Secure and Robust Wi-Fi Fingerprinting Indoor Localization

Wei Meng*, Wendong Xiao**, Wei Ni**, Lihua Xie*
*Nanyang Technological University, Singapore. Email: {meng0025, elhxie}@ntu.edu.sg
**Institute for Infocomm Research, A*Star, Singapore. Email: {wxiao, wni}@i2r.a-star.edu.sg

***Abstract*—Indoor positioning has emerged as a widely used application of Wi-Fi wireless networks. Fingerprinting techniques can provide a low-cost and high-accuracy localization solution by utilizing in-building communication infrastructures. However, existing fingerprinting localization algorithms are not resistant to outliers, for example, the accidental environment changes, access point (AP) attacks. Another drawback is that traditional *K* nearest neighbor (*KNN*) algorithm in the literature may not select the candidate reference points (RPs) correctly. In this paper, we propose a novel environmentally robust and attack resistant probabilistic fingerprinting localization method. In the offline phase, the distribution estimation of the signal strength is performed using probabilistic histogram method. Then in the online phase, a three-step location sensing method is proposed. In the first step, a simple and efficient outlier detection method named non-iterative "RANdom SAmple Consensus" (RANSAC) is run to detect and eliminate part of APs from which the signals measured are severely distorted by unexpected environment effects. In the second step, a novel region-based RP selection method which works like a "*family of probability*" is proposed to improve the possibility of the correctness of selection of the nearest RPs. In the final step, a simple weighted mean method is adopted for location determination. In the experiment section, we demonstrate the proposed method in our lab and find that the proposed strategies are resistant to outliers and can improve the localization accuracy effectively compared with existing methods.**

I. Introduction

The objective of this paper is to provide solutions to perform real-time indoor positioning with Wi-Fi techniques. Localization is an important topic in wireless networks. However, indoor positioning is challenging because of the non-line-of-sight (NLOS) transmission between emitters and receivers and the multi-path effect. There are various obstacles, for example, walls, cubicles, equipments, human beings which influence the propagating of the electromagnetic waves. As a solution, fingerprinting indoor positioning techniques can provides a low-cost and high-accuracy localization by utilizing in-building communication infrastructures. The fingerprinting localization techniques can be categorized into two broad categories: deterministic techniques [2] and probabilistic techniques [3]. Our work lies in the second category.

However, existing fingerprinting localization algorithms are not resistant to outliers, for example, the accidental environment changes, access point (AP) attacks. Due to dynamic environment, the measured fingerprint may deviate significantly from those stored in the RPs which are near the tag. This may lead to large localization errors. Another key issue in the fingerprinting localization algorithm is how to choose the candidate RPs to be compared with the tag and then to decide the region the target may reside in. However, the traditional *K* nearest neighbor (*KNN*) algorithm in the literature may not select the candidate RPs correctly which results in a large localization error.

To overcome the above two drawbacks, in this paper, we propose a probabilistic region-based fingerprinting method to reduce the outlier effect and improve the localization accuracy. In the offline training phase, the probability distributions of the signal strength received by each of the RPs from each AP are constructed using a probabilistic histogram method. Then in the online phase, we propose a three-step location sensing algorithm. In the first step which is named as *outlier detection and elimination*, we propose a simple non-iterative "RANdom SAmple Consensus" (RANSAC) method to detect and eliminate part of APs from which the signals measured by the tag are severely distorted by an unexpected environment effect. In the second step, a *region*-based reference point selection method is proposed to improve the robustness to the changes of the environment. In the final step, the unknown tag's coordinate is obtained by using weighted mean method.

II. Robust Region-based Fingerprinting Method

A. System Overview

The problem of interest is to localize or track the positions of the tags in the physical area of interest using the Wi-Fi technologies. We define the signal-strength vector received by a tag as $\mathbf{s} = (s_1, s_2, \ldots, s_n)$, where $s_j, 1 \leq j \leq n$ denotes the RSS value from the jth access point (AP) and n is the number of APs in the physical area of interest. There are m reference points used in the offline training phase, the RSS matrix received at the ith reference point, $1 \leq i \leq m$ can be denoted as $\theta_i = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^{p \times n}$, $\theta_j = [\theta_{j_1}, \ldots, \theta_{j_p}]^T$, p is the number of RSS samples received by a RP from the jth AP. Please note that some of θ_{j_t} may be equal to 0, which means that at time instant t, the RP does not detect the signal from the jth AP.

Our system can be divided into two phases: (1) *offline* phase in which we perform the distribution estimation and (2) *online* phase, in which we use a three-step location determination technique to infer the tag's location.

B. Offline Training Phase

During the offline phase, at each reference point with known location, the RSS measurements from access points (APs) are intensively sampled. Then we plot histograms to approximate their density functions. In [3], the authors have a result that probabilistic histogram method can lead to slightly lower location error on the average than other kinds of probabilistic method, such as nearest neighbor method and kernel method [3]. The histogram method is closely related to *discretization* of continuous values to discrete ones. For our case, we have one-dimensional variables and that the minimum and maximum of the RSS values are known. The method requires that we fix a set of bins, i.e., a set of non-overlapping intervals that cover the whole range of the variable from the minimum to the maximum. The number and widths of the bins are two adjustable parameters which will affect the resulting density estimate. For simplicity, we use equal-width bins. Generally the distribution of RSS values does not follow the Gaussian distribution due to the multi-path effect. The probability of each bin is calculated as follows:

$$P(RSS \in i\texttt{th interval}) = \frac{Count(i\texttt{th interval})}{\texttt{Size of Training Data}}. \tag{1}$$

C. Online Phase

In the online phase, we propose to use a three-step location sensing algorithm.

TABLE I
OUTLIER DETECTION

Rank	AP_1	AP_2	$\ldots$	AP_n
1	ID_{11}	ID_{21}	$\ldots$	ID_{n1}
2	ID_{12}	ID_{22}	$\ldots$	ID_{n2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
m	ID_{1m}	ID_{2m}	$\ldots$	ID_{nm}

1) Outlier Detection and Elimination: We assume that at each time instant, only the signal wave from a small portion of the access points (APs) are distorted dramatically due to the accidental environment changes, access point (AP) attacks. This is reasonable in most cases of real applications. On the other hand, it is known that if signals from most of the APs are influenced, the fingerprinting method does not work no matter whether the deterministic or probabilistic approaches are used. The main objective of this part is to find out and eliminate part of APs from which the signals measured by tags are severely distorted by the unexpected environment effect. As introduced before, in the *online* phase, the signal-strength vector received by a tag is $\mathbf{s} = (s_1, s_2, \ldots, s_n)$, where n is the number of the APs. Then we *map* these RSS values into the possibility measure according to (1). The sorting of these possibilities for APs is listed individually as shown in TABLE I, where each column presents the rank-list for an AP and ID_{ij} denotes the ID of the reference point which has jth largest value for ith AP. From the table, for each AP, we search the coordinates of the reference points which are in rank 1-q and then calculate their center point's coordinates $r_i, 1 \le i \le n$ as follows:

$$r_i = \frac{1}{q}\sum_{j=1}^{q} p_{ID_{ij}}, \tag{2}$$

where $p_{ID_{ij}}$ denotes the coordinates of the reference point which has ID number ID_{ij} as presented in TABLE I. Please note that q should be chosen carefully which can not be too large. Empirically, we choose q equal to 2 in our experiments.

After obtaining $r_i, 1 \le i \le n$, we run the outlier detection algorithm to find out and eliminate part of APs from which the signals measured by the tag are severely distorted by the unexpected environment effect. If the outliers exist, then some points (outliers) $r_i, 1 \le i \le n$ may have larger distances to most of the other points (inliers) as shown in Fig. 1.

The main idea of the outlier detection algorithm is adopted from "RANdom SAmple Consensus" (RANSAC) [4] which is widely used in image processing area. For our case, we use a simplified non-iterative RANSAC algorithm to save the time in signal processing. Alternately, we can use RANSAC or multi-level RANSAC, however, their computational complexities are high and may not be appropriate for online tracking applications. The pseudocode of outlier detection algorithm is addressed in Algorithm 1. Two parameters T and w need to be chosen carefully where T denotes the distance threshold for determining when a datum fits the model and w denotes the number of close data values required to assert that a model fits well to data. T can not be small or large, in our experiments, T is set to around 5 meter. For the parameter w, it should be no less than $n/2$ (n is even) or $(n + 1)/2$ (n is odd).

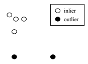


Fig. 1. Outlier Detection

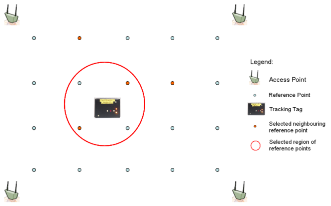

Legend:
Access Point
Reference Point
Tracking Tag
Selected neighbouring reference point
Selected region of reference points

Fig. 2. Region-based reference points selection.

2) Region-based Reference Points Selection: After eliminating the signals measured by reference points (RPs) and the tracking tag/target from the APs from which the outliers happened, we will select some candidate RPs to be compared with the tag. In the literature, the most commonly used method is K-nearest-neighbor (KNN) method. The KNN method is based on the assumption that if the reference point j is near

Algorithm 1: Outlier Detection

```
Input: [i] $r_i, 1 \leq i \leq n$ as in (2). [ii] Model: Euclidean
distance $\|r_i - r_j\|_2, 1 \leq \{i, j\} \leq n, i \neq j$. [iii] $T$, the distance
threshold for determining when a datum fits the model.
[iv] $w$, the number of close data values required to assert
that a model fits well to data.
Output: The IDs of the APs from which the
measurements by the tracking tag are seen as outliers.
Initialization: $t = 0$.
for $i = 1 : n$ do
    for $j = 1 : n,\ j \neq i$ do
        if $\|r_i - r_j\|_2 \leq T$ then
            $t \leftarrow t + 1$
        else
            $t \leftarrow t$
        end if
    end for
    if $t \geq w$ then
        Measurement from AP $i$ is an inlier
    else
        Measurement from AP $i$ is an outlier
    end if
end for
```

the tag, then the Euclidean distance in signal strength between them is small, and vice versa. Also the *KNN* method selects each candidate RP individually. Due to fluctuations in RSS measurements, this may result in that the chosen RPs may not be the actual K points which are nearest to the tag as shown in Fig. 2.

To solve this problem, we develop a region-based reference points selection method which works like a *family of probability* instead of single probability as in *KNN*-based method, to improve the robustness and accuracy. We form N reference points (N could be 4, 6, etc.) into a group, and each group of reference points covers a region in our Wi-Fi indoor positioning test-bed. From the viewpoint of fingerprinting mechanism, each fingerprint is a region-based group of reference points in our approach, instead of an individual reference point. To determine which fingerprint best matches the tag's RSS measurements, we calculate the sum of probability (SOP) in probability space, and the region with the minimum SOP will be selected as the matched region (see Fig. 2). The SOP is calculated as follows:

$$\mathsf{SOP} = \sum_{i=1}^{N} P_i = \sum_{i=1}^{N} \sum_{j=1}^{n'} P_{ij}, \tag{3}$$

where N is the number of RPs in the group and n' denotes the number of APs left after outlier detection. P_{ij} denotes the probability value mapped to ith RP and it is calculated according to (1) where $RSS = s_j$. Note that in (3), we use sum of probability instead of product of probability (POP). This is mainly because that in some case, P_{ij} may be equal to 0 which makes the product equal to 0.

Actually, in the candidate RPs selection process, we do not need to search the data through all the RP group list. We can just choose some groups in which the rank 1-4 reference points in TABLE I are involved which may speed up the localization process.

3) Location Estimation: The final location determination method is simple and we assume that the tracking tag's coordinates is a linear combination of the N reference points' coordinates $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ in the group selected in the last step.

$$\hat{\mathbf{x}}_0 = \sum_{i=1}^{N} w_i \mathbf{x}_i, \tag{4}$$

where $w_i = \frac{P_i}{\sum_{i=1}^{N} P_i}$, $P_i = \sum_{j=1}^{n'} P_{ij}$ as shown in (3).

III. Experimental Results

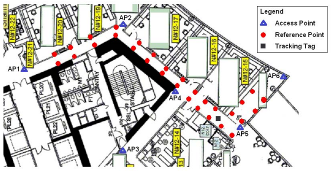


Fig. 3. The layout of the experimental test-bed.

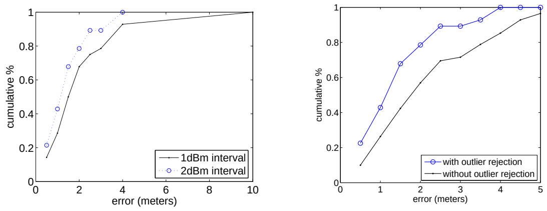
We evaluate the performance of our proposed environmental robust region-based fingerprinting algorithm using experimental studies in our lab environment as shown in Fig. 3. We assume each access point's communication radius is sufficient to cover the whole area of the network. The experiment setup is as follows. In the whole area, 6 access points (APs) are placed at different positions. Signal strengths are sampled at 30 reference points (RPs) and are stored into the calibration database. At each reference point, the sampling frequency is 10 Hz and the number of samples are about 1000. The distance between two neighboring RPs are set to about $2m$. Then in the online phase, a Wi-Fi tag is used as the tracking target. It measures a sample vector of RSSs from APs at its position and then the sample will be sent to a central server. We move it along a track in unit step, and calculate its position estimate at the center server. The two parameters used in the outlier detection, T and w are set to be 5m and 3 respectively. The performance metric of our experiment study is defined as the position estimate's error, i.e., $error = \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|_2$, where $\mathbf{x}_0$ and $\hat{\mathbf{x}}_0$ denote the tracking tag's real coordinates and the estimated one.

A. Impact of Width of Bins in Probabilistic Histogram

In our offline phase, we use a probabilistic histogram method to estimate the distributions of signal strength measurements. As we point out in section II-B, the width of bins is an important parameter which affects the final localization accuracy. In this part, we change the width between 1dB and 2dB and keep the number of reference points and access points equal to 30 and 6 respectively. The localization results of using different widths of bins is shown in Fig. 4(a). As seen from the figure, we can find that the performance of the 2-dB width histogram method is better than its 1-dB counterpart. The reasons for this result is that the width of bins is related to the interpolation of the distributions between a number of the most probable locations which is a challenging topic in the fingerprint-based localization problems. 2dB width of bins in some extent help each reference point cover larger area of signal strength distribution and not just the distribution of its own position. This is because the signal strength varies at different points in the sensing field.

B. *Outlier Rejection*

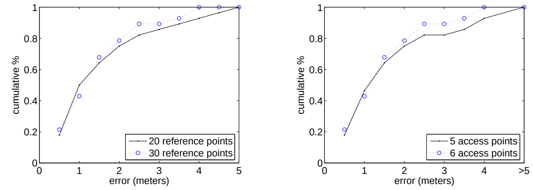
As we know, outlier, for instance, accidental environment change and AP attacks, is a big issue which makes the fingerprint based indoor localization difficult. In order to see how well our proposed method works in harsh environments, we artificially make two APs attacked. Then we run our experiments include and exclude the outlier rejection algorithm respectively. The result is shown in Fig. 4(b). As seen from the figure, with outlier rejection algorithm, the localization performance is better.



(a) (b)

Fig. 4. Cumulative percentile of error distance for (a) width=1dB and width=2dB; (b) with and without outlier detection.

C. *Impact of Density of RPs and APs*



(a) (b)

Fig. 5. Cumulative percentile of error distance for difference densities of RPs and APs.

In our experiments, we also change the density of the reference points. The localization result is shown in Fig. 5(a). The performance of using 30 RPs is slightly better than which using 20 RPs. However, at some points, the localization accuracy of using 20 RPs may be better than that of using 30 RPs. This is reasonable because with lower density of RPs, the probability of correct selection of group of nearest RPs will be higher. In order to investigate the impact of density of APs. Inversely, in our experiments, we shut down one AP, and with 5 APs left. Then one parameter used in outlier detection *w* is changed to be 2. The localization performance is tested and the result is shown in Fig. 5(b). From the figure, we can see that with higher density of APs, the localization performance is better. However, our method is not sensitive to the density of the APs. According to our experience, 4-5 APs is enough for the accuracy requirement of our clients.

D. *Comparison with other methods*

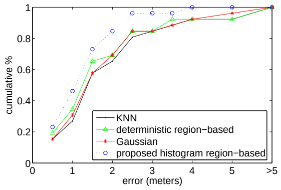


Fig. 6. Cumulative percentile of error distance for different methods.

In the experiments, as a benchmark, the performance of the proposed probabilistic region based method is compared against *KNN* method, deterministic region based method and Gaussian based probabilistic method. The performance comparison result between four different methods is presented in Fig. 6. From the figure, we can find that the proposed histogram region-based method has the best performance. Usually, a probabilistic method is better than a deterministic method. The reason for the fact that the proposed histogram method is better than Gaussian based probabilistic method is that in some cases, the real distribution of the signal strength has a significant bias with the Gaussian distribution. From the figure, we also find that deterministic region based method is slightly better than the *KNN* method. This is because that region based reference point selection method is more robust to the environment changes than *KNN* method.

IV. Conclusion

In this paper we have proposed an secure and environmentally robust region-based fingerprinting method for indoor positioning in Wi-Fi wireless networks. In the offline phase, a probabilistic histogram method is adopted to estimate the distribution of signal strength and then a radio map is built. In the online phase, we utilize a Wi-Fi tag to collect RSS information and localize the tag by using a three-step location sensing method. Implementation results showed that our proposed technique leads to a better localization performance.

References

[1] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, “LANDMARC: indoor location sensing using active RFID,” *Wireless Networks*, vol. 10, pp. 701-710, 2004.

[2] P. Bahl and V. N. Padmanabhan, “Radar: An in-building RF based user location and tracking system,” in proceedings of *IEEE Infocom*, 2000.

[3] T. Roos, P. Myllymaki, H. Tirri, P. Misikangas, and J. Sievanen, “A probabilistic approach to WLAN user location estimation, ” *Int'l J. Wireless Information Networks*, vol. 9, no. 3, pp. 155-164, July 2002.

[4] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, ” *Communications of the ACM*, vol. 24, no. 6, pp. 381-395, June 1981.