Sensors **2014**, *14*, 1850-1876; doi:10.3390/s140101850

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

A Spatial Division Clustering Method and Low Dimensional Feature Extraction Technique Based Indoor Positioning System

Yun Mo [1], Zhongzhao Zhang [1,*], Weixiao Meng [1], Lin Ma [1] and Yao Wang [1,2]

[1] Communication Research Center, School of Electronics Information Engineering, Harbin Institute of Technology, Harbin 150001, China; E-Mails: 11b905020@hit.edu.cn (Y.M.); wxmeng@hit.edu.cn (W.M); malin@hit.edu.cn (L.M.); wangyaowh2005@126.com (Y.W.)

[2] Communication Department, Shenyang Artillery Academy, Shenyang 110867, China

* Author to whom correspondence should be addressed; E-Mail: zzzhang@hit.edu.cn; Tel.: +86-186-1173-0500.

Received: 26 December 2013; in revised form: 14 January 2014 / Accepted: 20 January 2014 / Published: 22 January 2014

Abstract: Indoor positioning systems based on the fingerprint method are widely used due to the large number of existing devices with a wide range of coverage. However, extensive positioning regions with a massive fingerprint database may cause high computational complexity and error margins, therefore clustering methods are widely applied as a solution. However, traditional clustering methods in positioning systems can only measure the similarity of the Received Signal Strength without being concerned with the continuity of physical coordinates. Besides, outage of access points could result in asymmetric matching problems which severely affect the fine positioning procedure. To solve these issues, in this paper we propose a positioning system based on the Spatial Division Clustering (SDC) method for clustering the fingerprint dataset subject to physical distance constraints. With the Genetic Algorithm and Support Vector Machine techniques, SDC can achieve higher coarse positioning accuracy than traditional clustering algorithms. In terms of fine localization, based on the Kernel Principal Component Analysis method, the proposed positioning system outperforms its counterparts based on other feature extraction methods in low dimensionality. Apart from balancing online matching computational burden, the new positioning system exhibits advantageous performance on radio map clustering, and also shows better robustness and adaptability in the asymmetric matching problem aspect.

Keywords: clustering; outliers; GA-SVM; kernel PCA; asymmetric matching

1. Introduction

With the rapid development in the areas of mobile computing terminals and wireless techniques, indoor positioning systems have become unprecedentedly popular in recent years. Although the Global Positioning System (GPS) has been in service for decades, the indoor positioning ability of GPS is limited in indoor environments by the insufficient satellite coverage and poor positioning signals [1]. Not only does the indoor positioning draw attention from world famous academic research institutions but also large scale business activities have been deployed to solve this problem, such as the cooperation between Apple and WiFiSLAM, and the competition between Baidu and AutoNavi. As a consequence, several indoor positioning systems have been proposed in recent years, which are based on infrared [2], ultrasound and Radio Frequency (RF) [3], *etc.* Because the RF-based indoor positioning systems are capable of providing a wide range of coverage and using the existed WLANs as the fundamental infrastructure, fingerprinting methods [4–6] based on WLANs, as one of the most popular RF techniques, outperforms the other existing indoor positioning systems in civilian fields [7,8]. For instance, a convenient way based on propagation models for real-time indoor positioning without fingerprinting radio map basis is proposed in [9], but the Maximum Likelihood Estimation (MLE) and Least Square Optimization (LSO)-based probabilistic method used in the system would be time-consuming and computationally expensive in terms of mobile terminals. More importantly, the given confidence probability is lower than 10% under the condition that positioning accuracy is 2 m, which is sometimes insufficient for indoor positioning services, while fingerprinting positioning systems may normally provide confidence probabilities over 50% under the same conditions.

A typical fingerprinting indoor positioning system can be described as a situation where an end user takes RSS readings from available access points (AP) with a mobile terminal in an indoor environment. The positioning system then estimates the current location of the user according to a database, the so called fingerprint radio map, which contains pre-measured RSS values and the corresponding coordinates.

On the one hand, since a large indoor positioning region with a large fingerprint dataset could lead to high computational complexity and error margins, dividing it into several sub-regions is supposed to be able to improve the positioning performance [10]. Consequently clustering methods are widely applied to dividing the fingerprinting radio map into several sub-radio maps. However, the traditional clustering methods, e.g., K-Means, Fuzzy C-Means and Affinity Propagation [11,12], cannot theoretically process the outliers or singular points (an outlier means a sample point is assigned to a class by a cluster method but in physical space it is actually located in another class). This is a typical problem when deploying pattern recognition clustering methods in positioning systems. Most researchers simply ignore the outliers or delete those points, or artificially change the class label of the outlier to the one it is located in. Nevertheless, any of those solutions may lead to an increase in the positioning error rate. Furthermore, those methods for clustering the radio map essentially only depend on Received Signal Strength (RSS) values in signal space instead of considering their coordinate proximity in physical space. They actually generate the sub-radio maps in signal space, rather than in real sub-regions of the positioning area. Therefore, the coarse positioning in that case actually cannot prove that the terminal is located in a certain area, but only illustrate that the received RSS value may belong to one of the sub-datasets.

Besides, location privacy also should be taken into consideration sometimes [13]. For security reasons, sample points of certain areas such as confidential rooms within the radio map might be required to be clustered together, thereby providing the indoor positioning services of the dedicated area only to those authorized people. In this case, the traditional methods may not run well.

On the other hand, the deployment of feature extraction algorithms in the fingerprinting system is able to effectively process the radio map, *i.e.*, mapping it from the original signal space to a new feature space, thereby decreasing the noise interference and improving the location performance at the cost of increased computational complexity [14,15]. For instance, Reference [16] presents a positioning system based on Multiple (Linear) Discrimination Analysis (MDA or LDA) and Adaptive Neural Network (ANN). Though the Artificial Neural Network may suffer from the local minimum problem and over-fitting problems, the conception of Discriminant Components (DC) derived from MDA is efficiently introduced into the fingerprinting system. Parallel with DC, Principal Components (PC) derived from PCA is introduced in [17]. Apart from improved positioning accuracy, the proposed method also could reduce the number of training samples needed. Like the DC and PC used in [16–18], we pay attention to the aspect of dimensional reduction [19,20] (the original dimensionality of the radio map could be considered as the number of available APs) which is also a key factor for adjusting the available features of the feature extraction algorithm for indoor positioning. In fact, an appropriate algorithm can also enhance the robustness, balance the computational burden and save storage, which are all significant in terms of mobile computing.

Moreover, the number of APs received by a user in real-time phase may not always match the pre-stored radio map, e.g., one of those APs might be out of service or powered off at times. In that case, the traditional fingerprinting location method may not work out. Although some candidate options could deal with that, for instance set the RSS readings of the blocked AP as zero or remove the corresponding dimension of the radio map, the asymmetric matching problem still introduces severe systematic errors and reduces the positioning performance. However, by deploying an adaptive dimensional reduction technique, the impact of the missing APs could be strictly confined.

In this paper, for one thing, we propose the Spatial Division Clustering (SDC) method for reasonably dividing the radio map without singular points and the constraints presented above. After being integrated with optimized Support Vector Machine (SVM) technique [21,22], it is able to localize the test point (TP) into the sub region correctly during the so called coarse positioning process. To be specific, the SVM within the proposed positioning system is further optimized by a Genetic Algorithm (GA) [23], and generalized for multi-classification by the One *versus* One procedure. The proposed One *versus* One GA-SVM (OG-SVM) algorithm combined with the SDC method can reasonably cluster the radio map on the basis of coordinates and then classify the RSS sample into sub- regions for coarse positioning.

For another thing, we propose the Kernel PCA feature extraction algorithm based on Principal Component Analysis (PCA) [24–26] for dimensional reduction also as a solution for the asymmetric matching problem. Compared with other typical feature extraction methods such as Linear Discrimination Analysis (LDA) [27,28] and Local Discriminant Embedding (LDE) [29,30] used in positioning systems in our early works [14,15,20], the proposed method performs better in both low dimensional feature extraction and asymmetric matching accuracy when there is an AP outage.

The rest of this paper is arranged as follows: In Section 2, we will describe the structure of the traditional fingerprinting method for indoor positioning. After that, Section 3 starts with the introduction of the proposed new indoor positioning system, followed then by the theoretical analysis of the proposed SDC method with OG-SVM classification procedure and the Kernel PCA feature extraction method. In Section 4 we will provide experimental performances of the proposed methods and make comparisons with other typical algorithms. Section 5 finally presents the conclusions.

2. Fingerprinting Indoor Positioning System

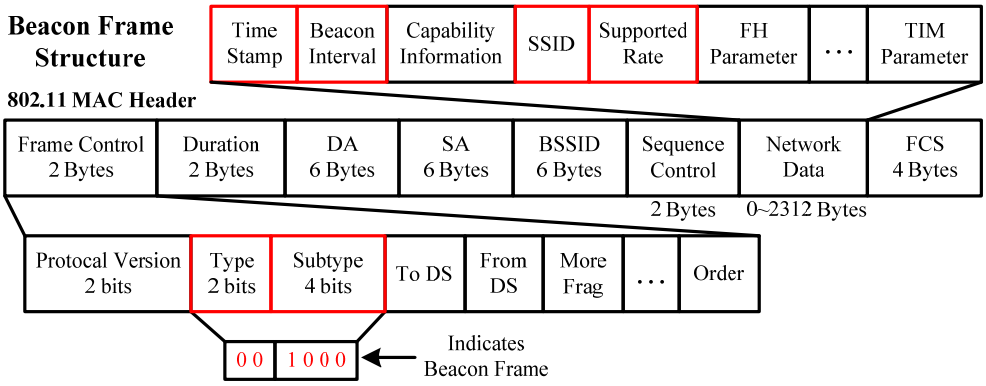
A typical fingerprinting indoor positioning system is introduced in this section. Firstly, an end user takes RSS readings from available APs with his/her (WLAN adapter equipped) device in an indoor environment. The positioning system then estimates the current location of the user based on the measured RSS values by matching the received values with the fingerprint database, which is the pre-stored table of RSS values over a grid of reference points (both their RSS values and location coordinates are recorded) on the positioning area. Therefore the traditional fingerprinting method mainly consists of two parts, which are radio map building and the online matching procedures, respectively.

2.1. Source of Received Signal Strength

It is significant and necessary to briefly introduce where and how the RSS derives, based on which we could better analyze the unstable factors and sources of noise for the radio map. Actually, the RSS values derived from different APs are mainly calculated based on the received beacon frames of the device.

The beacon frame is one of the management frames in IEEE 802.11-based WLANs and its structure is illustrated in Figure 1. It is periodically broadcast and terminal devices in passive scan mode can receive it without building a connection with any AP. The beacon frame is transmitted to announce the presence of a WLAN and includes all supported parameters. After receiving it, according to the information labeled with red rectangles in Figure 1, the terminal device is able to discriminate APs and calculate the RSS values over a sampling period. Specifically, The Beacon Interval is generally set to 100 microseconds; SSID identifies a specific WLAN; Supported Rate is a constant 1 Mbps and Time Stamp normally is used for compensation of interval inaccuracy [31]. Besides, the size of a beacon frame varies, depending on the instant transmitting status. Apart from the parameters presented above and the complexity of indoor propagation, the state of being in connection with an AP or not, the WLAN card, antenna and driver version of a terminal device (sensitivity of the adapter and the manufacturer) [32] also affect RSS values.

Figure 1. Main structure of a Beacon Frame.



2.2. Building Radio Map

Radio map actually is a dataset used to bridge RSS values with location information. By setting amounts of Reference Points (RP), it is able to statistically describe the electromagnetic environment of an indoor positioning area. It is similar to many published researches [12,33] about fingerprinting where building a radio map is composed of two parts, which are sampling RSS values and recording coordinates information, respectively.

Firstly, we sample and record RSS readings at known locations with a mobile terminal device. As presented above, the height and the direction of a device antenna affects the online signals quality which directly influences the system positioning accuracy. For simplicity and concentrating on the proposed algorithms, as a compromise resolution, we only take a holding-in-hand situation (a user is holding the mobile in hand for using the positioning service, therefore the height of the terminal normally is set to 1.2 m) into consideration and take four RPs in four directions (North, South, East and West), respectively, from the same location (the four RPs in four directions share the same coordinates). We denote the RSS values derived from APi at RPj as $\phi_{i,j}(\delta)$, δ = 1,2,...,q, $q \geq 1$ where q stands for the number of collected time samples, the average of the time samples thereby can be computed by:

$$\phi_{i,j} = \frac{1}{q}\sum_{1}^{q}\phi_{i,j}(\delta) \tag{1}$$

where $\phi_{i,j}$ is considered as actual RSS readings (in dBm) of APi at RPj. So the radio map of RSS part is denoted as $\mathbf{\Phi}$:

$$\mathbf{\Phi} = \begin{bmatrix} \phi_{1,1} & \phi_{1,2} & \cdots & \phi_{1,M} \\ \phi_{2,1} & \phi_{2,2} & \cdots & \phi_{2,M} \\ \vdots & \vdots & \cdots & \vdots \\ \phi_{N,1} & \phi_{N,2} & \cdots & \phi_{N,M} \end{bmatrix} \tag{2}$$

where M and N stand for the total number of available APs and RPs respectively. Therefore each row of $\mathbf{\Phi}$, the vector of the matrix, actually represents the RSS values of each RP, which is denoted as:

$$\phi_j = [\phi_{j,1}, \phi_{j,2}, \phi_{j,3}, \ldots, \phi_{j,M}], j = 1,2,\ldots,N \tag{3}$$

Then, the radio map can be denoted as $(P_{xy}{}^{j}, \phi_j)$, j = 1,2,...N, $\phi_j \in \mathbb{R}^M$, where the element $P_{xy}{}^{j}$ is the coordinates of the RPj, which is represented by (x_j, y_j). In the case when no RSS readings can be

detected from several APs at some RPs, the corresponding value is then set to be a minimal value instead of putting a zero because of the subsequent algorithm computation.

In addition, RSS should be collected systematically during different months or seasons which may cause evident RSS fluctuations. In this case, we could improve the system performance by enabling the radio map to store RSS samples of different periods and choose the corresponding database for the online matching process according to the current time which can be obtained from the timestamp of the beacon frame. Also, some extended Location Based Services (LBS) based on user gestures could be discriminated by built-in sensors of the mobile terminal firstly, and then the dedicated radio maps could be selected accordingly to provide the relative services.

2.3. WKNN for Online Matching

Many algorithms are widely used in fingerprinting method for matching the test points (TP) with the radio map, including *K*-Nearest Neighbors (KNN), Kernel Method [34], probabilistic approach [35] and Support Vector Regression (SVR) [15]. However, for simplicity and low complexity, we here take Weight *K*-Nearest Neighbors (WKNN) algorithm for the matching process in the proposed positioning system.

Specifically, in the online phase, a group of RSS readings is sampled by a terminal, and then it is matched with the most likely location by traversing all RPs of the radio map. For measuring the similarity between TP and each RP, WKNN algorithm calculates the distances between the TP and each RP by:

$$D_i = \left(\sum_{j=1}^{M} \left\| \phi_{test,j} - \phi_{i,j} \right\|^p \right)^{\frac{1}{p}}, i = 1,2,\dots,N, j = 1,2,\dots,M \tag{4}$$

where $\phi_{test,j}$ is the received RSS value from AP j of TP, D_i is the Manhattan distance and Euclidean distance when p=1 and 2, respectively. The first K RPs with the shortest distance are chosen to estimate the location of TP. Then the weight for each RP based on distance is defined as:

$$\omega_\zeta = \frac{\sigma}{D_\zeta + \mu}, \text{s.t.} \sum_{\zeta=1}^{K} \omega_\zeta = 1, \zeta = 1,2,\dots,K \tag{5}$$

where is the normalized parameter of the weight, μ is a minimal value set to prevent denominator becomes zero. Finally the output coordinates of the TP can be given by:

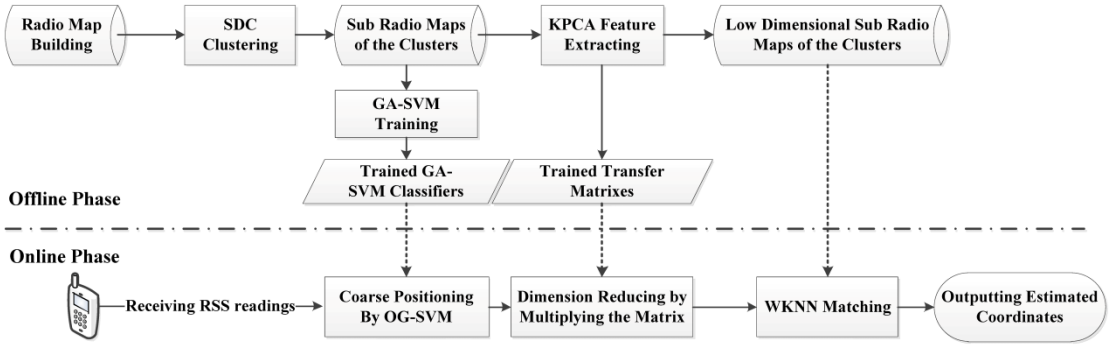
$$P_{xy}{}^{test} = \sum_{\zeta=1}^{K} \omega_\zeta P_{xy}{}^{\zeta}, \zeta = 1,2,\dots,K \tag{6}$$

It is obvious that the dimensionality of a radio map depends on both the number of RPs and quantity of deployed APs. Therefore, in the case of positioning a quite large area with many RPs needs to be set and numerous APs are required for dense coverage, so the size of radio map will be expanded considerably and the computational burden will be increased sharply. Besides, in case of some APs are broken down, the fingerprinting system may be severely damaged or even malfunction due to the missed dimension.

3. New Indoor Positioning System and the Proposed Methods Analysis

The process used by some positioning systems is designed to transmit the RSS to a central server first for subsequent computing and then download coordinates from the server [16]. Different from that, the proposed system is designed to be able to run independently on a mobile terminal without a requirement of being in connection with any AP. But in this case, the trained radio maps and models need to be stored on the mobile terminal. For the purpose of reducing the fingerprint dataset thereby facilitating the mobile terminal resource consumption and improving robustness, the proposed positioning system is designed with two phases, which are the offline and online phase, respectively, and the corresponding flow chart is presented in Figure 2.

Figure 2. Flow chart of the proposed indoor positioning system.



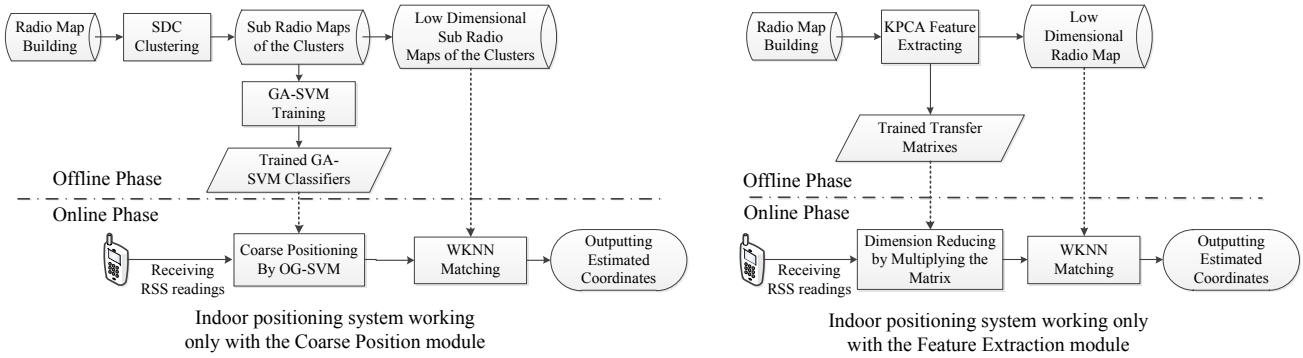
In the offline phase, RSS values are collected evenly on a grid with their coordinates as the radio map of the positioning area. After that the radio map is split into several sub-radio maps based on the SDC method. Then those sub-radio maps are trained by GA-SVM for building the classifiers. Thereafter the Kernel PCA algorithm is applied in each sub-radio map to extract the fingerprinting database into feature space and reduce the dimension of the radio maps. The low dimensional sub-radio maps for each cluster and corresponding trained transfer matrixes derived from the last step would be saved together with the GA-SVM classifiers and transferred to the mobile terminal for online real-time localization.

In the online phase, for real-time positioning, RSS values are measured by the mobile terminal user first. GA-SVM classifiers then will be used for locating the RSS value in the sub-region, which is also known as coarse positioning. Then, the transfer matrix of the sub-region is deployed to transfer the original received RSS values into corresponding low dimensionality in order to match with the low dimensional radio map of the sub-region. Afterwards, the WKNN algorithm is implemented as the precise location estimation method to match the RSS values with the low dimensional sub-radio map. Finally the positioning system outputs the estimated location coordinates.

Moreover, it is worth noting that the computational complexity, positioning error rate and the resource limitations of mobile phones are all comprehensively considered in our proposed system. Therefore most of the computational consumption is handled in the offline phase by a powerful computer processor (*i.e.*, clustering sub-radio maps, training SVM classifiers and generating transfer matrixes), thereby relieving the computational burden introduced by the proposed algorithms in the online stage. Furthermore, the proposed new indoor positioning system is designed to be well modularized

for conveniently adding other functionality modules. For instance, we could independently deploy the SDC with a OG-SVM coarse positioning module or Kernel PCA feature extraction module as two positioning systems, which are shown in Figure 3.

Figure 3. Flow charts of the indoor positioning system with a single module.



3.1. Spatial Division Clustering Method

As presented before, the outliers problem severely influences the coarse positioning accuracy and the integrity of sub-regions. Generally, the outliers only account for a small part of the radio map, but for a large scale radio map, getting rid of all the outliers may not be a reasonable way to proceed. Also, simply changing the class of those outliers to the nearest one may introduce unexpected errors, because, in terms of traditional cluster methods such as K-Means, the cluster centers would be changed accordingly as well.

The proposed SDC algorithm solves the problem by extracting the problem as a clustering process with distance constraints of physical location coordinates. The spatial division algorithm starts with defining the within-class scatter as:

$$S_w{}^c = \sum_{i=1}^{U} \frac{1}{U} (\phi_i^c - \bar{\phi}_c)(\phi_i^c - \bar{\phi}_c)^{\mathrm{T}}, i = 1,2,\dots,U, c = 1,2,\dots,G \tag{7}$$

where $S_w{}^c$ stands for the within-class scatter of the cluster c, and $c \leq G$ where G is the total number of possible clusters. $U, U \leq N$ is the total number of RPs that belongs to the cluster c. ϕ_i^c are those vectors (RSS values) of the RPs within the cluster c, and $\bar{\phi}_c$ is the mean value of the counterpart, which can be given by:

$$\bar{\phi}_c = \frac{1}{U} \sum_{i=1}^{U} \phi_i^c \tag{8}$$

After that the between-class scatter is defined as:

$$S_B{}^c = \frac{1}{G} \sum_{j=1}^{G} (\bar{\phi}_c - \bar{\phi}_j)(\bar{\phi}_c - \bar{\phi}_j)^{\mathrm{T}}, \quad c = 1,2,\dots,G, c \neq j \tag{9}$$

where S_B^c stands for the Between-class scatter of the cluster c, and $\bar{\phi}_j$ is the mean value of the RPs within the cluster j. Actually, $S_w{}^c$ is the covariance matrix of the zero mean vectors assigned to the cluster c while the $S_B{}^c$ is the covariance matrix of the cluster means, and the purpose of the proposed

clustering algorithm is to optimize the radio between the within-class scatter S_w and the between-class scatter S_B, which is denoted as Q, hence the objective function can be expressed as:

$$\text{argmin} \sum_{c=1}^{G} Q_c = \text{argmin} \sum_{c=1}^{G} \frac{S_w{}^c}{S_B{}^c} \tag{10}$$

The definitions of the within-class scatter and the between-class scatter are primarily derived from the Fisher Criterion which is used in LDA. The proposed clustering algorithm for indoor positioning employs the minimum radio between S_w and S_B as the criterion mainly because of the fact that the RPs nearby each other would share the same spatial structure, which means that RPs within same class are supposed to be nearby each other and a within-class scatter should be as small as possible, while on the contrary RPs in different classes are supposed to be far away from each other and the between-class scatters should be as large as possible.

Therefore maximizing the similarity meanwhile minimizing the difference may effectively cluster the RPs. Different from the traditional clustering methods, taking S_w / S_B as the measurement not only considers the distance between the independent RPs and updating the coefficient or cluster center, but also takes the similarity between classes into account. Instead of maximizing the value of the radio Q with classic convex optimization methods, the proposed algorithm previously assigns each two continuous RPs as a minimum class. It takes Q as the property of each class and runs clustering procedures in four steps as follows.

Step 1: Clustering centers determination

The radio Q of each class can be computed by:

$$Q_c = \frac{S_W{}^c}{S_B{}^c} = \frac{G\sum_{i=1}^{U}\left(\phi_i^c - \bar{\phi}_c\right)\left(\phi_i^c - \bar{\phi}_c\right)^{\mathrm{T}}}{U\sum_{J=1}^{U}\left(\bar{\phi}_c - \bar{\phi}_j\right)\left(\bar{\phi}_c - \bar{\phi}_j\right)^{\mathrm{T}}} \tag{11}$$

where G here equals to $N/2$ (in case of N is not divisible by 2, G equals to $(N-1)/2$ and the last 3 RPs assigned to a class). Then calculating the similarity of each pair of Q, hence the similarity between one class and all others is referred to as:

$$S_Q{}^c = \sum_{i=1}^{G} \left\| Q_c - Q_i \right\|, \qquad c = 1,2,\cdots,G, i \neq c \tag{12}$$

The Q of class c corresponding to the $max\ S_Q{}^c$ is chosen as the first cluster center which is denoted as Ctr^1. Then we compare all the other Q with the Ctr^1 and find the one with the lowest similarity (*i.e.*, to find $max\|Q_c - Ctr^1\|, c = 1,2,\ldots,G$) as the second cluster center Ctr^2. For the third center and so on, the similarity is calculated in advance, namely:

$$S_Q{}^{ij} = \left\| Q_i - Ctr^j \right\|, i = 1,2,\cdots,G, j = 1,2,\cdots,E \tag{13}$$

where E is the number of centers have been set. Therefore the next most suitable center with the least similarity can be set by min $S_Q{}^{i,j}$, hence the $(E+1)$th center is the Q of class i subjected to $max\{min(S_Q{}^{i,1}, S_Q{}^{i,2}, \ldots, S_Q{}^{i,E})\}. i = 1,2,\ldots,G$.

Step 2: Combination of clusters

Based on the centers derived from the previous step, the following process is to calculate similarity between each class and its centers, where Equation (13) is deployed here. Then the class is assigned to the most similar center in turn. Meanwhile, Q of the center will be updated by Equation (11) after each class is allocated in. If the total number of centers E is assigned, then E clusters will be formed consequently.

Step 3: Splitting of the clusters

In order to meet the condition that no outliers in positioning area after the radio map is clustered, RPs within a class is supposed to be subjected to the criterion:

$$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \le \varepsilon \tag{14}$$

where $x_i y_i, x_j y_j$ are any two sets of coordinates of RPs within a same class, and ε is the distance threshold based on the density of sampled RPs and location environment. Different from the combination process based on the signal features, the splitting process depends on the coordinates information (which is another part of radio map), namely:

$$\mathrm{Loc} = \left(\mathrm{P}_{xy}{}^{1}, \mathrm{P}_{xy}{}^{2}, \cdots, \mathrm{P}_{xy}{}^{N}\right) \tag{15}$$

Denoting the coordinates information of cluster C as:

$$\mathrm{Loc}^{\mathrm{C}} = \left(\mathrm{P}_{xy}{}^{\mathrm{C1}}, \mathrm{P}_{xy}{}^{\mathrm{C2}}, \cdots, \mathrm{P}_{xy}{}^{\mathrm{C}U}\right) \tag{16}$$

where $\mathrm{P}_{xy}{}^{C}$ stands for the coordinates information of the RPs belonging to the cluster C, and U here is the total number of RPs belong to the cluster C. Then the procedures of cluster splitting are addressed as follows:

a. **Initialization**: Initialize the $\mathrm{P}_{xy}{}^{C1}$ as an element of new cluster C_1, where C_1 is considered as the first sub cluster of C.

b. **IF** $\mathrm{P}_{xy}{}^{C2}$ satisfy the criterion Equation (14) with $\mathrm{P}_{xy}{}^{C1}$, **THEN** assign it to C_1.
ELSE set the $\mathrm{P}_{xy}{}^{C2}$ as an element belongs to a new cluster C_2.
End IF

c. **FOR** $\mathrm{P}_{xy}{}^{Ci}, i = 3,4, \ldots, U$
IF $\mathrm{P}_{xy}{}^{Ci}$ meets the criterion Equation (14) with $\mathrm{P}_{xy}{}^{Cj}, j = 1,2, \ldots, i-2$, **THEN** assign $\mathrm{P}_{xy}{}^{Ci}$ to the cluster which $\mathrm{P}_{xy}{}^{Cj}$ belongs to.
ElSE IF $\mathrm{P}_{xy}{}^{Ci}$ meets criterion Equation (14) with more than one $\mathrm{P}_{xy}{}^{Cj}$, **THEN** combine the clusters corresponding to those $\mathrm{P}_{xy}{}^{Cj}$ with the $\mathrm{P}_{xy}{}^{Ci}$ as a new cluster. $\mathrm{P}_{xy}{}^{Ci}$ works as bridge connection.
ELSE set a new cluster with $\mathrm{P}_{xy}{}^{Ci}$ as an element.
END FOR

For special requirements of the indivisible sub region, we could assign the RPs within that region as an independent cluster without participating in the combination and splitting steps.

Step 4: Outputs of clustering

Looping step1 to Step 3 until the number of output clusters comes to convergence, and then the clusters are formed. For some of the small clusters, they could be simply assigned into the nearest clusters. Finally the whole SDC method process is completed.

3.2. Classification by OG-SVM

3.2.1. Introduction of SVM in the Positioning System

OG-SVM is deployed to distinguish the TP to which cluster it belongs to, and locate it in the sub-region for the coarse location process. An introduction to SVM deployment in positioning is briefly given first. Denoting (ϕ_i,L_i), i = 1,2,...,N, $\phi_i \in \mathbb{R}^M$ (according to the experimental positioning environment, N here is the total number of RPs of two clusters) as the set of training samples, where ϕ_i is the vector of RP as mentioned before, and $L_i \in \{1, -1\}$ labels which class the vector belongs to. The purpose of SVM is to obtain the weight vector $\mathbf{w}$ and the scale b, such that:

$$L_i\left(\langle \mathbf{w} \cdot \phi_i \rangle + b\right) \geq 1 \tag{17}$$

where $\langle \mathbf{w} \cdot \phi_i \rangle$ stands for the inner product of the vectors $\mathbf{w}$ and ϕ_i. $\langle \mathbf{w} \cdot \phi_i \rangle + b$ is the so called hyper-plane that enables the training samples with the same label separate with others. In the case of nonlinear condition, a slack variable is introduced and denoted as $\zeta_i \geq 0$, i = 1,2,...,N, so Equation (17) is converted to:

$$L_i\left(\langle \mathbf{w} \cdot \phi_i \rangle + b\right) \geq 1 - \xi_i \tag{18}$$

The objective function is:

$$\min\left(\frac{1}{2}\langle \mathbf{w} \cdot \mathbf{w} \rangle + C\sum_{i=1}^{N} \xi_i\right) \tag{19}$$

where C is the key penalty parameter and element $\sum_{i=1}^{N} \xi_i$ defines maximum number of training errors. Also the inner product $\langle \phi_i \cdot \phi_j \rangle$ is replaced by kernel function, which is expressed as $K\langle \phi_i \cdot \phi_j \rangle$. The kernel methods are able to map the nonlinear dataset into a high (even infinite) dimensional feature space from which the dataset could be linearly separable. Radial basis function (RBF) is one of the kernel methods and is adopted in the proposed positioning system, which is defined as:

$$K\left(\phi_i, \phi_j\right) = \exp\left\{- g\left\|\phi_j - \phi_i\right\|^2\right\} \tag{20}$$

where g is another key parameter geometrically defining the width of the RBF. This might lead to the over-fitting problem if g is relatively small, while on the contrary, the flexibility and robustness might be weakened.

Lastly, the decision function or so called SVM classifier of the indoor positioning system can be obtained as:

$$f(x) = sign\left(\langle \mathbf{w}^* \cdot \phi \rangle + b\right) \tag{21}$$

Where **w*** is the solution of the optimal separating hyper-plane (OSH) that enables the samples with different labels to be most distinguishable, ϕ is the vector of a test point with unknown class label, and the output of the function will decide which class it belongs to (positive result decides one class and negative output decides another one).

3.2.2. Genetic Algorithm for SVM Optimization

Although SVM theoretically is a quadratic optimization problem and the optimal solution is given, the parameters C in Equation (19) and g in Equation (20) still need to be chosen properly due to reasons mentioned before. Therefore GA is integrated into the SVM training process to adjust the two parameters adaptively.

The Genetic Algorithm is derived from the bionic process in which a population evolves by competing with others and preserving its superiority in Nature. Each individual in a population would be eliminated for its weak adaptability or kept due to its strong performance. Consequently the new generation becomes more robust and adaptive.

GA is able to search a large solution space efficiently by adopting probabilistic transition procedure mechanics. It mainly includes three steps, which are selection, crossover and mutation. To be specific, selection is aimed at electing the optimal individuals for reproducing the next generation; Crossover is applied for exchanging information, thereby preserving and collecting the genetic advantage; Mutation is designed to introduce the variation for making new individuals. In terms of GA-SVM, the fitness function is defined as:

$$\min F(C,g) = \frac{1}{1+\kappa} \tag{22}$$

where κ is the classification accuracy rate. The searching space of the parameter g is defined by $\min\left\|\phi_j - \phi_i\right\|^2 \times 10^{-3}$, $\max\left\|\phi_j - \phi_i\right\|^2 \times 10^3$ while the counterpart of C is (0, 10). Generally, after randomly initializing the population, the fitness of each individual is calculated by Equation (22). Then a probability will be assigned to each individual according to the fitness (higher fitness value with higher probability). After that, new individuals are generated by the crossover and mutation operations. The whole process would be repeated until the new individual meets the preset values. Finally with N-fold cross validation (*i.e.*, training data is separated into N parts, one of which is deployed for validating accuracy while the remaining parts are the training sets, and the procedure is taken by N turns), the optimal combination of the parameters (C^*, g^*) can be obtained.

3.2.3. OG-SVM Method

Due to the fact that generally more than two clusters (or sub-regions) exist within an indoor positioning area, One *versus* One GA-SVM is adopted as the classification algorithm to deal with the multiple classes. Instead of deploying a multiple-class SVM, the OG-SVM method sets a group of binary-class SVM classifiers optimized by GA to perform the classification. To be specific, supposing that there are G clusters in the positioning region, there are $G(G-1)/2$ SVM classifiers that can be obtained after training each two clusters as a group with GA-SVM. In term of classifying a test point, it will be put into all SVM classifiers in turn. If it goes to the cluster c, c = 1,2,…,G, then cluster c gets

1 vote. Consequently the test point belongs to the cluster with most votes and thus the corresponding sub-region can be located.

3.3. Dimensionality Reduction by Kernel PCA

Kernel PCA is used in the proposed indoor positioning system to extract the features of the radio map and reduce its dimensionality. An analysis on Kernel PCA is presented below.

As denoted before in the proposed positioning system the RSS values of a cluster is given by $\mathbf{\Phi}_c = \{\phi_1, \phi_2, \ldots, \phi_U\}$, where U is the total number of vectors belong to the cluster c. In order to meet the constraint of PCA, vectors of $\mathbf{\Phi}_c$ has to be decentralized previously. Defining the nonlinear mapping $\partial: \mathbb{R}^M \to \mathcal{F}$ where $\mathbb{R}^M$ is the Euclidian space of samples and $\mathcal{F}$ is the feature space where inner product can be computed by a kernel function. Then the covariance matrix of the samples in feature space can be given by:

$$\overline{C} = \frac{1}{U}\sum_{i=1}^{U} \partial(\phi_i)\partial(\phi_i)^{\mathrm{T}} \tag{23}$$

Denoting λ and $\mathbf{v}$ as the eigenvalue and the eigenvector of $\bar{C}$ respectively, then the eigen-decomposition can be given as:

$$\lambda \mathbf{V} = \overline{C}\mathbf{V} \tag{24}$$

Based on the fact that the eigenvector $\mathbf{v}$ can be expressed in linear spanning space of $\partial(\phi_i),\ i = 1,2,\ldots U$, namely:

$$\mathbf{V} = \sum_{i=1}^{U} \eta_i \partial(\phi_i) \tag{25}$$

where η_i is the weight coefficient for each $\partial(\phi_i)$, we could substitute Equation (25) into Equation (24) and by pre-multiplying $\partial\left(\hat{\phi}_j\right)^T$, $j = 1,2,\ldots U$, then the equation can be given as:

$$\begin{gathered} \lambda\left(\sum_{i=1}^{U} \eta_i \partial(\phi_j)^{\mathrm{T}} \partial(\phi_i)\right) = \partial(\phi_j)^{\mathrm{T}} \cdot \frac{1}{U}\sum_{l=1}^{U} \partial(\phi_l)\partial(\phi_l)^{\mathrm{T}} \cdot \sum_{i=1}^{U} \eta_i \partial(\phi_i) \\ \lambda\left(\sum_{i=1}^{U} \eta_i K_{ji}\right) = \frac{1}{U}\sum_{l=1}^{U}\sum_{i=1}^{U} \eta_i K_{ji} K_{li} \end{gathered} \tag{26}$$

and the equation can be further expressed as $\lambda(\mathbf{K\eta})_j = \frac{1}{U}(\mathbf{K^2\eta})_j$, where $\mathbf{K} = \left[K(\phi_i, \phi_j)\right]_{U\times U}$, $\mathbf{\eta} = (\eta_1, \eta_2, \ldots, \eta_n)^{\mathrm{T}}$. Consequently it can be converted to:

$$\hat{\lambda}\mathbf{\eta} = \mathbf{K\eta} \tag{27}$$

Where λU is substituted by $\hat{\lambda}$. After eigen decomposition, denoting $\lambda^1, \lambda^2, \ldots, \lambda^U$ are the eigenvalues and $\mathbf{\eta}^1, \mathbf{\eta}^2, \ldots, \mathbf{\eta}^{\mathrm{U}}$ are the eigenvectors of $\mathbf{K}$ respectively, therefore the i-th eigenvalue and eigenvector can be given by:

$$\lambda_i = \frac{\hat{\lambda}^i}{U},\ \mathbf{V}_i = \sum_{j=1}^{U} \eta_j^i \partial(\phi_j) \tag{28}$$

where η_j^i is the *j*-th element of $\boldsymbol{\eta}^i$, $i = 1,2,...,U$. Hence, the projection of a test sample ϕ on *j*-th axis of the feature space is represented by:

$$\partial(\phi)^{\mathrm{T}}\mathbf{V}_j = \Delta^j \sum_{i=1}^{U} \eta_i^j \partial(\phi)^{\mathrm{T}} \partial(\phi_i) = \Delta^j \sum_{i=1}^{U} \eta_i^j K(\phi, \phi_i) \tag{29}$$

where Δ^i is a normalized factor computed by equation $(\mathbf{V})^{\mathrm{T}} \cdot \mathbf{V} = 1$. By adopting the maximum first *d* eigenvalues $\hat{\lambda}^1, \hat{\lambda}^2, \ldots, \hat{\lambda}^d$ and their corresponding *d* eigenvectors $\boldsymbol{\eta}^1, \boldsymbol{\eta}^2, \ldots, \boldsymbol{\eta}^d$ where $d \ll U$, the high dimensional dataset can be accordingly reduced to *d* dimension.

After defining the radio map of cluster *c* as $\boldsymbol{\Phi}_c^o = \left(\mathbf{P}_{xy}^i, \phi_i^o\right), i = 1,2,\ldots,U, \phi_i \in \mathbb{R}^M$ and its low dimensional counterpart as $\boldsymbol{\Phi}_c^o = \left(\mathbf{P}_{xy}^i, \phi_i^o\right),\ i = 1,2,\ldots,U,\ \phi_i^o \in \mathbb{R}^d$, the transfer matrix of the region can be expressed as:

$$\mathbf{M} = \left(\Delta^j \sum_{i=1}^{U} \eta_i^j K(\phi, \phi_i) \right), j = 1,2,\cdots,d \tag{30}$$

To conclude, in the offline phase of the positioning system, a low dimensional radio map for each cluster is generated by deploying the Kernel PCA algorithm with RBF aligned with the kernel function used in SVM. In the online phase, after a test point is located to a cluster by OG-SVM, the corresponding low dimensional radio map will be chosen accordingly. Therefore, a downsized test point after being decentralized can be computed by Equation (30) (*i.e.*, running Equation (29) *d* times for *d* axis or *d* dimensions) and expressed as $\phi_i^o = [\phi_{i,1}, \phi_{i,2}, \ldots, \phi_{i,d},]$. Moreover, the transfer matrix could be integrated or further compressed by mathematic methods [36,37]. The WKNN algorithm will finally be deployed as the measuring method for matching the ϕ_i^o throughout the radio map $\boldsymbol{\Phi}_c^o$ thereby obtaining the estimated coordinates.

4. Implementation and Performance Analysis

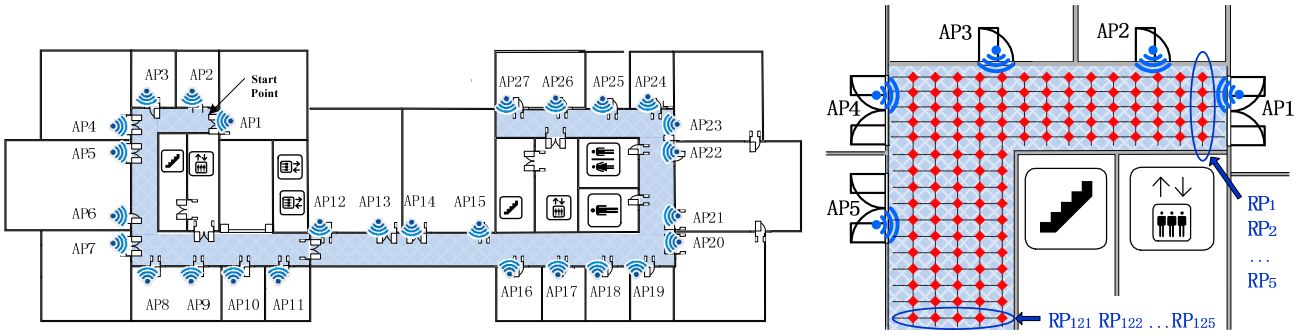
In general, the proposed indoor positioning system runs as following procedures: for the offline phase, firstly, we start by constructing the radio map. Secondly, we cluster it into several sub-radio maps by the SDC method. The third step is to train the sub-radio maps with OG-SVM, generating classifiers. Then, the following step is to reduce the dimension of each sub-radio map by Kernel PCA and generate the corresponding transform matrixes. For the online phase, firstly, we classify the test point to the sub-region by the OG-SVM method with those classifiers. After that the dimensions of the test point are reduced by the matrix generated offline. The final steps are matching the low dimensional test points with the low dimensional sub-radio maps by WKNN, and outputting the estimated coordinates. In this Section the experimental evaluations of the proposed method for indoor positioning system are elaborated in detail.

4.1. Indoor Positioning Environment

Figure 4 shows a floor plan of a research center. The fingerprint dataset was carefully measured in this typical office environment. The proposed indoor positioning system is built here with 27 Access Points (marked as AP1-AP27) located evenly in each room. Then we individually sample and record the RSS readings 100 times at each reference point (with a sampling rate of 2 times per second) with a

mobile terminal. The area of interest colored with blue is the corridor part (49.4 m × 14.1 m), where 828 locations are equally distributed as the experimental RPs.

Figure 4. Floor plan for the indoor positioning experiment and reference point setting.



4.2. Cluster Performance of SDC Method

In this subsection, the proposed SDC method is evaluated well in terms of both radio map division and positioning accuracy for indoor localization. K-Means and Fuzzy C-Means (FCM) algorithms are also implemented for verifying the analysis and testing the performance by comparison.

As shown in Figures 5 and 6, the Radio Map is clustered into six (marked as F1-F6 and K1-K6 respectively) sub-areas by deploying FCM and K-Means algorithm, where different colors represent different sub-regions and the black points stand for the outliers. In addition, the white blanks among the RPs are obstacles in the building where RSS cannot be tested.

Figure 5. Positioning area clustered by the FCM algorithm.

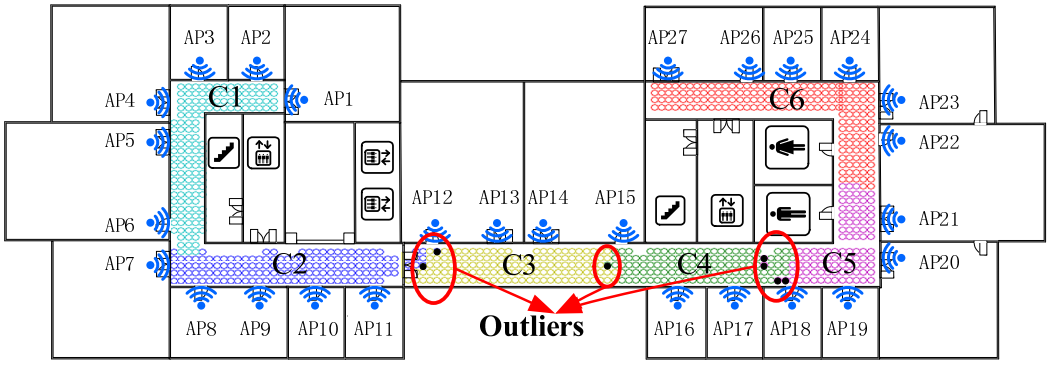
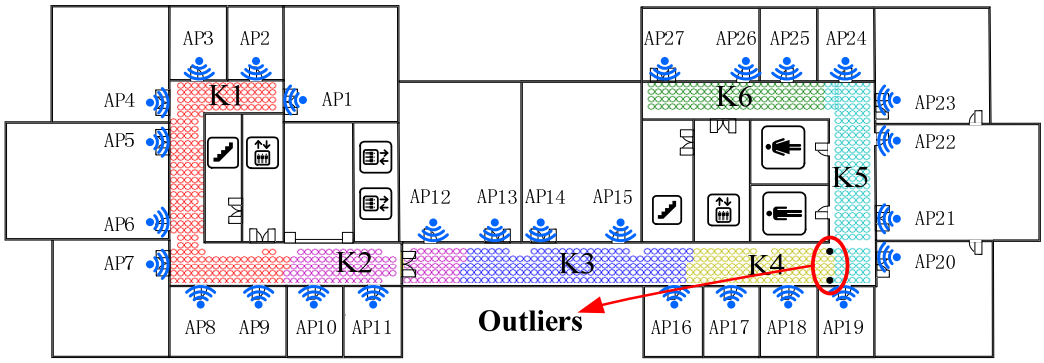
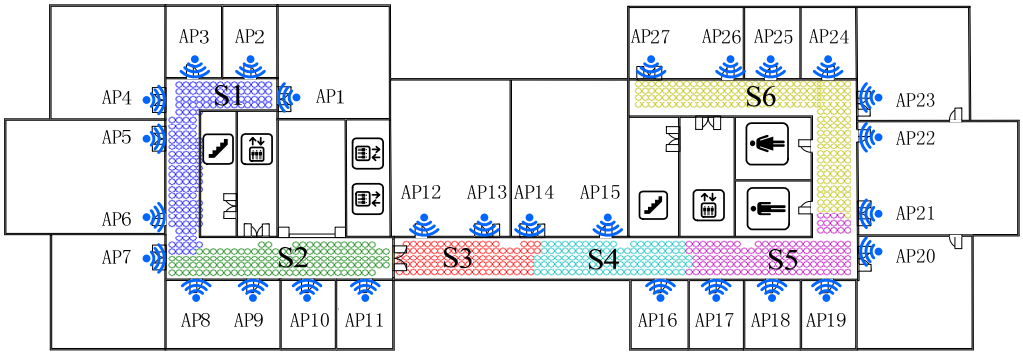


Figure 6. Positioning area clustered by the K-Means algorithm.



This demonstrates that, for clustering using FCM, the radio map is divided almost symmetrically but the outliers are distributed mainly in the middle three clusters and account for nearly 1% (7/828) of RPs, while for K-Means clustering, the divided sub-regions are slightly unbalanced in term of RP quantity, but few outliers exist in those regions. It is worth noting that the RPs are sampled on the grid evenly, and the experimental environment is relatively stable (few people walk around and all windows are closed). In this case, the outliers are supposed to be far less than in a practical environment. The proposed SDC method divides the interesting area as illustrated in Figure 7, where different regions are marked as S1-S6 with different colors. Compared with the other two algorithms, the SDC method is able to cluster the RPs more symmetrically without any outliers problems.

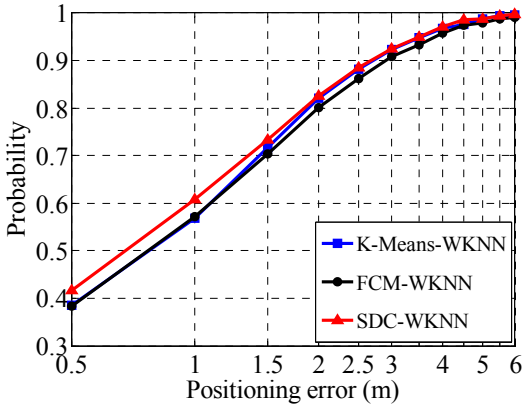
Figure 7. Positioning area clustered by the SDC algorithm.



Actually, dividing the radio map symmetrically may not prove that the clustering method is effective and suitable. Nevertheless, the structure of the experimental region is nearly balanced, building materials are almost uniform and all APs are arranged evenly. Therefore, in this case, clustering the RPs in a symmetric way is supposed to be more reasonable. Besides, the boundaries of each cluster are located near the corner or doors where RSS values normally fluctuate and are more distinguishable. It also demonstrates the reliability and effectiveness of the proposed SDC method based on the divided structure.

In order to verify the performance of different clusters in term of positioning accuracy, the WKNN method is directly deployed to all divided sub-regions for fingerprint localization based on the three clustering cases without considering coarse positioning (*i.e.*, assuming that which sub-region a TP belongs to is known). The fine positioning accuracy is shown in Figure 8, where the FCM method achieves a Confidence Probability (CP) over 80% with a positioning error (PE) within 2 m. For the K-Means algorithm, the CP is 2% better than the counterpart of the FCM. It is notable that the positioning accuracies are calculated for each region independently, and then added together with weights of RPs numbers of a cluster. The performance of the proposed SDC method is the same as that of the K-Means as PE equals 2 m too, but it is slightly superior to other algorithms when the PE is 1 or 1.5 m. Therefore, the proposed SDC method is better than other clustering methods for indoor localization due to its better positioning performance.

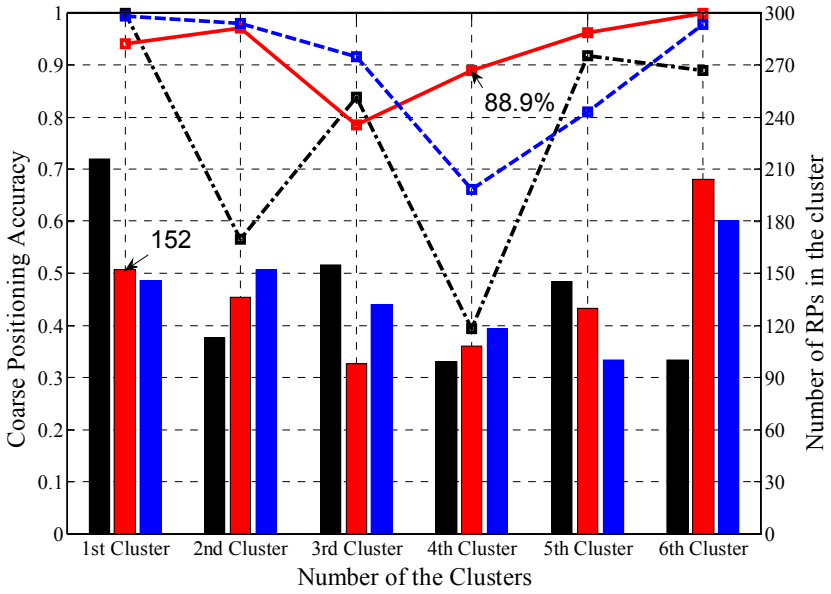
Figure 8. Positioning accuracies based on three different clustering methods.



4.3. Coarse Positioning Performance of the OG-SVM Method

Coarse positioning is responsible for allocating received RSS readings to the sub-regions where they belongs. The integrated information of the coarse position for the three clustering methods is demonstrated in Figure 9, where the black, red and blue bars represent the number of RPs in the regions clustered by K means, SDC and FCM, respectively, while the black, red and blue lines stand for the coarse positioning accuracies in the regions clustered by K Means, OG-SVM and FCM, respectively. For example, the first region (labeled as S1 before) clustered by SDC consists of 152 RPs, and OG-SVM coarse positioning accuracy of the S4 region is 88.9%. It clearly shows the distribution of RPs in all six regions and the classification accuracy for each cluster and each clustering method.

Figure 9. The different clustering results and the coarse positioning performances for the three methods.



To be more specific, the coarse positioning accuracy based on the FCM algorithm for each cluster is listed in Table 1, while the coarse positioning accuracy of the K-Means algorithm is shown in Table 2. The overall classification (*i.e.*, coarse positioning) accuracy of FCM is about 10% higher than the

K-Means (90.58% and 81.04%, respectively). Therefore, even if few outliers appear in the K-Means clusters which performs better than FCM, in terms of the coarse positioning accuracy it actually shows a reverse outcome.

Table 1. Coarse Positioning performance of FCM method.

Clusters	Number of RPs	Number of TPs Classified Correctly
C1	146	145
C2	152	149
C3	132	121
C4	118	78
C5	100	81
C6	180	176
Classification accuracy: 90.58%		

Table 2. Coarse Positioning performance of K-Means method.

Clusters	Number of RPs	Number of TPs Classified Correctly
K1	216	216
K2	113	64
K3	155	130
K4	99	39
K5	145	133
K6	100	89
Classification accuracy: 81.04%		

Besides, both tables show that the coarse positioning accuracy of the first and the last clusters are much higher than the clusters in the middle. According to the experimental results and previous analysis of the RSS database, it can be deduced that classification criterion based on the cluster centers, which is used by FCM and K-Means, runs well in the areas with distinguishable RSS values, but may not classify the TPs efficiently in the regions where RSS change stably or fluctuate within a narrow range.

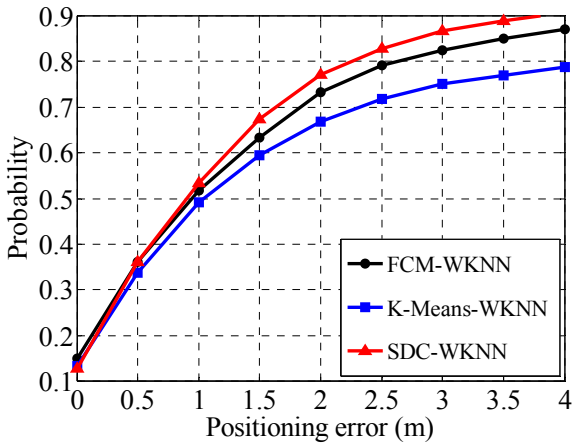
Compared with the two traditional clustering algorithms, K-Mean and FCM, the coarse positioning based on SDC with OG-SVM performs better, as shown in Table 3. Specifically, the classification accuracy of the proposed method is 93.84%, which is 12.80% greater than the result of K-Means and 3.26% higher than the FCM, while no outliers occur.

Table 3. Coarse Positioning performance of SDC method.

Clusters	Number of RPs	Number of TPs Classified Correctly
S1	152	143
S2	136	132
S3	98	77
S4	108	96
S5	130	125
S6	204	204
Classification accuracy: 93.84%		

Taking the coarse positioning procedure into the fingerprinting system (which actually is the single module system shown on the left of Figure 3), the advantage of the proposed SDC and OG-SVM method would be more apparent. As illustrated in Figure 10, the final estimated positioning accuracy of the proposed method is 77.4% under the condition that the positioning error is within 2 m. Compared with the 73.3% positioning accuracy of FCM and the 66.9% of K-Means under the same conditions, the proposed coarse positioning method is more effective and precise, thereby ensuring the following fine positioning procedure. Besides, according to extended experimental results, the coarse location accuracy of the proposed method can be further improved with more training samples in the OG-SVM, also clustering the radio map into a smaller number of regions by the proposed method may yield a better performance.

Figure 10. Positioning accuracies based on Coarse Positioning procedure.

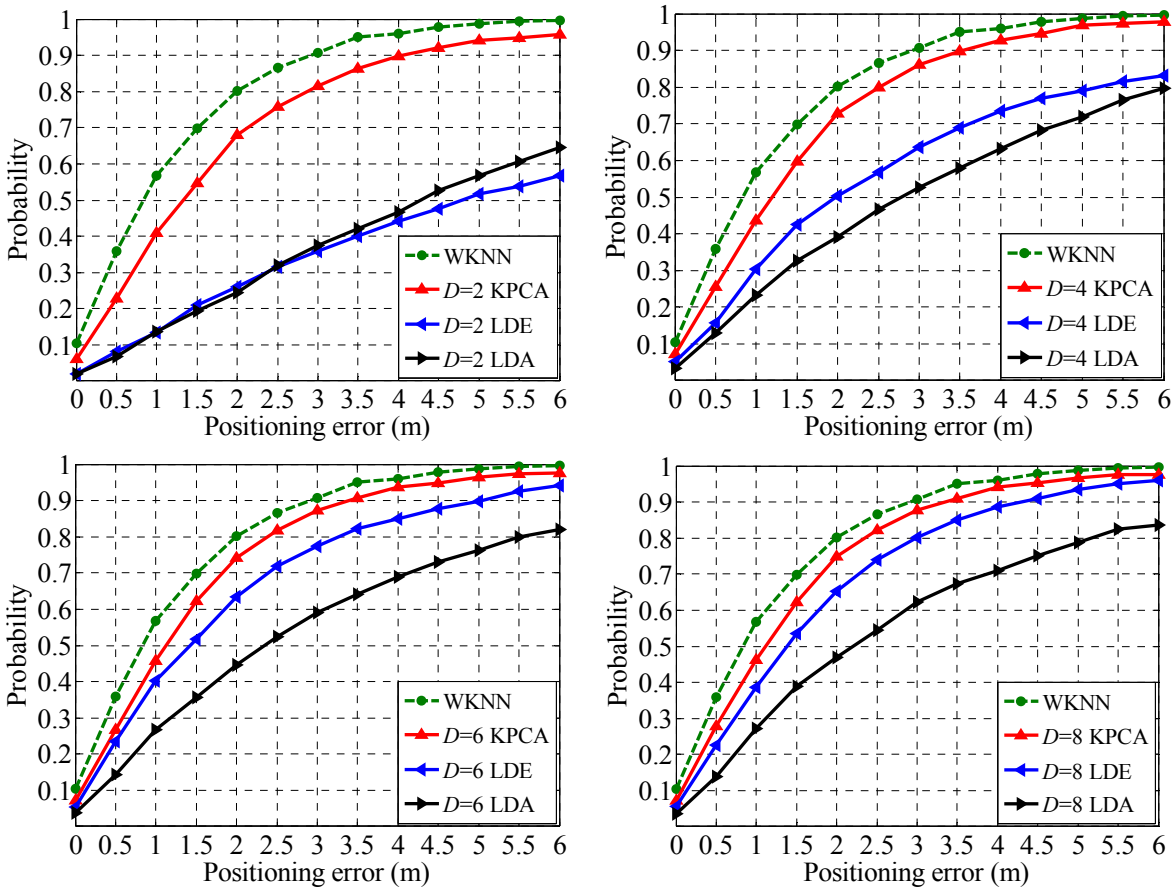


4.4. Low Dimensional Performance of Kernel PCA Method

Theoretically, feature extraction algorithms are able to improve the positioning accuracy by learning the inner structure of the dataset and eliminating part of the noises normally with a high dimension [14,15], but in this paper we focus on the capacities of different algorithms in very low dimensionality scenarios. As a direct evaluation of the low dimensionality performance of different feature extraction algorithms, Figure 11 demonstrates that the relationship between Confidence Probability (CP) and the Positioning Error (PE) distance. Specifically, the green dashed line represents the performance of the WKNN fingerprinting method with full dimensionality (27 dimensions for 27 APs), the red line stands for the performance of WKNN fingerprinting after dimensional reduction by the KPCA method. Similarly, the green and black lines represent the counterparte of the LDE and LDA methods, respectively.

As typical linear and manifold feature extraction methods, both LDE and LDA show significant properties in many pattern recognition aspects, however, in terms of extracting eigen-features within an indoor radio map, the Kernel PCA method reveals a better fitness, because of the fact that in the cases of D = 2, 4, 6, 8 where D stands for the dimensionality, the Kernel PCA method shows more outstanding performance according to the experimental result shown in Figure 11.

Figure 11. Positioning accuracies comparison between methods in the cases of $D = 2$, $D = 4$, $D = 6$ and $D = 8$, respectively.



As shown in Figure 11, the WKNN method achieves a CP of about 80% under the condition that PE is within 2 m. Compared with other algorithms, along with the increasing dimensionality, CP of the Kernel PCA approaches the WKNN faster. Therefore the proposed method outperforms other algorithms in a low dimensionality situation. For example, the CPs of LDA and LDE are 39.2% and 50.1%, respectively, under the condition that $D = 4$ and PE is within 2 m. the performance of the proposed Kernel PCA reaches up to 72.5%, which is less than the dimension-unreduced WKNN method, but far more competitive than others. Moreover, in this case the size of the radio map for online matching process is reduced 85% (calculated by $(1 - 4/27)$).

In addition, the number of nearest neighbors K also affects the WKNN positioning accuracy in this situation. We set the optimized value of K as 4 based on experiments. It is also worth noting that the WKNN method is supposed to perform best in an ideal experimental environment (small noise intensity) because compared with other dimension-reduced methods, it works on full dimensionality with all the radio map information. Dimensionality reduction actually implies that part of the information has to be lost though a comprehensive preprocessing has been done before in the feature extraction procedure.

4.5. Asymmetric Matching of the Kernel PCA Method

It is unavoidable that outages might occur occasionally, in which case the WKNN fingerprinting method is drastically affected and even fails to work. Taking the WKNN method as experimental counterpart, we assign the missed dimension as a group of minimum value. Then, according to Figures 12 and 13 below, under the condition that PE is within 2 m, the CPs of the WKNN method are 58.3%, 56.8% and 64.4% when the 6th AP, 12th AP and both 4th 24th APs is/are powered off, respectively. Generally, CP declines sharply about 20% compared with the case that all APs run well.

Figure 12. Positioning accuracies comparison when D = 6 and D = 8 respectively in the case of 6th AP outage.

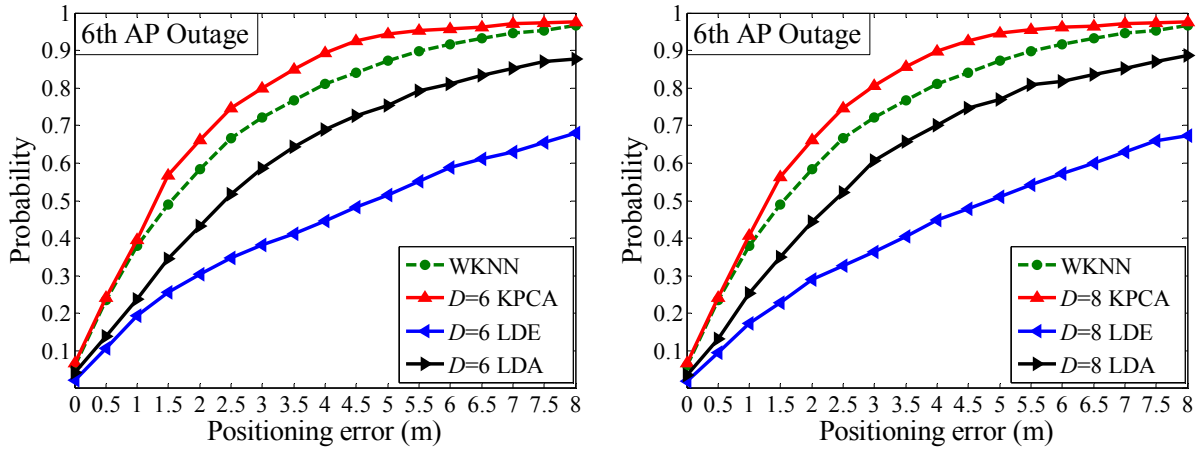
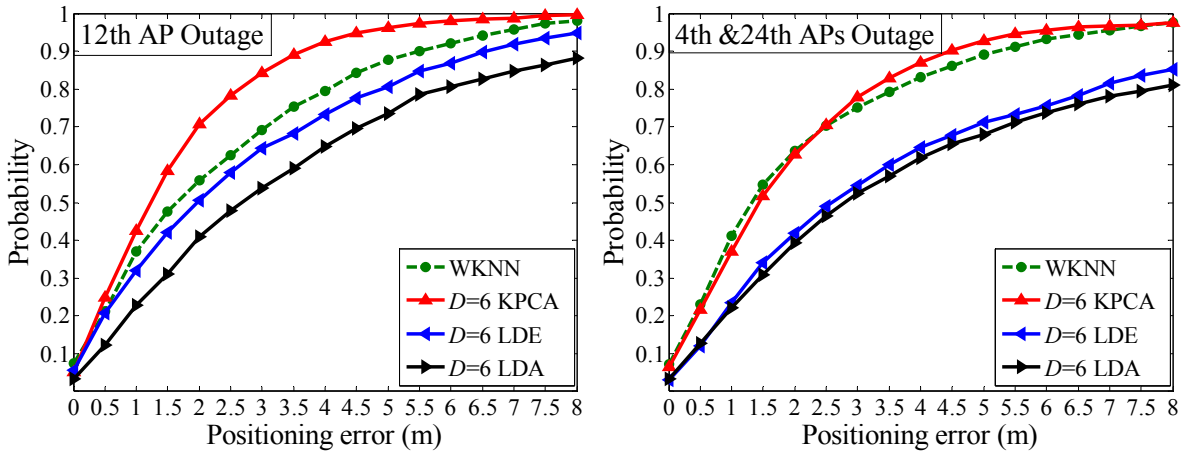


Figure 13. Positioning accuracies comparison in the cases of 12th AP outage and both 4th, 24th APs outage respectively.

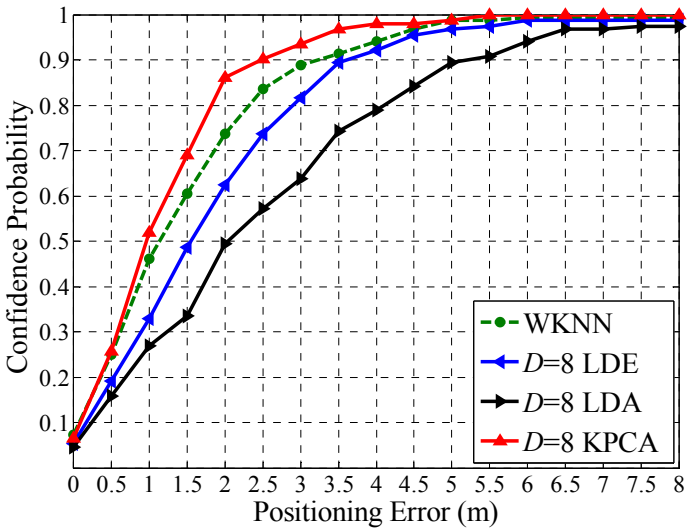


However, the proposed Kernel PCA method is far less affected by AP outages than the WKNN and other methods. For instance, with the situation that D = 6 and PE is within 2 m, it only declines 4% of CP when the 6th AP is powered off. Also, it keeps CP over 60% in all three cases (6th AP outage, 12th AP outage, both 4th and 24th APs outage). Specifically, under the condition that D = 6 and PE is within 2 m, the CPs of Kernel PCA method are 66.3%, 71.5% and 62.5%, respectively, which ranks top in the first two cases and slightly less than the WKNN method in the last case.

Besides, Figure 12 also illustrates that, in the case of one missing dimension (6th AP outage), the CPs are less affected by different target dimensionality (D = 6 or D = 8) in terms of the three feature extraction methods. This could mainly be attributed to the fact the lost information of one dimension is more significant, whereas the number of reduced dimensions plays a less important role. Moreover, in terms of the LDE and LDA methods, both of their CPs are less than either of the WKNN or Kernel PCA method, but it is worth noting that normally LDE performs better than LDA without AP outages, however the LDA surpasses the LDE in the case of 6th AP outage, and comes close to it when the 4th and 24th APs are powered off. Aside from instability and weak robustness of the two methods in low dimension situations, it is mainly due to the fact that different APs contribute to different information entropy in an indoor positioning environment, which was well analyzed in our previous work [20].

For testing the robustness and noise tolerance of the proposed positioning system, we set it in an unstable and more noisy circumstance, where we take S1 region shown in Figure 7 as the interesting area with 152 reference points and leave doors and windows open, and in addition people walk around and RSS values are sampled only 1 time as a test point. The performance of proposed algorithm is better than the full dimensional WKNN fingerprinting method and other positioning systems as illustrated in Figure 14. Besides, it is worth noting that the situation of APs outage as shown in Figures 12 and 13 could be considered as an extreme noisy environment case, which may firmly prove the effectiveness of the proposed method as well.

Figure 14. Positioning accuracies comparison in the noisy circumstance in S1 region.

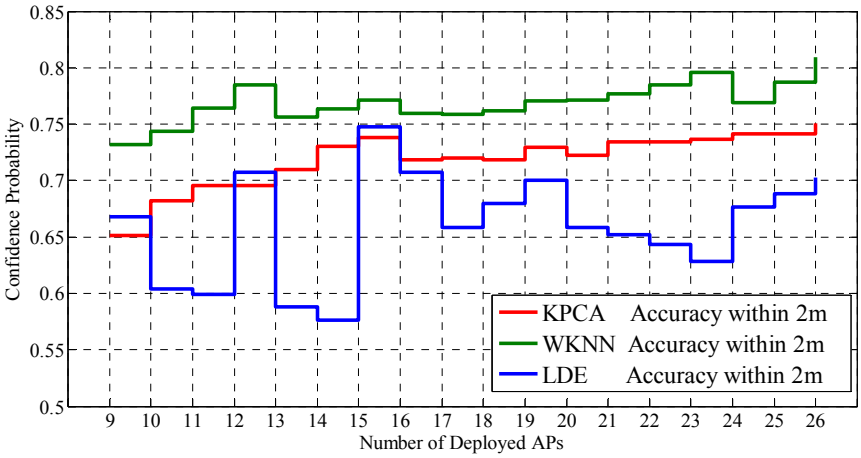


Moreover, environment dynamics including number of AP deployments and different sampling intervals are also taken into consideration. On the basis of ensuring all regions are being covered, performances of the proposed positioning system with different types of AP deployment are briefly evaluated as shown in Figure 15.

By and large, the confidence probability increases with the total number of deployed APs in terms of the WKNN method and the proposed system based on the KPCA method. However, the LDE shows outstanding positioning accuracy in some circumstances, e.g., fifteen APs are deployed in the building, though the instability of the method is obvious as well. Besides the reason that target dimension is unadjusted, the phenomenon can be partly attributed to the different discrimination of APs for different

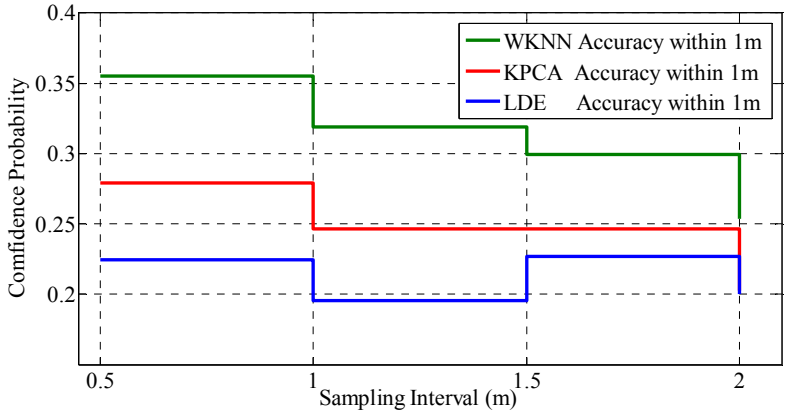
sample points, which is the reason that some researchers are concerned about AP selection schemes (to select most discriminating APs for positioning based on certain criterions, such as max mean, information entropy and joint entropy).

Figure 15. Performances of different positioning systems with different AP deployment under the condition that the positioning error distance is within 2 m and $D = 8$.



In terms of the relationship between sampling density and the system performance, according to the experimental results shown in Figure 16, the confidence probability goes down slowly as the sampling interval increases (density decrease). Compared with the influence of APs deployment, the positioning accuracy is less affected by the sampling interval.

Figure 16. Performances of different positioning systems with different sampling density under the condition that the positioning error distance is within 2 meters.



In sum, the Kernel PCA algorithm deployed in the proposed indoor positioning system is more capable of extracting the features of RSS with low dimensionality in an office environment, its robustness and generalization ability may provide higher positioning accuracy when dealing with asymmetric matching problem. The reduced dimension of the radio map may relieve the burden of the final online matching process, but it is undeniable that the computational complexity of the proposed method has increased in the previous feature extraction step. Specifically, the online computational complexity of the OG-SVM is $O(Cn_{sv})$, where C is the number of classes and n_{sv} is the number of

support vectors. The counterpart of KPCA is O(*dMN*), where *d* is the number of the (reduced) low dimensionality, *M* is the number of features (APs) and *N* is the number of reference points. Both of the LDE and LDA are O(*dM*). Besides, the computational complexity of WKNN method is O(*MN*). Therefore the computational complexity of the proposed positioning system is O(*dMN*) plus O(*dN*) and O(Cn_{sv}), so the other two systems share the same computational complexity, which is O(*dM*) plus O(*dN*). Compared with the two linear feature extraction methods (LDE and LDA), the proposed system underperforms others in terms of computational complexity due to the deployed kernel techniques. However, considering the contribution of dealing with unexpected AP outages and outstanding system robustness and stability, implementing the Kernel-PCA algorithm in the positioning system is still practical and effective.

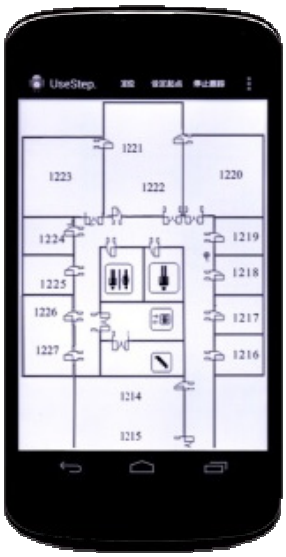
5. Conclusion

In this paper, firstly we propose the SDC method for clustering the radio map based on both RSS in signal space and coordinates in physical space. Compared with traditional clustering algorithms, the proposed method is more flexible and without outlier problems and constraints. Experimental results show that the fingerprinting method based on the sub-radio maps clustered by SDC outperforms its counterparts based on the FCM and K-Means clustering algorithms. After being integrated with OG-SVM, the coarse positioning accuracy of the proposed method is also better than that of the other algorithms.

Then we deploy the Kernel PCA method for reducing the dimensionality of the radio map, thereby enhancing the robustness and solving the asymmetric matching problem when AP outages occur. It turns out that the proposed Kernel PCA performs better than the LDA and manifold LDE methods in terms of extracting the features of an indoor radio map.

In addition, the structure of the proposed indoor positioning system is well modularized and mainly designed for mobile computing. It consists of the offline phase and online phase, respectively. The off-line phase is in charge of the main data computing process with a powerful PC server. All the computed data and trained functions derived from the offline stage would be stored and applied in the online module for the real time positioning procedure. We have validated the feasibility and effectiveness of the proposed indoor positioning system, and implemented it based on the Android OS as shown in Figure 17. Besides APs selection, inertial navigation and other approaches for indoor positioning are also under further development. The section of performance analysis might not be described in great detail, but a lot of experimental and implemental works on localization have been done in this study. Our future works will also keenly focus on WLAN- and WSN-based indoor positioning systems, information from sensors such as gyroscopes, accelerometers, thermometers and barometers available within mobile terminals will be further researched and deployed in our positioning system.

Figure 17. The proposed indoor positioning system running on Google Nexus 4.



Acknowledgments

This research is supported by National Natural Science Foundation of China (Grant No. 61101122 and No. 61302074), Postdoctoral Science-Research Development Foundation of Heilongjiang Province (Grant No. LBH-Q12080) and Specialized Research Fund for the Doctoral Program of Higher Education (Grant No. 20122301120004).

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Zhang, V.Y.; Wong, A. Kernel-based particle filtering for indoor tracking in WLANs. *J. Netw. Comput. Appl*. **2012**, *35*, 1807–1817.
2. Latif, S.; Tariq, R.; Haq, W.; Hashmi, U. Indoor Positioning System Using Ultrasonics. In Proceedings of International Bhurban Conference on Applied Sciences and Technology, Islamabad, Pakistan, 9–12 January 2012; pp. 440–444.
3. Wirola, L.; Laine, T.A.; Syrjarinne, J. Mass-Market Requirements for Indoor Positioning and Indoor Navigation. In Proceedings of International Conference on Indoor Positioning and Indoor Navigation, Zurich, Switzerland, 15–17 September 2010; pp. 1–7.
4. Kaemarungsi, K.; Krishnamurthy, P. Modeling of Indoor Positioning Systems Based on Location Fingerprinting. In Proceedings of IEEE International Conference on Computer Communications, 7–11 March 2004; pp. 1012–1022.
5. Kohoutek, T.K.; Mautz, R.; Wegner, J.D. Fusion of building information and range imaging for autonomous location estimation in indoor environments. *Sensors* **2013**, *13*, 2430–2446.
6. Guerrero, L.A.; Vasquez, F.; Ochoa, S.F. An indoor navigation system for the visually impaired. *Sensors* **2012**, *12*, 8236–8258.

7. Yucel, H.; Ozkir, T.; Edizkan, R.; Yazici, A. Development of Indoor Positioning System with Ultrasonic and Infrared Signals. In Proceedings of International Symposium on Innovations in Intelligent Systems and Applications, Trabzon, Turkey, 2–4 July 2012; pp. 2–4.
8. Raitoharju, M.; Fadjukoff, T.; Ali-Loytty, S.; Piche, R. Using Unlocated Fingerprints in Generation of WLAN Maps for Indoor Positioning. In Proceedings of 2012 IEEE/ION Position Location and Navigation Symposium, Myrtle Beach, SC, USA, 23–26 April 2012; pp. 23–26.
9. Mazuelas, S.; Bahillo, A.; Lorenzo, R.M.; Fernandez, P.; Lago, F.A.; Garcia, E.; Blas, J.; Abril,E.J. Robust indoor positioning provided by real-time RSSI values in unmodified WLAN networks. *IEEE J. Sel. Top. Signal Process*. **2009**, *3*, 821–831.
10. Liu, X.; Zhang, S.; Zhao, Q.; Lin, X. A Real-Time Algorithm for Fingerprint Localization Based on Clustering and Spatial Diversity. In Proceedings of International Congress on Ultra-Modern Telecommunications and Control Systems and Workshops, Moscow, Russia, 18–20 October 2010; pp. 74–81.
11. Chen, K.Y.; Yang, Q.; Yin, J.; Chai, X.Y. Power-efficient access-point selection for indoor location estimation. *IEEE Tran. Knowl. Data Eng*. **2006**, *18*, 877–888.
12. Feng, C.; Au, W.S.A.; Valaee, S.; Tan, Z. Received-signal-strength-based indoor positioning using compressive sensing. *IEEE Trans. Mob. Comput.* **2012**, *11*, 1983–1985.
13. Li, M.; Jiang, X.Y.; Guibas, L.J. Fingerprinting mobile user positions in sensor networks: Attacks and countermeasures. *IEEE Trans. Parallel Distrib. Syst*. **2012**, *23*, 676–683.
14. Deng, Z.A.; Xu, Y.B.; Ma, L. Indoor positioning via nonlinear discriminative feature extraction in wireless local area network. *Comput. Commun*. **2012**, *35*, 738–747.
15. Xu, Y.B.; Deng, Z.A.; Meng, W.X. An Indoor Positioning Algorithm with Kernel Direct Discriminant Analysis. In Proceedings of IEEE Global Telecommunications Conference, Miami, FL, USA, 6–10 December 2010; pp. 1–5.
16. Fang, S.H.; Lin, T.N. Indoor location system based on discriminant-adaptive neural network in IEEE 802.11 environments. *IEEE Trans. Neural Netw*. **2008**, *19*, 1973–1978.
17. Fang, S.H.; Lin, T.N. Principal component localization in indoor WLAN environments. *IEEE Trans. Mob. Comput.* **2012**, *11*, 100–110.
18. Fang, S.H.; Wang, C.H. A dynamic hybrid projection approach for improved Wi-Fi location fingerprinting. *IEEE Trans. Veh. Technol.* **2011**, *60*, 1037–1044.
19. Tenenbaum, J.; Silva, D.D.; Langford, J. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323.
20. Deng, Z.A.; X, Y.B.; Ma, L. Joint AP selection and local discriminant embedding for energy efficient and accurate Wi-Fi positioning. *KSII Trans. Internet Informa. Syst*. **2012**, *6*, 794–814.
21. Vapnik, V.N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, NY, USA, 1999.
22. Cai, Z.; Zhao, H.; Yang, Z.; Mo, Y. A modular spectrum sensing system based on PSO-SVM. *Sensors* **2012**, *12*, 15292–15307.
23. Sivanandam, S.N.; Deepa, S.N. Genetic Algorithms. *Introduction to Genetic Algorithms*, 1st ed.; Springer: Berlin, Germany, 2007.
24. Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*, 3rd ed.; Academic Press: London, UK, 2005.

25. Liu, C. Gabor-based kernel PCA with fractional power polynomial models for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell*. **2004**, *26*, 572–581.
26. Ham, J.; Lee, D.D.; Mika, S.; Scholkopf, B. A Kernel View of the Dimensionality Reduction of Manifolds. In Proceedings of International Conference on Machine Learning, Banff, AB, Canada, 4–8 July 2004; p. 47.
27. Zhu, M.L.; Aleix, M.A. Pruning noisy bases in discriminant analysis. *IEEE Trans. Neural Netw*. **2008**, *19*, 148–157.
28. Zhang, B.Y.; Sun, Y.M.; Bian, Y.L.; Zhang, H.K. Linear discriminant analysis in network traffic modeling. *Int. J Commun. Syst*. **2006**, *19*, 53–65.
29. Chen, H.T.; Chang, H.W.; Liu, T.L. Local Discriminant Embedding and Its Variants. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 846–853.
30. Wan, M.; Yang, G.; Lai, Z.; Jin, Z. Feature extraction based on fuzzy local discriminant embedding with applications to face recognition. *IET Comput. Vis*. **2011**, *5*, 301–308.
31. IEEE Standard for Information Technology—Telecommunications and Information Exchange Between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 11: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Amendment 8: Medium Access Control (MAC) Quality of Service (QoS) Enhancements; IEEE Amendment to IEEE Std 802.11; IEEE Computer Society, Washington, DC, USA, 11 November 2005.
32. Hossain, A.M.; Jin, Y.W.; Soh, W.S. SSD: A robust RF location fingerprint addressing mobile devices' heterogeneity. *IEEE Trans. Mob. Comput*. **2013**, *12*, 65–77.
33. Chen, L.; Li, B.; Zhao, K.; Rizos, C.; Zheng, Z. An improved algorithm to generate a Wi-Fi fingerprint database for indoor positioning. *Sensors* **2013**, *13*, 11085–11096.
34. Kushki, A.; Plataniotis, K.N.; Venetsanopoulos, A.N. Kernel-based positioning in wireless local area networks. *IEEE Trans. Mob. Comput*. **2007**, *13*, 689–705.
35. Castro, P.; Chiu, P.; Kremenek, T.; Muntz, R.R. A Probabilistic Room Location Service for Wireless Networked Environments. In Proceedings of 3rd Conference on Ubiquitous Computing, Atlanta, GA, USA, 30 September–2 October 2001; pp. 18–34.
36. Xu, Y.; Yang, J.Y.; Lu, J.F. An Efficient Kernel-Based Nonlinear Regression Method for Two-Class Classification. In Proceedings of International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; pp. 4442–4445.
37. Xu, Y.; Zhang, D.; Jin, Z.; Li, M.; Yang, J.Y. A fast kernel-based nonlinear discriminant analysis for multi-class classification. *Pattern Recogn.* **2006**, *39*, 1026–1033.