Peer-reviewed version available at *Appl. Sci.* **2017**, *7*, , 467; doi:10.3390/app7050467

Article

Graph-based Semi-supervised Learning for Indoor Localization Using Crowdsourced Data

Liye Zhang [1], Shahrokh Valaee [2], Yubin Xu [1,*], Lin Ma [1] and Farhang Vedadi [2]

1 Communication Research Center, Harbin Institute of Technology, Harbin 150001, China; wind_zhangliye@163.com (L. Z.); malin@hit.edu.cn (L. M.)
2 Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, M5S 3G4, Canada; valaee@ece.utoronto.ca (S.V.); farhang.vedadi@gmail.com (F.V.)
* Correspondence: ybxu@hit.edu.cn; Tel.: +86-451-8641-3513 (ext. 8203).

Abstract: Indoor positioning based on the received signal strength (RSS) of the WiFi signal has become the most popular solution for indoor localization. In order to realize the rapid deployment of indoor localization systems, solutions based on crowdsourcing have been proposed. However, compared to conventional methods, crowdsourced RSS values are more erroneous and can result in large localization errors. To mitigate the negative effect of the erroneous measurements, a graph-based semi-supervised learning (G-SSL) method is used to exploit the correlation between the RSS values at nearby locations to estimate an optimal RSS value at each location. Before using the G-SSL method, the Linear Regression (LR) algorithm is proposed to solve the device diversity problem in crowdsourcing system. Since the spatial distribution of the APs is sparse, the Compressed Sensing (CS) method is applied to precisely estimate the location of the APs. Based on the location of the APs and a simple signal propagation model, the RSS difference between different locations is calculated and used as an additional constraint to improve the performance of G-SSL. Furthermore, to exploit the sparsity of the weights used in the G-SSL, we use the CS method to reconstruct these weights more accurately and make a further improvement on the performance of the G-SSL. Experimental results show improved results in terms of the smoothness of the radio map and the localization accuracy.

Keywords: indoor localization; crowdsourcing; received signal strength; graph-based semi-supervised learning; linear regression; compressed sensing.

---

1. Introduction

Indoor *location-based services* (LBS) such as indoor positioning, tracking and navigation, have been receiving a lot of attention in recent years [1,2]. However, it remains a challenge to provide the users with an accurate and robust location estimation. Global Positioning System (GPS) is the most widely used localization system and provides precise positioning in outdoor environments. However, due to the lack of sufficient signal strength in most of the indoor areas, GPS is not a reasonable solution for indoor environments. Therefore, various alternatives to GPS have been proposed for indoor localization. Examples include but are not limited to the methods using Ultra-Wideband, Ultrasound, Infrared and Radio Frequency signals[2–6]. These alternatives provide a good localization accuracy for many applications, however, they require additional infrastructure that would be a disadvantage to their large-scale deployment.

With the growing deployment of WiFi access points in indoor environments and the widespread use of mobile devices such as smart phones, WiFi *received signal strength* (RSS)-based indoor localization methods are getting popular due to their low deployment cost and relatively high localization accuracy.

In general, there are two main categories of localization methods that use WiFi RSS readings. The first category comprises those methods that rely on the radio propagation model of the WiFi signal in indoor environments as well as the locations of the WiFi Access Points (AP). Specifically, the RSS readings from different access points are used to estimate the distance of a mobile device from those access points. Then a triangulation methods is used to estimate the location of the mobile device. The next category includes those methods that are based on WiFi RSS fingerprints also known as fingerprint-based methods. Originally proposed by P. Bahl et al [7], various fingerprint-based localization systems have been designed and developed during the last decade [7–9].

Typically, fingerprint-based methods consist of an offline phase followed by an online phase [7]. In the offline phase, RSS values from different WiFi access points are measured at some known locations throughout the indoor area. These location are referred to as Reference Points (RP) and the measured RSS vector for each RP is called a *fingerprints*. All fingerprints and their corresponding RPs are stored in a database called the *radio map*. In the online phase, a user's position can be estimated by comparing the RSS values measured by the user with the RSS fingerprints stored in the radio map.

A disadvantage to the offline phase of the fingerprint-based methods is the required time and labor to collect sufficient number of fingerprints throughout the indoor area. In addition, the RSS value of an AP at a certain location can change over time due to a number of reasons including but not limited to multipath fading, shadowing, moving objects and people [11]. To mitigate these RSS fluctuations, a large number of RSS measurements are collected at every reference point in the offline training phase. However, collecting more RSS measurements at any location, makes the offline phase even more time-consuming and labour-intensive. Several works have been proposed to reduce the workload of the offline phase[12–14]. The crowdsourcing method has been shown to be a promising approach to solving this problem[15–17]. In a crowdsourcing-based system, each user can contribute to the construction and updating of the radio map. Consequently, the number of RSS values collected in the offline training phase is greatly reduced. On other other hand, RSS measurements collected by the users moving in the environment is potentially more erroneous than those collected by the experts at the exact location of reference points.

One of the problems in the crowdsourcing localization system is that numerous of mobile devices is applied to build the radio map in the offline training phase and provide LBS for the device holders in the online phase. Due to the different WLAN adapter equipped in the mobile devices, the RSS values collected by the mobile device are subject to the difference of the WLAN adapter. As a result, different data collection devices may have different signal sensing capacities and yield different data distributions. Numerous studies show that, due to the hardware differences, the RSS differences collected by different devices exceeds more than 25dB [18–20]. Therefore, the localization accuracy is degraded significantly by the problem of RSS variations across different devices.

Another issue of indoor localization is the knowledge of the location of the access points. In most fingerprint-based methods, the location of the access points is considered to be unknown. This is a convenient simplifying assumption in many situations, especially when the signal strengths are measured in a passive mode. However, the knowledge of the location of the access points can enhance the localization accuracy. This is especially important since the location of an access point can be estimated using some signal processing techniques [10]. The location of an access point can then be used to correlate the received signal strength across neighbouring locations, as will be discussed in this paper.

In this paper, on the basis of graph-based semi-supervised learning (G-SSL) method, we propose RSS difference-aware G-SSL (RG-SSL) method and RSS difference-aware sparse graph SSL (RSG-SSL) method to smoothen the RSS values collected in the offline training phase and improve the localization results. Before smoothing RSS measurements using the G-SSL method, the problems discussed above is solved at the beginning. The Linear Regression (LR) method and the Compressed Sensing (CS)-based method of [10] are proposed to eliminate the device diversity problem and precisely estimate the AP locations, respectively. Then, using a signal propagation model, the RSS difference between two locations is calculated with respect to the locations of RPs and APs. Furthermore, RG-SSL method is proposed to smoothen the radio map in the offline training phase. By leveraging the RSS readings in the local neighbourhood, the effect of noise and erroneous measurements can be reduced to obtain a higher localization accuracy. Finally, the sparsity of the graph is discussed and RSG-SSL method is used to obtain a better RSS smoothing and localization result.

The rest of the paper is organized as follows. The related works are given and discussed in Section II. Section III formulates the indoor localization problem. In Section IV, the device diversity problem in crowdsourcing localization system is solved by linear regression method. The CS-based AP positioning method is explained in Section V. Section V also explains some experiments with the proposed CS-based AP positioning method. In Section VI, RG-SSL method is proposed with some experimental results. Finally, we explain the RSG-SSL method in Section VII and provide the localization results using RSG-SSL. Section VIII concludes the paper.

2. Background and Related Works

[2] and [12] proposed the CS-based method and the G-SSL method respectively, to reduce the workload of the radio map construction in the offline phase. Both methods, aim to reduce the number of reference points (RP) and RSS measurements. Also, [15–17] explore crowdsourcing-based methods to reduce the deployment workload by engaging the users to participate in radio map construction.

In [22], an RSS pre-processing method called the "sliding correlation time window filter" (SCTW) is used to reduce the noise in the measured RSS values. Similarly, in this paper, a sliding time window is used to average the RSS values collected in every RP to improve the accuracy of RSS measurements. However, this filter only uses a small number of the RSS values in the radio map and most of the information in the radio map is abandoned.

M. Hasani et al. [23] used a path-loss model to improve the reliability of the measured RSS values. In the offline phase of their method, a set of channel parameters are estimated for each access point. In the online phase, the user's location is found based on the calculated RSS values using the stored channel parameters. Their method results in a reliable localization thanks to the stability of the estimated channel parameters. In [24], S. Latif et al. proposed a D-model to estimate the radio signal strength in indoor areas. The experiments in their paper proved that the proposed D-model is caplable of estimating the RSS values with a high accuracy. Also their method models the wall attenuation more accurately compared to the method of [23]. Although the simulation result showed that the proposed method is fit for RFID positioning system, when this method is used in WiFi positioning system, the result is not satisfactory.

The signal propagation method gives us some inspiration, we proposed signal propagation-based outlier reduction technique (SPORT) to smooth the RSS collections in both the offline phase and the online phase and improve the localization accuracy [25]. In this method, we investigate the relationship of RSS values between adjacent locations using a signal propagation model and show that the outliers can be corrected using a signal propagation model. Experimental results show that SPORT greatly smoothens the radio map and improves the location accuracy.

In order to minimize the fluctuation of RSS values, M. S. Rahman Sakib et al. [26] developed a method using a Particle Filter (PF). Particle filters are used to perform non-linear and non-Gaussian estimations. However, in the online phase, a large number of particles have to be used in order to obtain a high positioning accuracy. Consequently, the computational cost is high which may be unacceptable for some indoor positioning applications.

L. Ma et al. [27] proposed a method based on the singular value thresholding (SVT) to recover the missing RSS values both in offline and online phases. In that paper, the authors argued that the positioning performance degrades significantly when some of the APs are occasionally turned off such as in a green WLAN system. Therefore, they proposed an SVT-based method to estimate the missing RSS values both in the radio map and the online RSS readings. They showed that their SVT-based method could achieve an acceptable positioning performance.

3. Problem Formulation

Suppose a set of ℓ RPs are selected throughout the indoor area and M APs are visible at each RP location. In the offline training phase, we collect the i-th *fingerprint* (c_i, r_i) at RP S_i, where $c_i = (x_i, y_i)^T$ is the geographical coordinates of S_i and r_i is an $M \times 1$ RSS vector. We refer to these fingerprints as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1

Peer-reviewed version available at *Appl. Sci.* **2017**, *7*, , 467; doi:10.3390/app7050467

labeled data. In the online phase, the user's location can be estimated by comparing the RSS value r_k collected at the unknown location of the user S_k with the fingerprints in the radio map. If r_k is similar to a particular r_i, then we reason that user's location S_k must be close to RP location S_i.

In practice, the RSS values measured by a mobile device are subject to multiple sources of noise, such as multi-path fading and shadowing. Fig. 1 illustrates the histogram of 100 RSS values from a single AP at a particular location inside the Bahen Building at the University of Toronto. The RSS values are distributed in a wide range of -70 dBm to -50 dBm. Occasionally, we cannot receive any power from this AP and a value of -110 dBm is used to denote the missing RSS value. Fig. 2 shows the RSS value from a single AP throughout the fourth floor of the Bahen Building after removing -110 dBm measurements and averaging over RSS values at each location.

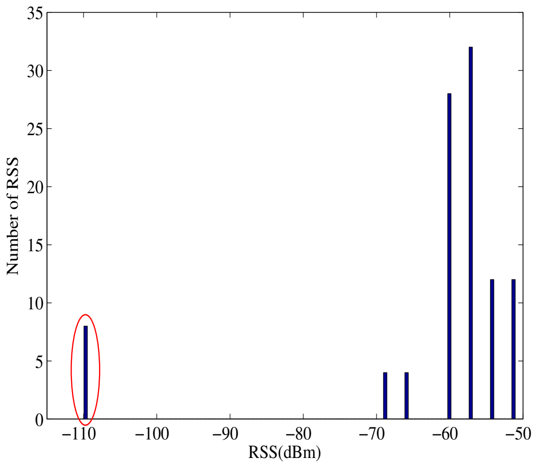


Figure 1. Histogram of 100 RSS values of a single AP measured at a location.

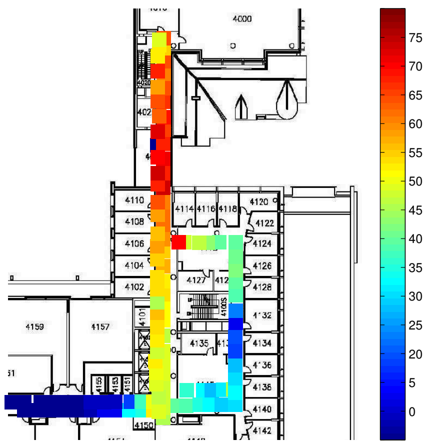


Figure 2. RSS values of an AP over the corridor area of the fourth floor of the Bahen Building, University of Toronto.

Next, we explain how we apply the G-SSL method to reduce the effect of noise in the radio map. Consider a set of u locations within the localization area that are not associated with RSS measurements hence we call them *unlabelled* data. In addition to these unlabelled locations, there are ℓ labelled RP locations as explained previously. Consequently, we have $\ell + u$ locations of labelled and unlabelled

Peer-reviewed version available at Appl. Sci. 2017, 7, , 467; doi:10.3390/app7050467

data. In the G-SSL method, a weighted graph is constructed using both labelled and unlabelled data. In this graph, the vertices represent the training data and all the vertices are connected by edges. The edge weight matrix which is calculated by the training data, represents the relationship between vertices in the graph by assigning a weight to each edge connecting two vertices in the graph. Each vertex on the graph corresponds to a location and the weighted edges between vertices represent the relationship between both RSS values and locations corresponding to those vertices. As mentioned earlier in this section, measured RSS values in an indoor environment are affected by different types of noise. However, in the graph representation of the G-SSL method, any two vertices on the graph are related not only by the RSS values measured at those vertices but also by the physical locations corresponding to those vertices. Therefore, the G-SSL is able to reduce the effect of noise in the measured RSS value by incorporating both RSS and location information. Next, we will explain the G-SSL method with more details.

Suppose $\Omega = (V, E)$ denotes the graph of the G-SSL method. The vertices of the graph, V, is defined as $\mathbf{V} = \{c_1, c_2, \ldots, c_\ell, c_{\ell+1}, \ldots, c_{\ell+u}\}$ where the first ℓ elements are the location coordinates of the labelled data and the next u elements are the location coordinates of the unlabelled data. For every edge between two vertices at S_i and S_j, we can calculate its weight w_{ij}. w_{ij} indicates the similarity between the two vertices and takes values in the range $[0, 1]$ with 0 indicating no similarity between the vertices. The result is an $(\ell + u) \times (\ell + u)$ weight matrix $\mathbf{W}$ containing all the calculated weights. The graph edges are usually undirected, so the edge (i, j) (weighted by w_{ij}) and the edge (j, i) (weighted by w_{ji}) are the same edge in the graph, which means $w_{ij} = w_{ji}$. In addition, the edge (i, i) does not exist, therefore, there are $\frac{1}{2}[(\ell + u) \times (\ell + u) - (\ell + u)]$ edges in the graph. In summary, only the corresponding number of graph weights are calculated which makes the weight matrix $\mathbf{W}$ a symmetric matrix. To calculate the weights, here we use the well-known *heat-kernel* function:

$$w_{ij} = \exp\left\{\frac{-\|c_i - c_j\|^2}{\tau}\right\}, \tag{1}$$

where $\|c_i - c_j\|^2 = d^2(S_i, S_j)$ is the square of the Euclidean distance between location S_i and S_j and τ is a parameter based on the application which controls how quickly the weight decreases.

The G-SSL uses $\mathbf{W}$ to estimate the labels of the unlabelled data using the relationship between different vertices in the graph. The result is a set of estimated labels $\hat{r}_i$ for $i \in \{1, 2, \ldots, \ell + u\}$. If c_i is close to c_j, the estimated label $\hat{r}_i$ is close to the given label r_j for all $j \in \{1, 2, \ldots, \ell\}$. The estimated labels $\hat{r}_i$ have to satisfy two conditions. First, for the labelled data, since the labels are already known, the estimated labels $\hat{r}_i$ must be close to the real labels. For the labelled data (c_i, r_i), we should have $\hat{r}_i = r_i$. This condition is enforced by minimizing the following loss function

$$\min_{\hat{R}} \sum_{i=1}^{\ell} \|\hat{r}_i - r_i\|^2, \tag{2}$$

where $\hat{R}$ is the $M \times (\ell + u)$ matrix of all estimated RSS values and $\| \bullet \|$ is the Euclidean distance.

The second condition is that the graph should be smooth. The smoothness of the graph comes from the fact that data points which are close to each other should have similar labels. To satisfy the smoothness condition, the estimated labels $\hat{r}_i$ and $\hat{r}_j$ should meet the following loss function

$$\min_{\hat{R}} \sum_{i=1}^{\ell+u} \sum_{j=1}^{\ell+u} w_{ij} \|\hat{r}_i - \hat{r}_j\|^2, \tag{3}$$

If c_i and c_j are close to each other, the weight w_{ij} would be large, and the labels $\hat{r}_i$ and $\hat{r}_j$ must be close in order for the whole term to be minimized. On the other hand, if c_i and c_j are far away from each other, the weight w_{ij} would be very small and the choice of the labels does not have much effect on the minimization.

Hence, the estimated labels that satisfies both conditions above can be estimated using:

$$\hat{\mathbf{R}}^* = \arg\min_{\hat{R}} \left\{ \sum_{i=1}^{\ell} \|\hat{r}_i - r_i\|^2 + \frac{\gamma}{2} \sum_{i=1}^{\ell+u} \sum_{j=1}^{\ell+u} w_{ij} \|\hat{r}_i - \hat{r}_j\|^2 \right\}, \tag{4}$$

where γ is a the weight of the smoothness term based on the application. γ is a design parameter used to enforce which term is of higher importance. In conclusion, the first term of the equation (4) penalizes the difference between the actual labels and the estimated labels and the second term ensures the smoothness of the graph.

In the offline phase, since the actual coordinates of S_i and S_j are already known, the RSS values can be calculated by G-SSL method. In the online phase, the data collected simultaneously from sensors on the mobile device can be used to estimate the relative displacement between S_i and S_j, that is, the distance $d(S_i, S_j)$. The RSS values can then be calculated using this additional information. The proposed G-SSL-based RSS smoothing method for crowdsourcing is summarized in the system diagram shown in Fig. 3.

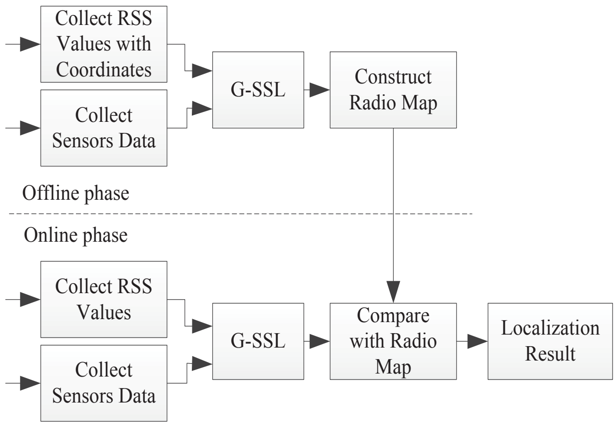


Figure 3. The system view of the proposed G-SSL based Localization.

4. Linear Regression Algorithm Against Device Diversity Problem

In the existing experimental systems, the same device is used to collect the RSS values in both the offline phase and the online phase. However, when the crowdsourcing method is widely applied to the indoor localization systems, a large number of different mobile devices have been used in the establishment of the radio map. In the online phase, a variety of mobile devices are also used by the users which are different from the device used to build the radio map. In this section, the linear regression (LR) algorithm is proposed to solve the device diversity problem in RSS-based crowdsourcing localization system.

We define $\mathcal{X}$ and $\mathcal{Y}$ are the signal space of different devices. Assume that the fingerprint $r_{\mathcal{X}}$ belongs to $\mathcal{X}$ is the nearest neighbor to the online point $r_{\mathcal{Y}}$ belongs to $\mathcal{Y}$. As described above, although they were collected at close physical locations, the RSS values have obvious difference. In order to solve the device diversity problem, the relationship between different devices have to be studied. Therefore, these RSS values collected by different devices could be processed to make the $r_{\mathcal{Y}}$ in closer to $r_{\mathcal{X}}$. Mathematically

$$\mathcal{X} \approx f(\mathcal{Y}), \tag{5}$$

By learning f, the radio map build by the training device could be used to localize any other devices.

Aiming to explore the mapping function between RSS values collected by distinct devices, the comparison results of RSS values across different training/tracking devices are plotted in Fig. 4. Every point on the figure represents RSS values from two different devices measured at the same location from the same AP at the same location. For example, the top right subplot in Fig. 4 represents

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1
Peer-reviewed version available at Appl. Sci. 2017, 7, , 467; doi:10.3390/app7050467

the RSS values measured by Lenovo laptop and Huawei mobile device. From Fig. 4, we can get a linear correlation between the RSS values measured by different devices. Hence, the following linear regression method can be employed as the mapping function.

$$r_{\mathcal{Y}} = a r_{\mathcal{X}} + b \tag{6}$$

where (a, b) are the coefficients in the mapping function.

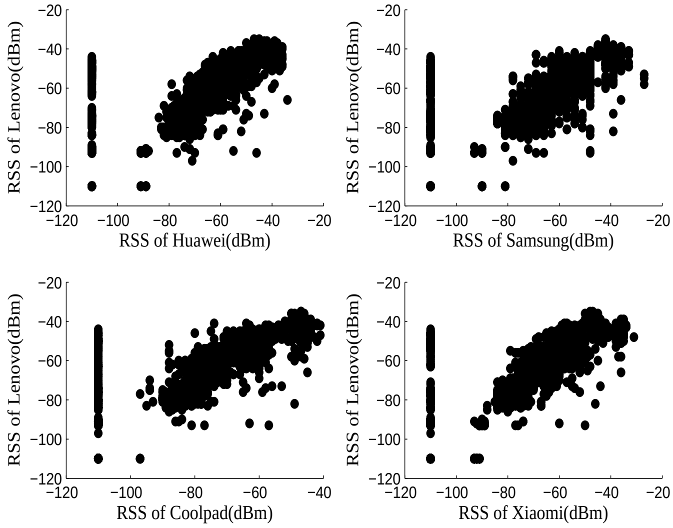


Figure 4. Linear correlation between RSS values for different devices.

4.1. Pre-processing of RSS Values

In the typical WLAN localization scenario, the RSS values collected by the mobile device are subject to multiple sources of noise, such as multi-path fading and shadowing. To mitigate these RSS fluctuations, a large number of RSS measurements are collected from each AP at every location. Let $\mathbf{RSS}_{li} = \{rss_1, rss_2, \ldots, rss_p\}$ be the set of RSS values collected at location l from the i-th AP. As shown in Fig. 4, if we cannot receive any power from the AP, a value of -110 dBm is used to denote the missing RSS value.

$$rss_{li} = \begin{cases} rss_{li}, & \text{if } rss_{li} > -110 \text{ dBm} \\ -110 \text{ dBm}, & \text{otherwise} \end{cases} \tag{7}$$

In order to obtain a high localization accuracy, the first step in localization system is to stabilize the collected RSS values prior to the localization process. Aiming to overcome the fluctuations, the average of the collected RSS values is calculated. In the calculation of the average value, the filled RSS values of -110 dBm could produce meaningless RSS values and will have a adverse impact. These filled RSS values could affect the localization process and produce erroneous location estimations. As a result, the average is calculated using the collected RSS values exclude the filled RSS values as the following equation:

$$r_{li} = \frac{\sum_{j=1}^{p} rss_j \mathbf{I}(rss_j \neq -110 \text{ dBm})}{\sum_{j=1}^{p} \mathbf{I}(rss_j \neq -110 \text{ dBm})} \tag{8}$$

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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1

Peer-reviewed version available at Appl. Sci. 2017, 7, , 467; doi:10.3390/app7050467

The average value r_{li} is used to build the radio map in offline training phase and estimate the current location in online localization phase.

4.2. Linear Regression Algorithm Against Device Diversity Problem

Before using the linear regression method, the parameters a and b in equation 6 should be computed. Since the outliers appear in the collected RSS values frequently and seriously affect the performance of the linear least squares (LLS) algorithm, the fast least trimmed squares (FAST-LTS) algorithm is used in this paper.

When the number of measured RSS values is c, the FAST-LTS solution for linear regression with intercept is given by

$$\min_{a,b} \sum_{i=1}^{h} d(i)^2 \tag{9}$$

where $h = int[(c+2)/2]$, $d(i) = \|r_{\mathcal{Y}} - (ar_{\mathcal{X}} + b)\|$ and $\| \bullet \|$ is norm 2 of a vector, $d(i)^2$ are the ordered squared residuals: $d(1)^2 \leq d(2)^2 \leq \cdots \leq d(i)^2 \leq \cdots \leq d(c)^2$.

Given the h-subset H_{old} of all nearest neighbors, the *C-step* is used to compute the a and b as follows [21]:

1. compute $\mathbf{a}_{old}$ and $\mathbf{b}_{old}$:= least squares regression estimator based on H_{old}
2. compute the residuals $d_{old}(i)$ for $i = 1, \cdots, c$
3. sort the absolute values of these residuals, $|d_{old}(1)| \leq |d_{old}(2)| \leq \cdots \leq |d_{old}(c)|$
4. arrange the absolute values of the residuals in ascending order, let H_{new} be a subset consisting of the nearest neighbors corresponding to the first h the absolute values of the residuals in the sequence
5. compute $\mathbf{a}_{new}$ and $\mathbf{b}_{new}$:= least squares regression estimator based on H_{new}

Repeating *C-step* with numerous H_{old}, a lot of regression coefficients will be gotten. The approximate solution is the coefficient corresponding to the least $\sum_{i=1}^{h} d(i)^2$. After getting the regression coefficient a and b, $r_{\mathcal{X}}$ is transformed as follows

$$r'_{\mathcal{X}} = ar_{\mathcal{X}} + b \tag{10}$$

where $r'_{\mathcal{X}} \in \mathcal{Y}$. As a result, both $r'_{\mathcal{X}}$ and $r_{\mathcal{Y}}$ belong to the same signal space, and a uniform radio map could be built using $r'_{\mathcal{X}}$ and $r_{\mathcal{Y}}$ in the offline training phase and a higher positioning accuracy could be obtained in online phase.

To verify the LR method, five distinct devices, namely Lenovo, Huawei, Samsung, Xiaomi and Coolpad, are used to collect RSS values at all RPs and the linear regression coefficients could be calculated based on the measured RSS values and the corresponding coordinates. When the regression coefficients are gotten, all the RSS values could be mapped into the same signal space by LR method and a uniform radio map could be built. Using the processed radio map, the user's location will be estimated with a high accuracy in online phase.

In our localization systems, we use the Lenovo device as the standard device, and all the RSS values collected by other devices are mapped into the signal space of Lenovo device. We take the (Huawei, Lenovo) pair as an example. As shown in Fig. 5, the collected data are more stable after pre-processing of RSS values, and the linear regression coefficients could be calculated by LTS method. Using the coefficients, the RSS values collected by Huawei device could be mapped into the signal space of Lenovo device. We compare the original RSS values and the transferred RSS values collected by Lenovo device with the RSS values collected by the Lenovo device, the comparison result is shown in Fig. 6. From the figure, we can see that the difference of signal distribution between different devices is reduced significantly. Accordingly, a uniform radio map can be built in the offline phase and the positioning performance could be improved in the online phase.

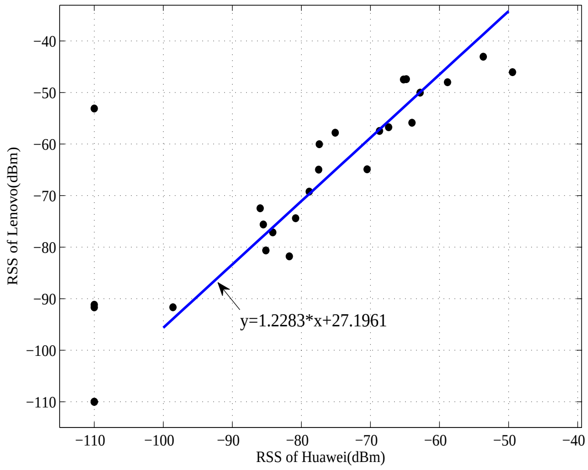


Figure 5. Linear correlation between Lenovo and Huawei.

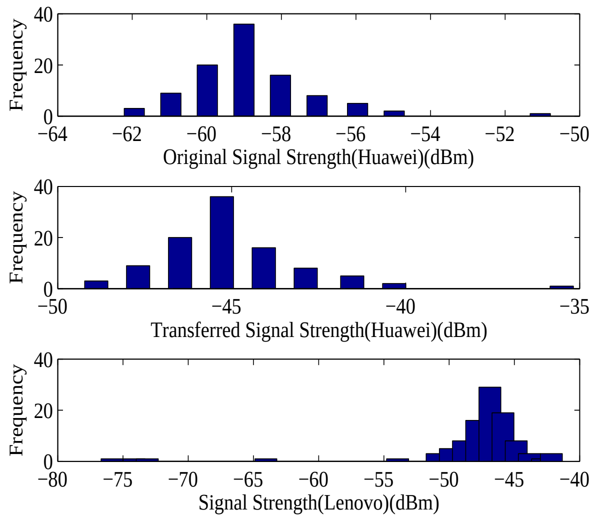


Figure 6. Comparison of signal distribution.

5. AP Localization Using Compressed Sensing Method

Typically, fingerprint-based localization methods don not rely on the location of the APs. In other words, the AP locations are assumed to be unknown. Nonetheless, better localization can be achieved if one could estimate the AP locations. Next, we discuss a compressed-sensing (CS)-based approach to estimating the AP locations.

Consider a set of N discrete locations throughout the indoor area. Suppose a set of M access points can be seen at each location. It is a practical assumption that the number of grid points is much larger than the number of access point in the indoor area i.e., $M \ll N$. We will use this assumption to apply a CS-based method to recover the location of the APs.

Compressed Sensing is a signal processing technique that can efficiently reconstruct a signal by exploiting the *sparsity* and *incoherence* properties of the signal [29–31]. Assume corresponding to the i-th AP, we define a vector θ_i of size N. θ_i is a vector that shows the location of the AP by assigning a one

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1

Peer-reviewed version available at Appl. Sci. 2017, 7, , 467; doi:10.3390/app7050467

to one the N element and zero for the rest of the element. For example, if $\theta_i(n) = 1$ then the location of the i-th AP is estimated to be the location of the n-th grid point in the indoor area. Concatenating all such vectors for all M APs results in a so-called index matrix, $\mathbf{\Theta}_{N\times M}$ as,

$$\mathbf{\Theta} = [\theta_1, \ldots, \theta_m, \ldots, \theta_M], \tag{11}$$

According to the CS theory, rather than measuring the M-sparse signal or its sparse representation Θ directly, compressive noisy RSS measurements in an ℓ-dimensional space are used. These compressive measurements are obtained by multiplying a random matrix by the original signal,

$$\mathbf{y} = \mathbf{\Phi\Psi\Theta} + \varepsilon, \tag{12}$$

where

1. $\mathbf{y}_{\ell\times M}$ are the compressive noisy RSS measurements.
2. $\mathbf{\Phi}_{\ell\times N}$ is the measurement matrix. Each row in this matrix represents the location of one RP, with an element of 1 to indicate the grid point at which the RP is located. Thus, only a few of RSS values are collected on the locations of RPs instead of measuring all the RSS values on the overall grid, which reduces the workload in the offline phase.
3. $\mathbf{\Psi}_{N\times N}$ is the sparsity basis on which the measured signals have sparse coefficients $\mathbf{\Theta}$. In this matrix, $\mathbf{\Psi}_{ij} = RSS(d_{ij})$ indicates the RSS values collected at grid point i from the AP located at grip point j, for all $1 \le i \le N$ and $1 \le j \le N$. Assume that the transmition power of an AP is $P_t(\text{dBm})$. Then $RSS(d)$ is calculated based on the empirical indoor propagation model of [10]:

$$RSS(d) = \begin{cases} P_t - 40.2 - 20\log(d), & \text{if } d \le 8 \\ P_t - 58.5 - 33\log(d), & \text{if } d > 8 \end{cases} \tag{13}$$

   where d is the physical distance from the transmitter (AP) to the receiver.
4. ε is the measurement noise.

The locations of the APs can be recovered by the following ℓ_0-minimization:

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_0, \ s.t. \ \mathbf{y} = \mathbf{\Phi\Psi\Theta}, \tag{14}$$

Unfortunately, solving (14) is both numerically unstable and NP-hard. Therefore, ℓ_1-minimization is used to recover the AP locations:

$$\hat{\mathbf{\Theta}} = \arg\min_{\mathbf{\Theta}} \|\mathbf{\Theta}\|_1, \ s.t. \ \mathbf{y} = \mathbf{\Phi\Psi\Theta}, \tag{15}$$

This is a convex optimization problem and various methods have been proposed to find the solution such as BP [32], OMP [33] and SP [34]. In this paper, we use OMP algorithm.

To evaluate the performance of the proposed CS-based AP localization algorithm, a few number of APs on the fourth floor of the Bahen Building at the University of Toronto have been localized. Fig. 7 shows the AP localization results. As seen in the figure, all the AP locations are estimated with a high level of accuracy. Although the localization results contain some errors, it brings limited effect to our RSS smoothing method proposed later.

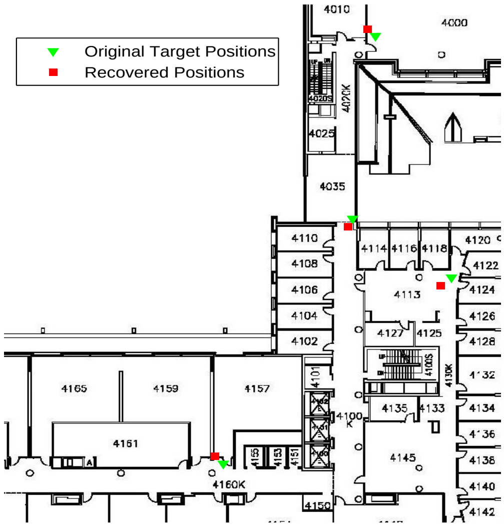


Figure 7. AP positioning results using CS method

6. RSS Difference-Aware Graph-Based Semi-Supervised Learning RSS Smoothing Method

The G-SSL method tries to set the same value for $\hat{r}_i$ and $\hat{r}_j$ if the coordinates c_i and c_j at locations S_i and S_j are similar. However, since the distance between each RP and each of the unlabelled locations is known, we can use this information to estimate the expected difference in RSS based on the known locations of the APs and the radio propagation model. Thus, we define $\hat{R}_d(S_i, S_j)$ as the estimated RSS difference between r_i and r_j at location S_i and S_j. We change the smoothing constraint to reflect that the difference $\|\hat{r}_i - \hat{r}_j\|$ of estimated RSS values $\hat{r}_i$ and $\hat{r}_j$ should be close to $\hat{R}_d(S_i, S_j)$. Accordingly, (4) can be written as:

$$\hat{\mathbf{R}}^* = \arg\min_{\hat{R}} \left\{ \sum_{i=1}^{\ell} \|\hat{r}_i - r_i\|^2 + \frac{\gamma}{2} \sum_{i=1}^{\ell+u} \sum_{j=1}^{\ell+u} w_{ij} \Big(\|\hat{r}_i - \hat{r}_j\| - \hat{R}_d(S_i, S_j) \Big)^2 \right\}. \tag{16}$$

6.1. Estimation of $\hat{R}_d(S_i, S_j)$

Consider one of the APs as shown in Fig. 8. The location of the AP, c_{AP}, can be estimated using the CS-based method in [10]. We use the indoor signal propagation model in [35]. Therefore, the RSS value at location S_i can be calculated as,

$$r_i = 10\log_{10} \frac{Ph_i}{d_i^{\alpha}} - 10\log_{10}(10^{-3}), \tag{17}$$

where d_i denotes the distance between the location of the i-th measurement and the AP, P is the transmission power of the AP, α is the propagation loss exponent and h is the combined effect of path loss, fading, and shadowing. Using this model and assuming $h_i = h_j$, we derive the following expression for $\hat{R}_d(S_i, S_j)$:

$$\hat{R}_d(S_i, S_j) = \|r_i - r_j\| = \|10\alpha\log_{10} \frac{d_j}{d_i}\| \tag{18}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1

Peer-reviewed version available at *Appl. Sci.* **2017**, *7*, , 467; doi:10.3390/app7050467

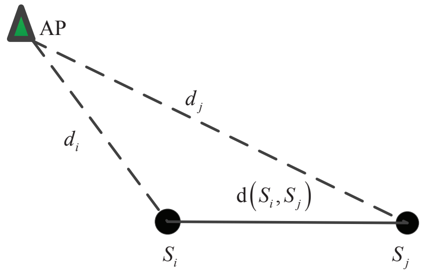


Figure 8. Mobile device is moving away from AP.

6.1.1. Offline Training Phase

In the offline training phase, the coordinates of RPs S_i and S_j, c_i and c_j, are already given in the radio map and the location of AP c_{AP} can be calculated precisely using the CS-based method. Thus, the Euclidean distance d_i and d_j between the RPs and AP can be obtained. Finally, the RSS difference $\hat{R}_d(S_i, S_j)$ can be calculated directly using (18).

6.1.2. Online Localization Phase

In the online localization phase, since the actual location of S_j is unknown, d_j cannot be calculated directly. However, $d(S_i, S_j)$ can be estimated using inertial sensor data and step counting algorithms and d_i can then be calculated. We can use $d_j = d_i - d(S_i, S_j)$ (the mobile device moves towards AP) or $d_j = d_i + d(S_i, S_j)$ (the mobile device moves away from AP) instead.

6.2. Finding the optimal solution

The cost function in (16) can be written as:

$$\begin{aligned} C = & \sum_{i=1}^{\ell} \|\hat{r}_i - r_i\|^2 + \frac{\gamma}{2} \sum_{i=1}^{\ell+u} \sum_{j=1}^{\ell+u} w_{ij} \|\hat{r}_i - \hat{r}_j\|^2 \\ & + \frac{\gamma}{2} \sum_{i=1}^{\ell+u} \sum_{j=1}^{\ell+u} w_{ij} \hat{R}_d^2(S_i, S_j) + \gamma \sum_{i=1}^{\ell+u} \sum_{j=1}^{\ell+u} w_{ij} \hat{R}_d(S_i, S_j) \|\hat{r}_i - \hat{r}_j\|, \end{aligned} \tag{19}$$

In order to find the optimal solution, we need to find the derivative of the cost function with respect to $\hat{\mathbf{R}}$. Since the cost function of (19) is not convex, we use the gradient descent method to solve the optimization problem. Next, we derive the derivative for each part of the cost function in (19). The first part of (19) can be written as:

$$\begin{aligned} C_1 & = \sum_{i=1}^{\ell} \|\hat{r}_i - r_i\|^2 \\ & = \mathrm{trace}\big((\hat{\mathbf{R}} - \mathbf{R})\mathbf{J}^{\mathrm{T}}\mathbf{J}(\hat{\mathbf{R}}^{\mathrm{T}} - \mathbf{R}^{\mathrm{T}})\big), \end{aligned} \tag{20}$$

where $\mathbf{R} = [r_1\ r_2\ \ldots\ r_{\ell+u}]$ is the RSS matrix and if the labels are not given, we use $\mathbf{0}_{M\times 1}$ instead. $\mathbf{J} = \mathrm{diag}(\delta_1, \delta_2, \ldots, \delta_{\ell+u})$ is a Hermitian indication matrix where $\delta_i = 1$ means that the corresponding i-th node in the graph is labelled and $\delta_i = 0$ otherwise. Using (20), $\frac{\partial C_1}{\partial \hat{\mathbf{R}}}$ can be written as:

$$\begin{aligned} \frac{\partial C_1}{\partial \hat{\mathbf{R}}} & = (\hat{\mathbf{R}} - \mathbf{R})(\mathbf{J} + \mathbf{J}^{\mathrm{T}}) \\ & = 2\mathbf{J}(\hat{\mathbf{R}} - \mathbf{R}), \end{aligned} \tag{21}$$

The second part of (19) can rearranged as:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1

Peer-reviewed version available at *Appl. Sci.* **2017**, *7*, , 467; doi:10.3390/app7050467

$$\begin{aligned}
C_2 &= \frac{\gamma}{2}\sum_{i=1}^{\ell+u}\sum_{j=1}^{\ell+u} w_{ij}\|\hat{r}_i - \hat{r}_j\|^2 \\
&= \gamma\sum_{i=1}^{\ell+u}\hat{r}_i^T\hat{r}_i\sum_{j=1}^{\ell+u} w_{ij} - \gamma\sum_{i=1}^{\ell+u}\sum_{j=1}^{\ell+u} w_{ij}\hat{r}_i^T\hat{r}_j \\
&= \gamma\text{trace}\big(\hat{\mathbf{R}}\mathbf{D}\mathbf{R}^T\big) - \gamma\text{trace}\big(\hat{\mathbf{R}}\mathbf{W}\mathbf{R}^T\big) \\
&= \gamma\text{trace}\big(\hat{\mathbf{R}}\mathbf{L}\mathbf{R}^T\big),
\end{aligned} \tag{22}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacian and $\mathbf{D} = \text{diag}(\mu_1, \mu_2, \ldots, \mu_{\ell+u})$ where $\mu_i = \sum_{j=1}^{\ell+u} w_{ij}$ for all $i \in \{1, 2, \ldots, \ell+u\}$. Differentiating C_2 yields,

$$\frac{\partial C_2}{\partial \hat{\mathbf{R}}} = 2\gamma\hat{\mathbf{R}}\mathbf{L}, \tag{23}$$

The derivative of the third part of the cost with respect to $\hat{\mathbf{R}}$ is equal to 0. The last part of (19) is:

$$\begin{aligned}
C_4 &= \gamma\sum_{i=1}^{\ell+u}\sum_{j=1}^{\ell+u} w_{ij}\hat{R}_d(S_i, S_j)\|\hat{r}_i - \hat{r}_j\| \\
&= \gamma\sum_{i=1}^{\ell+u}\sum_{j=1}^{\ell+u} \kappa_{ij}\|\hat{r}_i - \hat{r}_j\|,
\end{aligned} \tag{24}$$

where $\kappa_{ij} = w_{ij}\hat{R}_d(S_i, S_j)$. In order to find $\frac{\partial C_4}{\partial \hat{\mathbf{R}}}$, first we find $\frac{\partial C_4}{\partial \hat{\mathbf{r}}_n}$ for $1 \le n \le \ell+u$. Using [35],

$$\begin{aligned}
\frac{\partial C_4}{\partial \hat{r}_n} &= \frac{\partial}{\partial \hat{r}_n}\left(\gamma\sum_{i=1}^{\ell+u}\sum_{j=1}^{\ell+u}\kappa_{ij}\|\hat{r}_i - \hat{r}_j\|\right) \\
&= \frac{\partial}{\partial \hat{r}_n}\left(\gamma\sum_{j=1, j\neq n}^{\ell+u}\kappa_{nj}\|\hat{r}_n - \hat{r}_j\|\right) + \frac{\partial}{\partial \hat{r}_n}\left(\gamma\sum_{i=1, i\neq n}^{\ell+u}\kappa_{in}\|\hat{r}_i - \hat{r}_n\|\right)
\end{aligned} \tag{25}$$

Since $\kappa_{ij} = w_{ij}\hat{R}_d(S_i, S_j)$, $\kappa_{ni} = \kappa_{in}$. Therefore:

$$\begin{aligned}
\frac{\partial C_4}{\partial \hat{r}_n} &= 2\times\frac{\partial}{\partial \hat{r}_n}\left(\gamma\sum_{j=1, j\neq n}^{\ell+u}\kappa_{nj}\|\hat{r}_n - \hat{r}_j\|\right) \\
&= 2\times\gamma\sum_{j=1, j\neq n}^{\ell+u}\kappa_{nj}\frac{\hat{r}_n - \hat{r}_j}{\|\hat{r}_n - \hat{r}_j\|} \\
&= 2\gamma\mathrm{g}_n,
\end{aligned} \tag{26}$$

where $\mathrm{g}_n = \sum_{j=1, j\neq n}^{\ell+u}\kappa_{nj}\frac{\hat{r}_n - \hat{r}_j}{\|\hat{r}_n - \hat{r}_j\|}$ and $\frac{\partial C_4}{\partial \hat{\mathbf{R}}}$ is obtained using:

$$\frac{\partial C_4}{\partial \hat{\mathbf{R}}} = 2\gamma\mathbf{G}, \tag{27}$$

where $\mathbf{G} \triangleq [\mathrm{g}_1\ \mathrm{g}_2\ \ldots\ \mathrm{g}_{\ell+u}]$. Finally, in order to find the optimal solution, we set $\frac{\partial C}{\partial \hat{\mathbf{R}}} = 0$:

$$\frac{\partial C_1}{\partial \hat{\mathbf{R}}} + \frac{\partial C_2}{\partial \hat{\mathbf{R}}} - \frac{\partial C_4}{\partial \hat{\mathbf{R}}} = 0. \tag{28}$$

Using (21), (23) and (27):

$$\hat{\mathbf{R}} = (\mathbf{R}\mathbf{J} + \gamma\mathbf{G})(\mathbf{J} + \gamma\mathbf{L})^{-1}. \tag{29}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1

Peer-reviewed version available at Appl. Sci. 2017, 7, , 467; doi:10.3390/app7050467

In summary, to find the optimal solution, initialize $\mathbf{G} = \mathbf{0}_{M\times(\ell+u)}$. Then use an iterative procedure as follows: First, find $\hat{\mathbf{R}}$ as the solution of (29). Second, update $\mathbf{G}$ based on the result of the first step and the definition of $\mathbf{G}$. Repeat the two steps until convergence.

6.3. Experimental Results

In order to verify our method, we collected RSS data from the 4th floor of the Bahen Building at the University of Toronto. The radio map was constructed using a step-counter-assisted RSS measurement method. Sensor information from the accelerometer is used to estimate the distance between the RPs. Using this system, a radio map consisting of 251 RPs thorughout the entire 4th floor of the Bahen building has been created in less than 30 minutes. However, the resulting radio map has only 5 RSS measurements at each RP. Consequently, it is more error-prone compared to the traditionally generated radio maps in which for each RP hundreds of measurements are collected. The Proposed localization procedure is tested on a sequence of 35 test points collected on a path from Room 4000 (top of Fig. 9) to Room 4148 (bottom of Fig. 9).

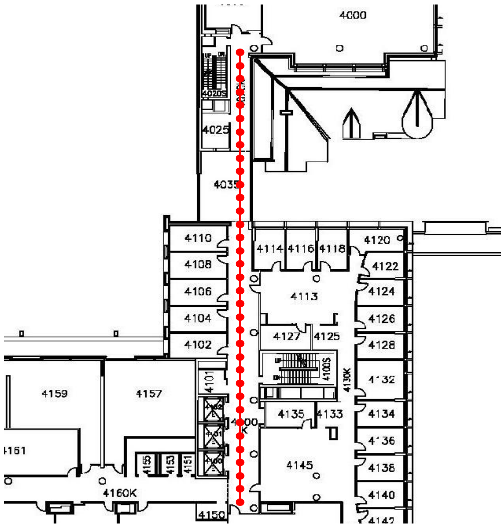


Figure 9. Actual locations of Test Points.

The RSS values from a single AP in the original radio map and the test points are shown in Fig. 10(a) and 11(a) respectively. As can be seen, although the RSS values are generally consistent with the signal propagation model, there are some large fluctuations at some RPs. To eliminate the negative effects caused by this fluctuation, the proposed RG-SSL method is applied to smooth the RSS values. In order to obtain more accurate results, 125 unlabelled data throughout the whole 4th floor of the Bahen Building are considered. Following steps are repeated until all the labelled points are smoothed:

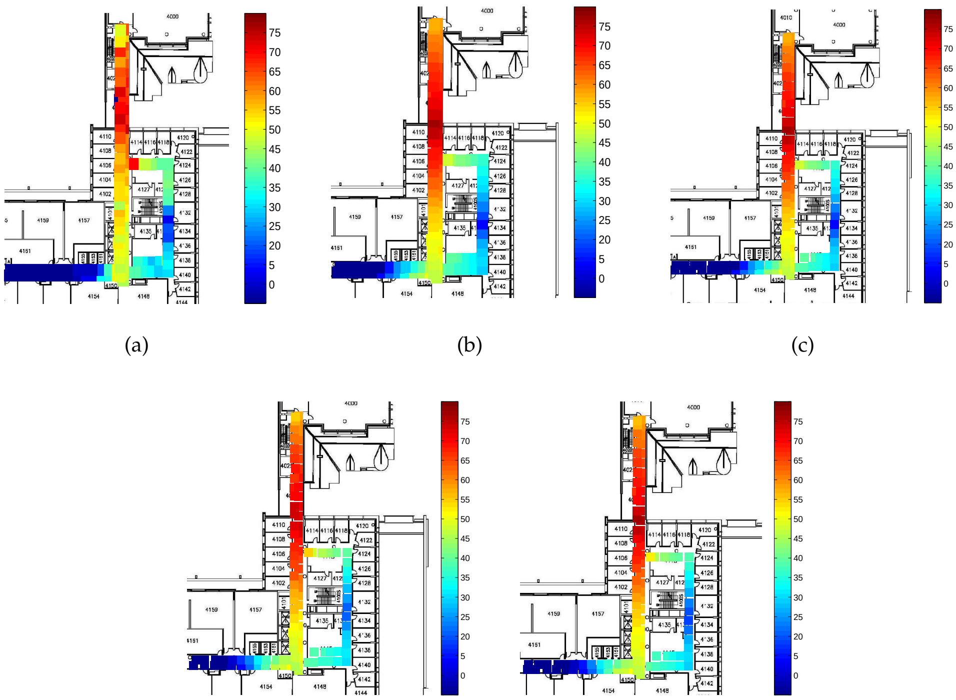
1. Set one of the labelled points as unlabelled.
2. Use the rest of the labelled points, 125 unlabelled points and RG-SSL method to estimate the RSS value of the above unlabelled point.

As comparisons, the G-SSL method, SCTW method and SPORT method are also simulated in this paper. The simulation results are shown in Fig. 10 and 11. We can see that the RG-SSL method successfully smooths out the radio map and the effect of signal fluctuation is mitigated. The localization result from directly using the original radio map and test point data are shown in Fig. 12(a). Compared with the actual locations in Fig. 9, there are some significant errors in the localization results as certain distinct test points have been localized erroneously to a single location. The localization result using

Peer-reviewed version available at Appl. Sci. 2017, 7, , 467; doi:10.3390/app7050467

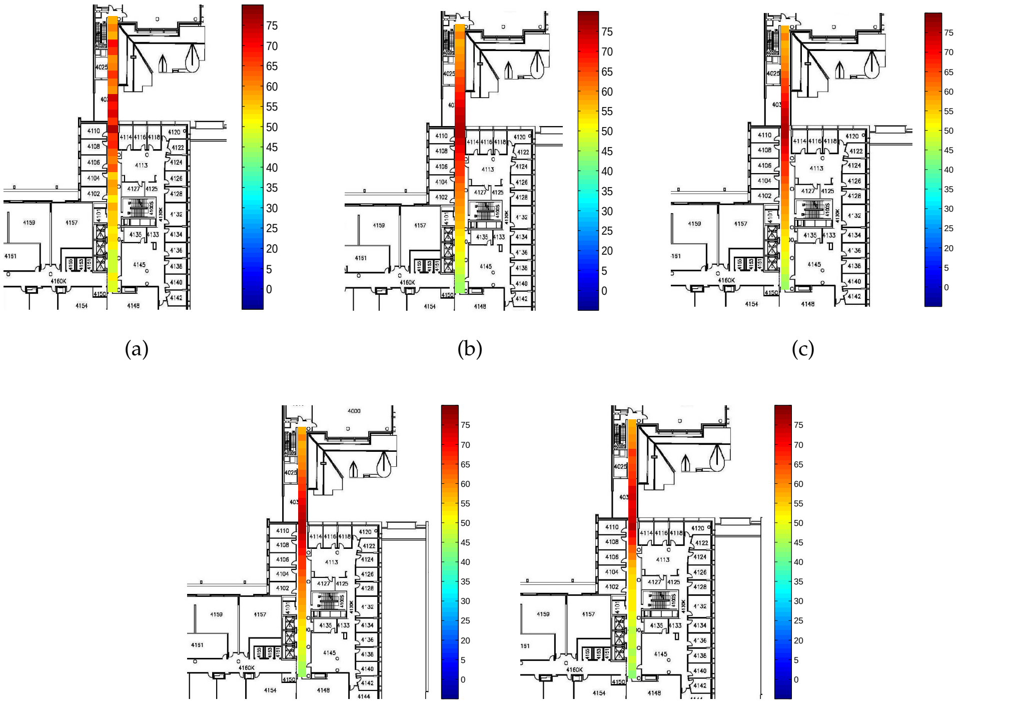
the modified radio map can be seen in Fig. 12(b) to Fig. 12(e). We see that the localization results are improved compared to the results in Fig. 12(a). Most of the test points that were erroneously localized to one location in Fig. 12(a) are now localized to correct distinct locations. These incorrect estimates were causing a large amount of localization error in the original method however are greatly reduced using both the RG-SSL method and the other methods. Clearly, the localization results of the proposed RG-SSL method are closer to actual locations than the results calculated by the other methods. Furthermore, the trajectory obtained in Fig. 12(b) is clearly smoother than those in Fig. 12(c) to Fig. 12(e).

We can readily see the performance gain of the RG-SSL method in the cumulative distribution function (CDF) of the localization error for the RG-SSL method and the other methods, as shown in Fig. 13. It is clear that the proposed localization method outperforms the other methods. The average localization error has been reduced from 2.89m to 2.07m, and notably, the maximum localization error has been reduced from 10m to 4m.



(a) (b) (c) (d) (e)

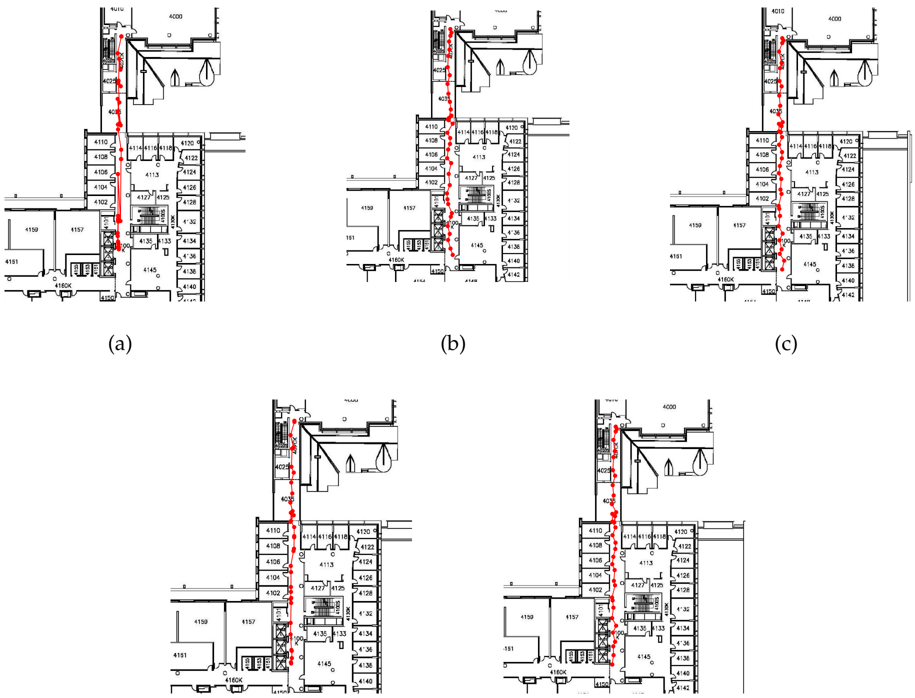
Figure 10. Comparison of signal distribution of radio map. (a) Original radio map (b) RG-SSL method (c) G-SSL method (d) SCTW method (e) SPORT method



(a) (b) (c)

(d) (e)

Figure 11. Comparison of signal distribution of test data. (a) Original radio map (b) RG-SSL method (c) G-SSL method (d) SCTW method (e) SPORT method



(a) (b) (c)

(d) (e)

Figure 12. Comparison of Localization results. (a) Original radio map (b) RG-SSL method (c) G-SSL method (d) SCTW method (e) SPORT method

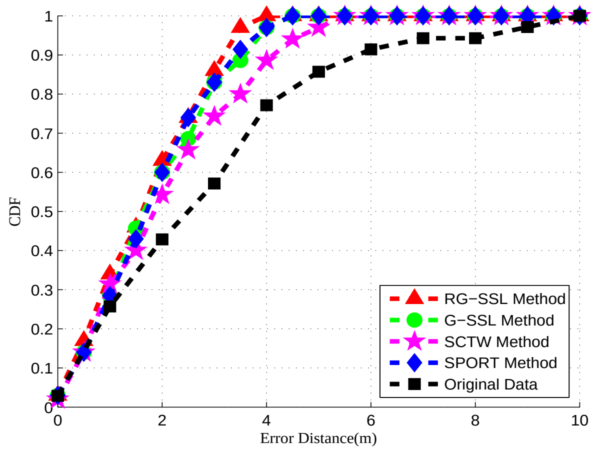


Figure 13. Cumulative distribution function of the localization error.

7. RSS Difference-aware Sparse Graph-based Semi-Supervised Learning Method and Experimental Results

7.1. Sparse Graph Construction for RG-SSL using CS Method

Since the radio map is constructed using a step-counter-assisted RSS measurement method, the coordinate of each RP calculated by this method contains a lot of noise. In the proposed RG-SSL method, the *heat-kernel* function is used to construct the graph and calculate the edge weights based on the Euclidean distance. However, the Euclidean distance and consequently the weights, are very sensitive to noise.

The accuracy of the generated graph will greatly affect the positioning performance. When the vertices in the graph are far away from each other, the graph weight is much smaller than the graph weight calculated for neighboring vertices. Therefore, the graph weight matrix is sparse. Since the CS method is robust to noisy data, we can use it to estimate the graph weight matrix[36]. As mentioned in Section II, we denote the vertex set $V = \{c_1, c_2, \ldots, c_\ell, c_{\ell+1}, \ldots, c_{\ell+u}\}$. Given the measurement matrix $\mathbf{A}$ and the matrix for unknown reconstruction coefficients $\mathbf{W}$ we can reconstruct a sparse $\mathbf{W}$ from $\mathbf{V} = \mathbf{AW}$ using:

$$\min_{\mathbf{W}} \|\mathbf{W}\|_0, \; s.t. \; \mathbf{V} = \mathbf{AW}, \tag{30}$$

where $\| \bullet \|_0$ denotes the ℓ_0-norm. The ℓ_0-norm minimization is NP-hard. However, if the solution is sparse enough, the following convex ℓ_1-norm minimization can be used to solve the sparse representation problem:

$$\min_{\mathbf{W}} \|\mathbf{W}\|_1, \; s.t. \; \mathbf{V} = \mathbf{AW}, \tag{31}$$

Suppose the noise in the collected RSS is denoted by ξ. Then,

$$\mathbf{V} = \mathbf{AW} + \xi = [\mathbf{A}\ \mathbf{I}] \begin{bmatrix} \mathbf{W} \\ \xi \end{bmatrix} = \mathbf{BW}', \tag{32}$$

where $\mathbf{B} = [\mathbf{A}\ \mathbf{I}]$ and $\mathbf{W}' = \begin{bmatrix} \mathbf{W} \\ \xi \end{bmatrix}$. Thus the ℓ_1-norm minimization can be rewritten as:

$$\min_{\mathbf{W}'} \|\mathbf{W}'\|_1, \; s.t. \; \mathbf{V} = \mathbf{BW}', \tag{33}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1

Peer-reviewed version available at Appl. Sci. 2017, 7, , 467; doi:10.3390/app7050467

For each c_i in the vertex set, the measurement matrix $\mathbf{B}_i$ is constructed as $\mathbf{B} = [c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_{\ell+u}, I]$ and w'_i is calculated using ℓ_1-norm minimization:

$$\min_{w'_i} \|w'_i\|_1, \; s.t. \; c_i = \mathbf{B}_i w'_i, \tag{34}$$

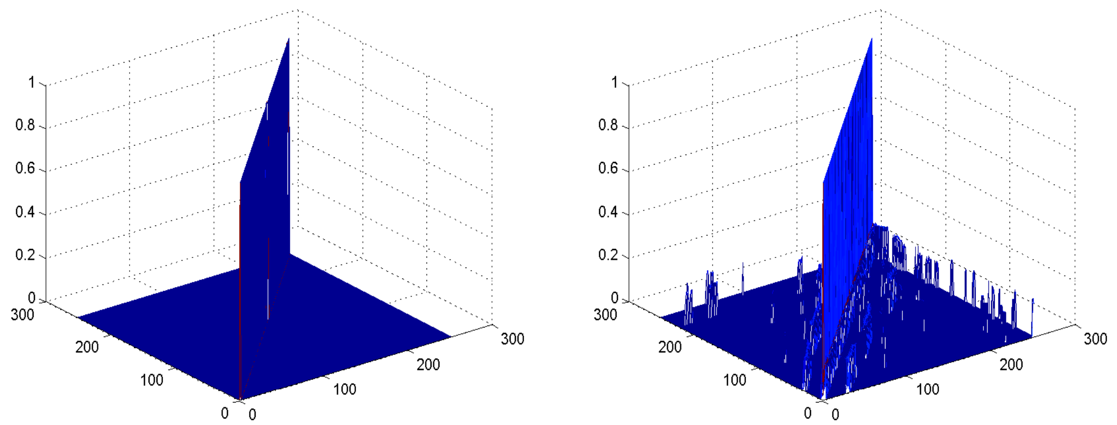
where w'_i is the i-th column of the matrix $\mathbf{W}$. Then the graph weights w_{ij} are obtained using:

$$w_{ij} = \begin{cases} w'_i(j), & \text{if } j < i \\ w'_i(j-1), & \text{if } j > i \\ 0, & \text{if } j = i \end{cases}, \tag{35}$$

where $i, j \in \{1, 2, \ldots, \ell + u\}$ and $w'_i(j)$ denotes the j-th element of vector w'.

7.2. Experimental Results

Since the labels of all the vertices in the graph are necessary for sparse reconstruction of the graph weight matrix, CS method can only be used in the offline phase. The weighted matrices calculated by the *heat-kernel* function and CS method are shown in Fig. 14(a) and Fig. 14(b), respectively. Each pixel in the figure represents the weight value w_{ij} between two vertices and $0 \le w_{ij} \le 1$. A larger value of w_{ij} between vertex S_i and S_j means a stronger correlation between them. If the vertices around the vertex S_i have strong correlations with the vertex S_i, we can get a more accurate RSS estimates for the vertex S_i. As we can see from Fig. 14(a), since the measurements are noisy, the weight matrix contains some errors. In the weight matrix, the weight values are very small between different vertices, which means the relationship between different vertices is very weak. Therefore, the information transferred between different vertices is inaccurate and the estimated RSS values using this weight matrix are not accurate enough. As a result, the localization accuracy is reduced by the inaccurate relationship between different locations.

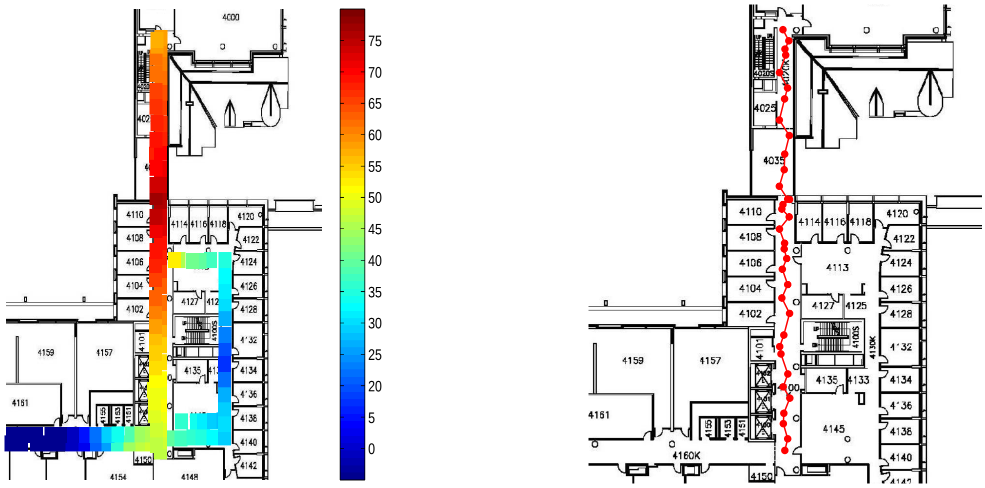


(a) (b)

Figure 14. Comparison of Weighted graph. (a) Weighted graph calculated by heat kernel method. (b) Weighted graph calculated by CS method

Due to the sparsity of the graph and robustness to noise, the weight matrix is recovered more precisely than the traditional *heat-kernel* function. The relationship between different vertices in Fig. 14(b) is much clearer than Fig. 14(a). Comparing Fig. 14(b) with Fig. 14(a), the graph weight values calculated by the CS method are much larger than those obtained using the *heat-kernel*. As a result, it is possible to get more useful information between different vertices using the matrix in Fig. 14(b).

Therefore, the estimated RSS values are more accurate than those calculated by the *heat-kernel* as shown in Fig. 15(a) and Fig. 10(b). Based on the matrix calculated by the CS method, the localization results are more accurate in Fig. 15(b). From Fig. 16 we can learn that the maximum localization error has been reduced from 4 to 3.5m thanks to the more accurate radio map. Meanwhile, the average localization error has been reduced from 2.07m to 1.98m.



(a) (b)

Figure 15. Smoothed signal distribution of radio map and localization results using RSG-SSL. (a) Smoothed signal distribution of radio map. (b) Localization results.

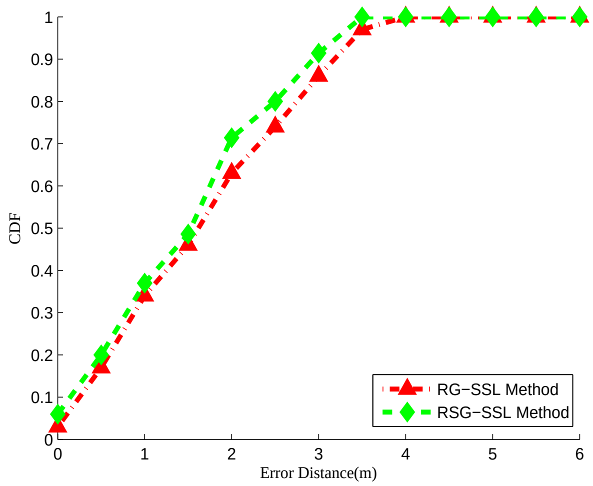


Figure 16. Cumulative distribution function of localization error.

8. Conclusion

In this paper, the effect of noise and erroneous measurements caused by the crowdsourced data are reduced using the relationship between RSS values of different locations. The RG-SSL method is proposed to improve the localization accuracy by smoothing the RSS values and using label propagation to better estimate the radio map. The relationship between the RSS values is represented using a weighted graph connecting different locations. Additionally, the RSS difference is introduced in the traditional G-SSL method to achieve a better performance. In order to obtain the RSS difference, a CS-based method is used to precisely localize the location of the APs. Noisy RSS values can be corrected using the proposed RG-SSL method, resulting in a higher localization accuracy. Due to the sparsity of the weighted graph in the G-SSL, the weighted graph is reconstructed more accurately by the CS method compared to the traditional heat-kernel function which is the idea of the proposed RSG-SSL method. The experimental results performed at the University of Toronto show that a smoothed radio map and online RSS values are obtained by RG-SSL method and the localization accuracy is improved. The RSG-SSL method applied in the offline phase also resulted in an improved performance.

Acknowledgments: This paper is supported by National Natural Science Foundation of China (61571162), Natural Science Foundation of Hei Longjiang Province China (F2016019), Postdoctoral Science-Research Development Foundation of Hei Longjiang Province China (LBH-Q12080), and Science and Technology Project of Ministry of China Public Security Foundation (2015GABJC38).

Author Contributions: Liye Zhang, Shahrokh Valaee, Yubin Xu and Lin Ma conceived the idea of the paper; Liye Zhang designed and performed the experiments; Shahrokh Valaee, Yubin Xu and Lin Ma analyzed the data; Liye Zhang wrote the paper; Shahrokh Valaee, Farhang Vedadi, Yubin Xu and Lin Ma revised the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	linear dichroism

References

1. Sorour, S.; Lostanlen Y.; Valaee S.; Majeed K. Joint indoor localization an radio map construction wiht limited deployment load. *IEEE Transactions on Mobile Computing*. **2015**, *14*, 1031-1043.
2. Feng C.; Valaee S.; Au A. W. S.; Tan Z. Received-signal-strength-based indoor positioning using compressive sensing. *IEEE Transactions on Mobile Computing*. **2012**, *11*, 1983-1993.
3. Gu Y.; Lo A.; Niemegeers I. A survey of indoor positioning systems for wireless personal network. *IEEE Communications Surveys and Tutorials*. **2009**, *11*, 13-32.
4. Alarifi, A.; Al-Salman, A.; Alsaleh, M.; Alnafessah, A.; Al-Hadhrami, S.; Al-Ammar, M.A.; Al-Khalifa, H.S. Ultra Wideband Indoor Positioning Technologies: Analysis and Recent Advances. *Sensors*. **2016**, *16*, 707.
5. Feng C.; Valaee S.; Tan Z. Localization of wireless sensors using compressive sensing for manifold learning. In Proceedings of 2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Tokyo, Japan, 13-16 September 2009; pp. 2715-2719.
6. Harle R. A survey of indoor inertial positioning systems for pedestrians. *IEEE Communications Surveys and Tutorials*. **2013**, *15*, 1281-1293.
7. Bahl P.; Padmanabhan V. Radar: an in-building RF-based user location and tracking system. In Proceedings of 2000 Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), Tel Aviv, Israel, 26-30 March 2000; pp. 775-784.
8. Kushki A.; Plataniotis K. N.; Venetsanopoulos A. N. Kernel-based positioning in wireless local area networks. *IEEE Transactions on Mobile Computing*. **2007**, *6*, 689-705.
9. Rai A.; Chintalapudi K. K.; Padmanabhan V. N.; Sen R. Zee: Zero-effort crowdsourcing for indoor localization. In Proceedings of 18th Annual International conference on Mobile Computing Network, Istanbul, Turkey, 22-26 August 2012; pp. 293-304.
10. Feng C.; Valaee S.; Tan Z. Multiple Target Localization Using Compressive Sensing. In Proceedings of 2009 IEEE Global Communications Conference (GLOBECOM), Hawaii, USA, 30 November-4 December 2009; pp. 1-6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 April 2017 doi:10.20944/preprints201704.0114.v1

Peer-reviewed version available at *Appl. Sci.* **2017**, *7*, , 467; doi:10.3390/app7050467

11. Viel B.; Asplund M. Why is fingerprint-based indoor localization still so hard?. In Proceedings of 2014 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), Budapest, Hungary, 24-28 March 2014; pp. 443-448.
12. Pan J. J.; Pan S. J.; Yin J.; Ni L. M.; Yang Q. Tracking mobile users in wireless networks via semi-supervised colocalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. **2012**, *34*, 587-600.
13. Au A. W. S.; Feng C.; Valaee S.; Reyes S.; Sorour S.; Markowitz S. N.; Gold D.; Gordon K.; Eizenman M. Indoor tracking and navigation using received signal strength and compressive sensing on a mobile device. *IEEE Transactions on Mobile Computing*. **2013**, *12*, 2050-2062.
14. Redzic M.; Brennan C.; O'Connor N. SEAMLOC: seamless indoor localization based on reduced number of calibration points. *IEEE Transactions on Mobile Computing*. **2014**, *13*, 1326-1337.
15. Yang S.; Dessai P.; Verma M.; Gerla M. FreeLoc: calibration-free crowdsourced indoor localization. In Proceedings of The 32nd IEEE International Conference on Computer Communications (INFOCOM), Turin, Italy, April 14-19 2013; pp. 2481-2489.
16. Wu C.; Yang Z.; Liu Y. Smartphones based crowdsourcing for indoor localization. *IEEE Transactions on Mobile Computing*. **2015**, *14*, 444-457.
17. Yu N.; Xiao C.; Wu Y.; Feng R. A Radio-Map Automatic Construction Algorithm Based on Crowdsourcing. *Sensors*. **2016**, *16*, 504.
18. Kaemarungsi K., Distribution of wlan received signal strength indication for indoor location determination. in Wireless Pervasive Computing, 2006 1st International Symposium on, 16-18 Jan. 2006; pp. 2952-2957.
19. Zheng V. W.; Pan S. J.; Yang Q. and Pan J. J., Transferring Multi-device Localization Models using Latent Multi-task Learning. In Proceedings of the 23rd national conference on Artificial intelligence, Chicago, USA, 13-17 July 2008; pp. 1427-1432.
20. Tsui A. W.; Chuang Y. H. and Chu H. H. Unsupervised Learning for Solving RSS Hardware Variance Problem in WiFi Localization. *Mobile Networks and Applications*, **2009**, *14*, 677-691.
21. Rousseeuw P. J. and Driessen K. V. Computing LTS Regression for Large Data Sets. *Data Mining and Knowledge Discovery*, **2006**, *12*, 29-45.
22. Liu X. D.; He W.; Tian Z. S. The improvement of RSS-based location fingerprint technology for cellular networks. In Proceedings of 2012 International Conference on Computer Science and Service System (CSSS), Nanjing, China, 11-13 August 2012; pp. 1267-1270.
23. Hasani M.; Lohan E.-S.; Sydanheimo L.; Ukkonen L. Path-loss model of embroidered passive RFID tag on human body for indoor positioning applications. In Proceedings of 2014 IEEE RFID Technology and Applications Conference (RFID-TA), Tampere, Finland, 8-9 September 2014; pp. 170-174.
24. Latif S.; Memon A.; Chawdhry B.; Zielinski R.; Khan G. Accuracy Assessment of D-Model for Modeling Wall Attenuation in Indoor Environment. In Proceedings of 2014 Sixth International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), Tetovo, Macedonia, 27-29 May 2014; pp. 71-76.
25. Zhang L.; Shahrokh V.; Zhang L.; Xu Y.; Ma L. Signal propagation-based outlier reduction technique (SPORT) for crowdsourcing in indoor localization using fingerprint. Personal, Indoor, and Mobile Radio Communications (PIMRC), 2015 IEEE 26th Annual International Symposium on, Hong Kong, China, 30 Aug.-2 Sept. 2015; pp. 2008-2013.
26. Rahman Sakib M. S.; Quyum M. A.; Andersson K.; Synnes K.; Korner U. Improving Wi-Fi based indoor positioning using Particle Filter based on signal strength. In Proceedings of 2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Singapore, 21-24 April 2014; pp. 1-6.
27. Ma L.; Xu Y. Received signal strength recovery in green WLAN indoor positioning system using singular value threholding. *Sensors*. **2015**, *15*, 1292-1311.
28. Zelnik-Manor L. Self-tuning spectral clustering.*Advances in Neural Information Processing Systems*. **2004**, *17*, 1601-1608.
29. Candès E. J. Compressive sampling. In Proceedings of the international congress of mathematicians, Madrid, Spain, 22-30 August 2006; pp. 1433-1452.
30. Baraniuk R. G. Compressive sensing. *IEEE Signal Processing Magazine*. **2007**, *24*, 118-121.
31. Candès E. J.; Wakin M. B. An introdution to compressive sampling. *IEEE Signal Processing Magazine*. **2008**, *25*, 21-30.

32. Chen S. S.; Donoho D. L.; Saunders M. A. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing,* **1998**, *20*, 33-61.
33. Tropp J.; Gilbert A. Signal Recovery from Random Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory.* **2008**, *53*, 4655-4666.
34. Dai W.; Milenkovic O. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Signal Processing Magazine.* **2009**, *55*, 2230-2249.
35. Pourahmadi V.; Valaee S. Indoor Positioning and distance-aware graph-based semi-supervised learning method. In Proceedings of 2012 IEEE Global Communications Conference (GLOBECOM), Anaheim, USA, 3-7 December 2012; pp. 315-320.
36. Cheng B.; Yang J.; Yan S.; Fu Y.; Huang T. S. Learning with ℓ^1-Graph for image analysis. *IEEE Transactions on Image Processing.* **2010**, *19*, 858-866.