressing mobile devices' heterogeneity," *IEEE Transactions on Mobile Computing*, vol. 12, no. 1, pp. 65–77, 2013.

[27] C. Laoudias, P. Kolios, and C. Panayiotou, "Differential signal strength fingerprinting Revisited," in *2014 International Conference on Indoor Positioning and Indoor Navigation*, pp. 30–37, Busan, Korea, 2014.

[28] C. Cai, L. Deng, and S. Li, "CSI-based device-free indoor localization using convolutional neural networks," in *2018 IEEE 4th International Conference on Computer and Communications (ICCC)*, pp. 753–757, Chengdu, China, 2018.

[29] I. Ashraf, M. Kang, S. Hur, and Y. Park, "MINLOC:Magnetic field patterns-based indoor localization using convolutional neural networks," *IEEE Access*, vol. 8, pp. 66213–66227, 2020.

[30] S. J. Sadiq and S. Valaee, "Automatic device-transparent RSS-based indoor localization," in *IEEE Global Communications Conference*, San Diego, CA, USA, 2015.

[31] Y. Li, S. Williams, B. Moran, and A. Kealy, "A probabilistic indoor localization system for heterogeneous devices," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6822–6832, 2019.

[32] Y. Ye and B. Wang, "RMapCS: radio map construction from crowdsourced samples for indoor localization," *IEEE Access*, vol. 6, pp. 24224–24238, 2018.

[33] P. J. Rousseeuw and K. V. Driessen, "Computing LTS regression for large data sets," in *Data Mining and Knowledge Discovery*, pp. 29–45, Springer, 2006.