Amirkabir University of Technology
(Tehran Polytechnic)
Vol. 46, No. 1, Spring 2014, pp. 59- 66

Amirkabir International Journal of Science& Research
(Electrical & Electronics Engineering)
(AIJ-EEE)

Indoor Positioning and Pre-processing of RSS Measurements

V. Pourahmadi [1*] and S. Valaee[2]

1- Assistant Professor, Electrical Engineering Department, Amirkabir University of Technology, Tehran, Iran
2- Professor, ECE department, University of Toronto, Toronto, Canada

ABSTRACT

Rapid expansions of new location-based services signify the need for finding accurate localization techniques for indoor environments. Among different techniques, RSS-based schemes and in particular one of its variants which is based on Graph-based Semi-Supervised Learning (G-SSL) are widely-used approaches The superiority of this scheme is that it has low setup/training cost and at the same time it leads to low localization error. Analyzing the G-SSL method we can observe that its performance is highly dependent on its inputs (RSS measurements). The main objective of this work is to further improve the accuracy of G -SSL based schemes by performing multiple RSS measurements and then passing them through pre-processing blocks to improve the reliability of the corresponding RSS vector at each Sample Points (SPs). Experimental results are then followed to show the performance of the proposed method compared to what we get with the original G-SSL approach.

KEYWORDS

Recovered Hidden Markov Model, Electric Arc Furnace, Voltage Flicker, Power Quality parameters.

* Corresponding Author, Email: v.pourahmadi@aut.ac.ir

I. INTRODUCTION

Smartphones, tablets, and generally new generation of portable devices are usually equipped with many sensors to collect more information regarding the user or the environment that the device is operating. These extra information can be used to enhance quality or provide new services to the users.

Location-based services is an example of such applications and Global Positioning System (GPS) sensors are usually the primary source of location information for outdoor users. The only limitation of GPS sensors is that they require line-of-sight visibility of GPS satellites and this requirement is the main reason that the GPS location estimation does not have a good performance in indoor environments.

Several alternative mechanisms have been proposed to present indoor user localization with good accuracy [1], [2]. For instance, one approach is to use the current deployment of cellular networks. In this method the user location is determined using the cell-ID of the cell that the user is associated with and also may use enhanced observed time difference (E-OTD) of the network, [1], [3]. One advantage of cellular-based positioning systems is that this implementation does not require additional hardware installation and has low setup cost, but, the accuracy of this method, especially in rural areas, is not very high. RFID-based localization is another alternative, which despite its high accuracy, it has high deployment cost which is the main drawback of this approach. RFID-based scheme relies on initial deployment of some RFID tags over the area. Knowing the location of RFID tags, a mobile device can estimate its location by finding its closest RFID tag(s) [4], [5].

Another group of localization techniques are based on the users' Received Signal Strength (RSS) from the environment. One common selection is to use the signal that a user receives from the WLAN Access Points (APs). These approaches usually have a training phase in which the RSS is measured in some known grid points of the map (environment). These points are usually referred to as the Reference Points (RPs) or Finger-Prints (FPs). Having these RPs and their corresponding RSS, a user can find its location by comparing its measured RSS with the RSS of the known RPs [6], [7], [8], [9].

As Wi-Fi APs are usually available in indoor environments, RSS-based localization approaches do not require installation of new devices in the network and therefore they have a low deployment cost. One limiting factor of this method is that the initial training of the system is time consuming and usually a costly step.

Several ideas have been proposed trying to reduce the time/cost of the training phase. For instance, reference [10] suggests a scheme in which the RFs are collected during the operation of the system therefore it reduces the training phase. Some other methods like [11], [12], [13] propose a localization algorithm which tries to get to the same localization accuracy by using fewer number of RFs which consequently simplifies the training phase. This approach is based on the concept of Graph-based Semi-Supervised Learning (G-SSL). The details of G-SSL scheme will be described in Section II but the main idea is that we have not many RPs but instead we we have access to a large number of unlabelled RSS measurements (collected from different locations but we do not know their location). In other words, we have many Sample Points (SPs), i.e., RSS measurements, where only a few of them have labels (x and y coordinates). Based on G-SSL, we first estimate a label for SPs without labels and then use the whole database for localization of new users.

Several algorithms have been proposed to improve the performance of G-SSL based scheme. For example, in [14], the authors suggest a modification in the G-SSL cost function and show that how this new cost function can reduce the localization error. In this work, we want to look at the G-SSL procedure from different perspective and will try to improve the accuracy of location estimation by improving the quality of our RSS measurement. More specifically, we will try to increase the reliability of the measured RSS vector by first letting the device to perform multiple RSS measurements and then we pass the collected data through pre-processing blocks using which we try to identify more reliable components which ultimately improve the quality of the input information of the G-SSL scheme. We will also present some experimental results to show the performance of the proposed schemes.

This paper is organized as follows. First, we review the G-SSL approach in Section II and then in Section III we present G-SSL formulation when we have access to multiple RSS measurement at each SP. The proposed pre-processing schemes are discussed in Section IV and their performance are then evaluated with some real data in Section V. Finally, Section VI concludes the paper.

II. Review of G-SSL Approach

Having some labeled and some unlabeled SPs, in this paper we use G-SSL to first find an estimation of the location of the unlabeled readings and then use the resulted map to estimate the location of the newly measured RSSs from different APs. As we want to improve the accuracy of the estimation through pre-processing of the input data, we need to have a good understanding of G-SSL algorithm that we review in this section.

Consider an indoor environment with M APs, where we have $\ell+u$ measurements (ℓ labeled and u unlabeled). As labeled SPs are costly (should be collected by an expert user during the training phase), in practice, ℓ is usually not a large number. On the other hand, it is possible that we have large number of unlabeled SPs (it is cheep to collect and we can even collect them during the operation of the system, e.g., use some portable Wi-Fi devices to periodically report their RSS measurements while their are moving in the environment). With this explanation, a common assumption in practical localization techniques is that that $\ell \ll u$.

For each sample point $\mathcal{S}_i$ we use $(\mathbf{r}_i, \mathbf{b}_i)$ to denote pair of *(RSS measurement, Position)*. The RSS measurement $\mathbf{r}_i$ is an $M \times 1$ vector where its mth element represents the power level (in dBm) that $\mathcal{S}_i$ receives from the mth AP, and $\mathcal{S}_i[m]$ is set to -110 if $\mathcal{S}_i$ does not receive any signal from the mth AP. The position of measurement $\mathcal{S}_i$ is shown by $\mathbf{b}_i = \begin{bmatrix} x_i \\ y_i \end{bmatrix}$. Note that we only have access to the values of $\mathbf{b}_i$ that correspond to the labeled sample points.

We also use $\hat{\mathbf{b}}_i$ to represent our estimation of the position of the measurement $\mathcal{S}_i$. Initially, we do not have any estimation regarding the location of unlabeled data; therefore, we can set $\hat{\mathbf{b}}_i = \mathbf{b}_i$ if $i \in \{1, \cdots, \ell\}$, and $\hat{\mathbf{b}}_i = 0$ otherwise. Furthermore, $\mathsf{d}(\mathcal{S}_i, \mathcal{S}_j)$ and $\hat{\mathsf{d}}(\mathcal{S}_i, \mathcal{S}_j)$ denote the actual and the estimated distance between sample points $\mathcal{S}_i$ and $\mathcal{S}_j$, respectively.

A weight graph $\Omega = (V, E)$ associated with the collected SPs is constructed by adding a vertex to set V for each of the labeled and unlabeled RSS measurements, i.e., $V = \{\mathbf{r}_1, \mathbf{r}_2, \cdots, \mathbf{r}_L, \mathbf{r}_{\ell+1}, \cdots, \mathbf{r}_{\ell+u}\}$, and set E represents the set of all edges in this graph. For each labeled SP, the initial label of its corresponding vertex is set to $\mathbf{b}_j$.

In G-SSL, the label propagate from labeled vertices to other vertices is based on the similarity between different vertices. The similarity of two i and j vertices is denoted by w_{ij}, the weight of the edge connecting vertices i and j. Matrix $\mathbf{W}$ (with w_{ij} as its (i,j) entry) will be then an $(\ell+u) \times (\ell+u)$ matrix. The way to calculate w_{ij} is not unique and can be modified for different applications. One common estimation method is based on the *heat-kernel* function:

$$w_{ij} = \exp\left\{\frac{-\|\mathbf{r}_i - \mathbf{r}_j\|^2}{\tau}\right\}, \tag{1}$$

where τ is a parameter that can be set based on the application.

The aim of G-SSL is to find a label for all vertices (labeled and unlabeled) such that they fit the structure of the graph implied by the similarity measures. In this process, there is a possibility that we let G-SLL change the labels of all vertices, even ones correspond to labeled SPs. This may help improving the accuracy of the estimation in scenarios that there is a chance of error in the labels of the labeled SPs [15].

The output of G-SSL scheme is a set of estimated labels ($\hat{\mathbf{b}}_i$ for $i \in \{1, 2, \cdots, \ell+u\}$) satisfying: (i) for all $i \in \{1, 2, \cdots, \ell\}$, the estimated labels $\hat{\mathbf{b}}_i$ are close to the given labels $\mathbf{b}_i$; (ii) the estimated labels are smooth based on the similarity measure defined by $\mathbf{W}$ [15]. In other words, we want to determine $\hat{\mathbf{B}}^*$ such that:

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

where $\hat{\mathbf{B}}$ is $2 \times (\ell+u)$ matrix, columns of which represent the estimated labels for different SP.

The objective function in (2) has two terms: (i) the first term ensures that the estimated labels of the labeled data remain close to the co-ordinates reported during the training phase and (ii) the second term is used to ensure the smoothness of the results. Meaning that it increases the cost function by the value of $w_{ij}\|\hat{\mathbf{b}}_i - \hat{\mathbf{b}}_j\|^2$ if the estimated labels of nodes i and j are not equal.

III. Problem Formulation

RSS-based indoor localization schemes can use G-SSL to make a better localization map (using unlabeled data) which later improves the accuracy of location estimation. Furthermore, as discussed in Section I, in this study, we wanted to see how we can take advantage of multiple RSS measurement to improve the accuracy of G-SSL approach, and if any pre-processing scheme can be helpful in this regards.

Before we go to the details of the proposed pre-proccessing schemes, this section presents the mathematical formulation that will be used in the rest of the paper. For this study, we assume that we want to perform localization in an indoor environment with M APs and for each of the $\ell+u$ SPs, say S_l, we have measured the received signal strength for K time slots. The received RSS is denoted by $\mathfrak{R}_l$:

$$\mathfrak{R}_l = \left[\mathfrak{r}_l^1, \mathfrak{r}_l^2, \cdots, \mathfrak{r}_l^K\right], \tag{3}$$

where $\mathbf{R}_l$ is an $M \times K$ matrix and $\mathfrak{r}_l^k$ for $k \in \{1, 2, \cdots, K\}$ represents the vector of RSS measurment at S_l during the kth sampling period. In other words, $\mathfrak{r}_l^k(m, 1)$ (the mth row of $\mathfrak{r}_l^k$) represents the RSS of the mth AP at S_l during the kth measurement.[2] Ideally, if there were no noise and fading, all the measured RSS vectors should be the same, i.e., $\mathfrak{r}_l^1 = \mathfrak{r}_l^2 = \cdots = \mathfrak{r}_l^K$. However, in practical scenarios, $\mathfrak{r}_l^k$ for $k \in \{1, 2, \cdots, K\}$ are not equal due to the existence of noise and multi-path fading.

In G-SSL approach we have one RSS measurement per SP which is used in calculation of the weight matrix. As discussed before, in this paper, we aim to propose some schemes that use AP measurements in different time slots, $\mathfrak{R}_l$, and come up with a measurement vector for each SP, i.e., $\mathbf{r}_l$, which is more suitable for G-SSL.

One common approach which is helpful in canceling out the noise and the fading effect is to set $\mathbf{r}_l$ as the average of all RSS measurements:

$$\mathbf{r}_l = \frac{\sum_{k=1}^{K} \mathfrak{r}_l^k}{K}. \tag{4}$$

The resulted $\mathbf{r}_l$ can then be used in heat-kernel function to estimate the similarity matrix. Therefore, the entries of $\mathbf{W}$, i.e., w_{ij} can be evaluated as:

$$w_{ij} = \exp\left\{\frac{-\mathsf{d}(S_i, S_j)}{\tau}\right\}, \tag{5}$$

where τ is a design parameter and $\mathsf{d}(S_i, S_j)$ is a function that should be defined such that it represents the closeness of the two measurements at sample points S_i and S_j. Euclidian distance between $\mathbf{r}_i$ and $\mathbf{r}_j$ is one possible selection of $\mathsf{d}(S_i, S_j)$, i.e.:

$$\mathsf{d}(S_i, S_j) = \|\mathbf{r}_i - \mathbf{r}_j\|^2. \tag{6}$$

The localization error of this approach can be considered as a baseline performance when we want to study the performance of other schemes. In the next section we propose two pre-processing methods using which we intend to increase the localization accuracy.

IV. Pre-processing of RSS measurements

Analyzing RSS readings of APs in different time-slots, we can identify three areas that we might be able to improve our estimation by better handling of: I) Outliers (some RSS readings which are too far from the correct reading), II) large variation in RSS readings of some APs, 3) low RSS reading for some APs which are far from the SP. In the following, we introduce two schemes which try to reduce the effect of these non-idealities.

A. *Outliers removal using RPCA*

Having some noisy observations of a set of parameters, Principle Component Analysis (PCA) is probably one of the mains tools to remove the contribution of the noisy elements and reconstruct

[2]We set $\mathfrak{r}_l^k(m, 1) = -110$ if we do not receive any signal from the mth AP during the kth measurement.

the original data. However, in some applications with sparse noises (with probably large magnitude), the direct application of PCA leads to a result which is potentially very far from the actual data. Robust Principle Component Analysis (RPCA) is a technique that is proposed for noise reduction in these scenarios. In RPCA, we have a matrix $\mathbf{M}$ which is a grossly corrupted observation of low rank matrix $\mathbf{L}$, i.e., $\mathbf{M} = \mathbf{L}+\mathbf{S}$, where $\mathbf{S}$ is a sparse matrix. the results of [16] shows that under some week assumptions, the Principal Component Pursuit (PCP) estimate solving:

$$\begin{aligned} \min \quad & \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \\ s.t. \quad & \mathbf{L} + \mathbf{S} = \mathbf{M} \end{aligned} \tag{7}$$

exactly recovers the low-rank $\mathbf{L}$ and the sparse $\mathbf{S}$. In (7), $\|\mathbf{L}\|_*$ is the nuclear norm of matrix $\mathbf{L}$.

As defined in Section III, matrix $\mathfrak{R}_l$ represents the RSS measurements at S_l during K sampling periods. As we mentioned, in an ideal case $\mathfrak{R}_l$ should be of rank-1; however, due to the noise and fading, the eigenvalues of $\mathfrak{R}_l$ has more than one non-zero values (still one of the eigenvalues are much larger that the others). Apart from the variation caused by the environment, we can also see a number of big jumps in the RSS reading of one APs. These readings might be due to some malfunctioning of the sensor device or due to a deep fading so we can treat them as outliers. Usually there are only a few outliers in one reading matrix; however, since their values are very different from the actual value they have a significant effect on the statistics of the data, e.g., the average of the RSS.

Based on this observation, one idea of improving the accuracy of RSS readings is to try to decompose $\mathfrak{R}_l$ into an outlier matrix ($\mathfrak{E}_l$) and a low-rank matrix representing the original RSS reading plus noise ($\hat{\mathfrak{R}}_l$), i.e.,:

$$\mathfrak{R}_l = \hat{\mathfrak{R}}_l + \mathfrak{E}_l, \tag{8}$$

An important property of (8) is that since $\mathfrak{E}_l$ represents outliers, it can be modelled as a sparse matrix.

With this formulation, we are now able to apply RPCA techniques by substituting $\mathbf{M}$, $\mathbf{L}$, and $\mathbf{S}$ in (7) with $\mathfrak{R}_l$, $\hat{\mathfrak{R}}_l$, and $\mathfrak{E}_l$, respectively. Having $\mathfrak{E}_l$, we can identify the rows of $\mathfrak{R}_l$ which has some outliers; therefore, it is possible remove the contribution of their corresponding APs in calculation of $\mathsf{d}(S_i, S_j)$. For each point S_l we define vector $\mathbf{t}_l$ such that:

$$\mathbf{t}_l(m,1) = \prod_{k=1}^{K} \max\Big(\operatorname{sgn}\big(\nu_{th} - |\mathfrak{E}_l(m,k)|\big), 0\Big), \tag{9}$$

where ν_{th} is a threshold value (if the absolute value of $\mathfrak{E}_l(m,k)$ is larger that ν_{th}, that entry is considered as an outlier) and $\operatorname{sgn}(.)$ returns the sign of its input argument. Intuitively, $\mathbf{t}_l(m,1)$ is one only if there is no large estimated noise on the mth row of $\mathfrak{E}_l$ (corresponds to the mth AP).

As presented in Section III, the similarity measure between two sample points is usually evaluated based on (5) and (6). However, due to the existence of the outliers some entries of $\mathbf{r}_i$ and $\mathbf{r}_j$ are not accurate. To remove the contribution of these readings we first use (9) to compute $\mathbf{t}_i$ and $\mathbf{t}_j$ and define F_{ij} as:

$$F_{ij} = M \times \frac{<\mathbf{t}_i, \mathbf{t}_j>}{\|<\mathbf{t}_i, \mathbf{t}_j>\|_1}, \tag{10}$$

where $<\mathbf{t}_i, \mathbf{t}_j>$ is the inner product of $\mathbf{t}_i$ and $\mathbf{t}_j$, and $\|.\|_1$ returns the l_1-norm of its input argument. Given F_{ij}, we modify (6) as:

$$\mathsf{d}(S_i, S_j) = \|<\mathbf{r}_i - \mathbf{r}_j, F_{ij}>\|^2, \tag{11}$$

which in fact cancel out the effect of an AP if RPCA estimates that there is at least one outlier between the RSS readings of that AP. In Section V, we conduct an experiment and study the performance of this outlier removal technique.

B. Weighted Euclidian Distance

Based on its definition, each row of $\mathfrak{R}_l$ represents the signal strength that the device can receive from one of the APs at location S_l. Furthermore, the variation of RSS measurements in one row implies the strength of noise and fading power that exists on the link between the device and that particular AP. We define V_l as the vector representing the standard deviation of the RSS measurements corresponding to each AP, i.e.,:

$$\mathbf{v}_l = \left(\frac{1}{K}\sum_{k=1}^{K}(\mathfrak{r}_l^k - \overline{\mathfrak{r}_l})^2\right)^{\frac{1}{2}}, \tag{12}$$

where $\overline{\mathfrak{r}_l} = \frac{1}{K}\sum_{k=1}^{K}\mathfrak{r}_k^l$.

Another property of RSS measurement is that, due to the structure of the environment and the location of the APs, it is possible that at some locations, we receive a very week signal (or even no signal) from some of the APs. To identify these far away or not accessible APs at each location, we define $\mathbf{b}_l$ as a binary $M \times 1$ vector where:

$$\text{if } \overline{\mathfrak{r}_l}(m,1) > P_{th} \quad (13)$$
$$\text{then} \quad \mathbf{b}_l(m,1) = 1 \quad (14)$$
$$\text{otherwise} \quad \mathbf{b}_l(m,1) = 0, \quad (15)$$

where P_{th} is a threshold value.

Given $\mathbf{v}_l$ and $\mathbf{b}_l$, we can define vector Γ_l as:

$$\Upsilon_l = <\mathbf{b}_l, \mathbf{v}_l>, \quad (16)$$

where $<\mathbf{b}_l, \mathbf{v}_l>$ is the inner product of $\mathbf{b}_l$ and $\mathbf{v}_l$. Intuitively Υ_l represents the accuracy of the RSS readings of the APs which are *in range of location* S_l. In other words, the larger entries of Υ_l corresponds to RSS values which experience larger variation between different RSS readings. This implies that the signals that we receive from these APs experience a higher distortion from noise or fading.

Looking back into (5) and (6), we see that we have defined $\mathsf{d}(S_i, S_j) = \|\mathbf{r}_i - \mathbf{r}_j\|^2$ to find the similarity between the two sample points S_i and S_j. With this similarity measure, we give the same weights to the RSS reading that we receives from different APs. However, due to the noise/ fading effect we do not have the same confidence for all entries of $\mathbf{r}_i$ and $\mathbf{r}_j$, i.e., the RSS reading of the mth AP is more reliable if its corresponding value at $\Upsilon_i(m,1)$ and $\Upsilon_j(m,1)$ are smaller. Based on this observation, we can define $\mathsf{d}(S_i, S_j)$ such that it relies more on the RSS values which have lower variations (for S_i and S_j) and instead give lower weights to the measurements which are not very stable. To this end we define Γ_{ij} such that:

$$\Gamma_{ij} = M \times \frac{\mathbf{g}_{ij}}{\|\mathbf{g}_{ij}\|_1}, \quad (17)$$

where $\|.\|_1$ returns the l_1-norm of its input argument and:

$$\mathbf{g}_{ij}(m,1) = \begin{cases} \frac{1}{\sqrt[4]{\Upsilon_i(m,1)^2 + \Upsilon_j(m,1)^2}} & \Upsilon_i(m,1) > 0 \text{ or } \\ & \Upsilon_j(m,1) > 0 \\ 0 & \text{Otherwise} \end{cases} \quad (18)$$

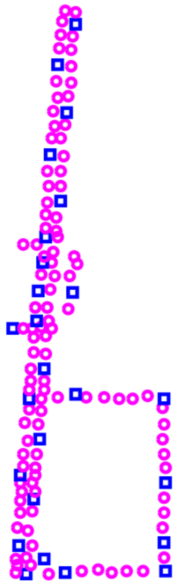

Fig. 1. The Actual Locations of All Sample Points.

The weighted Euclidian distance can then be computed as:

$$\mathsf{d}(S_i, S_j) = \| <\mathbf{r}_i - \mathbf{r}_j, \Gamma_{ij}> \|^2 \quad (19)$$

We then uses (5) to find the corresponding $\mathbf{W}$ matrix. As will be shown in Section V, the application of the modified distance function can significantly improve the performance of the G-SSL based localization algorithm.

V. Experimental Results

To study the performance of the propose schemes in practical environment, we have setup the following experiment where we collect 128 Sample Points (with an average distance of 1.5m) on the second floor of the Canadian National Institute for the Blind (CNIB) an area of approximately 18m×36m with 23 installed APs (we do not know the location of the APs). At each SP we have 10 measurements of the RSS from all APs. The device that we have used for this measurement is HP iPAQ hx4700. To evaluate the performance of G-SSL method, we partition the available SPs into two groups: (i) first, we select ℓ of the SPs (and their corresponding coordinates and RSS readings)

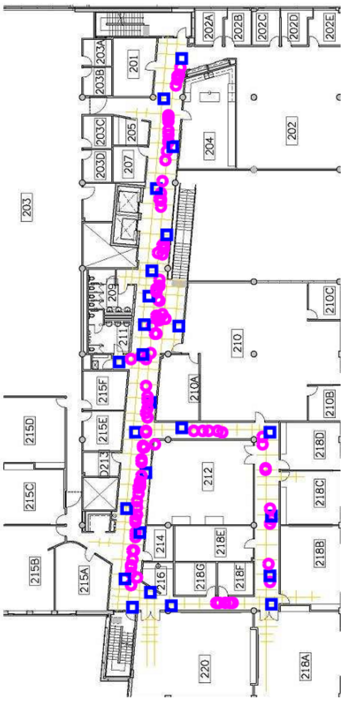

Fig. 2. The Localization Result Using the Graph-based Semi-supervised Method.

and set them as the labeled data; (ii) then, select the remaining $u = 128 - \ell$ SPs as unlabeled data[3]. As an example, Fig. 1 shows a case with $\ell = 25$ where the labeled data are represented by blue squares and the pink circles show the actual locations of unlabeled data.

As mentioned, in this experiment, for all SPs we have measured the RSS for 10 times; however, to have a more practical model, we assume that only the first three readings are available if a SP is selected as an unlabeled SP. This assumption is due to that generally unlabelled data is collected as a user walking through the environment so there is not much time for many measurements of the RSS. Clearly, when we have less RSS readings, we loose the accuracy of noise and outliers detection.

To see the performance of the proposed pre-processing techniques, first we evaluate the localization error if we apply the G-SSL approach where we use average of all RSS readings, equations (4) and (6), and do not perform any outlier removal and we have used the simple Euclidian distance to measure the similarity between two SPs. Figure 2 presents an example where $\ell = 25$, $u = 103$. In this figure, the blue squares represent the labeled data and the pink circles show what we have estimated as the locations of unlabeled data (the G-SSL does not have access to the coordinates of the unlabelled data it is supposed to find them). The localization error is computed as the average of the distance between the estimated and the actual locations of the unlabelled SPs. Repeating this process for different number of labeled SP, we evaluate the corresponding localization error in each case. The solid line in Fig. 3 depicts this localization error.

Having the performance of the baseline G-SSL method, we use RPCA to first remove outliers and then apply the G-SSL algorithm. As expected, we get higher localization accuracy when we reduce the contribution of the outliers on the RSS measurements. The dotted line in Fig. 3 shows the performance of this scheme. Note that in both schemes (Original baseline G-SLL and the RPCA method), we give the same weight for the RSS measurements from all APs.

To see the performance of *weighted Euclidian distance* scheme, we have evaluated the localization error of the same network when we follow the procedure described in Section IV-B. The result is depicted in Fig 3 (the dashed line). As can seen, considering the *weighted Euclidian distance* instead of the simple *Euclidian distance* significantly improve the localization accuracy.

As another example, we repeat the above experiment with another device. In this case we have used a Samsung Omnia II smartphone to collect the RSS measurements. We then applied the three localization algorithms for different number of labeled SPs. The results are shown in Fig. 4, where we can observe the performance of the RPCA and the weighted Euclidian distance methods compared with the original G-SSL approach.

VI. Conclusion

Collecting many Reference Points (RPs) during the training phase is one of the limiting factors of RSS-based indoor localization schemes. Graph-based Semi-Supervised Learning (G-SSL) is an scheme which tries to reduce the number of required RPs by substituting them with some unlabeled RSS measurements. In this paper, we first

[3]For these SPs, we assume that we only have access to their RSS readings and not their corresponding locations.

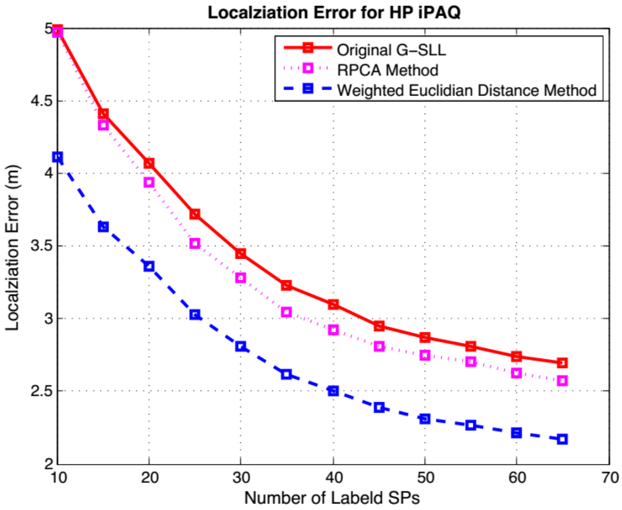

Fig. 3. The G-SSL Based Localization Error for Data Collected with HP iPAQ .

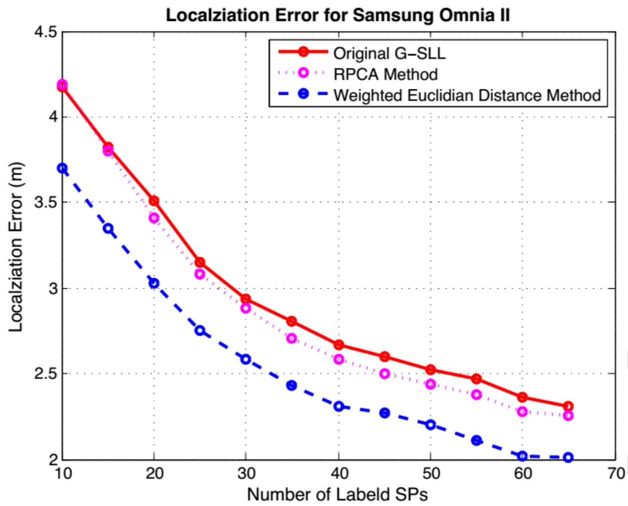

Fig. 4. The G-SSL Based Localization Error for Data Collected with Samsung Omnia II.

discuss how the accuracy of the collected RSSs (at RPs and test points) impact the performance of the G-SSL, and then, present some schemes that gets raw RSS measurements as input and provides RSS vectors with statistics which are more suitable for G-SSL based localization schemes. The performance of the proposed scheme have been evaluated using some data collected in a real-world scenario. The resutls show a significant improvement compared to the original G-SSL.

REFERENCES

[1] H. Liu, H. Darabi, P. Banerjee, and J. Liu, “Survey of wireless indoor positioning techniques and systems,” *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 37, no. 6, pp. 1067–1080, 2007.

[2] F. Seco, A.R. Jiménez, C. Prieto, J. Roa, and K. Koutsou, “A survey of mathematical methods for indoor localization,” in *Intelligent Signal Processing, 2009. WISP 2009. IEEE International Symposium on*. IEEE, 2009, pp. 9–14.

[3] J.J. Caffery and G.L. Stuber, “Overview of radiolocation in CDMA cellular systems,” *Communications Magazine, IEEE*, vol. 36, no. 4, pp. 38–45, 1998.

[4] JR Guerrieri, MH Francis, PF Wilson, T. Kos, LE Miller, NP Bryner, DW Stroup, and L. Klein-Berndt, “RFID-assisted indoor localization and communication for first responders,” in *Antennas and Propagation, 2006. EuCAP 2006. First European Conference on*. IEEE, 2006, pp. 1–6.

[5] X. Zhang, J. Peng, and X. Cao, “RFID indoor localization algorithm based on dynamic netting,” in *Computational and Information Sciences (ICCIS), 2010 International Conference on*. IEEE, 2010, pp. 428–431.

[6] C. Feng, W.S.A. Au, S. Valaee, and Z. Tan, “Compressive sensing based positioning using RSS of WLAN access points,” in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[7] C. Feng, W. Au, S. Valaee, and Z. Tan, “Received signal strength based indoor positioning using compressive sensing,” *Mobile Computing, IEEE Transactions on*, , no. 99, 2011.

[8] K. Kaemarungsi and P. Krishnamurthy, “Modeling of indoor positioning systems based on location fingerprinting,” in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, march 2004, vol. 2, pp. 1012 – 1022 vol.2.

[9] Jun Ma, Xuansong Li, Xianping Tao, and Jian Lu, “Cluster filtered KNN: A WLAN-based indoor positioning scheme,” in *World of Wireless, Mobile and Multimedia Networks, 2008. WoWMoM 2008. 2008 International Symposium on a*, june 2008, pp. 1 –8.

[10] K. Chintalapudi, A. Padmanabha Iyer, and V.N. Padmanabhan, “Indoor localization without the pain,” in *Proceedings of the sixteenth annual international conference on Mobile computing and networking*. ACM, 2010, pp. 173–184.

[11] B. Yang, J. Xu, J. Yang, and M. Li, “Localization algorithm in wireless sensor networks based on semi-supervised manifold learning and its application,” *Cluster Computing*, vol. 13, no. 4, pp. 435–446, 2010.

[12] J. Pan, S. Pan, J. Yin, L. Ni, and Q. Yang, “Tracking mobile users in wireless networks via semi-supervised colocalization,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, , no. 99, pp. 1–1, 2012.

[13] T. Pulkkinen, T. Roos, and P. Myllymäki, “Semi-supervised learning for WLAN positioning,” *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 355–362, 2011.

[14] Vahid Pourahmadi and Shahrokh Valaee, “Indoor Positioning and Distance-Aware Graph-Based Semi-Supervised Learning Method,” in *Global Communications Conference (GLOBECOM), 2012 IEEE*. IEEE, 2012, pp. 315–320.

[15] X. Zhu and A.B. Goldberg, “Introduction to semi-supervised learning,” *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1–130, 2009.

[16] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, “Robust Principal Component Analysis?,” *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.