Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Received signal strength based indoor positioning algorithm using advanced clustering and kernel ridge regression*

Yanfen LE†, Hena ZHANG, Weibin SHI, Heng YAO†‡

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

†E-mail: leyanfen@usst.edu.cn; hyao@usst.edu.cn

Received Mar. 3, 2020; Revision accepted Oct. 25, 2020; Crosschecked Dec. 24, 2020; Published online Feb. 5, 2021

Abstract: We propose a novel indoor positioning algorithm based on the received signal strength (RSS) fingerprint. The proposed algorithm can be divided into three steps, an offline phase at which an advanced clustering (AC) strategy is used, an online phase of approximate localization at which cluster matching is used, and an online phase of precise localization with kernel ridge regression. Specifically, after offline fingerprint collection and similarity measurement, we employ an AC strategy based on the *K*-medoids clustering algorithm using additional reference points that are geographically located at the outer cluster boundary to enrich the data of each cluster. During the approximate localization, RSS measurements are compared with the cluster radio maps to determine to which cluster the target most likely belongs. Both the Euclidean distance of the RSSs and the Hamming distance of the coverage vectors between the observations and training records are explored for cluster matching. Then, a kernel-based ridge regression method is used to obtain the ultimate positioning of the target. The performance of the proposed algorithm is evaluated in two typical indoor environments, and compared with those of state-of-the-art algorithms. The experimental results demonstrate the effectiveness and advantages of the proposed algorithm in terms of positioning accuracy and complexity.

Key words: Indoor positioning; Received signal strength (RSS) fingerprint; Kernel ridge regression; Cluster matching; Advanced clustering

https://doi.org/10.1631/FITEE.2000093 **CLC number:** TN92

1 Introduction

Recent advances in information science and wireless network technology have made it practical and accessible to provide indoor positioning services, such as indoor personal navigation (Li LQ et al., 2015), healthcare monitoring (Rodriguez et al., 2004; Honeine et al., 2011), and personalized information delivery (Harroud et al., 2003), to consumers. Achieving a satisfactory positioning accuracy in a complicated indoor environment for these applications has become an attractive research topic (Al Nuaimi and Kamel, 2011; Shi et al., 2018; Kumar and Rajawat, 2019; Zhang et al., 2019).

Localization methods based on the received signal strength (RSS) have been extensively studied in recent years for their advantages of high flexibility, low cost, and no additional hardware (Wu et al., 2016; Chen C et al., 2018; Fang XM et al., 2018). Among these methods, the ones employing the locations of access points (APs) or anchors have been proposed using the propagation model to describe the relationship between RSS and the distance from the receiver to the transmitter. However, it is quite difficult to obtain the exact distance between the target node and APs because of the dynamic and unpredictable nature of radio channels which are troubled by shadowing, multipath, and blocking. Thus, a more feasible method has been developed, comparing the online RSS measurement with a pre-built radio map,

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 51705324 and 61702332)

ORCID: Yanfen LE, https://orcid.org/0000-0001-5792-8676; Heng YAO, https://orcid.org/0000-0002-3784-4157

to estimate the position of the node (Fang SH and Lin, 2012; Xue et al., 2018). This type of method is referred to as fingerprint localization, which usually comprises two phases: offline and online. During the offline phase, a number of particular locations or reference points (RPs) are set throughout the monitoring area. For each RP, RSS measurements from different APs or anchors are gathered and recorded. Note that the locations of APs are not necessarily known at this point. Then, a radio map is constructed that contains RSS measurements for each RP associated with its position. During the online phase, the real-time RSS measurement of the target is collected to estimate its position. In the K-nearest neighbor (KNN) criterion (Bahl and Padmanabhan, 2000), the Euclidean distance between the online RSS measurement and the fingerprints on the radio map is considered for selecting RPs, in which K nearest RPs are selected and their convex hull is counted as the position estimate of the target. It is one of the most convenient and accessible positioning methods. A weighted KNN (WKNN) method (Niu et al., 2015) was developed by assigning a weight to each RP position. The main concept of these methods is as follows: the closer the target to an RP, the higher the similarity between RSSs in two positions of the RP and the closest target. However, during the offline and online phases, RSS measurements are prone to random fluctuations because of the complexity of indoor environments with irregular personal activity, thereby making the RSS readings in each RP insufficiently stable.

To solve the above-mentioned problem, efforts have been made to build a robust and adaptive model that could accurately describe the relationship between the positions and RSSs. Sparse recovery algorithms based on the compressive sensing (CS) theory (Feng et al., 2012; Al-Moukhles et al., 2016) and the least absolute shrinkage and selection operator (LASSO) algorithm (Khalajmehrabadi et al., 2017a) have been applied to model the localization problem. Because the target is located at a specific point in space at each time interval, online measurements are associated with a unique subset of RPs. Based on this theory, the minimization problem, including the residuals between the radio map and online measurements and the weighted L_2-norm of groups of RPs, has been solved by the positioning method (Khalajmehrabadi et al., 2017b). These methods improved the positioning accuracy to some extent; however, the inevitability of a non-linear relationship between RSS distribution and positions of RPs makes it difficult to further improve the performance of these approaches.

Kernel-based methods, which are effective ways to extend linear algorithms to non-linear problems for machine learning (Maalouf and Homouz, 2014), have been used in localization problems in recent years (Mahfouz et al., 2013, 2016; Huang and Manh, 2016). In these methods, non-linear functions, i.e., a linear combination of kernels, are trained with sampled RSSs to minimize the error between the model outputs and the actual ones. Other machine learning techniques have been applied to fingerprinting (Dai et al., 2016; Yan et al., 2018). Neural networks based on deep learning and extreme learning machines have been used to obtain optimal weights by fully exploring the RSS features (Lu et al., 2016; Wang et al., 2017). These methods improved the localization accuracy in complex indoor environments. However, a large amount of training data is required to perform proper training, thus resulting in high computational complexity.

Recently, clustering strategies have been applied to fingerprinting-based methods to reduce the maximum positioning error. During the offline phase, RPs are divided into a number of clusters according to specific features of RSS measurements. During the online phase, approximate localization is followed by precise localization. Different methods are used to decompose the RSS readings. One method of generating clusters is based on the affinity propagation algorithm (Feng et al., 2012; Hu et al., 2018), and generates exemplars and corresponding clusters by recursively transmitting real-valued messages between the pairs of RPs based on the similarity of RSS measurements. The Euclidean distance between the online measurement and individual exemplar's RSS or the weighted average RSS of cluster members is used as the similarity criterion for online cluster matching. However, during the online phase, a set of APs might be lost or weakened because of some unforeseen reasons. This may affect the accuracy of cluster matching. In addition, some AP selection schemes, such as the highest information gain selection (Chen YQ et al., 2006), the strongest set selection

(Youssef et al., 2003), the Fisher criterion (Khalajmehrabadi et al., 2017a), and random selection (Feng et al., 2012), have been applied to approximate localization. Unlike the cluster criterion, the similarity between coverage vectors of APs has been proposed to group RPs (Kushki et al., 2007). This is essentially spatial filtering based on the premise that RPs that are geographically close to each other can receive signals from the same subset of APs. The Hamming distance between the binary AP coverage vectors is used to measure the difference between the RSSs. Usually, the distance between the coverage vectors of an RP and the online RSS is adjusted as a filter threshold to select PRs that pass through the filter for precise localization. However, the efficiency of these clustering methods is based on a necessary condition that AP has a stable distribution of the RSS signals at a certain location, which is almost impossible in practice for the time variation of the signal in the propagation environment. Our experimental results showed that a fluctuation range of 6 dBm can be observed over a sample period of 1 s at a certain RP, which means a distance of 1.4 to 2.0 m of signal propagation in a typical indoor environment (Mahfouz et al., 2016). It must be mentioned that although clustering helps improve indoor positioning accuracy, wrong cluster matching in the approximate localization phase would lead to an unacceptable positioning error during precise localization.

Considering the challenges mentioned above, we use kernel-based ridge regression (KRR) (Saunders et al., 1998; Mahfouz et al., 2016) to define the positioning function with the training fingerprint data collected during the offline phase. Ridge regression conquers the over-fitting and multicollinearity disadvantageous to the least squares method without consideration of the assumptions or prior knowledge of the model. KRR extends the RR method to a nonlinear problem. However, KRR is not sparse and has a time complexity of $O(N^3)$, where N is the number of pieces of training data, and the computation time increases with the increase in the density of the matrices for all the required training data (Maalouf and Homouz, 2014). Thus, the clustering scheme is applied before KRR in our algorithm to reduce the RPs involved in the calculation to a specific set. Considering that the members of a cluster may have similar RSS characteristics but a dispersed geographic distribution, we expand the members of each cluster by adding the RPs located on the outer boundary of the clusters. This process reduces the online computation time of the regression algorithm and improves the positioning accuracy.

Contributions of our work can be summarized in three aspects: First, we introduce an advanced clustering (AC) strategy to solve the problem that RSS signals with similar features may have disparate geographic locations. The location is jointed with the RSS to improve the clustering efficiency via the addition of a small number of cluster members. Second, we propose an online kernel-based positioning model within the scope of a cluster for fast and accurate localization. Third, different clustering metrics and cluster matching schemes are investigated to further improve the positioning accuracy.

2 Indoor positioning algorithm

In a typical indoor positioning scenario, a target that carries the signal receiver obtains RSS measurements from available APs. Note that it is not necessary to know the locations of these APs. The target uses the positioning algorithm to estimate its current position on the map using merely online RSS readings. By comparing the current RSS readings with fingerprints pre-stored in the radio map, the target determines its position relative to the fixed points on the map. In this study, we propose a new algorithm to locate the target by incorporating the techniques of cluster matching and KRR. Specifically, AC is applied to refine the positioning accuracy, and kernel-based regression transfers approximately the nonlinear relationship between the RSS readings and positions to a linear program.

The proposed algorithm with two phases is shown in Fig. 1. In the offline phase, we set up several cluster radio maps based on the similarity between features of the RSSs. In the online phase, coarse localization is first performed to reduce the possible position area where the target may be located to a smaller one. In this process, methods A and B are applied to select the cluster to which the target belongs. The former method uses the distance between the features of the cluster head (CH) and the online vector for cluster matching, while the latter considers

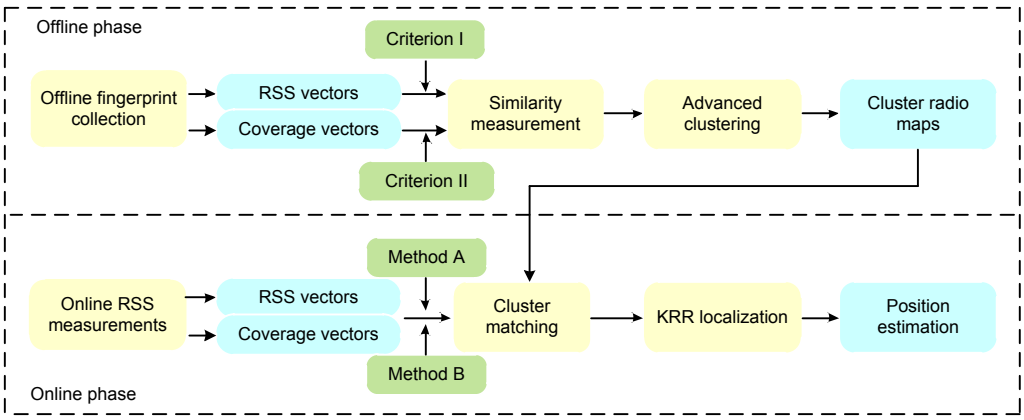

Fig. 1 The proposed indoor positioning algorithm

all the members of each cluster. Then the KRR approach is used to locate the target in the selected cluster. Details will be described in this section.

2.1 Offline phase

During the offline phase, suppose that the indoor area is divided into a set of N RPs, denoted as $P=\{\boldsymbol{p}_i\}$ (i=1, 2, …, N), where $\boldsymbol{p}_i$ represents the position of the i^{th} RP with the coordinates of (x_i, y_i). Then the entire radio map can be represented as

$$\boldsymbol{F}=\begin{bmatrix}\boldsymbol{f}_1\\ \boldsymbol{f}_2\\ \vdots\\ \boldsymbol{f}_N\end{bmatrix}=\begin{bmatrix}x_1 & y_1 & \boldsymbol{\psi}_1\\ x_2 & y_2 & \boldsymbol{\psi}_2\\ \vdots & \vdots & \vdots\\ x_N & y_N & \boldsymbol{\psi}_N\end{bmatrix}, \tag{1}$$

where $\boldsymbol{f}_i$ is the fingerprint of $\boldsymbol{p}_i$ and $\boldsymbol{\psi}_i=[\psi_{i,1}, \psi_{i,2}, \ldots, \psi_{i,M}]$ is the average of the RSS readings for M APs or anchors sampled in a specific time interval.

In an indoor environment, it often happens that the target at some RPs cannot receive radio signals from some APs or anchors. Thus, we record −100 dBm in the corresponding entity on the radio map for the unavailability of RSS readings. Note that this is generally set experimentally.

Generally, the complexity of the positioning algorithm is primarily determined by the number of samples, that is, the number of RPs in the radio map. In addition, in an extensive monitoring area, the complex indoor layout tends to make the radio signal have regional features. So, offline clustering is usually performed to provide a trade-off between algorithm complexity and positioning accuracy. In this study, two clustering schemes are recommended to measure the similarity of the RSS features of RPs in different indoor environments, denoted as criteria I and II. These two criteria are briefly defined as follows:

Criterion I: Based on the similarity between the offline RSS readings of RPs, the similarity metric in this criterion, denoted as $S_1(\boldsymbol{f}_i, \boldsymbol{f}_{i'})$, is defined as the inverse of the Euclidean distance of two radio map vectors $\boldsymbol{f}_i$ and $\boldsymbol{f}_{i'}$:

$$S_1(\boldsymbol{f}_i,\boldsymbol{f}_{i'})=\left\|\boldsymbol{\psi}_i-\boldsymbol{\psi}_{i'}\right\|^{-1},\ \forall i,i'\in\{1,2,...,N\},\ i\neq i'. \tag{2}$$

Criterion II: We choose the RPs with the highest similarity for categorization in a cluster using binary coverage vectors. The coverage vector is denoted as $\boldsymbol{I}_i=[I_{i,1}, I_{i,2}, \ldots, I_{i,M}]$ ($i\in\{1, 2, \ldots, N\}$), where $I_{i,j}$=1 ($j\in\{1, 2, \ldots, M\}$) if a target at $\boldsymbol{p}_i$ can receive the radio signal from anchor j for 90% of the sampling time, and $I_{i,j}$=0 otherwise. The threshold of 90% is experimentally set. The main reason for using coverage vectors is the none-line-of-sight (NLOS) propagation in indoor environments. Adjacent points may possess distinct coverage vectors owing to the interference of obstructions, such as walls and elevators. The similarity based on the coverage vectors of fingerprints $\boldsymbol{f}_i$ and $\boldsymbol{f}_{i'}$ can be measured by the inverse of their Hamming distance:

$$S_2(\boldsymbol{f}_i,\boldsymbol{f}_{i'})=\frac{1}{\sum_{m=1}^{M}\left|I_{i,m}-I_{i',m}\right|},\ \forall i,i'\in\{1,2,...,N\},\ i\neq i'. \tag{3}$$

Based on $S_1(\boldsymbol{f}_i, \boldsymbol{f}_{i'})$ or $S_2(\boldsymbol{f}_i, \boldsymbol{f}_{i'})$, the entire radio map $\boldsymbol{F}$ can be divided into L cluster radio maps $\boldsymbol{F}_l$ (l=1, 2, …, L) with a K-medoids clustering algorithm. In our algorithm, criterion I or criterion II is used for clustering depending mainly on the indoor environment. In a complex indoor environment, where coverage vectors can show the most distribution features of radio signals but with relatively low computational complexity, criterion II is a better option than criterion I. No matter which similarity criterion is adopted, a higher value of S_1 or S_2 corresponds to a higher similarity of the features between the RSS signals. Here, we take criterion II as an example. The pseudocode of offline clustering with criterion II is listed in Algorithm 1, in which lines 1–22 describe the clustering process based on Eq. (3). However, a complex indoor environment sometimes causes the RSSs to behave inconsistently. For example, an RP that is far from an AP receives a stronger radio signal than the RP that is closer to the AP, leading to scattered locations of cluster members. Here, we improve the clustering effect using transboundary RPs. This improvement is referred to as AC in this study. Consider a scenario where some RPs are geographical neighbors of members of a particular cluster to which they do not belong. Then in our algorithm, we add these RPs into this cluster and rebuild the cluster radio maps (lines 23–29 in Algorithm 1). Note that the number of added RPs should be thoroughly considered to seek a balance between the size of the cluster radio map and the computational complexity. This value is set as 2 in our experiments. Though the number of members of a cluster increases slightly, the positioning accuracy is greatly improved, as can be observed in the experimental results.

From Algorithm 1, once all the cluster members are selected, a set of L AC radio maps is created. Each map $\boldsymbol{F}_l=\{\boldsymbol{f}_{l(i)}\}$ (l=1, 2, …, L, i=1, 2, …, $\tilde{N}_l$, where $\tilde{N}_l$ is the number of fingerprints or members of the l^{th} cluster) is a subset of the entire radio map $\boldsymbol{F}$. The CH of $\boldsymbol{F}_l$ is denoted as $\boldsymbol{H}^{\text{o}}(l)$, which essentially is an RP in the corresponding cluster radio map whose RSS represents the most features of the cluster.

Algorithm 1 Offline clustering with criterion II

Input: entire radio map $\boldsymbol{F}$
Output: AC maps $\boldsymbol{F}_l$ (l=1, 2, …, L)

1 create a head set $H^{\text{o}}=\{\boldsymbol{H}^{\text{o}}(l)\}=\{\boldsymbol{f}_l\}$ (l=1, 2, ..., L) and denote the remaining RPs in $\boldsymbol{F}$ as D^{o}
2 **while** flag=1 **do**
3 initialize all $\boldsymbol{F}_l$ with $\boldsymbol{F}_l \leftarrow \boldsymbol{H}^{\text{o}}(l)$ and D_l with $D_l \leftarrow 0$
4 **for** all $\boldsymbol{p}_i \in D^{\text{o}}$ **do**
5 **for** all $\boldsymbol{p}_l \in H^{\text{o}}$ **do**
6 compute $S_2(\boldsymbol{f}_i, \boldsymbol{f}_l)$ according to Eq. (3)
7 **end for**
8 $l^* = \arg\max_l S_2(\boldsymbol{f}_i, \boldsymbol{f}_l)$
9 $\boldsymbol{F}_{l^*} \leftarrow \{\boldsymbol{F}_{l^*}, \boldsymbol{f}_i\}$
10 $D_{l^*} \leftarrow D_{l^*} + S_2(\boldsymbol{f}_i, \boldsymbol{f}_{l^*})$
11 **end for**
12 randomly select a number l and $\boldsymbol{p}_r \in \boldsymbol{F}_l$
13 initialize D_r with $D_r \leftarrow 0$
14 **for** all $\boldsymbol{p}_i \in \boldsymbol{F}_l$ **do**
15 compute $S_2(\boldsymbol{f}_i, \boldsymbol{f}_r)$ according to Eq. (3)
16 $D_r \leftarrow D_r + S_2(\boldsymbol{f}_i, \boldsymbol{f}_r)$
17 **end for**
18 **if** $D_r > D_l$ **then**
19 $\boldsymbol{H}^{\text{o}}(l) \leftarrow \boldsymbol{f}_r$
20 **else** flag$\leftarrow$0
21 **end if**
22 **end while**
23 **for** all $\boldsymbol{F}_l$ **do**
24 **for** all $\boldsymbol{f}_i \in \boldsymbol{F}$ **do**
25 **if** $\boldsymbol{f}_i \in \boldsymbol{F}_l$ and $\boldsymbol{f}_{i+1} \notin \boldsymbol{F}_l$ **then**
26 $\boldsymbol{F}_l \leftarrow \{\boldsymbol{F}_l, \boldsymbol{f}_{i+1}\}$
27 **end if**
28 **end for**
29 **end for**

2.2 Online phase

2.2.1 Approximate localization by cluster matching

During the online phase, an RSS measurement vector collected by the target is used for localization:

$$\boldsymbol{\psi}_{\text{t}} = [\psi_{\text{t},1}, \psi_{\text{t},2}, ..., \psi_{\text{t},M}]. \tag{4}$$

We use the corresponding coverage vector $\boldsymbol{I}_{\text{t}}=[I_{\text{t},1}, I_{\text{t},2}, \ldots, I_{\text{t},M}]$ to indicate the available APs. First, the approximate localization is implemented by cluster matching, which refines the region of interest to a subset of the entire database. This process not only decreases the complexity of the localization algorithm, but also limits the maximum localization error to this subset. Specifically, the approximate localization is performed by comparing the similarity between the online measurement and CHs to identify to which cluster the target of online measurement belongs. For indoor environments with different layout characteristics, we use criteria I and II to perform cluster matching. For each criterion, there are two matching strategies recommended in this study, denoted as

methods A and B. Taking the similarity metric defined in Eq. (2) as an example, method A determines the similarity between $\boldsymbol{\psi}_{\mathrm{t}}$ and $\boldsymbol{H}^{\mathrm{o}}(l)$ as

$$S_{\mathrm{l}}^{\mathrm{A}}(\boldsymbol{\psi}_{\mathrm{t}}, \boldsymbol{H}^{\mathrm{o}}(l))=\left\|\boldsymbol{\psi}_{\mathrm{t}}-\boldsymbol{\psi}_{\boldsymbol{H}^{\mathrm{o}}(l)}\right\|^{-1}, \ \forall l \in\{1,2,...,L\}, \quad (5)$$

where $\boldsymbol{\psi}_{\boldsymbol{H}^{\mathrm{o}}(l)}$ is the offline RSS reading corresponding to the index number of $\boldsymbol{H}^{\mathrm{o}}(l)$. Instead of using the RSS of the CH for cluster matching, method B calculates the similarity by considering all the members of each cluster:

$$S_{\mathrm{l}}^{\mathrm{B}}(\boldsymbol{\psi}_{\mathrm{t}}, \boldsymbol{F}_l)=\frac{1}{\tilde{N}_l}\sum_{u=1}^{\tilde{N}_l}\left\|\boldsymbol{\psi}_{\mathrm{t}}-\boldsymbol{\psi}_{l(u)}\right\|^{-1}, \ \forall l \in\{1,2,...,L\}, \quad (6)$$

where $\boldsymbol{\psi}_{l(u)}$ represents the RSS readings of one RP in the l^{th} cluster. No matter which method is applied during this phase, the cluster with the smallest distance (i.e., with the highest similarity to the target) is selected as the potential region. According to our experiments, these two methods for cluster matching lead to a little variation of positioning accuracy. Thus, the remaining experimental results in this study are obtained with method A to select the matching cluster. Finally, the likeliest matching cluster is denoted as $\hat{l}$.

2.2.2 Precise localization by kernel-based ridge regression

Ridge regression is the linear regression intended to overcome the sparse estimation problem of the regression coefficients in the least squares method (Maalouf and Homouz, 2014). The solution in our application is achieved within the scope of the matched cluster. For each cluster $\boldsymbol{F}_l$, let $\boldsymbol{X}_l=\{\boldsymbol{\psi}_{l(i)}\}$ (i=1, 2, …, $\tilde{N}_l$) in $\mathbb{R}^{\tilde{N}_l\times M}$ be the training RSS data set and $\boldsymbol{Y}_l=\{(x_{l(i),}\, y_{l(i)})\}$ (i=1, 2, …, $\tilde{N}_l$) in $\mathbb{R}^{\tilde{N}_l\times 2}$ be the output set of $\boldsymbol{X}_l$. Each row vector in $\boldsymbol{X}_l$ denotes a sample in the input space with a corresponding output coordinate vector in $\boldsymbol{Y}_l$. The general linear model in a matrix form can be represented as

$$\boldsymbol{Y}_l=\boldsymbol{X}_l\boldsymbol{w}_l+\boldsymbol{\varepsilon}_l, \quad (7)$$

where $\boldsymbol{\varepsilon}_l$ is the residual vector in $\mathbb{R}^{\tilde{N}_l\times 2}$ and $\boldsymbol{w}_l$ the weight vector of the regression hyperplane in $\mathbb{R}^{M\times 2}$, which can be determined by minimizing the sum of the squared residuals. To avoid an overly large, unstable estimate, a regularization parameter is added to shrink the least-squares coefficients. The objective function of the ridge regression is then defined as

$$f(\boldsymbol{w}_l)=(\boldsymbol{Y}_l-\boldsymbol{X}_l\boldsymbol{w}_l)^{\mathrm{T}}(\boldsymbol{Y}_l-\boldsymbol{X}_l\boldsymbol{w}_l)+\lambda_l\boldsymbol{w}_l{}^{\mathrm{T}}\boldsymbol{w}_l, \quad (8)$$

where $f(\boldsymbol{w}_l)$ denotes the objective function and $\lambda_l\geq 0$ is a regularization parameter to make a trade-off between the bias and variance. The gradient with respect to $\boldsymbol{w}_l$ is computed to find the solution to the minimization problem (8), and we can obtain

$$\boldsymbol{w}_l=(\boldsymbol{X}_l^{\mathrm{T}}\boldsymbol{X}_l+\lambda_l\boldsymbol{I}_M)^{-1}\boldsymbol{X}_l^{\mathrm{T}}\boldsymbol{Y}_l, \quad (9)$$

where $\boldsymbol{I}_M$ is an $M\times M$ identity matrix.

However, ridge regression is not an ideal solution for indoor localization with the sampled RSS data. This is because in most practical indoor environments, the relationship between RSSs and positions is much more complex than that in an empty room because of multipath, shadowing conditions, and NLOS propagation of radio signals. Instead, a more general nonlinear mapping function is used to map the data from a lower-dimensional space into a higher-dimensional one, where the relationship becomes linear. In our algorithm, a kernel function κ satisfying Mercer's condition is employed to solve this problem (Saunders et al., 1998). Let $\phi(\cdot)$ be a general non-linear mapping function:

$$\phi:\boldsymbol{\psi}_{l(m)}\in\mathbb{R}^M\rightarrow\phi(\boldsymbol{\psi}_{l(m)})\in\mathbb{R}^{\Lambda}, \ \forall m\in\{1,2,...,\tilde{N}_l\}, \quad (10)$$

where superscript Λ stands for a higher-dimensional space. It is not necessary to know $\phi(\cdot)$ as long as the kernel function $\kappa(\boldsymbol{\psi}_{l(m)}, \boldsymbol{\psi}_{l(n)})=\phi(\boldsymbol{\psi}_{l(m)})\phi(\boldsymbol{\psi}_{l(n)})$ is introduced as a format of dot product. According to the matrix inversion lemma, $\boldsymbol{w}_l$ can be rewritten as

$$\boldsymbol{w}_l=\boldsymbol{X}_l^{\mathrm{T}}(\lambda_l\boldsymbol{I}_M+\boldsymbol{X}_l\boldsymbol{X}_l^{\mathrm{T}})^{-1}\boldsymbol{Y}_l. \quad (11)$$

So, in a high-dimensional feature space, let

$$\boldsymbol{\alpha}_l=\left(\boldsymbol{K}_l+\lambda_l\boldsymbol{I}_{\tilde{N}_l}\right)^{-1}\boldsymbol{Y}_l, \quad (12)$$

where $\boldsymbol{I}_{\tilde{N}_l}$ is an $\tilde{N}_l \times \tilde{N}_l$ identity matrix and $\boldsymbol{K}_l$ is an $\tilde{N}_l \times \tilde{N}_l$ kernel matrix with the elements of $\kappa(\boldsymbol{\psi}_{l(m)}, \boldsymbol{\psi}_{l(n)})$. In our algorithm, the radial basis function kernel is used:

$$\kappa(\boldsymbol{\psi}_{l(m)}, \boldsymbol{\psi}_{l(n)}) = \exp\left(\frac{\| \boldsymbol{\psi}_{l(m)} - \boldsymbol{\psi}_{l(n)} \|^2}{-2\varepsilon^2}\right), \quad (13)$$

where ε is the width of the kernel. Then Eq. (11) can be rewritten as

$$\boldsymbol{w}_l = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\alpha}_l. \quad (14)$$

Note that $\boldsymbol{\alpha}_l$ of each cluster is trained after the AC procedure during the offline phase. In our algorithm, KRR is used to predict the position based on online RSS measurement and pre-trained vector $\boldsymbol{\alpha}_l$. Thus, the position estimate of the target, denoted as $\hat{\boldsymbol{p}}_{\mathrm{t}}$, can be determined as

$$\hat{\boldsymbol{p}}_{\mathrm{t}} = \boldsymbol{w}_l^{\mathrm{T}} \boldsymbol{\psi}_{\mathrm{t}} = \sum_{n=1}^{\tilde{N}_l} \boldsymbol{\alpha}_{l(n)} \boldsymbol{\psi}_{l(n)}^{\mathrm{T}} \boldsymbol{\psi}_{\mathrm{t}} = \sum_{n=1}^{\tilde{N}_l} \boldsymbol{\alpha}_{l(n)} \kappa(\boldsymbol{\psi}_{\mathrm{t}}, \boldsymbol{\psi}_{l(n)}), \quad (15)$$

where $\boldsymbol{\alpha}_{l(n)}$ stands for the n^{th} row of $\boldsymbol{\alpha}_l$. Note that according to Eq. (13), the complexity of solving $\boldsymbol{\alpha}_l$ is $O(\tilde{N}_l^3)$, which results in overwhelming computational complexity with a large data set. Cluster matching, as presented in Section 2.2, can limit the size of the data set and improve the real-time performance of the proposed algorithm.

3 Experimental results and discussion

Experimental evaluation of the proposed positioning algorithm was carried out in real typical indoor environments. Stationary nodes (TI CC2430 with TinyOS) worked as anchors to broadcast signals in the network, and RSS measurements were obtained by a moving target (TI CC2430 with sniffer software based on TinyOS) with a maximum sampling rate of two samples per second.

3.1 Fingerprinting setup and RSS observation

Real data was measured from an office building and a school building. Specifically, the experiments were carried out on the ninth floor of the Optical Engineering and Technology Building and the first floor of the Third Teaching Building at the University of Shanghai for Science and Technology, China. For the convenience of description, two surveying sites were shortened to LA and LB. The respective dimensions of these two sites were 45 m×15 m and 20 m×16 m. A total of 18 anchors were arranged in LA, with 90 RPs in an average grid spacing of 1.8 m. In LB, there were 12 anchors and 93 RPs in the same grid spacing. Fig. 2 shows the layouts of the experimental sites.

Note that because of the layouts of the sites, anchors were arranged on the sides of each site without an even distribution. Data collection was performed over several days during the offline phase, including office hours and class time, to capture more non-linear RSS features. RSS readings from anchors at every RP were averaged and recorded over 120 s as $\boldsymbol{\psi}_i = [\psi_{i,1}, \psi_{i,2}, \ldots, \psi_{i,18}]$ (i=1, 2, …, 90) for LA and

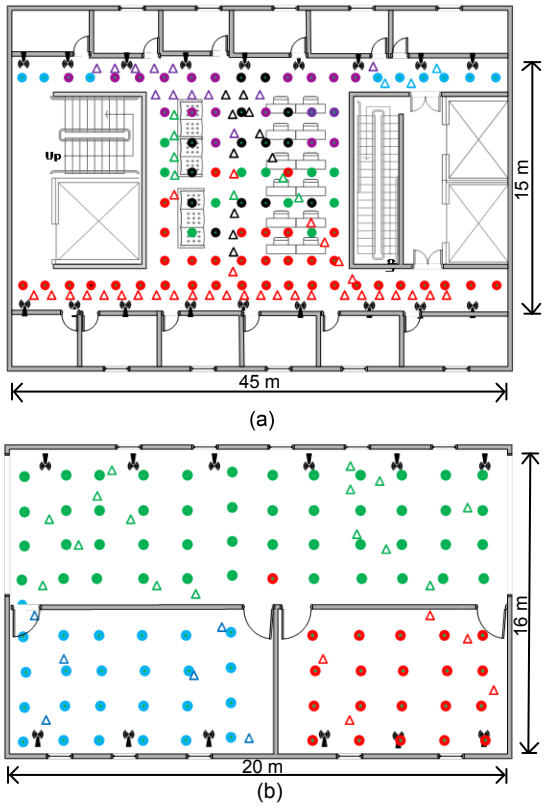

Fig. 2 Layouts of the experimental sites: (a) LA; (b) LB
Each dot denotes one RP, and different colors for the RPs indicate different clusters. Each triangle represents the location of each online test point with the corresponding cluster. References to color refer to the online version of this figure

$\boldsymbol{\psi}_i=[\psi_{i,1}, \psi_{i,2}, \ldots, \psi_{i,12}]$ (i=1, 2, …, 93) for LB. The online RSS observations were performed on different days with selected locations shown in Fig. 2. Because of the distinct layouts of these two sites, RSS measurements were distributed differently. Figs. 3a and 3b show examples of RSS for a certain anchor over all RPs in LA and LB, respectively.

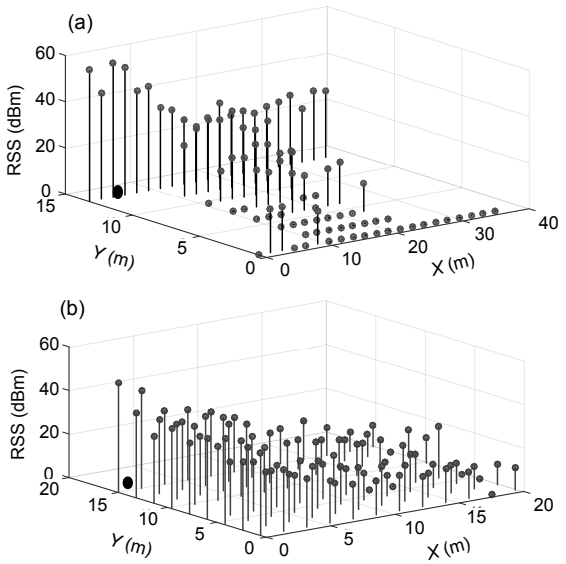

Fig. 3 Examples of RSS distribution: (a) LA; (b) LB (Black circles represent the positions of anchors)

Fig. 3 shows different radio propagation patterns due to the different indoor environments of each site. The reason for this difference may be that the lift wells and fire ladders in LA almost entirely block the radio signals, leading to the failure of the target to receive signals in several RPs. LB, on the other hand, contains only classrooms and corridors, making the signal more stable. This can also be seen from Fig. 2. It shows that most of the RPs belonging to the same cluster are geographically close to each other. However, things seem to be trickier in LA. Thus, AC is proposed to solve this problem in a complicated indoor environment such as LA.

3.2 Experimental results

The positioning error is commonly measured by the Euclidean distance between the actual and estimated locations of the test points. We used the average error, known as the mean absolute error (MAE), and the empirical cumulative distribution function (CDF) of errors to evaluate the proposed positioning algorithm. MAE is defined as

$$\mathrm{MAE}=\frac{1}{V}\sum_{v=1}^{V}e_v=\frac{1}{V}\sum_{v=1}^{V}\left\|\boldsymbol{p}_v-\hat{\boldsymbol{p}}_v\right\|, \tag{16}$$

where V is the number of test points, e_v the positioning error of the v^{th} test point, $\boldsymbol{p}_v$ the actual position of the v^{th} test point, and $\hat{\boldsymbol{p}}_v$ the estimated position.

The performance of the proposed algorithm is influenced by several factors, such as the criterion used for clustering, the number k of transboundary RPs added to the cluster, the number L of the clusters generated, and the kernel width ε used in the kernel function. Fig. 4 shows the CDFs of the positioning error with or without offline clustering at LA and LB.

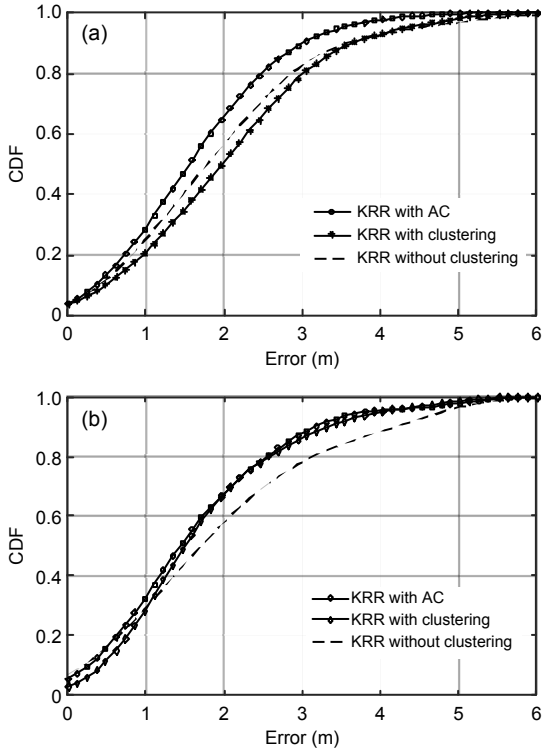

Fig. 4 Cumulative distribution function (CDF) of the positioning error with or without offline clustering: (a) LA; (b) LB

KRR: kernel-based ridge regression; AC: advanced clustering

Here, for the KRR with AC in Fig. 4, criterion II was used to generate five clusters in LA and criterion I was used to generate three clusters in LB. The algorithm of KRR with clustering involves identical steps to the algorithm of KRR with AC except for the clustering, and it adopts the traditional clustering method rather than AC. In our algorithm, two RPs in the cluster boundary were added to each cluster. All

clusters in the same site shared the same values of parameters ε and λ ($\varepsilon_A^2 = 50$, $\lambda_A = 0.0001$, $\varepsilon_B^2 = 10$, λ_B=0.000 01, where subscripts A and B indicate LA and LB, respectively). The results showed that the effects of clustering varied with the localization sites. In LA, the AC strategy led to an enhancement on the positioning accuracy. In LB, both clustering strategies outperformed the algorithm without clustering. However, the AC method slightly improved the localization accuracy compared with the traditional clustering in LB. The reason may be that the environment in LB is relatively neat, and that RSS signals in LB are more evenly distributed, making it possible to adopt traditional clustering without introducing significant cluster matching errors.

3.3 Localization performance analysis

Since clustering plays an important role in approximate localization, we first investigated the effects of these two similarity criteria on clustering. Similarity criteria I and II for AC were applied in LA and LB in the offline phase and three clusters were generated, while the same online positioning algorithms were used for final position estimation. Fig. 5 provides the CDF of the positioning error.

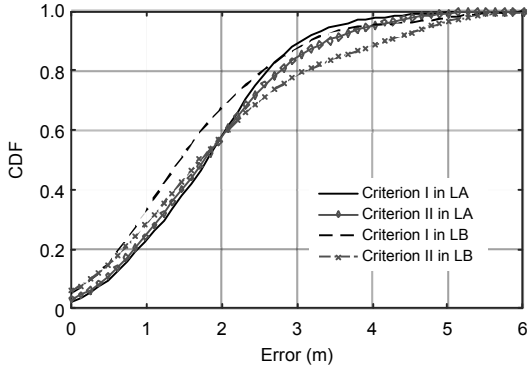

Fig. 5 Cumulative distribution function (CDF) of the positioning error with different similarity criteria in offline reference point clustering

AC with criterion I achieved better performance than AC with criterion II in LB, while these two criteria returned comparable localization errors in LA (MAE=1.75 m for criterion I and MAE=1.77 m for criterion II). Considering that it takes less time using criterion II to calculate the similarity, we adopted criterion II in LA and criterion I in LB in the following experiments. However, these two cluster-matching methods defined in Eqs. (5) and (6) generated almost identical results regardless of sites in the online phase. Accordingly, we used the former (i.e., method A) in the subsequent experiments.

In RSS-based fingerprint localization, the number of RPs is an important factor. In general, as the density of RPs increases, the positioning accuracy will be improved. Of course, larger density may be accompanied by other problems, such as the increase of the time in the offline phase and the high computation cost in the online phase. Table 1 shows the localization performance of the proposed algorithm with different numbers of RPs in LA. The change in the RP number signified a change in the training set. In Table 1, for each set of RP numbers, we consider two cases of using and not using the AC strategy during the offline phase. Note that all the results were obtained using KRR (L=3, $\varepsilon_A^2 = 50$, λ_A=0.0001). As can be observed in Table 1, the AC-based approximate localization algorithm provided larger positioning accuracy than the algorithm without AC, regardless of the number of RPs. The same comparison made in experimental site LB led to a consistent conclusion. AC decreased the algorithm size and provided an improvement in the positioning accuracy.

Table 1 Positioning error statistics in LA

Condition	Average error (m)	Positioning error (m)		Variance (m^2)
		87% test points	100% test points	
90 RPs without AC	1.92	3.04	5.10	1.69
46 RPs without AC	2.03	3.27	5.43	2.02
34 RPs without AC	2.18	3.79	5.51	2.13
90 RPs with AC	1.77	2.76	4.48	1.31
46 RPs with AC	1.82	3.05	5.02	1.79
34 RPs with AC	2.04	3.32	5.11	2.18

Based on the above observation, we have evaluated the localization performance when the cluster number changed. Taking the 90 RPs in LA as an example, Fig. 6 illustrates the performance comparison with different cluster numbers when AC was used in the offline phase and KRR was applied within the range of each cluster. As shown in Fig. 6, there was no linear relationship between the number of clusters and the positioning accuracy. On one hand, when more clusters were generated, a smaller region for precise localization was determined after the approximate localization. This caused the training set to be too

small to provide reliable regression parameters and high possibility of choosing the wrong cluster. On the other hand, fewer clusters usually mean more members of each cluster, and this induces higher complexity of KRR in precise localization. Therefore, it is recommended to experimentally set the number of clusters to obtain the desired performance.

Usually, the performance of the RSS-based localization approach is highly related to the numbers of RPs and anchors as reported. As shown in Table 1, the number of RPs had no significant effect on the KRR-based positioning accuracy. However, MAEs of the proposed algorithm in LA and LB, as shown in Fig. 7, were highly dependent on the number of anchors when all RPs (e.g., 90 RPs in LA and 93 RPs in LB) were used. Other parameters were consistent with those in previous experiments.

As shown in Fig. 7, increasing the number of anchors helped improve the average localization accuracy no matter whether there was cluster or not. In addition, when there were only a few anchors arranged in the experimental site, clustering had no significant influence on the positioning accuracy. In LA, when 10 or fewer anchors were used for positioning, an average localization error of 2.13 m was achieved with or without clustering. In LB, the corresponding anchor number and average localization error were 6 and 2.06 m, respectively. Despite this, considering the computational complexity of KRR, the AC method is still recommended in the offline phase.

3.4 Comparison with prior work

We compared the proposed localization algorithm with other fingerprint approaches, including KRR without clustering (Mahfouz et al., 2016), LASSO-based localization (Khalajmehrabadi et al., 2017a), CS-based positioning (Feng et al., 2012), and the WKNN-based method (Niu et al., 2015). Fig. 8 describes the CDFs of these algorithms implemented in LA and LB. The entire radio map, including all the fingerprints, was used for clustering and localization. It can be observed that our proposed algorithm had better performance in both experimental sites. Further analysis showed that the positioning performances of these algorithms varied in LA, while it was an opposite case in LB. The reason may be that in LA, affected by the complex indoor environment, there

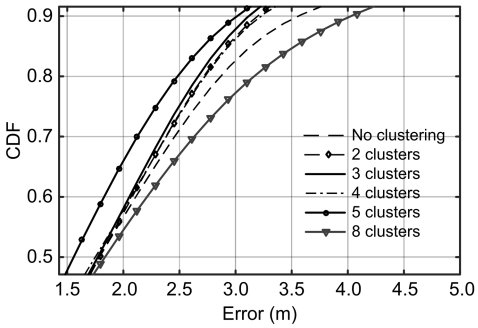

Fig. 6 Cumulative distribution function (CDF) of the localization error in LA with different numbers of clusters

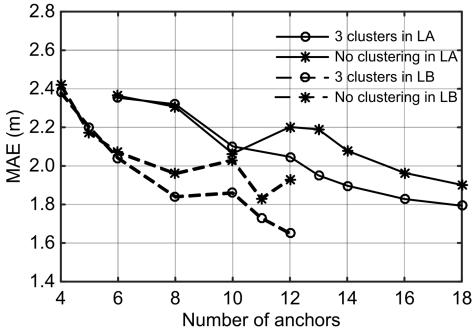

Fig. 7 Mean absolute error (MAE) for the proposed localization algorithm

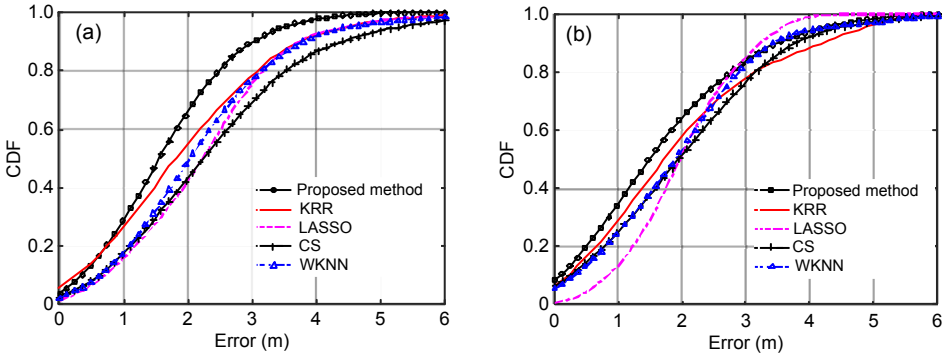

Fig. 8 Cumulative distribution function (CDF) for the KRR, LASSO-, CS-, and WKNN-based methods and the proposed algorithm: (a) LA; (b) LB

were some outliers making these approaches suffer from more substantial localization errors. The performance shown in Fig. 8a indicates the robustness of the proposed algorithm.

4 Conclusions

In this paper, we have introduced a new algorithm based on KRR with AC for indoor RSS based fingerprint positioning. In the offline phase, an AC method has been proposed to take into account the fact that the RSS signals of adjacent sample points do not necessarily have the similar value. Reference points, which are not the members of a cluster but are adjacent to the members of the cluster, have been added to the cluster in this phase. A set of cluster radio maps has been set up for approximate localization in the online phase. The cluster matching significantly reduced the computational complexity of the KRR-based localization algorithm. Our experimental results showed that the parameters, including the regularization parameter in the ridge regression and the width of the kernel function, are consistent between different clusters in a single environment, which makes KRR easy to implement and requires no additional memory space for storing the cluster parameters. To demonstrate the efficiency of the proposed algorithm, we have implemented the algorithm in two scenarios. Although the performance improvements varied in different scenarios, experimental results demonstrated that the proposed algorithm always leads to effective localization of the target with its RSS signals.

Contributors

Yanfen LE and Heng YAO designed the research. Yanfen LE and Hena ZHANG processed the data. Yanfen LE drafted the manuscript. Weibin SHI helped organize the manuscript. Yanfen LE and Heng YAO revised and finalized the paper.

Compliance with ethics guidelines

Yanfen LE, Hena ZHANG, Weibin SHI, and Heng YAO declare that they have no conflict of interest.

References

Al-Moukhles H, Jaber AK, Abdel-Qader I, 2016. Impact of APs selection scheme on compressive sensing-fingerprinting based IPS performance. Proc IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conf, p.1-7. https://doi.org/10.1109/uemcon.2016.7777849

Al Nuaimi K, Kamel H, 2011. A survey of indoor positioning systems and algorithms. Proc Int Conf on Innovations in Information Technology, p.185-190. https://doi.org/10.1109/innovations.2011.5893813

Bahl P, Padmanabhan VN, 2000. RADAR: an in-building RF-based user location and tracking system. Proc IEEE Conf on Computer Communications and 19th Annual Joint Conf of the IEEE Computer and Communications Societies, p.775-784. https://doi.org/10.1109/infcom.2000.832252

Chen C, Wang YJ, Zhang Y, et al., 2018. Indoor positioning algorithm based on nonlinear PLS integrated with RVM. *IEEE Sens J*, 18(2):660-668. https://doi.org/10.1109/jsen.2017.2772798

Chen YQ, Yang Q, Yin J, et al., 2006. Power-efficient access-point selection for indoor location estimation. *IEEE Trans Knowl Data Eng*, 18(7):877-888. https://doi.org/10.1109/TKDE.2006.112

Dai H, Ying WH, Xu J, 2016. Multi-layer neural network for received signal strength-based indoor localisation. *IET Commun*, 10(6):717-723. https://doi.org/10.1049/iet-com.2015.0469

Fang SH, Lin T, 2012. Principal component localization in indoor WLAN environments. *IEEE Trans Mob Comput*, 11(1):100-110. https://doi.org/10.1109/tmc.2011.30

Fang XM, Jiang ZH, Nan L, et al., 2018. Optimal weighted K-nearest neighbour algorithm for wireless sensor network fingerprint localisation in noisy environment. *IET Commun*, 12(10):1171-1177. https://doi.org/10.1049/iet-com.2017.0515

Feng C, Au WSA, Valaee S, et al., 2012. Received-signal-strength-based indoor positioning using compressive sensing. *IEEE Trans Mob Comput*, 11(12):1983-1993. https://doi.org/10.1109/tmc.2011.216

Harroud H, Ahmed M, Karmouch A, 2003. Policy-driven personalized multimedia services for mobile users. *IEEE Trans Mob Comput*, 2(1):16-24. https://doi.org/10.1109/tmc.2003.1195148

Honeine P, Mourad F, Kallas M, et al., 2011. Wireless sensor networks in biomedical: body area networks. Proc Int Workshop on Systems, Signal Processing and Their Applications, p.388-391. https://doi.org/10.1109/WOSSPA.2011.5931518

Hu JS, Liu HL, Liu DW, et al., 2018. Reducing Wi-Fi fingerprint collection based on affinity propagation clustering and WKNN interpolation algorithm. Proc 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conf, p.2463-2468. https://doi.org/10.1109/IMCEC.2018.8469697

Huang CC, Manh HN, 2016. RSS-based indoor positioning based on multi-dimensional kernel modeling and weighted average tracking. *IEEE Sens J*, 16(9):3231-3245. https://doi.org/10.1109/JSEN.2016.2524537

Khalajmehrabadi A, Gatsis N, Pack DJ, et al., 2017a. A joint

indoor WLAN localization and outlier detection scheme using LASSO and elastic-net optimization techniques. *IEEE Trans Mob Comput*, 16(8):2079-2092. https://doi.org/10.1109/tmc.2016.2616465

Khalajmehrabadi A, Gatsis N, Akopian D, 2017b. Structured group sparsity: a novel indoor WLAN localization, outlier detection, and radio map interpolation scheme. *IEEE Trans Veh Technol*, 66(7):6498-6510. https://doi.org/10.1109/TVT.2016.2631980

Kumar C, Rajawat K, 2019. Dictionary-based statistical fingerprinting for indoor localization. *IEEE Trans Veh Technol*, 68(9):8827-8841. https://doi.org/10.1109/tvt.2019.2929360

Kushki A, Plataniotis KN, Venetsanopoulos AN, et al., 2007. Kernel-based positioning in wireless local area networks. *IEEE Trans Mob Comput*, 6(6):689-705. https://doi.org/10.1109/TMC.2007.1017

Li LQ, He Z, Nielsen J, et al., 2015. Using Wi-Fi/ magnetometers for indoor location and personal navigation. Proc Int Conf on Indoor Positioning and Indoor Navigation, p.1-7. https://doi.org/10.1109/ipin.2015.7346764

Lu XX, Zou H, Zhou HM, et al., 2016. Robust extreme learning machine with its application to indoor positioning. *IEEE Trans Cybern*, 46(1):194-205. https://doi.org/10.1109/tcyb.2015.2399420

Maalouf M, Homouz D, 2014. Kernel ridge regression using truncated Newton method. *Knowl-Based Syst*, 71:339-344. https://doi.org/10.1016/j.knosys.2014.08.012

Mahfouz S, Mourad-Chehade F, Honeine P, et al., 2013. Kernel-based localization using fingerprinting in wireless sensor networks. Proc IEEE 14th Workshop on Signal Processing Advances in Wireless Communications, p.744-748. https://doi.org/10.1109/SPAWC.2013.6612149

Mahfouz S, Mourad-Chehade F, Honeine P, et al., 2016. Non-parametric and semi-parametric RSSI/distance modeling for target tracking in wireless sensor networks. *IEEE Sens J*, 16(7):2115-2126. https://doi.org/10.1109/JSEN.2015.2510020

Niu JW, Wang BW, Shu L, et al., 2015. ZIL: an energy-efficient indoor localization system using ZigBee radio to detect WiFi fingerprints. *IEEE J Sel Areas Commun*, 33(7):1431-1442. https://doi.org/10.1109/jsac.2015.2430171

Rodriguez MD, Favela J, Martinez EA, et al., 2004. Location-aware access to hospital information and services. *IEEE Trans Inform Technol Biomed*, 8(4):448-455. https://doi.org/10.1109/titb.2004.837887

Saunders C, Gammerman A, Vovk V, 1998. Ridge regression learning algorithm in dual variables. Proc 15th Int Conf on Machine Learning, p.515-521.

Shi LF, Wang Y, Liu GX, et al., 2018. A fusion algorithm of indoor positioning based on PDR and RSS fingerprint. *IEEE Sens J*, 18(23):9691-9698. https://doi.org/10.1109/jsen.2018.2873052

Wang XY, Gao LJ, Mao SW, et al., 2017. CSI-based fingerprinting for indoor localization: a deep learning approach. *IEEE Trans Veh Technol*, 66(1):763-776. https://doi.org/10.1109/TVT.2016.2545523

Wu Z, Fu KC, Jedari E, et al., 2016. A fast and resource efficient method for indoor positioning using received signal strength. *IEEE Trans Veh Technol*, 65(12):9749-9758. https://doi.org/10.1109/tvt.2016.2530761

Xue WX, Yu KG, Hua XH, et al., 2018. APs' virtual positions-based reference point clustering and physical distance-based weighting for indoor Wi-Fi positioning. *IEEE Intern Things J*, 5(4):3031-3042. https://doi.org/10.1109/jiot.2018.2829486

Yan J, Zhao L, Tang J, et al., 2018. Hybrid kernel based machine learning using received signal strength measurements for indoor localization. *IEEE Trans Veh Technol*, 67(3):2824-2829. https://doi.org/10.1109/TVT.2017.2774103

Youssef MA, Agrawala A, Shankar AU, 2003. WLAN location determination via clustering and probability distributions. Proc 1st IEEE Int Conf on Pervasive Computing and Communications, p.143-150. https://doi.org/10.1109/PERCOM.2003.1192736

Zhang Y, Li DP, Wang YJ, 2019. An indoor passive positioning method using CSI fingerprint based on Adaboost. *IEEE Sens J*, 19(14):5792-5800.