A Comparative Study on Gaussian Process Regression-based Indoor Positioning Systems

Md. Sakib Anwar, Fariha Hossain, Nusrat Mehajabin, Md. Mamun-Or-Rashid, Md. Abdur Razzaque
Green Networking Research Group, Department of Computer Science and Engineering,
University of Dhaka, Dhaka-1000, Bangladesh
{2014-016-619,2014-216-608}@student.cse.du.ac.bd, nusrat@cse.du.ac.bd, mamun@cse.univdhaka.edu, razzaque@du.ac.bd

***Abstract*—Gaussian Process Regression (GPR) has been proved to be one of the most accurate ways of predicting online radio map for fingerprinting based localization, as it can better mimic the characteristics of wireless radio signals. However, the accuracy of the GPR model depends on the mean function used and most of the functions perform poorly while being used in localization. This paper presents a thorough comparative analysis on different Indoor Positioning Systems (IPS) exploiting GPR with different mean functions, among which zero mean and linear mean are the most commonly used ones. This paper also introduces two new mean functions-Single Hidden Layer Neural Network (NN) and Multiple Hidden Layer NN which outperforms traditional mean functions.**

***Index Terms*—Indoor Positioning System, Gaussian Process Regression, Neural Network, Fingerprinting**

I. Introduction

The omnipresence of smartphone and high speed Internet has drastically increased the need for localizing devices in both indoor and outdoor conditions for location based services (LBS) such as locating rooms in office or a product in a supershop. Localization can be defined as the process of finding the physical location on a 2D or 3D space of a device or an object with respect to some predefined reference points. Localization in outdoor environment has come to a satisfactory accuracy level *i.e.* 1 to 5 meter [1] with the introduction of Advanced Global Navigation System (AGNSS) [2] along with the contemporary Global Positioning System (GPS). Indoor positioning system (IPS) is the same process as outdoor positioning - localizing an object or a device, except the search space is housed indoor. The traditional methods used for outdoor localization is unsuited even after continuous probing by the research community, leaving room for further improvements in terms of accuracy and efficiency.

The main challenge of indoor positioning system is lack of Line of Sight (LOS) in addition to dynamics of the environment [2, 3]. Hence, new technologies and new methodologies are needed to localize in indoor condition. Though the challenge starts with this single issue, it extends to more obstacles like choosing a suitable technology for localization, availability of the chosen technology, algorithms used and their efficiency. Most of the technologies that can be used for indoor localization mainly uses radio wave. So the properties of radio waves in dynamic indoor environment with large obstacles like walls and furniture have to be taken into consideration as they affect radio waves significantly [2, 3].

In literature, technologies such as camera, infrared, ultra wide band and wireless access points have been used to localize in indoor conditions [2]. Among the different technologies wireless access point has shown the most promising outcomes due to its availability and use of different measures of radio signal which can be used as features for localization. Along with different technologies many types of algorithms starting from simple similarity function to complex machine learning algorithms have been deployed to increase the accuracy of IPS [3–5]. Although incremental, the development in terms of accuracy and efficiency has not come to a satisfactory level.

Fingerprinting is the process of saving a particular feature of Wi-Fi radio wave for a particular location which is now-a-days used as a reference for localization of a new device. Although accurate, it has some problems such as being labour intensive, time consuming and unable to adapt to dynamic environment. Thus, an innovative way to build a radio map online dynamically has to be established that can act as a standard using predictive model or a deterministic model which can predict the radio map based on some training data can be possible solutions. In the literature, this problem is addressed over and over again with the most common approach being Gaussian Regression Process (GPR) [3–5]. However, different mean functions and different similarity functions have been used with little performance improvement over other existing solutions. Furthermore, limited research has been done on the number of access points (AP) required to build an accurate model for a given area. This paper aims at showing how different mean functions work under different conditions and introduces two new mean functions which are Single Hidden Layer Neural Network (NN) and Multiple Hidden Layer NN. The performance study shows that these two mean functions perform better than the existing mean functions and generate higher correlation coefficient which signifies that our mean functions create a better association between the variables involved.

The rest of the paper is organized as follows, Section II describes the different works on the literature for localization. Section III describes our network model, environment of the test bed and our proposed method. Section IV shows the performance evaluation of the different models that we have compared. Finally, section VI concludes the paper with a summary of the finding.

978-1-5386-5229-9/18/$31.00 ©2018 IEEE

II. Related Works

In this section, we discuss the recent works on Indoor Positioning Systems (IPS). The existing works in the literature can be divided into two major phases *i.e.* radio map construction and localization. The recent works, improvements and the remaining challenges in these sections are discussed below.

Radio maps are of two types: Deterministic and Probabilistic models [6]. In deterministic models a fingerprint at a location is represented by a list of average received signal strength indicator (RSSI). Whereas in probabilistic models the fingerprint of a location is predicted based on some training data [3–5]. Although, the probabilistic approach gives better and more precise radio map than the deterministic one [7], both methods have some drawbacks. The main drawbacks of the existing fingerprinting based localization systems are mentioned earlier. In addition, to build a fine grained radio map we need to measure a large number of RSSI values for different APs for a large number of virtual reference points (VRP). Besides, RSSI values fluctuate a lot due to dynamic environments and so real time RSSI may differ a lot from the one saved in the database and lead to a major localization error. All in all, these expensive and error prone radio map construction method interferes with further development of the model.

Plethora of methods have been introduced to reduce this manual effort of offline survey and update the radio map online when changes occur. Some of the novel methods are point by point calibration [8], fixed reference point methods, learning based methods [9] and crowd sourcing methods [10]. Some of these methods *e.g.* reference anchors [2] need extra hardware which is hard to deploy in a large environment. In addition, crowd sourcing requires a large amount of data which needs a lot of time to be collected. Furthermore, it requires continuous inertial measurement unit (IMU) monitoring that drains a lot of Mobile device's (MD) battery which is an impractical solution. Different general regression based methods have been used e.g. polynomial fitting, exponential fitting logged model [11] etc which only predicts the RSSI mean not the variance-making data extrapolation very difficult. Gaussian Process Regression use both the posterior mean and the variance. Zero Mean Gaussian for localization [5] predicts zero at the location which is far from the training points which is impractical. Log Distance Mean GPR for localization [12] works better in open space with no obstacles but performs poorly in environment having obstacles as RSSI value attenuates due to multi path effect and shadow fading.

In the localization part, the similarities between the fingerprints in the database with the real time data from mobile devices are compared. Different similarity functions have been developed to find this similarity e.g. Cosine Similarity, Euclidean Distance, Manhattan Distance etc [4, 5].

III. System Design

In this section, we describe the problem, the environment of our sample area and finally develop an algorithm which compares among the different types of models and similarity functions.

A. Problem Definition

The main point of concern is that various works in literature have addressed the problem with fingerprinting and have used GPR to solve these problems using different mean and similarity functions. However, to the best of our knowledge, a comparison among these approaches have not been carried out yet. This leads to a decision problem while building a new model to solve other related problems. Thus, a comparative study of the existing methods have to be made which can be further referenced for new algorithms or models using different approaches.

The first decision problem is choosing a mean function which will be integrated with the GPR model, that further affects the accuracy of the system. Choosing a mean function is vital because not all mean functions can capture the dynamic nature of RSSI changes. We have used four mean functions where two are mentioned in the literature and two are novel to this article.

Secondly, while matching the new fingerprint with the stored fingerprint database we need to choose a similarity function which gives a satisfying level of accuracy. The problem is similarity functions are not built for measuring the similarity between RSSI vectors. Thus, a method needs to be devised to get better accuracy with existing similarity functions and a comparison is needed so that the right one can be chosen.

Finally, we need to determine the number of APs required to localize in a given room, that is we have to know the point of saturation for a given area. This can be achieved by using enough APs in a room and observing how changing the number of APs affect the correlation coefficient. The point at which correlation starts to drop can also be accurately identified as the point of saturation.

B. System Environment

We assume a temperature controlled room which is well furnished having the setup of a classroom or office. A variable number of people may access the Wi-Fi facility in the room. n APs are assumed to be present in the domain. The set of APs is represented by $AP = (AP_1, AP_2, \ldots, AP_n)$. We splitted the room into $X \times Y$ grid so that we can locate each grid using (x, y) co-ordinate. All of these locations or a subset of these locations can be used as virtual reference points (VRPs). Let the number of VRPs be m denoted by the set $VRP = (VRP_1, VRP_2, \ldots, VRP_m)$. Locations of APs and VRPs are stored in matrix $P_{AP}(N \times D)$ and $P_{VRP}(M \times D)$, respectively, where D is the dimension of the location information meaning it can take the value of 2 in case of single storey setup and 3 in case of multi storied setup.

$$P_{AP} = \begin{bmatrix} p_{AP_{1,0}} & p_{AP_{1,1}} \\ \vdots & \vdots \\ p_{AP_{n,0}} & p_{AP_{n,1}} \end{bmatrix} P_{VRP} = \begin{bmatrix} p_{VRP_{1,0}} & p_{VRP_{1,1}} \\ \vdots & \vdots \\ p_{VRP_{m,0}} & p_{VRP_{m,1}} \end{bmatrix}$$

Here, $(p_{AP_{i,0}}, p_{AP_{i,1}})$ is the co-ordinate of the i^{th} AP and $(p_{VRP_{i,0}}, p_{VRP_{i,1}})$ is the co-ordinate of the i^{th} VRP.

C. Proposed Method

Now, we develop an algorithm that can compare the different performance variables we want to evaluate. The whole algorithm can be divided into three parts-

- Online Radio Map Construction
- Localization and
- Comparison Based on the Output.

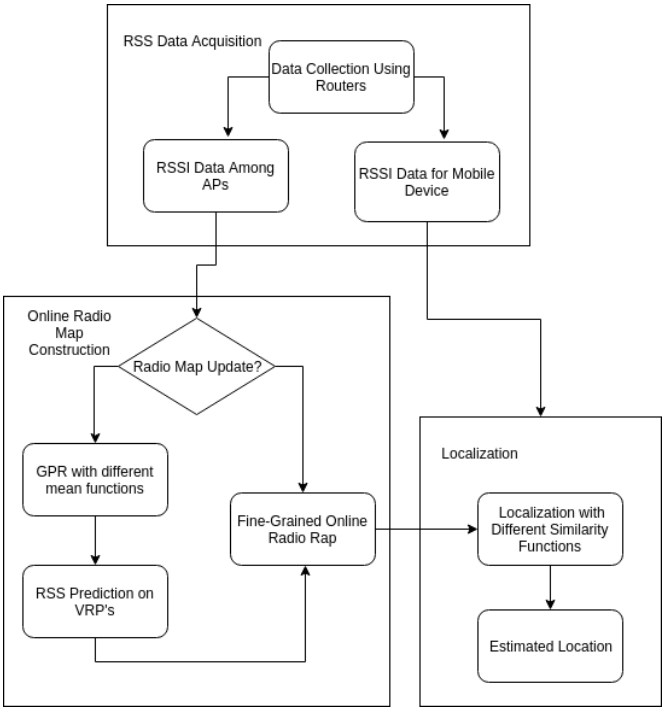

Fig. 1. GPR based Indoor Localization

1) Online Radio Map Construction: In this phase, we build a fine grained radio map that can be used later to localize a new mobile device in the vicinity. We have used Gaussian Process Regression (GPR) to build a predictive online radio map. However, we have used different mean functions while building the model. The different mean functions that we have considered are:

- Zero Mean
- Linear Mean
- Single Layer Neural Network (NN) Mean
- Multiple Layer Neural Network Mean.

The data has been collected from the APs using a special firmware described later which allows us to run custom programs in the APs. The APs capture packets that are being transmitted by other APs in the vicinity. APs always transmit beacon frames transmitting their Basic Service Sets Identifier (BSSID) for other devices that want to connect to them. In our custom program that run in the APs, we have scanned for such packets and measure the RSSI an AP is getting for other APs in the vicinity. This works as the training data of our GPR model.

Once the model has been trained with a particular kind of mean function we need to predict the fingerprint for the virtual reference points(VRPs). As mentioned earlier, we have n APs and m VRPs in the vicinity which are contained in the sets AP and VRP whose locations are also recorded in the database. After sniffing packets with AP we will have a matrix

$$\mathbf{Y_{train}} = \begin{bmatrix} y_{train_{1,1}} & y_{train_{1,2}} & y_{train_{1,3}} & \cdots & y_{train_{1,n}} \\ y_{train_{1,1}} & y_{train_{2,2}} & y_{train_{2,3}} & \cdots & y_{train_{2,n}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{train_{n,1}} & y_{train_{n,2}} & y_{train_{n,3}} & \cdots & y_{train_{n,n}} \end{bmatrix}$$

where, $y_{train_{i,j}}$ is the RSSI of i^{th} AP is receiving for j^{th} AP where $AP_i, AP_j \in AP$. The location of the APs along with the matrix $\mathbf{Y_{train}}$ is used to train the model. Then, we predict the RSSI a device will get for all the $AP_i \in AP$ for each $VRP_j \in VRP$. This means we will get a matrix

$$\mathbf{Y_{pred}} = \begin{bmatrix} y_{pred_{1,1}} & y_{pred_{1,2}} & y_{pred_{1,3}} & \cdots & y_{pred_{1,n}} \\ y_{pred_{2,1}} & y_{pred_{2,2}} & y_{pred_{2,3}} & \cdots & y_{pred_{2,n}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_{pred_{m,1}} & y_{pred_{m,2}} & y_{pred_{m,3}} & \cdots & y_{pred_{m,n}} \end{bmatrix}$$

where, $y_{pred_{i,j}}$ is the RSSI that we should get at i^{th} VRP for j^{th} AP where $VRP_i \in VRP$ and $AP_j \in AP$. With this prediction we have successfully built our radio map which we will use later to localize a new mobile device in the vicinity.

2) Localization: In this phase, we have a new device in the domain and we want to calculate its location in the test bed area. This can be done using a number of similarity measuring functions. The functions we have considered for localization are:

- Cosine Similarity (CS)
- Manhattan Distance (MD)
- Euclidean Distance (ED) and
- Weighted K Nearest Neighbour (WKNN).

Firstly, we use the APs to measure the RSSI value the device is getting for each AP present in the experimental area. We save this value as a vector denoted as

$$\mathbf{Y_{test}} = \begin{bmatrix} y_{test_1} & y_{test_2} & \cdots & y_{test_n} \end{bmatrix}$$

where y_{test_i} is the RSSI the mobile device is getting for the i^{th} AP where $AP_i \in AP$.

We take this value and measure its similarity with each row of the matrix $\mathbf{Y_{pred}}$. However, matching directly often generates a very poor output so we represent the vector with z-score to generate greater accuracy [13]. Suppose row i generates the most similarity score when compared with the RSSI vector $\mathbf{Y_{test}}$ then the output will be the location of i^{th} VRP. This method of calculating location is implemented in case of CS, MD and ED. However, in case of WKNN, we choose location of K VRPs based on the similarity score and then we calculate the output based on these K locations.

Once we have all the results, we have compared the results based on two performance metric, accuracy and correlation.

For measuring accuracy we have calculated the euclidean distance between the predicted location and the actual location. The mean of these errors is the average error of that particular approach. For correlation coefficient, we have calculated it between our generated location and the actual location using cosine similarity function. That is the correlation coefficient shows how much our calculated location is related to the original location.

IV. Performance Evaluation

This paper aims at showing how different mean functions affect the accuracy of GPR model based localization methods. The performance metrics that we have considered are:

- Localization Accuracy and
- Correlation Coefficient.

In the literature, zero mean [5] has been used in most cases but it can not attain a great accuracy and often requires a huge amount of data to operate well enough. However, we have used three different mean functions along with regular zero mean GPR and have shown their localization accuracy. A well known localization method is WKNN but it is highly dependent on the value of K that we choose, sometimes it is overfitting and sometimes it is underfitting and thus is a point of concern. Lastly, once we have a radio map, we have a plethora of similarity measuring functions which ultimately affect the accuracy of the model.

A. Experimental Setup

We conducted our experiment in a 5.75 m $\times$ 5.5 m classroom environment. Three (3) wireless access points have been used with some specific design advantages that are needed to conduct our experiment. We have used Gl-AR150 access points which come with OpenWrt pre-installed in them. They have 64MB storage with 16MB flash storage which is sufficient for our custom applications and are small in size making them easily deployable. These routers were used to demonstrate that Commercial of the Shelf (COTS) routers are becoming portable and easily customizable now-a-days and can be used to localize in indoor conditions [2, 3, 5]. We divided the test bed into 4 $\times$ 4 *i.e.* 16 blocks of equal size which acted as virtual reference points (VRPs) of our system. One server was used to collect data from APs, extract and pre-process them for further use. Then different mean GPRs were used to predict RSSI for VRPs and different distance similarity functions were used to find the location of the mobile devices.

B. Localization Accuracy

The performance variable that we have considered to measure localization accuracy are mean functions of the GPR model and different values of K for WKNN based localization.

In Fig. 2 different iteration of GPR was used for optimization of the GPR model. As shown in the figure, with the increase of the number of iterations the mean error comes to a stable position and when the number of iteration is 5 all the mean functions *i.e.* Zero, Linear, Single Layer Neural Network (NN) and Multiple Layer NN give the least error. It can also be observed from the figure that Multiple Layer NN performs better than all other mean function.

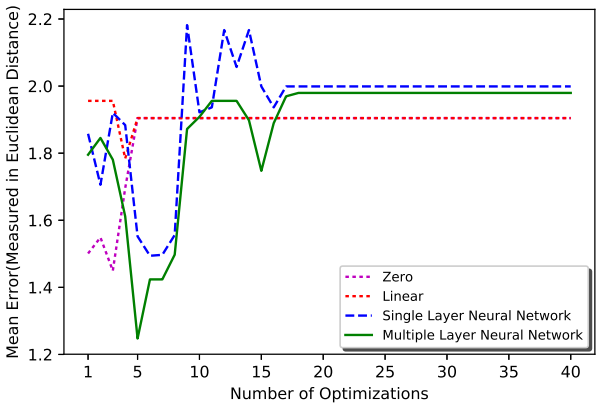

Fig. 2. Localization accuracy for varying number of optimizations

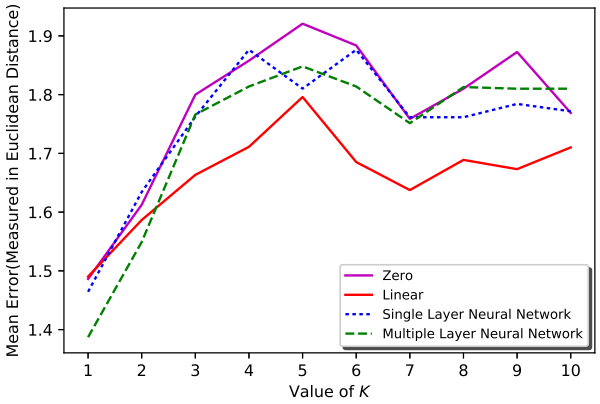

Fig. 3. Localization accuracy for varying number of K Values

WKNN ranks the similarity scores and takes the mean of the top K values. But from Fig.3 it can be recognized that in a saturated environment if the value of K is increased, the amount of error increases. It performs the best when the value of K is 1 in a saturated environment. So, it can be said that it is better to use a simple similarity function rather than WKNN in a saturated environment.

C. Correlation Coefficient

Correlation co-efficient describes how strong the relation is among data points. The range of correlation coefficient(CC) is -1 to 1. Correlation coefficient 1 means there is a strong positive correlation, 0 means no correlation and -1 means is a strong negative correlation among the data. Change of correlation coefficient for different mean functions using different number of APs and different similarity functions is shown below.

The Fig. 4 presents the CC performances of the studied similarity functions. It can be clearly observed that Single Layer NN performs the best. Also, the correlation of Zero

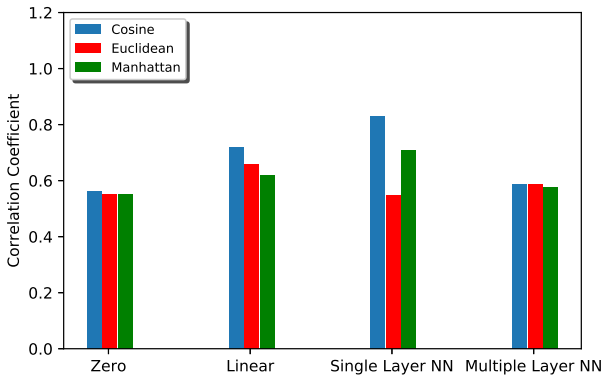

Fig. 4. Correlation co-efficient for different similarity functions

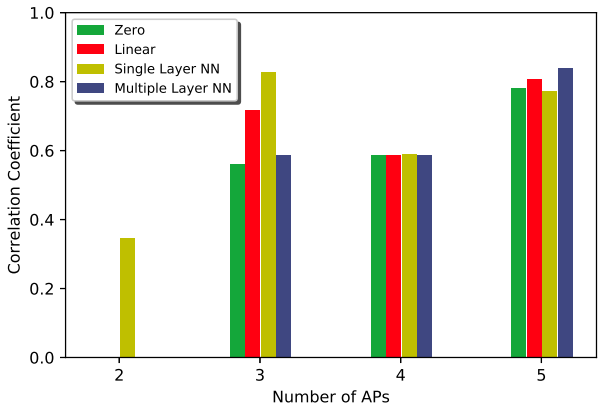

Fig. 5. Correlation co-efficient using various number of APs

Mean and Multiple Layer NN does not depend that much on the similarity function used. Nonetheless, cosine similarity function works the best in all cases.

The Fig. 5 shows the change of CC for different number of APs. It is observed from the graph that when the number of AP is 2 only Single Layer NN can provide some level of correlation. Although, with the increase of number of APs the CC increases, it can also be observed that the CC start decreasing after a threshold. In our case, the point of saturation was 3 for Single Layer NN, which performs as good as others with 5 APs.

V. Conclusion

This paper compared the existing mean functions with two new mean functions for Gaussian Process Regression (GPR) based localization methods, using different similarity measuring functions. The performance evaluation showed that Multiple Hidden Layer Neural Network gave the least amount of error in a saturated environment and Single Hidden Layer NN provided the best correlation coefficient using minimum number of APs. This paper also showed that cosine similarity is the best for measuring the similarity between two RSSI vectors. Overall, the result showed that the neural network based mean functions perform better than the existing mean functions. However, given plenty of APs, any algorithm can provide desired accuracy which is a costly solution. Thus, research is needed to calculate the minimum number of APs required to localize precisely using an efficient mean and similarity measuring function.

References

[1] Hakan Koyuncu and Shuang Hua Yang. “A survey of indoor positioning and object locating systems”. In: *IJCSNS International Journal of Computer Science and Network Security* 10.5 (2010), pp. 121–128.

[2] Rainer Mautz. “Indoor positioning technologies”. In: ETH Zurich, Department of Civil, Environmental, Geomatic Engineering, Institute of Geodesy, and Photogrammetry, 2012.

[3] Han Zou et al. “WinIPS: WiFi-based non-intrusive indoor positioning system with online radio map construction and adaptation”. In: *IEEE Transactions on Wireless Communications* 16.12 (2017), pp. 8118–8130.

[4] Brian Ferris Dirk Hähnel and Dieter Fox. “Gaussian processes for signal strength-based location estimation”. In: *Proceeding of Robotics: Science and Systems*. Citeseer. 2006.

[5] Sudhir Kumar, Rajesh M Hegde, and Niki Trigoni. “Gaussian process regression for fingerprinting based localization”. In: *Ad Hoc Networks* 51 (2016), pp. 1–10.

[6] Mikkel Baun Kjærgaard. “A taxonomy for radio location fingerprinting”. In: *International symposium on location-and context-awareness*. Springer. 2007, pp. 139–156.

[7] Petri Kontkanen et al. “Topics in probabilistic location estimation in wireless networks”. In: *Personal, Indoor and Mobile Radio Communications, 2004. PIMRC 2004. 15th IEEE International Symposium on*. Vol. 2. IEEE. 2004, pp. 1052–1056.

[8] Dongsoo Han et al. “Building a practical Wi-Fi-based indoor navigation system”. In: *IEEE Pervasive Computing* 13.2 (2014), pp. 72–79.

[9] He Wang et al. “No need to war-drive: Unsupervised indoor localization”. In: *Proceedings of the 10th international conference on Mobile systems, applications, and services*. ACM. 2012, pp. 197–210.

[10] Qideng Jiang et al. “A probabilistic radio map construction scheme for crowdsourcing-based fingerprinting localization”. In: *IEEE Sensors Journal* 16.10 (2016), pp. 3764–3774.

[11] Jie Yang and Yingying Chen. “Indoor localization using improved rss-based lateration methods”. In: *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*. IEEE. 2009, pp. 1–6.

[12] Hyuk Lim et al. “Zero-configuration indoor localization over IEEE 802.11 wireless infrastructure”. In: *Wireless Networks* 16.2 (2010), pp. 405–420.

[13] Zheng Yang, Zimu Zhou, and Yunhao Liu. “From RSSI to CSI: Indoor localization via channel response”. In: *ACM Computing Surveys (CSUR)* 46.2 (2013), p. 25.