Received July 2, 2017, accepted August 2, 2017, date of publication August 9, 2017, date of current version August 29, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2737651

Achieving Cost-Efficient Indoor Fingerprint Localization on WLAN Platform: A Hypothetical Test Approach

MU ZHOU, (Senior Member, IEEE), YACONG WEI, ZENGSHAN TIAN, XIAOLONG YANG, AND LINGXIA LI

Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Yacong Wei (2540462563@qq.com)

This work was supported in part by the Program for Changjiang Scholars and Innovative Research Team in University under Grant IRT1299, in part by the Special Fund of Chongqing Key Laboratory (CSTC), Fundamental and Frontier Research Project of Chongqing under Grant cstc2017jcyjAX0380 and Grant cstc2015jcyjBX0065, in part by the University Outstanding Achievement Transformation Project of Chongqing under Grant KJZH17117, and in party by the Postgraduate Scientific Research and Innovation Project of Chongqing under Grant CYS17221.

ABSTRACT Received signal strength (RSS) is a typical type of measurements used for indoor fingerprint localization on wireless local area network platform. To make good use of RSS information, we rely on the hypothetical test approach to perform localization with the optimized access points (APs). Specifically, in offline phase, the operating characteristics function is used to minimize the sample capacity of fingerprints at each reference point, and meanwhile the APs are optimally selected based on the concept of information gain criterion. Then, in online phase, the F-test and T-test approaches are used to conduct the RSS variance and mean test, respectively, with the purpose of achieving RPs matching, namely coarse localization. After that, the density-based spatial clustering of applications with noise is developed to realize fine localization with the improved accuracy performance. The extensive experimental results demonstrate that the proposed system is able to avoid the blindness of fingerprints collection as well as improve the effectiveness of fingerprints matching especially under the small sample capacity of fingerprints.

INDEX TERMS Indoor localization, fingerprints matching, hypothetical test, OC function, cost efficiency.

I. INTRODUCTION

Since the performance of Global Positioning System (GPS) degrades significantly in indoor or underground environment, various indoor localization systems have been applied in many kinds of Location-based Services (LBSs), such as parking cars in underground parking lot, navigating passengers in unfamiliar airport, and pushing advertisements to customers in large shopping mall. The imperious demand of system flexibility leads to the diversity of measurements used for indoor localization [1], such as Ultrasonic Wave (UW) [2], [3], Infrared Ray (IR) [4], Ultra Wide Band (UWB) [5], Radio Frequency Identification (RFID) [6], [7], and Wireless Local Area Network (WLAN) [8]–[11]. Among them, the WLAN fingerprint localization is more favored on account of its good environmental adaptability and low infrastructure cost. Specifically, in offline phase, the Received Signal Strength (RSS) at each Reference Point (RP) is collected to construct a fingerprint database, and then in online phase, the newly-collected RSS is matched against the pre-constructed fingerprint database to obtain the optimal location estimate of the target. To achieve this goal, Small *et al.* [12] propose to use the Manhattan distance to measure the similarity between the fingerprints, and the location corresponding to the most similar fingerprint to the new RSS is selected as the estimate of the target. Similarly, Saha *et al.* [13] rely on the Euclidean distance to depict fingerprints similarity. Torres-sospedra *et al.* [14] claim that the Sorensen distance plays better than the Euclidean distance since the former one not only takes the relative distance between the RSS data into account, but also distributes relatively larger weights to the RSS data collected at the locations closer to any AP. Kushki *et al.* [15] use the kernel function to map the fingerprints (with the dimensions equaling to the number of APs) into the vectors in a higher dimensional feature space with the purpose of improving the separability of fingerprints.

However, the above-mentioned studies have not considered the sample capacity of fingerprints and RSS variation property, which will deteriorate the effectiveness and efficiency of fingerprint localization. To deal with this problem, a new cost-efficient indoor localization approach is proposed to optimize the sample capacity as well as AP number and locations. In concrete terms, our system consists of the offline and online phases. In offline phase, the Operating Characteristics (OC) function and information gain criterion are considered to minimize the sample capacity at each RP and find the optimal APs for localization respectively, and then in online phase, after conducting the F-test and T-test on the RSS variance and mean respectively, the Density-based Spatial Clustering of Applications with Noise (DBSCAN) is developed to perform fine localization.

All the experiments are conducted in an actual indoor environment based on WLAN platform to demonstrate the system practicability. The three main contributions of this paper can be summarized as follows.

a) The OC function and information gain criterion are considered to construct a cost-efficient fingerprint database.

b) The DBSCAN is developed for fine localization based on the coarse localization results of conducting the F-test and T-test on the RSS variance and mean respectively.

c) The extensive experiments in an actual indoor WLAN environment demonstrate that our system is capable of providing the effective and efficient location estimate of the target.

The rest of the paper is organized as follows. Section 2 shows some related works. In Section 3, the sample capacity of fingerprints is minimized to construct a cost-efficient fingerprint database. Section 4 illustrates the process of AP optimization based on the concept of information gain criterion. In Section 5, the coarse localization is performed by conducting the F-test and T-test on the RSS variance and mean respectively. Section 6 presents the idea of DBSCAN-based fine localization. The experimental results are provided in Section 7. Finally, we conclude the paper in Section 8.

II. RELATED WORKS

In this section, we collect by rich literatures on the existing indoor localization systems, which can be classified into two representative categories, triangulation- and scene analysis-based localization systems.

A. TRIANGULATION-BASED LOCALIZATION SYSTEMS

Triangulation-based localization systems depend on the geometrical property of signal propagation to perform localization by using the Angle of Arrival (AOA) [16]–[18], Time of Arrival (TOA) [19], and Time Difference of Arrival (TDOA) [20] measurements. Although the angle measurement is featured with high location resolution, it is required to use specially designed antennas to distinguish different signal paths. By using the time measurement for localization, the distance from each AP to the target is calculated by multiplying the propagation speed with travel time. To guarantee the calculation effectiveness, the time synchronization between the APs and target is strictly required, and thereby the system robustness is limited.

B. SCENE ANALYSIS-BASED LOCALIZATION SYSTEMS

Scene analysis-based localization works by using the characteristics of RSS data [21]. As one of the most representative scene analysis-based localization systems, the UW localization system [2] is with high accuracy performance, but it is easily suffered by the multi-path fading. The IR localization system [4] can only be applied to the point-to-point mode since it is vulnerable to the interference of unexpected light. The UWB localization system [5] is difficult to be popularized due to the problems of the short ranging and requirement of extra contact lines. The performance of RFID localization system [6] may deteriorate seriously when the signal is blocked by the metal. The WLAN localization system is based on the fingerprint database construction, which is mostly time consuming and labor intensive.

To solve the overhead problem of fingerprint database construction, many studies attempt to avoid the process of manual fingerprinting like using the propagation models. To achieve this goal, the Log-distance Path Loss (LDPL) model [21], attenuation factor model [22], Motley Keenan (MK) model [23], and RADAR model [24] are employed to construct the fingerprint database. The LDPL model assumes the logarithmic loss of RSS with the distance. The attenuation factor model is featured with strong flexibility under the well-tuned loss coefficient. The MK model takes the RSS attenuation by the walls and floors into account to a great extent. The RADAR model well depicts the RSS fluctuation which is caused by the obstacles between each AP and the target. On the other hand, the signal interpolation approach can also help a lot in reducing the overhead of fingerprint database construction [25], [26]. As an example, the Gaussian Process Latent Variable Model (GPLVM) [26] constructs a topological connectivity graph to estimate the corresponding locations of unlabeled RSS data. The Model-based Radio Interpolation (MRI) [25] infers the RSS based on the propagation model for each virtual AP. The Kriging algorithm [27] enriches the raw fingerprint database by using the linear trend approach to estimate the RSS at unmarked locations. The compressive sensing [28] increases the granularity of fingerprint database to achieve good refinement of fingerprints. The manifold alignment [29] predicts the RSS fluctuation based on the mapping relations between the signal and physical spaces.

III. SAMPLE CAPACITY MINIMIZATION

A. RSS PROPERTY

In target environment, we define the RSS sequence collected at the i-th$(i = 1, \cdots, n)$ RP as $\mathrm{RSS}_i = \{\mathrm{rss}_{i1}, \cdots, \mathrm{rss}_{im}\}$, where $\mathrm{rss}_{ij} = (\mathrm{rss}_{ij1}, \cdots, \mathrm{rss}_{ijr})(j = 1, \cdots, m)$, n and r stand for the number of RPs of APs respectively, and m is the length of RSS sequence. According to the central limit theory of

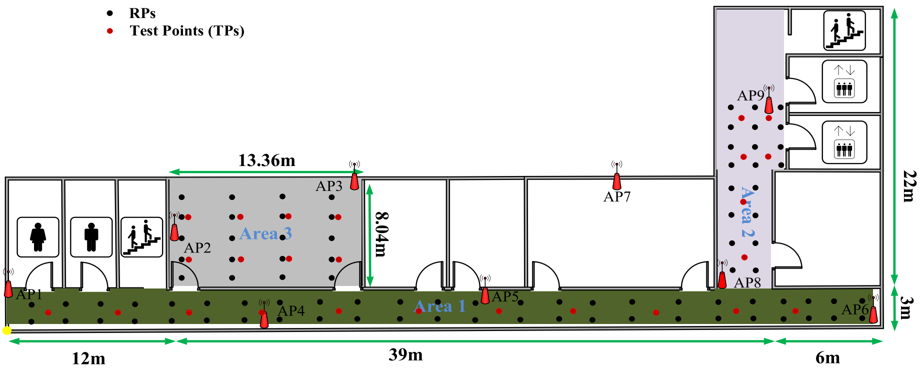
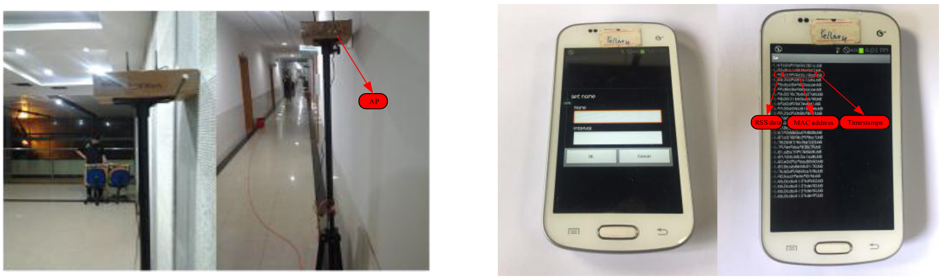


FIGURE 1. Experimental layout.



(a)

(b)

FIGURE 2. Hardware on WLAN platform. (a) Deployed APs. (b) Developed software.

great numeral [30], the distribution of random variable tends to be the normal one when the number of experiments is sufficiently large. Based on this, we set the mean and variance of RSS data from the l-th$(l = 1, \cdots, r)$ AP, $\text{rss}_{i1l}, \cdots, \text{rss}_{iml}$, as $E(\text{rss}_{ijl}) = \mu_1$ and $D(\text{rss}_{ijl}) = \sigma^2$ respectively, and then the corresponding standardization of the sum of RSS data can be written as

$$Y_m = \frac{\sum_{j=1}^{m} rss_{ijl} - E(\sum_{j=1}^{m} rss_{ijl})}{\sqrt{D(\sum_{j=1}^{m} rss_{ijl})}} = \frac{\sum_{j=1}^{m} rss_{ijl} - m\mu_1}{\sqrt{m}\sigma} \tag{1}$$

From (1), the distribution function of Y_m, $F_m(x)$, satisfies

$$\lim_{m\to\infty} F_m(x) = \lim_{m\to\infty} P\{Y_m \leq x\} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\frac{-t^2}{2}} dt = \phi(x) \tag{2}$$

The equation above indicates that when the value m is sufficiently large, $F_m(x)$ will tend to the normal distribution. Since the mean of RSS data from the l-th AP equals to $\bar{X} = \sum_{j=1}^{m} \text{rss}_{ijl}/m$, the distribution of $\bar{X}$ approaches the normal one with the mean μ_1 and variance σ^2/m.

B. HYPOTHETICAL TEST

Setting μ, μ_1, and $\bar{X}$ as the idealistic, population, and sample mean of RSS data at a given RP, we define the error acceptance range as $|\mu_1 - \mu| \leq \sigma$. Based on this, we construct the hypothesis test model as

$$H_0 \colon |\mu_1 - \mu| \leq \delta \quad H_1 \colon |\mu_1 - \mu| > \delta \tag{3}$$

where δ (> 0) is a given threshold. Then, the OC function is defined as

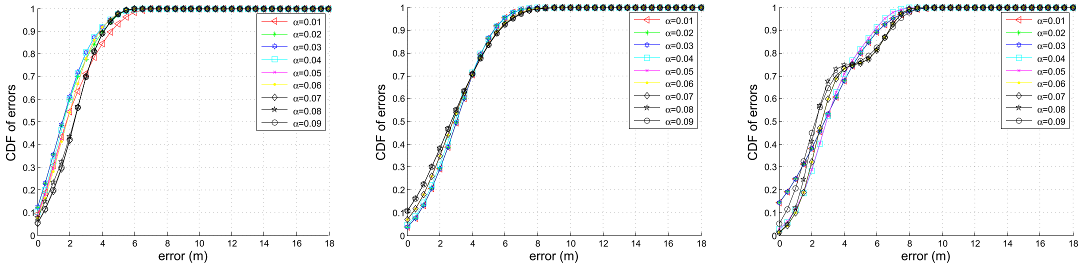
$$\begin{aligned} \beta(\mu_1) &= P_{\mu_1}(\text{accept } H_0) \\ &= P_{\mu_1}\left\{-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma}/\sqrt{m} < Z_{\alpha/2}\right\} \\ &= P_{\mu_1}\left\{-\lambda - Z_{\alpha/2} < \frac{\bar{X} - \mu_1}{\sigma}/\sqrt{m} < -\lambda + Z_{\alpha/2}\right\} \\ &= \phi(Z_{\alpha/2} - \lambda) - \phi(-Z_{\alpha/2} - \lambda) \\ &= \phi(Z_{\alpha/2} - \lambda) + \phi(Z_{\alpha/2} + \lambda) - 1 \end{aligned} \tag{4}$$

where $\lambda = \sqrt{m}(\mu_1 - \mu)/\sigma$.

When the truth is $|\mu_1 - \mu| \leq \delta$ but the wrong decision $|\mu_1 - \mu| > \delta$ is made by the hypothesis test, the mistake of discarding the truth will occur, notated as $P\{H_0$ is true, but reject $H_0\}$. To control this mistake to a certain extent, we require the probability of making such wrong decision not over a given threshold α ($\in (0, 0.1)$) [31].

We assume that $\beta(\mu_1) = P_{\mu_1}(\text{accept } H_0) \leq \beta_1$, and then

$$\beta_1 = \phi(Z_{\alpha/2} - \sqrt{m}\delta/\sigma) + \phi(Z_{\alpha/2} + \sqrt{m}\delta/\sigma) - 1 \tag{5}$$



(a)

(b)

(c)

FIGURE 3. Heat maps of RSS with different sample capacity. (a) In area 1. (b) In area 2. (c) In area 3.

By setting β_1 as 0.01, which is also used in [32], the value $\phi\,(x)$ will approach 1 when $x > 3.5$. In addition, the sufficiently large value m leads to $\phi\,(Z_{\alpha/2} + \sqrt{m}\delta/\sigma) \to 1$, which indicates that $\beta\,(\mu_1) \approx \phi\,(Z_{\alpha/2} - \lambda)$. Since $\beta\,(\mu_1)$ monotonously decreases with λ, we obtain $\beta\,(\mu_1) \le \beta_1 \approx \phi\,(Z_{\alpha/2} - \sqrt{m}\delta/\sigma)$ when $\lambda = \sqrt{m}\,(\mu_1 - \mu)/\sigma \ge \sqrt{m}\delta/\sigma$. To summarize, under the conditions of $Z_{\alpha/2} - \sqrt{m}\delta/\sigma \le -Z_{\beta_1}$ and $\sqrt{m} \ge (Z_{\alpha/2} + Z_{\beta_1})\sigma/\delta$, we can calculate $m \ge 21.6$.

IV. AP OPTIMIZATION

Considering the propagation property of the signal and geometrical structure in target environment, we use the concept of information gain criterion [33] to perform AP optimization. Specifically, we label each RP with r features, where the l-th$(l = 1, \cdots, r)$ feature is the mean of RSS data from the l-th AP. Then, according to the location resolution of each AP, we select the APs corresponding to the highest location resolution as the ones for localization. In concrete terms, we first calculate the desired information with respect to the target environment as

$$H(V) = -\sum_{j=1}^{n} p_j \log_2(p_j) \tag{6}$$

where p_i is the prior probability of the i-th RP, which equals to $1/n$ when the target is assumed to be equally likely to be located at each RP.

Second, for each AP, we divide the RPs into v categories, notated as $V_{l1}, \cdots, V_{lv}$, where V_{li} $(i = 1, \cdots, v)$ is the i-th set of RPs with the same RSS for the l-th AP, and then calculate the corresponding conditional entropy as

$$H(V/AP_l) = \sum_{i=1}^{v} \frac{H(V_{li})card\,(V_{li})}{n} \tag{7}$$

where $H(V_{li}) = -\sum_{j=1}^{card\,(V_{li})} p_{lij}\log_2(p_{lij})$ and $card\,(V_{li})$ represents the number of RPs in V_{li}.

Third, we define the information gain with respect to each AP as

$$\text{Infogain}(\text{AP}_l) = H(V) - H(V/AP_l) \tag{8}$$

Finally, the larger information gain indicates the higher location resolution of the corresponding AP. Based on this, the first w APs with the largest information gain are selected.

V. COARSE LOCALIZATION

To perform coarse localization, we conduct the F-test and T-test on the RSS variance and mean respectively. The purpose of F-test is to examine the difference of RSS variance at two different locations under the situation that their population mean and variance of RSS data are unknown. For the locations with the same RSS variance, we continue to conduct the T-test on the RSS mean, and then the RPs with the same RSS mean to the target are selected as the matched RPs.

A. F-TEST

1) TEST MODEL

By setting m_1 and m_2 as the sample capacity of RSS data at two different locations, their RSS mean and unbiased estimate of RSS variance can be calculated as $\bar{X} = \sum_{j=1}^{m_1} \text{rss}_{ijl}/m_1$ and $\bar{Y} = \sum_{j=1}^{m_2} rss_{i'jl}/m_2$ and $S_1^2 = \sum_{j=1}^{m_1} (\text{rss}_{ijl} - \bar{X})^2/(m_1 - 1)$ and $S_2^2 = \sum_{j=1}^{m_2} (\text{rss}_{i'jl} - \bar{Y})^2/(m_2 - 1)$. Then, the two-side hypothesis test model is constructed as

$$H_0: \sigma_1^2 = \sigma_2^2 \quad H_1: \sigma_1^2 \ne \sigma_2^2 \tag{9}$$

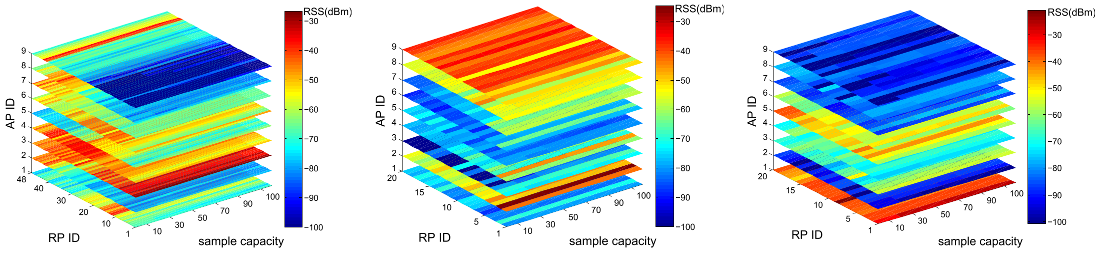
2) F-TEST STATISTIC

Based on the property of χ^2 distribution, we have

$$\begin{cases} \dfrac{(m_1 - 1){S_1}^2}{\sigma_1^2} \sim \chi^2(m_1 - 1) \\ \dfrac{(m_2 - 1){S_2}^2}{\sigma_2^2} \sim \chi^2(m_2 - 1) \end{cases} \tag{10}$$

Using the sample variance of RSS data to test the corresponding population one, we construct the F-test statistic as

$$F = \frac{\frac{(m_1-1){S_1}^2}{\sigma_1^2}/(m_1 - 1)}{\frac{(m_2-1){S_2}^2}{\sigma_2^2}/(m_2 - 1)} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} \tag{11}$$

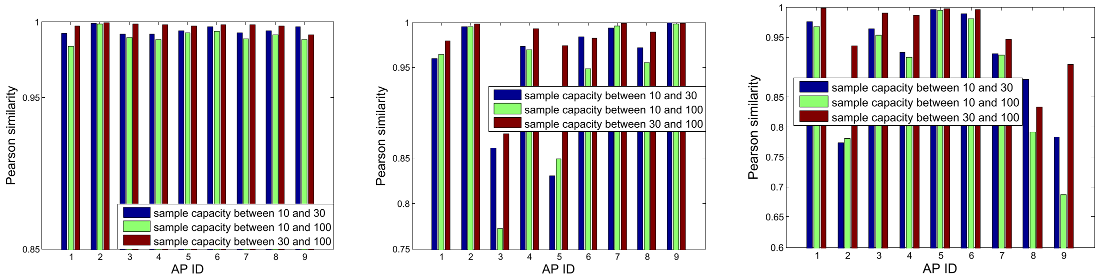


(a)

(b)

(c)

FIGURE 4. Pearson similarity of RSS distributions between different sample capacity. (a) In area 1. (b) In area 2. (c) In area 3.



(a)

(b)

(c)

FIGURE 5. CDFs of errors with different sample capacity.

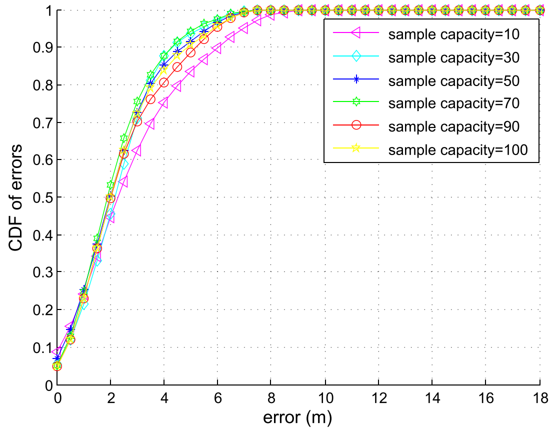


FIGURE 6. CDFs of errors by different number of APs.

Under the condition of $\sigma_1^2 = \sigma_2^2$, which indicates that these two RSS data sets are with the same population variance, we can obtain that the F-test statistic obeys the F-distribution.

$$F = \frac{S_1^2}{S_2^2} \sim F(m_1 - 1,\ m_2 - 1) \tag{12}$$

3) REJECT REGION

Based on (12), the reject region is constructed as

$$F = \frac{S_1^2}{S_2^2} \geq k_1 \cup F = \frac{S_1^2}{S_2^2} \leq k_2 \tag{13}$$

where the values of k_1 and k_2 are determined by the probability of making the wrong decision of rejecting the truth, such that

$$P\{H_0 \text{ is true, but reject } H_0\} = P\left\{\frac{S_1^2}{S_2^2} \geq k_1 \cup \frac{S_1^2}{S_2^2} \leq k_2\right\} \tag{14}$$

By requiring that the probability of making the wrong decision in (14), namely "reject the truth" mistake, is not over α, the values of k_1 and k_2 can be calculated by

$$\begin{cases} k_1 = F_{\alpha/2}(m_1 - 1,\ m_2 - 1) \\ k_2 = F_{1-\alpha/2}(m_1 - 1,\ m_2 - 1) \end{cases} \tag{15}$$

where $F_{\alpha/2}$ and $F_{1-\alpha/2}$ stand for the $\alpha/2$ and $1 - \alpha/2$ percentiles in F-distribution. Thus, the reject region is finally obtained as

$$\frac{S_1^2}{S_2^2} \geq F_{\alpha/2}(m_1-1, m_2-1) \cup \frac{S_1^2}{S_2^2} \leq F_{1-\alpha/2}(m_1 - 1, m_2 - 1) \tag{16}$$

Based on (16), the RPs with the same RSS variance to the target are obtained as $U = \{U_1, \cdots, U_r\}$, where U_l is the set of RPs without falling into the reject region for the l-th AP.

B. T-TEST

1) TEST MODEL

We continue to construct the test model for T-test as

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 \neq \mu_2 \tag{17}$$

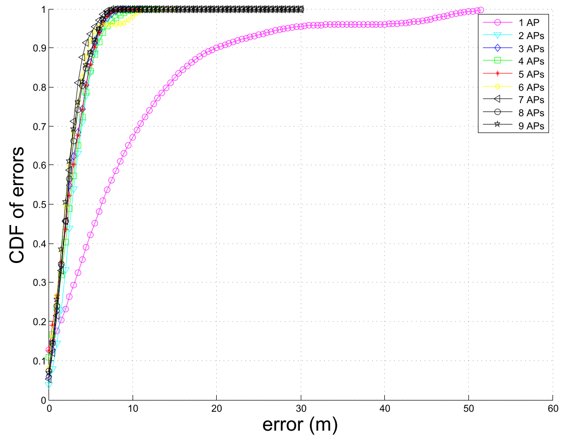


FIGURE 7. Mean of errors by different number of APs.

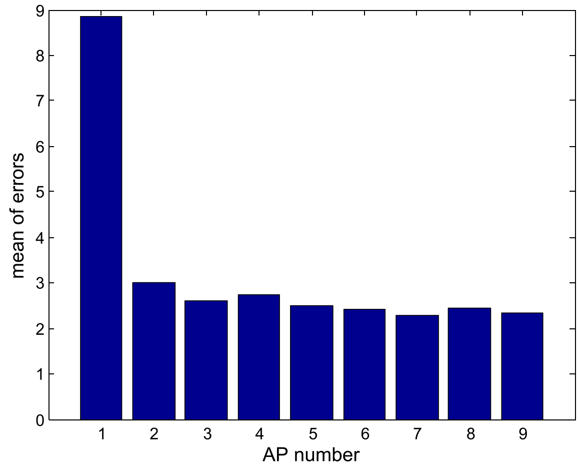


FIGURE 8. CDFs of errors with different value α. (a) In area 1. (b) In area 2. (c) In area 3.

2) T-TEST STATISTIC

After the F-test, we obtain $\sigma_1{}^2 = \sigma_2{}^2 = \sigma^2$ for the RPs belonging to U, and then $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2,\ \frac{\sigma^2}{m_1} + \frac{\sigma^2}{m_2})$. According to the transformation property of normal distribution, we further obtain

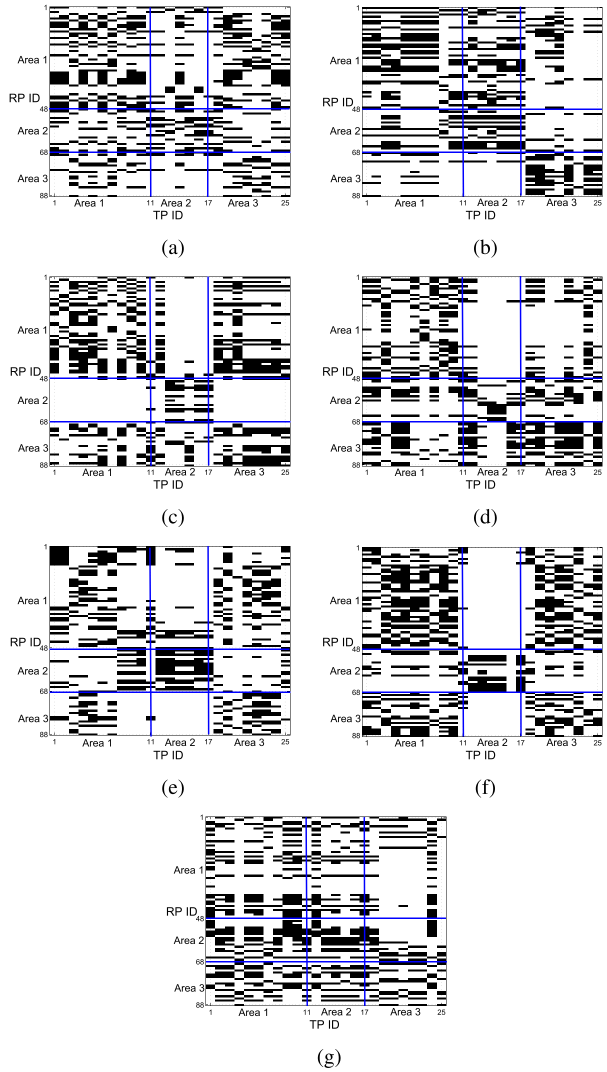
$$U = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}} \sim N(0,\ 1) \tag{18}$$

Based on the property of χ^2 distribution, we have

$$\begin{cases} \dfrac{(m_1 - 1){S_1}^2}{\sigma^2} \sim \chi^2(m_1 - 1) \\ \dfrac{(m_2 - 1){S_2}^2}{\sigma^2} \sim \chi^2(m_2 - 1), \end{cases} \tag{19}$$

and then

$$V = \frac{(m_1 - 1){S_1}^2}{\sigma^2} + \frac{(m_2 - 1){S_2}^2}{\sigma^2} \sim \chi^2(m_1 + m_2 - 2) \tag{20}$$



(a)

(b)

(c)

(d)

(e)

(f)

(g)

FIGURE 9. Result of F-test by different APs. (a) For AP5. (b) For AP8. (c) For AP7. (d) For AP3. (e) For AP1. (f) For AP4. (g) For AP6.

According to the definition of t-distribution, one has

$$\frac{U}{\sqrt{V/m_1+m_2-2}} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{S_w\sqrt{\frac{1}{m_1}+\frac{1}{m_2}}} \sim t(m_1+m_2-2) \tag{21}$$

where $S_w = \sqrt{\frac{(m_1-1)S_1^2+(m_2-1)S_2^2}{m_1+m_2-2}}$. Under the condition of $\mu_1 = \mu_2$, which means these two RSS data sets are with the same population mean, the T-test statistic can be expressed as

$$t = \frac{(\bar{X} - \bar{Y})}{S_w\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}} \tag{22}$$

3) REJECT REGION

Based on (22), the reject region is constructed as

$$|t| = \left| \frac{(\bar{X} - \bar{Y})}{S_w\sqrt{\frac{1}{m_1} + \frac{1}{m_2}}} \right| \geq k_3 > 0 \tag{23}$$

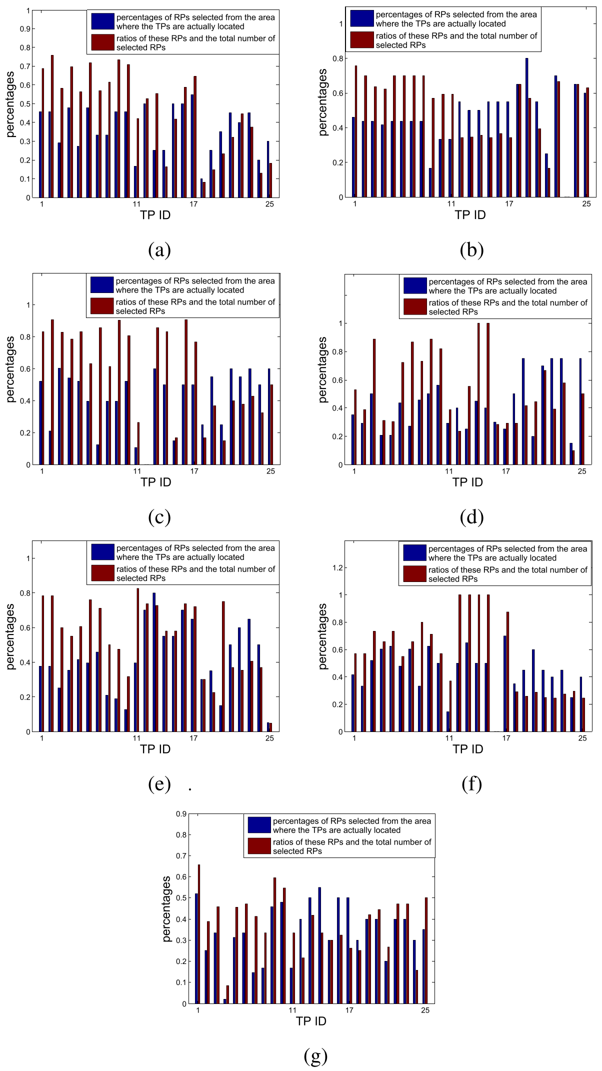


FIGURE 10. Result of RPs selection. (a) For AP5. (b) For AP8. (c) For AP7. (d) For AP3. (e) For AP1. (f) For AP4. (g) For AP6.

where the value k_3 is determined by

$$P\left\{H_0 \text{ is true, but reject } H_0\right\} = P\left\{\left|\frac{\bar{X}-\bar{Y}}{S_w\sqrt{\frac{1}{m_1}+\frac{1}{m_2}}}\right| \geq k_3\right\} \tag{24}$$

Similar to the F-test, we require the probability of making the ‘‘reject the truth’’ mistake in (24) not over α, and then the value k_3 can be calculated by

$$k_3 = t_{\alpha/2}\left(m_1+m_2-2\right) \tag{25}$$

where $t_{\alpha/2}\left(m_1+m_2-2\right)$ is the $\alpha/2$ percentile in t-distribution. Thus, the reject region is finally obtained as

$$|t| = \left|\frac{\left(\bar{X}-\bar{Y}\right)}{S_w\sqrt{\frac{1}{m_1}+\frac{1}{m_2}}}\right| \geq t_{\alpha/2}\left(m_1+m_2-2\right) \tag{26}$$

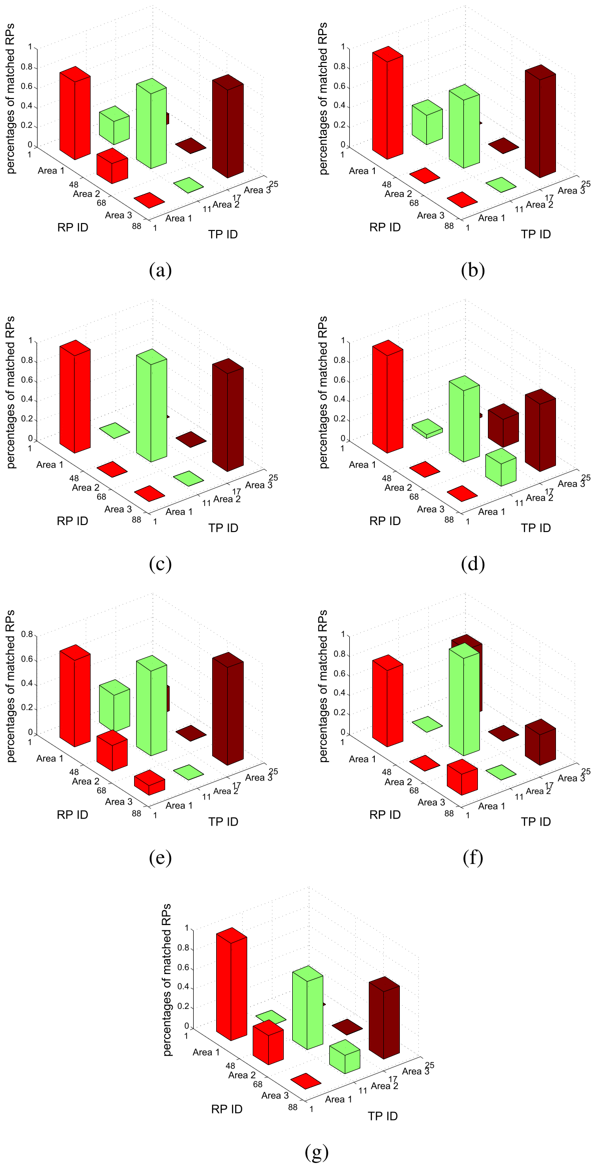


FIGURE 11. Result of T-test by different APs. (a) For AP5. (b) For AP8. (c) For AP7. (d) For AP3.(e) For AP1. (f) For AP4. (g) For AP6.

4) RPs MATCHING

Based on (26), the result of coarse localization can be described as

$$\mathrm{B}_{\mathrm{AP}_l} = \begin{cases} 0, & \text{accept } H_0 \\ 1, & \text{reject } H_0 \end{cases} \tag{27}$$

In (27), $\mathrm{B}_{\mathrm{AP}_l} = 0$ indicates that these two RSS data sets are much likely to be collected at neighboring RPs or even at the same one due to their same population mean, but on the contrary, they are probably corresponding to the far-away RPs. For each localization query, a subset of RPs, G_l ($\subset U_l$), is selected as the set of matched RPs for the l-th AP.

VI. FINE LOCALIZATION

In our system, we achieve fine localization by conducting the DBSCAN [34] on the matched RPs. The pseudo-code of this process is shown below.

Algorithm 1 Pseudo-Code of DBSCAN for Fine Localization

```
Input: $G = \{G_1, \cdots, G_r\}$, adjacent radius $\varepsilon$, and adjacent density threshold $Minpts$
Output: Estimated location of the target
All the RPs in $G$ are labeled as the unvisited ones;
Initialize the cluster ID as $o = 1$;
do
Randomly select an unvisited RP $p (\in G)$ as the visited one;
if Number of RPs in the $\varepsilon$-domain of $p \geq Minpts$ then
    Set $p$ as a new cluster $M_o = \{p\}$;
    Construct a set $Q = \{$RPs in the $\varepsilon$-domain of $p\}$;
    for each $s \in Q$ do
        if $s$ is unvisited then Label $s$ as the visited one;
            if Number of RPs in the $\varepsilon$-domain of $s \geq Minpts$ then add the RPs in the $\varepsilon$-domain of $s$ into $Q$;
                if $s$ does not belong to any cluster then add $s$ into $M_o$;
                end if
            end if
        end if
    end for
    Label the cluster ID of the RPs in $M_o$ with $o$;
    Count the number of RPs in $M_o$;
    $o := o + 1$;
else
    Label $p$ as an outlier which does not belong to any cluster;
end if
until all the RPs in $G$ have been visited;
Find the cluster containing the largest number of RPs, namely maximum cluster;
Select the geometrical center of the RPs in maximum cluster as the location estimate of the target.
```

VII. EXPERIMENTAL RESULTS

All the experiments are conducted on a floor with the dimensions of 57 m by 25 m, as shown in Fig. 1. Some photos of the deployed APs and developed software for WLAN RSS recording are shown in Fig. 2.

A. RESULT OF PARAMETERS DISCUSSION

1) PARAMETER α

The Cumulative Density Functions (CDFs) of errors with different value α are shown in Fig. 3, from which we can find that when $0.02 \leq \alpha \leq 0.05$, the high localization accuracy can be well preserved. In addition, considering that the over-large value α cannot effectively constrain the probability of making the "reject the truth" mistake, we set $\alpha = 0.02$ in the results that follow.

2) PARAMETER m

Based on the heat maps of RSS with different sample capacity in Fig. 4, Fig. 5 compares the corresponding Pearson similarity [35] of RSS distributions in each area. By taking the AP3 in area 2 as an example, the Pearson similarity of RSS distributions between m = 30 and 100 is 87.68%, but it decreases to 77.21% when the sample capacity for the former RSS distribution is reduced to 10. Considering both the effectiveness and efficiency of fingerprint database construction, we set $m = 30$ in our system.

The CDFs of errors with different sample capacity is shown in Fig. 6, from which we can find that with the increase of sample capacity, the localization accuracy generally improves as expected, while when the sample capacity is over 30, the increase of sample capacity has slight impact on localization performance. For example, we calculate that the mean error decreases from 2.65 m to 2.28 m (by 13.96%) when the sample capacity increases from 10 to 30, while it continues to decrease by only 7.02% when the sample capacity increases from 30 to 70.

B. RESULT OF AP OPTIMIZATION

In target environment (see Fig. 1), the selected 9 APs are sequenced as AP5, AP8, AP7, AP3, AP1, AP4, AP6, AP9, and AP2 in the decreasing order of information gain. Fig. 7 and 8 show the CDFs and mean of errors by using different number of APs for localization. As can be seen from these results, we observe that with the increase of AP number, the localization accuracy generally improves as expected, whereas it shows an decreasing trend when the AP number is over 7, which can be interpreted by the fact that adding the new APs with small information gain may result in the decrease of location resolution. Therefore, we select the first $w = 7$ APs corresponding to the largest information gain for localization.

C. RESULT OF HYPOTHETICAL TEST

1) F-TEST

Fig. 9 shows the result of F-test by using different APs. From this figure, we can find that most of the RPs belonging to the same or adjacent area to the one where the corresponding Test Points (TPs) are located are selected by the F-test. To show this result clearer, Fig. 10 illustrates the percentages of RPs selected from the area where the TPs are actually located and the corresponding ratios of these RPs and the total number of selected RPs for each AP.

2) T-TEST

Fig. 11 shows the result of T-test by using different APs. By taking the AP6 as an example, the percentages of the matched RPs belonging to the same area to the one where the corresponding TPs are located with respect to area 1, 2,

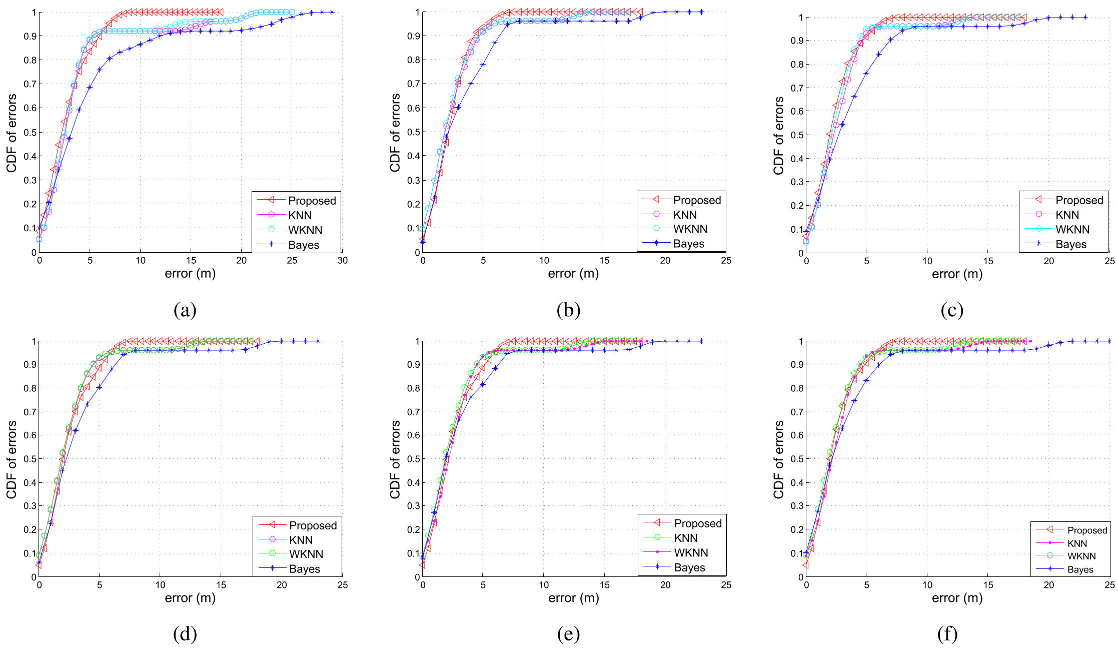


FIGURE 12. Result of fine localization by different localization algorithms. (a) $m = 10$. (b) $m = 30$. (c) $m = 50$. (d) $m = 70$. (e) $m = 90$. $m = 100$.

and 3 are 78.57%, 100%, and 31.25% respectively. The small percentage of matched RPs in area 3 is due to the fact that many RPs in adjacent area 1 are with the same population mean to the ones in area 3.

D. RESULT OF FINE LOCALIZATION

Finally, Fig. 12 compares the CDFs of errors by the proposed and three existing localization algorithms, namely KNN [24], WKNN [24], and Bayes [36]. From this figure, we can find that when the sample capacity is over 30, the localization performance of the proposed one varies slightly with the increase of sample capacity, and meanwhile it performs better in constraining the "tail error" (e.g., the errors over 10 m) than the others. In addition, when the sample capacity is much small (e.g., $m = 10$), the proposed one can well preserve the high localization accuracy.

VIII. CONCLUSION

In this paper, we propose a new cost-efficient indoor WLAN fingerprint localization system by using the hypothetical test approach with the reduced time and labor cost for fingerprint database construction. In addition, the OC function is considered to minimize the sample capacity of fingerprints, and meanwhile the APs are optimized by employing the concept of information gain criterion. After that, the F-test and T-test are used to conduct the RSS variance and mean test respectively with the purpose of realizing fine localization by the DBSCAN. Finally, the experiments demonstrate that the proposed system is featured with high localization accuracy as well as good robustness to sample capacity of fingerprints.

REFERENCES

[1] L. Wang, W. Liu, N. Jing, and X. Mao, "Simultaneous navigation and pathway mapping with participating sensing," *Wireless Netw.*, vol. 21, no. 8, pp. 1–19, 2015.

[2] K.-W. Kim, J. Kwon, C.-L. Lee, and J. Han, "Accurate indoor location tracking exploiting ultrasonic reflections," *IEEE Sensors J.*, vol. 24, no. 16, pp. 9075–9088, Dec. 2016.

[3] J. Yoon and T. Parkn, "Maximizing localization accuracy via self-configurable ultrasonic sensor grouping using genetic approach," *IEEE Trans. Instrum. Meas.*, vol. 7, no. 65, pp. 1518–1529, Jul. 2016.

[4] S. Wu, N. Zhang, H. Zhou, and X. S. Shen, "High precision ranging with IR-UWB: A compressed sensing approach," *Wireless Commun. Mobile Comput.*, vol. 17, no. 16, pp. 3015–3031, Dec. 2016.

[5] R. Maalek and F. Sadeghpour, "Accuracy assessment of ultra-wide band technology in locating dynamic resources in indoor scenarios," *Autom. Construct.*, vol. 63, pp. 12–26, Mar. 2016.

[6] N. Aldin, E. Ercelebi, and M. Aykac, "Advanced boundary virtual reference algorithm for an indoor system using an active RFID interrogator and transponder," *Analog Integr. Circuits Signal Process.*, vol. 3, no. 88, pp. 415–430, 2016.

[7] M. Zhou, Q. Zhang, Z. Tian, Y. Liu, and Z. Zhang, "Simultaneous pathway mapping and behavior understanding with crowdsourced sensing in WLAN environment," *Ad Hoc Netw.*, vol. 58, pp. 160–170, Apr. 2017.

[8] I. Bisio *et al.*, "A trainingless WiFi fingerprint positioning approach over mobile devices," *IEEE Antennas Wireless Propag. Lett.*, vol. 13, pp. 832–835, 2014.

[9] S. H. Fang, T. N. Lin, and K. C. Lee, "A novel algorithm for multipath fingerprinting in indoor WLAN environments," *IEEE Trans. Wireless Commun.*, vol. 7, no. 9, pp. 3579–3588, Sep. 2008.

[10] A. Sciarrone, C. Fiandrino, I. Bisio, F. Lavagetto, D. Kliazovich, and P. Bouvry, "Smart probabilistic fingerprinting for indoor localization over fog computing platforms," in *Proc. IEEE Int. Conf. Cloud Netw.*, Oct. 2016, pp. 39–44.

[11] M. Zhou, Q. Zhang, Y. Wang, and Z. Tian, "Hotspot ranking based indoor mapping and mobility analysis using crowdsourced Wi-Fi signal," *IEEE Access.*, vol. 5, no. 1, pp. 3594–3602, 2017.

[12] J. Small, A. Smailagic, and D. Siewiorek, "Determining user location for context aware computing through the use of a wireless LAN infrastructure," Carnegie Mellon Univ., Tech. Rep., 2000.

[13] A. Saha and P. Sadhukhan, "A novel clustering strategy for fingerprinting-based localization system to reduce the searching time," in *Proc. IEEE Int. Conf. Recent Trends Inf. Syst.*, Jul. 2015, pp. 538–543.
[14] J. Torres-sospedra *et al.*, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat.*, 2014, pp. 261–270.
[15] A. Kushki, K. N. Plataniotis, and A. N. Venetsanopoulos, "Kernel-based positioning in wireless local area networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 6, pp. 689–705, Jun. 2007.
[16] B. Jachimczyk, D. Dziak, and J. Kulesza, "Using the fingerprinting method to customize RTLS based on the AOA ranging technique," *Sensors.*, vol. 6, no. 16, p. 876, 2016.
[17] X. Jie and J. Kyle, "ArrayTrack: A fine-grained indoor location system," in *Proc. Usenix Conf. Netw. Syst. Des. Implement.*, 2013, pp. 71–84.
[18] K. Swarun, G. Stephanie, K. Dina, and D. Rus, "Accurate indoor localization with zero start-up cost," in *Proc. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 483–494.
[19] H. Liu, H. Darabi, P. Banerjee, and J. Liu, "Survey of wireless indoor positioning techniques and systems," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 37, no. 6, pp. 1067–1080, Nov. 2007.
[20] R. Wang, X. Yu, S. Zheng, and Y. Ye, "Design of a TDOA location engine and development of a location system based on chirp spread spectrum," *Springerplus*, vol. 5, no. 1, p. 1963, 2016.
[21] S. Jung, C. Lee, and D. Han, "Wi-Fi fingerprint-based approaches following log-distance path loss model for indoor positioning," in *Proc. Int. Microw. Workshop Ser. Intell. Radio Future Pers. Terminals*, 2011, pp. 1–2.
[22] Y. Shih, Y. Hsu, C. Chen, C.-C. Tseng, and E. Sha, "Adaptive attenuation factor model for localization in wireless sensor networks," *Int. J. Pervasive Comput. Commun.*, vol. 4, no. 3, pp. 257–267, 2008.
[23] K. Kaustav, S. Datta, M. Pal, and R. Ghatak, "Motley Keenan model of in-building coverage analysis of IEEE 802.11 n WLAN signal in electronics and communication engineering department of National Institute of Technology Durgapur," in *Proc. Int. Conf. Microelectron., Comput. Commun.*, 2016, pp. 1–6.
[24] B. Paramvir and P. Venkata, "RADAR: An in-building RF-based user location and tracking system," in *Proc. Int. Conf. Comput. Commun.*, vol. 2. 2000, pp. 775–784.
[25] H. Shin, Y. Chon, Y. Kim, and H. Cha, "MRI: Model-based radio interpolation for indoor war-walking," *IEEE Trans. Mobile Comput.*, vol. 14, no. 6, pp. 1231–1244, Jun. 2015.
[26] B. Ferris, D. Fox, and N. Lawrence, "WiFi-SLAM using Gaussian process latent variable models," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 2480–2485.
[27] S. Jan, S. Yen, and Y. Liu, "Received signal strength database interpolation by Kriging for a Wi-Fi indoor positioning system," *Sensors.*, vol. 15, no. 9, pp. 21377–21393, 2015.
[28] Z. Gu, Z. Chen,Y. Zhang, Y. Zhu, M. Lu, and A. Chen, "Reducing fingerprint collection for indoor localization," *Comput. Commun.*, vol. 83, pp. 56–63, Jun. 2015.
[29] M. Zhou, Y. Tang, Z. Tian, and X. Geng, "Semi-supervised learning for indoor hybrid fingerprint database calibration with low effort," *IEEE Access.*, vol. 5, no. 99, pp. 4388–4400, 2017.
[30] A. de Moivre, *The Doctrine of Chances*, vol. 8, no. 3. New York, NY, USA: Springer, 2001, p. xiv and 816.
[31] K. Tout, R. Cogranne, and F. Retraint, "Fully automatic detection of anomalies on wheels surface using an adaptive accurate model and hypothesis testing theory," in *Proc. Signal Process. Conf.*, 2016, pp. 508–512.
[32] L. Gan "On the control of errors of the second kind in hypothesis test," *Stat. Decision*, vol. 346, no. 22, pp. 35–37, 2011.
[33] Y. Chen, Q. Yang, J. Yin, and X. Chai, "Power-efficient access-point selection for indoor location estimation," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 7, pp. 877–888, Jul. 2006.
[34] R. Kisore and C. Koteswaraiah, "Improving ATM coverage area using density based clustering algorithm and Voronoi diagrams," *Inf. Sci.*, vol. 376, pp. 1–20, Jan. 2016.
[35] R. Wu, J. Wang, and K. Yuan, "Monte Carlo simulation of poly choric correlation and Pearson correlation coefficient," *J. Beijing Univ. Aeronautics Astron.*, vol. 35, no. 12, pp. 1507–1510, 2009.
[36] M. Youssef and A. Agrawala, "The Horus WLAN location determination system," in *Proc. Int. Conf. Mobile Syst., Appl., Services*, 2005, pp. 205–218.

MU ZHOU (SM'17) received the Ph.D. degree in communication and information systems from the Harbin Institute of Technology, China, in 2012. He was a Joint-Cultivated Ph.D. Student with the University of Pittsburgh, USA, and a Post-Doctoral Research Fellow with The Hong Kong University of Science and Technology, Hong Kong. He joined the Chongqing University of Posts and Telecommunications, China, where he has been a Full Professor with the School of Communication and Information Engineering since 2014. His main research areas include wireless localization and navigation, signal reconnaissance and detection, and convex optimization and deep learning. He has served on technical program committees of the IEEE ICC, GLOBECOM, WCNC, IWCMC, VTC, IWCMC, and so on.

YACONG WEI is currently pursuing the M.S. degree with the Chongqing University of Posts and Telecommunications. Her research areas include WLAN indoor localization and statistical analysis theory.

ZENGSHAN TIAN received the Ph.D. degree from the University of Electronic Science and Technology of China. He is currently a Full Professor with the Chongqing University of Posts and Telecommunications. His main research areas include personal communication, precise localization and attitude measure, and data fusion.

XIAOLONG YANG received the Ph.D. degree from the Harbin Institute of Technology. He is currently a Lecturer with the Chongqing University of Posts and Telecommunications. His main research areas include power allocation, cognitive radio networks, and wireless localization.

LINGXIA LI received the M.S. degree from the Chongqing University of Posts and Telecommunications (CQUPT). She is currently an Associate Professor with CQUPT. Her main research areas include future mobile communications and broadband wireless access technologies.