WIFI FINGERPRINT INDOOR POSITIONING SYSTEM USING PROBABILITY DISTRIBUTION COMPARISON

Nicolas Le Dortz *

Supélec
Département SSE
91192 Gif-Sur-Yvette, France
Email: nld@kth.se

Florian Gain, Per Zetterberg

School of Electrical Engineering
KTH Royal Institute of Technology
100 44 Stockholm, Sweden
Email : {gain, perz}@kth.se

ABSTRACT

Positioning services are increasingly used for applications such as navigation, advertising and social media. While outdoor navigation based on GPS and/or cellular systems works well, indoor navigation is a much tougher challenge. This paper presents a new indoor positioning method based on Wi-Fi fingerprints, i.e. RSSI measurements from multiple Wi-Fi access points. During an offline phase, fingerprints are collected at known positions in the building. This database of locations and the associated fingerprints are called the radio map. In the online mode, the current Wi-Fi fingerprint probability distributions are compared with those of the radio map. The user location is estimated by calculating a weighted average of the three offline positions that best match the online measurements. Experiments show that our technique is superior to other proposed methods and reaches a median error of 2.4m.

Index Terms— Indoor positioning, Wireless LAN, Signal strength fingerprint, Probability distribution, Bhattacharyya distance

1. INTRODUCTION

Recently, positioning systems such as GPS [1] have become very popular and have made possible a large range of applications. However, most actual techniques, especially those requiring satellite coverage, are not suitable for indoor positioning. And, as nearly all modern buildings are equipped with Wi-Fi access points, indoor positioning using IEEE 802.11 standard has now become a realistic alternative. Moreover, recent smartphones are commonly equipped with Wi-Fi sensors, which makes them adequate devices to implement such an indoor positioning system. The range of potential applications is very large. Indoor positioning systems could be used to give access to an interactive map of a building. For example, they could orientate a person through an airport to the boarding gate, help a student find his classroom or facilitate the way of finding items of a shopping list in a supermarket.

One successful approach for indoor positioning is based on Wi-Fi fingerprints [2]. It is applicable to scenarii with severe multipath unlike triangulation techniques where the distance to the base-stations need to be estimated based on time-of-arrival, round-trip-time or signal strength attenuation [3]. Moreover, those techniques often require uninterfered propagation paths to work well. The fingerprint-based algorithms work differently and contain two phases: an offline and an online phase. The purpose of the offline phase is to collect information about the Wi-Fi access points signal strengths at different locations. During the online phase, the measured signal strengths are compared to the offline measurements in order to estimate the user position.

For example, the positioning system RADAR [4] uses the Euclidean distance between vectors of strengths as a similarity criterion while the conditional joint probabilities are suggested in [5] and [6].

In an attempt to improve the accuracy of fingerprint-based indoor positioning systems, we propose a new method that compares online and offline signal strength probability distributions in order to find the nearest offline locations. Contrary to other techniques for which the signal strengths are averaged, we take advantage of the signal strength variations by considering the whole probability distributions. When applying the RADAR [4] and LOCATOR [6] methods to our testing data, we find that our method is about 1m more accurate at the 50% level of the CDF of the positioning error.

This article is organized as follows: in Section 2, the principles of the method are described. Section 3 gives the main experimental results and Section 4 proposes a comparison with other techniques.

2. METHODOLOGY

2.1. Estimation of signal strength probability distributions

In both offline and online phases, the signal strength probability distribution of each detected access point (AP) is estimated from a set of samples collected at the location of interest. A sample is a set of instantaneous signal strengths measured in dBm, denoted $\{s_i\}_{i\in O}$, where i is the identifier (MAC address) of the AP and O is the set of detected APs.

One way to proceed is to use histogram estimation. It approximates the probability of AP i to have signal strength s as the relative frequency of occurrences of s over the total number of samples L:

$$P_i(s) = \frac{N_i(s)}{L} \tag{1}$$

where $N_i(s)$ counts the number of samples for which the signal strength of AP i is s. $P_i(s)$ is a function of discrete values of s in the interval $[s_{min}, s_{max}]$ (determined by the sensor bounds) and constitutes an estimation of the probability distribution of AP i. It is important to notice that the minimum step between two discrete values of s is limited by the resolution of the measuring device. For instance, the smartphone we used could only measure integer values.

Figure 1 shows an example of such a distribution. We observe that the signal strength has strong variations even if the measurement

*The author performed the work while studying in the School of Electrical Engineering at KTH

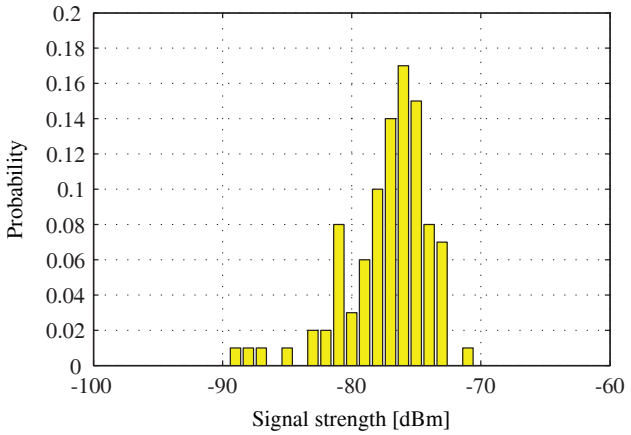

Fig. 1. Estimated signal strength probability distribution of one AP ($L = 100$)

was taken statically. They can be caused by fading or perturbations such as persons walking in the building.

2.2. Offline phase

During the offline phase, the signal strength probability distributions $\left\{P_i^l(s)\right\}_{l\in\{1\cdots N\}}$ are estimated at N evenly located positions, with L_{off} samples per location. They are stored to constitute the offline map, which serves as a reference database for the positioning system.

However, it is necessary to be cautious when taking the offline measurements. Indeed, the user's body acts as a barrier for the signal and can therefore perturb the measurements. For each location, it is preferable to estimate the probability distributions in several orientations and then combine them together by averaging the probabilities.

2.3. Online phase

The purpose of the online phase is to find the position of the user by measuring the signal strengths of the detected APs. Similarly to the offline phase, a set of L_{on} samples is collected at the user location and is used to estimate the signal strength probability distributions $Q_i(s)$ of the detected APs. These probability distributions are then compared to those of the offline map in order to find the most similar offline locations.

For that purpose, we keep, for each offline location l, the set O_l^q of the q strongest APs ($q \geq 1$) with respect to their average strength and we calculate for each of them the Bhattacharyya coefficient [7]

$$B_{i,l} = \sum_{s\in[s_{min},s_{max}]} \sqrt{P_i^l(s)\cdot Q_i(s)} \tag{2}$$

which is a common measure of the overlap between two distributions. If AP i is part of the set O_l^q but not detectable at the user location, we assume $B_{i,l} = 0$.

Then, the coefficients $B_{i,l}$ are averaged over the q strongest APs and we define the average Bhattacharyya distance between the current user position and the offline location l as

$$d_l = \begin{cases} -\ln\left(\frac{1}{q}\sum_{i\in O_l^q} B_{i,l}\right) & \text{if } \sum_{i\in O_l^q} B_{i,l} > 0 \\ -\infty & \text{otherwise} \end{cases} \tag{3}$$

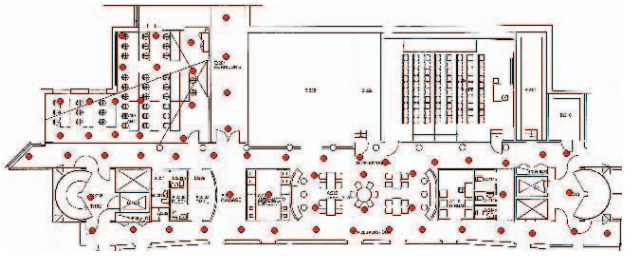

Fig. 2. Layout of the floor and positions of the 62 offline locations

The set C_k of the k nearest neighbors is chosen among the offline locations by taking those with the smallest Bhattacharyya distances.

Finally, the user position is estimated by calculating a weighted average of the k nearest neighbors positions. And, to give more weight to the neighbors with the highest similarity, the weighting values are chosen to be $w_l = 1/d_l$. Thus, the user coordinates $\hat{\mathbf{x}} = [\hat{x}, \hat{y}]$ are estimated from the neighbors coordinates $\mathbf{x_l} = [x_l, y_l]$ as

$$\hat{\mathbf{x}} = K \sum_{l\in C_k} w_l \cdot \mathbf{x_l} \tag{4}$$

where the normalization factor K is

$$K = \frac{1}{\sum_{l\in C_k} w_l}$$

3. EXPERIMENTAL RESULTS

3.1. Experimental environment

The experiment was performed in the floor located next to our lab which area has a dimension of 65m by 25m. During the offline phase (see 2.2), we collected measurements at 62 offline positions located 3 meters away from each other. The layout of the floor and the positions of the offline points are shown in figure 2.

The measurements were performed using a specifically-designed Java API implemented on an Android smartphone. At each offline location, a set of $L_{off} = 100$ samples was collected in order to estimate the strength probability distribution of each AP as described in 2.2. We could detect a maximum of about 15 different APs depending on the locations[1].

To analyze the performances of our system, we used a test set of online measurements statically taken at different known locations of the floor. For each online point we collected up to 20 samples. The effects of the number of online samples is later discussed in 3.2.2.

3.2. Effects of the parameters on performance

3.2.1. Number of strongest APs

In this section, we analyze how changing the number of APs affects the accuracy. Figure 3 shows the 1st quartile (Q_1), the median and the 3rd quartile (Q_3) of the positioning error in meters.

First, we observe that Q_1 decreases slowly and reaches a floor level (about 2.1m) when q increases. In other words, it means that a small positioning error can hardly be improved by considering more

[1] In fact, it corresponds physically to 5 or 6 different devices with multiple MAC addresses

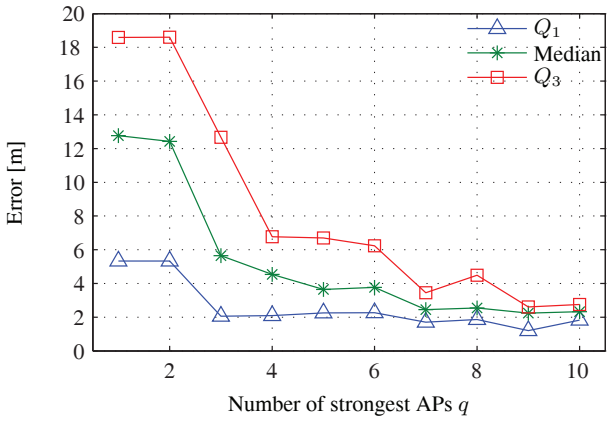

Fig. 3. Effects of the number of strongest APs on the accuracy: 1st quartile, median and 3rd quartile of the error ($k = 3$, $L_{on} = 5$)

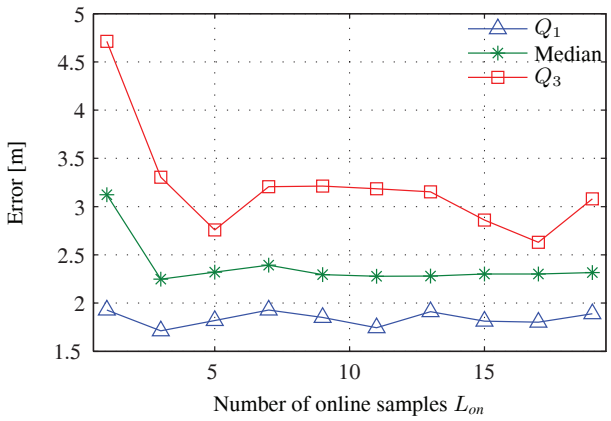

Fig. 4. Effects of the measurement time on the accuracy: 1st quartile, median and 3rd quartile of the error ($k = 3$, $q = 10$)

APs. Second, we see that the median and the 3rd quartile of the error decrease much faster than Q_1 does. It means that the risk of making a big positioning error is attenuated when we use information from more APs. Thus, the main advantage of using more APs is to improve the stability of the positioning. However this improvement comes at higher computational cost since the number of probability distribution comparisons scales with the number of APs.

In the remainder of the paper, we use $q = 10$ as it gives the best accuracy with $Q_1 = 1.9$m, $Q_3 = 3.2$m and a median error of 2.4m.

3.2.2. *Measurement time*

In this part, we show the effects of the measurement time – i.e. the number of online samples – on the accuracy. It is an important parameter since it determines the refreshing rate of the estimation. It is therefore a critical factor for the implementation of a real-time tracking application.

To analyze its influence, we proceed similarly as in the previous section by looking at the different quartiles of the positioning error. We see in Figure 4 that only a few samples are sufficient to determine the user position correctly. Indeed, the accuracy is not much improved by taking more than 5 samples. With a modern smartphone, it corresponds to a measurement time of about 5 seconds, which is acceptable for real-time positioning.

Nevertheless, taking less than 5 samples does not badly deteriorate the accuracy even if the risk of making a larger positioning error is slightly increased .

3.2.3. *Number of nearest neighbors*

The number of neighbors also strongly affects the performance of the system and the value of k must be chosen carefully. Figure 5 points out different tendencies for each quartile of the error. Whereas the 1st quartile of the error increases almost constantly with k, the median and the 3rd quartile have a minimum respectively for $k = 4$ and $k = 3$.

It means that taking only one neighbor into consideration can enable to reach a very good accuracy for the locations close to an offline point but can also lead to a bigger positioning error if it is not the case or/and if a 'wrong' closest neighbor is chosen. On the other hand, considering too many neighbors, even if it limits the

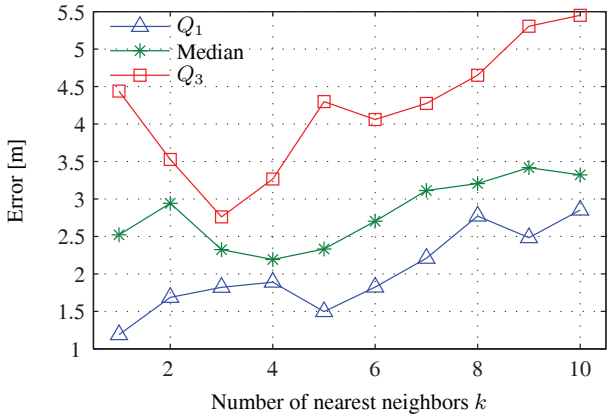

Fig. 5. Effects of the number of nearest neighbors on the accuracy: 1st quartile, median and 3rd quartile of the error ($L_{on} = 5$, $q = 10$)

risk of 'wrong' neighbor, widens the potential area for the estimated position and therefore leads to a lower accuracy.

In our experimental conditions, the optimal value would be $k = 3$. Indeed, as shown in Figure 5, it limits the risk of a large error and enables to reach a median error of 2.4m.

3.2.4. *Application to real-time positioning*

Previously, we presented the performances of our system for a static user. However, it was also tested in a more realistic situation when the user is moving across the floor and wants his position to be updated frequently. For that purpose, we slightly modified the system by implementing a sliding window. That is, the position is estimated from a buffer containing the 5 most recent measured samples. When a new sample is available, it replaces the oldest one and the user position is re-estimated from the updated buffer. Thereby, we have access to a position estimate for each new sample coming up from the smartphone.

The empirical tests showed very good results using this method

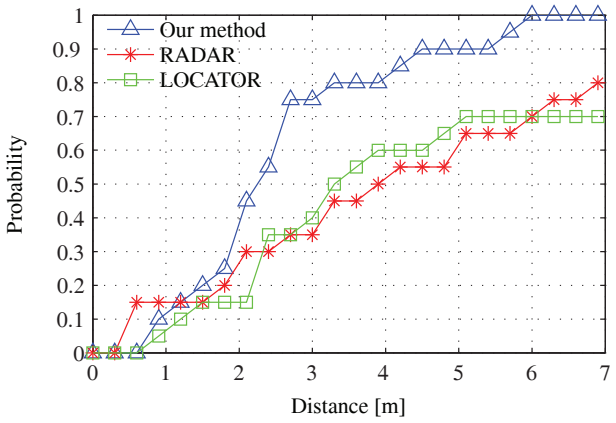

Fig. 6. CDF of the positioning error for different positioning systems (truncated to 7m)

since we were able to track our position very accurately (positioning error <3m) when walking across our building. Furthermore, the positioning accuracy was not affected much in a noisy environment (walking people nearby, opening and closing doors, etc.).

4. PERFORMANCE COMPARISON WITH OTHER METHODS

In this section, the performances of our method are compared to two other popular systems using WiFi fingerprints. The first one, RADAR [4], uses the Euclidean distance between online and offline vectors of signal strengths to determine the k nearest neighbors. The second one, LOCATOR [6], requires as in our method, to estimate the signal strength probability distribution of each AP during the offline phase. The nearest neighbor is chosen to be the offline point with the lowest conditional joint probability given the received vector of signal strengths.

In order to propose a fair comparison between the methods, we used the same offline map and the same testing data to calculate the positioning error. Figure 6 shows the CDF (Cumulative Distributive Function) of the error in function of the distance for each technique. It is defined as the probability of the positioning error ϵ to be lower than a certain distance d:

$$\text{CDF}_\epsilon(d) = P(\epsilon \leq d) \quad d \geq 0 \tag{5}$$

We observe that our method outperforms the two other ones. Although the minimum error is similar for the three methods, we see that our maximum error is around 5.5m where LOCATOR and RADAR can lead to positioning errors above 7m.

In terms of computational complexity, our method is slightly more demanding. Indeed, computing Bhattacharyya distances between probability distribution requires more operations than computing a simple Euclidean distance or conditional joint probabilities.

5. CONCLUSION AND FUTURE WORK

The system presented in this paper has proved to be suitable to accurately locate an user in a building. It is based on the existing Wi-Fi infrastructure and can be implemented on portable devices like smartphones. We have investigated on the effects of different parameters on the accuracy. With the optimal number of neighbors and number of APs and for a measurement time of 5 seconds, our system has a median error of 2.4m and a maximum error of 5.5m. These are performances that a user can expect from a localization service. Besides it outperforms the Wi-Fi fingerprint-based techniques RADAR [4] and LOCATOR [6] and can be used as a real-time indoor tracking system.. It is therefore adequate for many interactive uses and we can easily imagine it to be integrated to existing navigation systems or in commercial smartphone applications.

In a future work, we want to add a prediction system that would take into account a model of the user movement. We also consider dividing the search area into smaller clusters. It would reduce the computational complexity, especially when the system is used in large buildings.

6. REFERENCES

[1] P. Enge and P. Misra, "Special issue on global positioning system," *Proceedings of the IEEE*, vol. 87, no. 1, pp. 3–15, 1999.

[2] A. Taheri, A. Singh, and A. Emmanuel, "Location fingerprinting on infrastructure 802.11 wireless local area networks (wlans) using locus," in *Local Computer Networks, 2004. 29th Annual IEEE International Conference on*. IEEE, 2004, pp. 676–683.

[3] 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.

[4] P. Bahl and V.N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*. Ieee, 2000, vol. 2, pp. 775–784.

[5] M.A. Youssef, A. Agrawala, U. Shankar, et al., "Wlan location determination via clustering and probability distributions," in *Pervasive Computing and Communications, 2003.(PerCom 2003). Proceedings of the First IEEE International Conference on*. IEEE, 2003, pp. 143–150.

[6] A. Agiwal, P. Khandpur, and H. Saran, "Locator: location estimation system for wireless lans," in *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*. ACM, 2004, pp. 102–109.

[7] S.H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," in *International Journal of mathematical models and methods in applied sciences*. Citeseer, 2007.