Abstract—WiFi localization based on fingerprinting method became popular during last decade. However, collecting information to construct a WiFi signal map is challenging due to high cost. We assume that it is possible to gather such information with cheap cost using crowd sourcing with smartphones. Each user’s current location is estimated by either GPS, WiFi, radio cell signals or manually set by the user and the WiFi scan information is collected into a server. In this paper, we focus on outdoor geolocalization based multichannel WiFi Map where each channel conveys different accuracy of the reference position. The proposed approach optimizes importance weight on the different channels of WiFi map using Least Squares. Gradient Descent algorithm optimizes weight parameter to minimizes localization error on the training set. The experiment shows the proposed approach outperforms localization with equal importance.

I. INTRODUCTION

It is well known that wireless signal strength including WiFi decays as the device move further from an AP. One can estimate distance from each AP scanned by the device from RSSI [1]. However, due to multipath phenomenon on non-line-of-sight (NLOS), trilateration which uses distances from known position is ineffective for WiFi based geo-localization. Alternatively, during the recent decade, localization based on fingerprinting method became popular by providing resolution to the multipath issue with matching current RSSI observations to the database collection of WiFi signal map [2].

In this paper, we focus on outdoor geolocalization based on WiFi signals. In general, GPS provides accurate outdoor location but it takes tens of second to obtain GPS signals and calculate user’s current location, especially in the city surrounded by skyscrapers. WiFi based localization is a good alternative which only takes a few second to scan APs and provides geolocation with compatible accuracy.

As noted earlier, fingerprinting method is prevalent for WiFi based localization but it requires construction of WiFi signal map in prior to localization. It is very challenging to gather WiFi signal environment covering whole outdoor area since one needs to scan for WiFi APs everywhere. We assume that it is possible to gather such information with cheap cost using crowd sourcing with smartphones.

Crowd sourcing data contains reference position from each user’s current location estimated by either GPS, WiFi, radio cell signals or manually set by the user. The data also contains WiFi AP scan log at the location is collected and stored in the server. Each estimation type among GPS, WiFi, Cell or manually set has different level of innate accuracy. We can aggregate all such data into one WiFi map but it may help to give higher weights to the WiFi signal information collected with higher accuracy of the referenced position. For the convenience, we may form a WiFi map with each location estimation type separately.

We are interested in Fingerprinting based localization which fuses multichannel WiFi signal map where each channel of the map is collected with the different accuracy of reference position. While each source has different level of accuracy, fusing such information will enhance accuracy. Manually annotated position will be the most accurate while the cost of collecting such data will be the highest. On the contrary, automatically collected WiFi scan log with WiFi or cell based estimation will have larger collection but less accurate.

Both information can be fused to enhance localization accuracy as well as database coverage of the WiFi map. During the localization phase after the map construction, if the WiFi AP matches to the WiFi map of the higher accuracy, it should have higher weights to the one that matches to lower accuracy. However, if the none of the AP matches to the accurate map, one should use less accurate map instead.

The remainder of this paper is organized as follows. Section II reviews WiFi localization based on Fingerprint method. The proposed method is described in Section III. Quantitative and qualitative analysis of the experiment results are evaluated in Section IV. Merits of the proposed framework are discussed in Section V, followed by concluding remarks in Section VI.

II. WIFI LOCALIZATION METHOD BASED ON RSSI FINGERPRINT

In this section, we introduce an improved WiFi Fingerprinting method which includes AP matching rate between the scanned AP list and constructed WiFi map and additional weight parameters giving higher weights to the APs with strong signals.

Current location of a person can be estimated by comparing WiFi environment at the location to the WiFi signal map. WiFi environment is scanned from a smartphone built-in scanner and written in a log with MAC address and RSSI of each observed AP. Outdoor geolocations are partitioned into grid cells and one compares WIFI signal strength pattern which is referred as Fingerprint at the current location to ones in the...
WiFi signal map. In principle, location can be estimated from the location of the grid cell that bears the most similar signal pattern. Nearest neighbor (NN) model is the best fit for such approach, given similarity metric of the WiFi environment.

\[ \hat{y} = \frac{1}{Z} \sum_{c \in C} w_c p_c \tag{1} \]
\[ Z = \sum_{c \in C} w_c \tag{2} \]
\[ w_c = m_c \sum_{a \in c} \log(r(a) + S)(\theta - \Delta r_c(a)) \tag{3} \]

Our Fingerprinting based grid cell weight model is described in (3). The weight model of a grid cell \( c \) consists of three terms, matching ratio \( m_c \), logarithm term and \( \Delta r \) term. For each AP \( a \) in a cell \( c \), RSSI difference between the one observed in the device and the one in the WiFi map is represented in \( \Delta r_c(a) \) where \( \theta \) converts difference to the similarity. While exponential function is the one widely used for the purpose, we adopt linear function for convenience when deriving gradient.

The logarithm term gives higher weights to the APs with strong signal. So et al. [3] mentioned in their paper that localization accuracy can be improved by giving higher weights to the APs with the stronger signals. Inspired by their work, the logarithm term provide the same effect. RSSI observed in the device is captured by \( r(a) \) and a positive constant \( S \) is added to ensure positive input to the logarithm function.

Zhang et al. [4] introduced localization based on cosine similarity of the WiFi scanned log of the device and the database. However, their approach assumes all APs are observed at the time of localization, thus, only considers signal vector difference of the matched APs. It is trivial that not all the APs will be observed in the real world environment. If the both grid cells have the same similarity score, one with higher matching ratio should be chosen because it gives less uncertainty which may caused by unobserved APs. The term \( m_c \) in (3) represents the matching ratio.

As described in section I, we gather information to construct WiFi map by crowd sourcing with various source of reference position. For example, type of reference position source includes GPS, WiFi, LTE cell signal based localization and manual annotation of a user. Each source forms different channel of WiFi map, and using the multichannel map, the weight model is expanded as follows, where \( \Delta r^s_c(a) \) calculates RSSI difference of a AP \( a \) in the source channel \( s \):

\[ w_c = m_c \sum_{a \in c} \sum_{s \in c} \log(r(a) + S)(\theta^s - \Delta r^s_c(a)) \tag{4} \]

Importance weight of each channel is described in \( \theta_s \). Source channel of higher accuracy should have higher \( \theta_s \) values so that the similarity of the same RSSI difference will be boosted more compared to the one in low accurate channel.

### III. PARAMETER OPTIMIZATION FOR MULTICHANNEL WiFi LOCALIZATION

For the proposed weights model described in the previous channel, we optimize importance weights \( \theta_s \) of each source \( s \) by gradient descent method. Loss function is defined from Least Squares as in (5).

\[ \min_{\theta_s} \mathcal{J}(\theta^s) = \sum_{n=1}^{N} ||y_n - \hat{y}_n||^2 \tag{5} \]

where \( y_n \) is the true position and \( \hat{y}_n \) is estimated position for the n-th training sample. The reference cell position \( p_c \) is consisted of longitude \( p^x_c \) and latitude \( p^y_c \). For computational convenience, we can shift the coordinate system of each sample by \( -y_n \) so that true position reside on the origin and the localization error is represented as norm-2. Importance of the source channel will not be affected.

The objective function (5) of the \( -y \) shifted version for a sample is expanded as follows;

\[ ||\hat{y}||^2 \]
\[ = \left| \left| \frac{1}{Z} \sum_{c \in C} w_c p_c \right| \right|^2 \]
\[ = \left| \left| \sum_{c} \left( \sum_{s} \left( \sum_{a} \log(r(a) + S)(\theta^s - \Delta r^s_c(a)) \right) \right) m_c \right| \right|^2 \]
\[ = \left| \left| \sum_{c} \left( \sum_{s} \left( \sum_{a} \log(r(a) + S)(\theta^s - \Delta r^s_c(a)) \right) \right) m_c \right| \right|^2 \]
\[ = \left| \left| \sum_{c} \left( \sum_{s} \left( \sum_{a} \log(r(a) + S)(\theta^s - \Delta r^s_c(a)) \right) \right) m_c \right| \right|^2 \]

where \( p_c \) is the coordinate of the pCells shifted by \( -y \), \( m_c \) is the matching ratio of APs observed from the device and those in the pCell, \( r(a) \) represents RSSI observed from the device and \( \Delta r^s_c(a) \) is the RSSI difference between the device RSSI and the mean RSSI in the pCell. From this point, we omit second term with \( p^y_c \) to enhance readability.

We further replace terms as follows to increase readability:

\[ L^s_c = \sum_{a} \log(r(a) + S) \tag{10} \]
\[ D^s_c = \sum_{a} \Delta r^s_c(a)(\log(r(a) + S)) \tag{11} \]
Algorithm 1 Gradient Descent Algorithm

1: procedure GradDesc($D, y)$
2: for $s \in S$ do
3: $\theta^s \leftarrow 40$
4: while $\sum_{k=1}^{\lVert D \rVert} \|y_k - \hat{y}_k\|^2_2 \geq \epsilon$ do
5: $\theta \leftarrow \theta - \alpha \frac{\partial J(\theta)}{\partial \theta}$
6: return $\theta$

Then, we can rewrite (9) as:

$$\frac{\text{Numerator}}{\text{Denominator}} = \frac{\sum_s (\sum_c L^s_c m^s_c p^s_c)^2 \theta^s - \sum_s \sum_c D^s_c m^s_c}{\sum_s (\sum_c L^s_c m^s_c)^2 \theta^s - \sum_s \sum_c D^s_c m^s_c}$$

We optimize (12) using Gradient Descent algorithm described in Algorithm 1, where the gradient is derived as follows:

$$\frac{\partial J(\theta^s)}{\partial \theta^s} = \text{Numerator} \quad \text{Denominator}$$

where the Numerator and Denominator are derived as follows:

Numerator

$$= 2 \left[ \sum_s (\sum_c L^s_c m^s_c p^s_c) \theta^s - \sum_s \sum_c D^s_c m^s_c \right]$$

$$- (\sum_c L^s_c m^s_c p^s_c) \cdot \left( \sum_s (\sum_c L^s_c m^s_c) \theta^s - \sum_s \sum_c D^s_c m^s_c \right)$$

$$- \sum_c L^s_c m^s_c \cdot \left( \sum_s (\sum_c L^s_c m^s_c p^s_c) \theta^s - \sum_s \sum_c D^s_c m^s_c p^s_c \right)$$

Denominator

$$= \left[ \sum_s (\sum_c L^s_c m^s_c) \theta^s - \sum_s \sum_c D^s_c m^s_c \right]^3$$

IV. EXPERIMENTS

The proposed method is evaluated by the dataset we collected ourselves for 6 weeks. Three areas of different population is selected, for example, downtown city area, residential population and rural area.

Two different sets of WiFi AP scan logs are collected. The first set is collected with four different sources of reference position which are to construct WiFi maps. The four sources of AP scan logs are manually annotated, collected in the building with known position, referred from GPS position and WiFi localization result.

The other data set consists of manually annotated WiFi AP scan logs which are to simulate the proposed localization method as in training and test samples. The number of simulation samples collected from downtown, residential and rural area are 8821,10918 and 12100 alternately. We ran 5-fold cross-validation to evaluate the proposed localization with optimized weight parameters. For each area, the data set is partitioned into 5 folds to form train and test sets. The optimized parameters are set to 40 for our experiments. Using (1) for fingerprinting localization, we compare two different sets of $\theta^s$ in (4). The first set uses equal weights $\theta^s$ across all source channel whereas the second set is optimized using the proposed algorithm.

Fig. 1 illustrates localization error as well as the corresponding weight parameters $\theta^s$ as the gradient descent algorithm iterates for the downtown area. The error decreases along the iteration and converges after 2000 iterations. The optimized weight parameters show that manually annotated channel is the most critical source of information to minimize localization error. On the contrary, the weight for in building logs are the least important. It is because the simulation WiFi scan logs are collected from outdoor environment. Channels of GPS and WiFi localization does not change from initial values. The quantitative results are shown in the Table I. For all three different population environments, localization with optimized $\theta^s$ outperforms the equal weights significantly both in mean and standard deviation in localization errors in meters. Table II shows optimized parameters for each area. Each area has different population which leads to different WiFi signal environment. Hence, it is natural that the optimized parameter shows different patterns across different areas. Although the optimized value varies, manually annotated channel is the most important one across all areas. It implies that localization error can be decreased when the Wifi signals are manually collected at the target area because the reference position is accurate. On the other hand, the other abundant amount of automatic data sources may be less reliable but yet required to localize successfully. The parameters are averaged out across 5 cross-validation folds.

Qualitative result is illustrated in the Fig 2 showing how the proposed approach improves the localization result. The figure illustrates $4 \times 4$ grid cells with WiFi similarity scores.

![Localization error in meters](image1)

![Importance weights](image2)

Fig. 1: Localization error and the corresponding importance weights over iterations.

![Localization error comparison](image3)

**TABLE I: Localization error comparison between two different weight parameters across different environments. All errors are in meters.**

<table>
<thead>
<tr>
<th>Environment</th>
<th>Equal weights</th>
<th>Optimized weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downtown</td>
<td>30.92 ± 2.28</td>
<td>22.80 ± 0.47</td>
</tr>
<tr>
<td>Residential area</td>
<td>79.10 ± 7.29</td>
<td>17.98 ± 0.66</td>
</tr>
<tr>
<td>Rural</td>
<td>112.78 ± 18.98</td>
<td>21.57 ± 3.49</td>
</tr>
</tbody>
</table>
Fig. 2: A localization example for equal weights and optimized weights

(a) Localization with equal weights
(b) Optimized weights
(c) Legends

Fig. 3: Localization improvements in meters

Fig. 2 (a) takes equal weights for $\theta^s$ and the grid cell (2,4) achieved high scores of 0.4 which drags localization toward that cell. On the contrary, optimized weights has low scores for the same cell due to low weights of GPS channel and the localization result is closer to the true position.

V. DISCUSSION

In the previous section, we showed that the proposed approach improves results in all areas. The localization example in Fig. 2 illustrates how it improves the localization result. However, there may be cases when the results get worse due to WiFi logs used for WiFi map construction are imperfect. Specifically, some manually annotated logs will have incorrect position as reference due to annotation error. The proposed approach assumes that although each logs contains error each channel has innate accuracy that results in improvement overall when distinguished weights are adopted.

Fig. 3 shows overall improvements in the residential area. Vertical axis represents improvement of the optimized localization from the equal weights, while horizontal axis is the sample indices sorted over improvements. The blue area on the positive vertical axis depicts total improvements of the localization when using optimized weights, whereas blue area on the negative area shows total error increase. It illustrates that while only the half of the all samples are improved, the amount of improvement is much more significant than the deterioration. It is evident that overall performance got significantly improved by comparing blue areas between positive and negative vertical axis.

VI. CONCLUSION

During last few years, WiFi based localization is moving toward adopting Gaussian Process Regression(GPR) to build WiFi map [5]. Using GPR, one can construct maps of finer grid based on the observation logs. It also provides uncertainty estimation of each reference position, which may improve localization accuracy. The proposed approach optimizes weight parameter over a set of difference channels and the same approach can be applied to the GPR based WiFi map which will be left for future works.

REFERENCES