To appear: IEEE 2015 Global Communications Conference, Exhibition & Industrial Forum.

Robust, Cost-Effective and Scalable Localization in Large Indoor Areas

Tong Guan, Wen Dong, Dimitrios Koutsonikolas, Geoffrey Challen, Chunming Qiao
Computer Science and Engineering, the State University of New York at Buffalo
Email: {*tongguan, wendong, dimitrio, challen, qiao*}@buffalo.edu

***Abstract*—Indoor location information plays a fundamental role in supporting various interesting location-aware indoor applications. Widely deployed WiFi networks make it feasible to perform indoor localization by first establishing a received signal strength (RSS) map covering the whole area based on a signal propagation model, then determining a location from an online RSS measurement given the RSS map. However challenges remain in practical deployments, due to inaccurately estimated RSS values in the RSS map and insufficient number of access points (AP) in large indoor areas. To address these challenges, we develop a robust, cost-effective and scalable localization system (REAL). Our approach takes the error from the indoor radio signal propagation model into consideration. It also exploits information about APs which are not visible at a given location and an optimal clustering method in the location searching phase. Our real-world experimental results demonstrate that REAL achieves considerable localization accuracy at a very low training cost even for a large indoor area. In addition, the results show that our scheme can also be effectively applied to Bluetooth networks with sparse signal coverage.**

I. Introduction

Indoor location information is a fundamental part of mobile and ubiquitous computing, supporting a variety of interesting applications such as location-aware advertisements for shoppers and indoor navigation for the blind. Although the satellite-based global positioning system (GPS) provides efficient and scalable services to mobile users in outdoor cases, it is not suitable for establishing indoor location because the signal is attenuated, reflected, and scattered by complicated indoor environments. Therefore, how to design and implement a scalable and cost-effective indoor localization system has been extensively studied over the last decade.

Positioning algorithms can be coarsely classified into two categories: trilateration and fingerprinting. The basic idea of trilateration is to estimate the position of an object by measuring its distance from at least three known reference points. However, the accuracy achieved through trilateration is unacceptable without deploying extra infrastructure such as ultra wideband (UWB) signals. In this paper, we focus on the fingerprinting approach to avoid such overhead. Fingerprinting usually contains two phases: an off-line training phase of collecting features (e.g. a vector of RSS measurements from various APs) at known locations to establish an RSS map; and an online localization phase of estimating a location based on the RSS map and an online RSS measurement.

Despite considerable progress in this area, many challenges remain—particularly for deployments spanning large areas. More specifically, most validation experiments are either carried out in small areas [9][12] or done under the assumptions that more than three reference base stations can be seen at all locations [13]. In contrast, data collected in our building shows that at 20% of the locations, the number of WiFi APs seen by a device is fewer than three[1]. The case is worse for a Bluetooth network deployed in the same area with similar density: more than three Bluetooth beacons are only seen at 45% of measured locations due to a shorter communication range. Another major challenge is that in order to improve location accuracy, some schemes require collecting a large set of training locations, which is labor and time intensive. Such schemes may be feasible for a small indoor area but not practical for a large one. The works in [9][10][12] try to reduce the training effort by predicting RSS values instead of taking measurements manually. However, they rely on inaccurately predicted RSS values in the location searching phase and ignore the error introduced from propagation model. Finally, existing schemes take two extreme directions in searching for a closest match from the RSS map: they either find only one RSS vector from the map and use the corresponding location as the output [9]; or they find a largest cluster including all the candidate locations and return the weighted average coordinates as the result [12]. Both scenarios do not consider the modeling error and no prior studies have addressed what is the optimal cluster size.

To address these challenges, we design a low-cost and robust indoor localization system called REAL. REAL builds up an RSS map with only a few training locations. It estimates the location based on the RSS map by considering the modeling error, utilizing information of APs which are invisible at a given location, and an optimized clustering method. The major novelties and contributions of this work are:

- We propose a probabilistic approach to consider the error introduced from modeling signal propagation. Our evaluation results show that our approach outperforms the case when estimated RSS is directly used for comparison. We also develop a new location searching algorithm by utilizing both redundant AP information and a clustering method, to reduce the influence of the modeling error.

- We evaluate the effectiveness of our system in a real environment, the $3750m^2$ 3rd floor of Davis Hall at University at Buffalo. In this WiFi network, only 80% locations see more than three APs. The evaluation results show that even with a small set of 5 training locations, we are still able to achieve a median error of $4.78m$.

- We show that our scheme can effectively work with 12 Bluetooth beacons sparsely deployed in the same area, where

[1] the number of distinct physical APs

55% locations see fewer than three beacons. Our results show that with only 39 training locations, we can achieve a median error of $5.56m$.

The rest of the paper is organized as follows. In Section II, we provide the background about RSS mapping techniques. In Section III, we introduce our localization system REAL, and its mapping technique and improved location searching algorithm. We present our performance evaluation in a real building in Section IV. Finally, we conclude our work in Section V.

II. Related Work

Trilateration techniques depend on an accurate estimation of distance, however, estimating distance by utilizing RSS in 802.11 networks is quite inaccurate due to multipath, refraction or shadow fading. The work in [1] has shown that distance estimation falls between 2/3 and 3/2 of the actual distance only 69% of the time. The huge ranging error will greatly affect the accuracy of trilateration solutions. In order to acquire accurate ranging information, research has focused on deploying additional infrastructure to estimate distances. UbiSense [2] and the work in [4] relied on UWB signals to calculate distance by estimating the time of arrival (TOA) of direct signal path. Although remarkable performance has been achieved in such techniques, the complexity and specific hardware requirement make such systems impractical to deploy.

Previous indoor fingerprinting systems have exploited the possibility of various technologies independent of 802.11 networks. For example, Active Badge [8] and EIRIS [7] employed infrared signals to label rooms in an office environment. Cricket [5] and the work in [6] built an infrastructure that deploys beacons at various locations utilizing ultrasound signals. Although such research has shown promising results, their limited applicability to small indoor environment and significant cost during the learning phase makes them poorly-suited for scaling to larger areas. However, with the growing smartphone market and parallel massive deployment of 802.11 networks, performing localization based on RSS is a promising approach because it exploits existing wireless networks and saves the cost of deploying other infrastructure. RADAR [9] system is a pioneering work in establishing an RSS map where RSS measurements from three APs are collected at known locations. RADAR proposed two approaches to establish the RSS map: manually measured, or theoretically estimated based on a propagation model. The first strategy relies on manual resources to collect the ground truth in order to acquire an accurate mapping, while the second strategy exploits a simple propagation model with empirical configured parameters. Manual mapping slightly outperforms theoretical estimation in terms of localization error but theoretical estimation dramatically reduces the cost to perform the site survey, which makes practical deployment over a large space possible.

Several schemes have since improved upon RADAR, in either the map-generating phase or the searching phase. Horus [10] utilized stochastic interpretation in building the RSS map and a probabilistic technique to search for the best-matched location, however, the training cost is high, with measurements taken every 1.5 meters in a 68 by 26 meter space. To reduce the pre-deployment effort, TIX [11] modified APs to measure the RSS from neighboring APs, and linear interpolation was then applied to recover the RSS from each AP at every location. However, this system requires knowledge of AP transmission power and modification of commercial APs. Chintalapudi *et al.* [13] proposed a localization system called EZ which requires fewer RSS measurements than most existing schemes. However, EZ depends on boundary information collected from GPS, which is time-consuming especially in a large indoor area. ARIADNE [12] exploited a more sophisticated ray-tracing model and simulated annealing to learn the parameters, but this approach depends on the placement of APs and assumption that all APs can be detected at any location in a small space.

In terms of location searching, one typical comparison metric is the least mean square error (LMSE) of the RSS measurement vector. Intuitively, a smaller MSE between two RSS measurement vectors indicates a shorter distance between them. Pandey *et al.* [15] use the second smallest MSE as the comparison metric. ARIADNE proposed a clustering approach that groups locations based on MSE—the cluster with the largest size is returned as the final result. However, these approaches do not work effectively when there are multiple locations with similar MSEs. Additionally, they rely on inaccurate estimated RSS values and ignore the error introduced from modeling the relationship between RSS and distance.

To summarize, although considerable improvements have been achieved regarding indoor localization over 802.11 networks, none has achieved an impressive performance for large indoor cases while preserving acceptable accuracy, low pre-deployment cost, and robustness. Our goal is to design an indoor localization system that overcomes these shortcomings through a set of key features: a) low complexity and cost, the system does not require additional infrastructure, reconfiguration of the existing 802.11 networks, or installation of any other hardware (such as UWB) on commercial-off-the-shelf (COTS) devices such as smartphones, tablets or laptops, b) robustness, the system is adaptive to environmental changes such as movement of furniture or people, c) scalability, the system is easily adaptable to even larger areas and works under infrastructure using similar radio signals such as Bluetooth and ZigBee, d) accuracy, the system offers excellent performance compared with existing approaches in terms of deployment cost and size of the localization space.

III. Localization with REAL

We utilize the same propagation model as in [14], which is defined as follows:

$$RSS_d = \mathcal{F}(d, N_{ob}) \tag{1}$$

$$= RSS_0 + \alpha log_{10}(d) + \sigma N_{ob} \tag{2}$$

where RSS_d represents the RSS reading at distance d from the AP, RSS_0 is the RSS right at the AP, α and σ correspond to the signal fading coefficient due to direct path distance and walls respectively, and N_{ob} represents the number of walls on the direct path. This model assumes that only walls contribute to signal attenuation caused by obstacles, and does not consider shadow fading factors from refraction and reflection. The simplification in (2) offers several advantages. First, formula (2) does not contain any path loss coefficient which is sensitive to environmental variants; Second, this propagation model includes both line-of-sight (LOS) and non-LOS cases where

RSS is determined only by log distance and fading from walls, and no more efforts are need to determine which path loss scheme should be applied; Third, the training cost in terms of both off-line data collection and computation is greatly reduced because this model is simple and linear. The system's adaptivity is enhanced because its training cost is low even when environmental changes occur. We discuss this in greater depth in the evaluation section. Finally, this model increases the scalability and extendability of our system, since it is feasible to acquire distance information and the number of walls for a large site.

REAL contains an off-line training phase and an online localization phase. During the off-line phase, only a few RSS measurements are collected at known locations, they are used to train a simple propagation-based distance-to-RSS model and build an RSS map covering the whole area of interest. During the online phase, we utilize the constructed RSS map to determine the location given an online RSS measurement. Specifically, we rely on a probabilistic method to find candidate locations with high posteriori probability, and then we exploit redundant AP information and a clustering algorithm to robustly estimate the target's location from these candidates.

A. Establishing RSS Map (Off-Line Phase)

1) Training the propagation based model: In this step, RSS samples and their corresponding location coordinates are collected in order to learn the parameters RSS_0, α, and σ. Mathematically, given a training set $\{rss_{train}, Location(x,y)\}_M$ of size M, where rss_{train} indicates an RSS measurement and $Location(x,y)$ is the corresponding 2D coordinates, we obtain the parameters RSS_0, α, and σ in formula (2) by minimizing the following objective function:

$$E(RSS_0, \alpha, \sigma) = \sum_M \varepsilon^2 \tag{3}$$

$$= \sum_M (rss_{train} - \mathcal{F}_{RSS_0,\alpha,\sigma}(d, N_{ob}))^2 \tag{4}$$

where ε is the error introduced by the simple log distance propagation model considering walls. We assume that the coordinates of all APs are already known, and therefore the distance d to each AP can be calculated. To compute the number of walls on the direct path from a receiver to an AP, we have written a program to analyze graphic floor plan. This program searches for any line element with thickness and length above a certain threshold which is recognized as a wall. A list of all walls specified by their length and location is returned. Every time when two locations are given, we are able to calculate N_{ob} between them by going through this list.

We use the "BFGS" [16] algorithm to solve this optimization problem, and add restrictions for RSS_0, α, and σ to guarantee that the solutions are physically meaningful. Specifically, RSS_0 should not be greater than the maximum RSS ever received, and the value of α and σ are less than zero because RSS is attenuated due to the increase of distance and existence of walls. After RSS_0, α, and σ are obtained, the modeling error is computed and then used to calculate the probability of seeing an online RSS measurement given a trained propagation model. Additional details are discussed in the online phase section.

2) Building the radio map: After the propagation model is learned, we are able to utilize RSS_0, α, and σ to calculate the estimated RSS from a certain AP at any location, as long as the coordinates of this location and target AP are available. First, we divide the floor map into G small grids evenly distributed over the map. Then, at each grid point location with its coordinates known, for each AP, we compute the distance d to it and N_{ob} on the direct path. Finally, we plug d and N_{ob} into formula (2) to estimate the RSS vector at every grid point. Note that the radio cards currently installed on smartphones cannot detect signals below a certain threshold. We utilize a simple piecewise function which sets the RSS value to $minRSS$ when it is below the threshold. Here, we use the minimum RSS value collected in the training set as $minRSS$. In other words, our distance-to-RSS relationship is defined as follows:

$$\mathcal{F}(d, N_{ob}) = \begin{cases} RSS_0 + \alpha log_{10}(d) + \sigma N_{ob}, & if > minRSS \\ minRSS, & if \leq minRSS \end{cases} \tag{5}$$

Therefore, an RSS map denoted by $\{\boldsymbol{RSS_g}\}_G$ is established covering the entire area of interest without further manual effort. Specifically, for each grid point g, its estimated RSS vector is defined as $\boldsymbol{RSS_g} = [rss_g^1, rss_g^2, \ldots, rss_g^N]$, $rss_g^i, i \in N$ represents the predicted RSS value from AP_i, where N is the number of total APs in the area.

B. Location Searching (Online Phase)

During the online phase, our objective is to estimate the current location (x,y) given its online RSS measurement vector: $\boldsymbol{RSS_{measure}} = [rss_m^1, rss_m^2, \ldots, rss_m^n]$, based on the RSS map $\boldsymbol{RSS_g}$ of size G built in the off-line phase. Specifically, $\boldsymbol{RSS_{measure}}$ contains signal readings from n different APs, rss_m^i is the actual RSS reading from AP i, and it is obvious that $n \leq N$. A mathematical expression for making a decision is therefore the following:

$$find\ g \in G,\ arg\max P(\boldsymbol{RSS_{measure}} | \boldsymbol{RSS_g}) \tag{6}$$

1) Computing likelihood for each grid point: At grid point $g \in G$, the likelihood of seeing an RSS vector given the propagation model is the following:

$$P(\boldsymbol{RSS_{measure}} | \boldsymbol{RSS_g}) = \prod_i^{i \in apset} p(rss_m^i | rss_g^i) \tag{7}$$

where rss_m^i represents the online RSS value from AP_i and rss_g^i is the predicted RSS from the same AP at grid point g, $apset$ is the set of APs utilized to compute the likelihood. Since each AP is independent from the others, the likelihood of seeing an RSS measurement vector is the product of likelihood for obtaining a single RSS reading for each AP in the $apset$.

To determine the probability $p(rss_m^i | rss_g^i)$, we define that rss_m^i follows a Gaussian distribution with mean rss_g^i and standard deviation ε_i at each AP i:

$$rss_m^i \sim \mathcal{N}(rss_g^i, \varepsilon_i^2) \tag{8}$$

$$p(rss_m^i | rss_g^i) = \frac{1}{\varepsilon_i \sqrt{2\pi}} e^{-\frac{(rss_m^i - rss_g^i)^2}{2\varepsilon_i^2}} \tag{9}$$

where $\mathcal{N}$ represents Gaussian distribution, and ε_i denotes the error between the estimated RSS value from the propagation

model and the real measurement for AP i. Specifically, ε_i can be calculated after RSS_0, α, and σ are learned in the training phase. By minimizing the objective function (4), we are actually finding the propagation model that is best tuned with the training set. Note that most previous research has relied on reconstructed RSS vectors and compared online RSS readings with them directly, ignoring the error introduced from the RSS-to-distance model. In contrast, our probabilistic approach provides a robust way to capture both the shadow fading effect from other resources and the error from modeling.

2) Refining the estimated location: When the number of observed APs is fewer than three, it is highly likely that potential candidate grid points will be symmetric around an AP or to a symmetric line (e.g. a line connecting two APs), making it difficult to reach a final decision. Fig 1 shows a heat map representing the likelihood (normalized to 1 over all the grid points) of seeing a certain RSS measurement at all the grid points in the WiFi network; the lighter the grid point's color, the more likely the target is located at such grid location. The blue triangles represent the APs observed in this online RSS measurement and the black cross is the actual location. This figure indicates that when only two APs are observed, there are multiple areas with yellow color in the map and they are symmetric around the line connecting the two observed APs. To reduce the influence from ambiguous areas, we exploit redundant AP information and a new clustering method to obtain a better estimation of the target's real location.

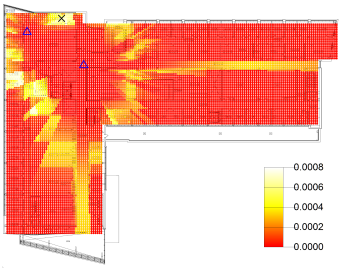

Fig. 1: Heat map showing possible locations when using O-apset

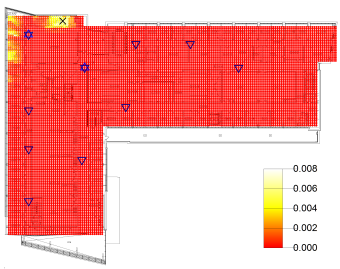

Fig. 2: Heat map showing possible locations when using R-apset

a) Adding redundant AP information: Here, we propose a method to reduce influence from symmetry by adding some redundant APs to the original observable O-apset. A redundant AP helps to eliminate one or more symmetric areas that are not in accordance with the constraint applied by it. In Fig. 2, the light colored zones are considerably reduced after extra APs (blue flipped triangles) are taken into consideration. Specifically, we first find the candidate grid points whose likelihood is above a certain threshold. With the coordinates of APs already known, we then can find the top three geometrically-closest APs for each candidate grid point, referred to as localizable AP set. Finally, all these localizable AP sets are combined with the O-apset and used as a new $apset$ to recalculate likelihood in formula (7). In order to compute (9) when no rss_m^j from an invisible AP_j ($AP_j \notin \boldsymbol{RSS_{measure}}$) is actually received, we perform a small trick by padding this missing value with $minRSS$. Naturally, there also exists an extreme case in which the whole AP set N will be utilized to recalculate the likelihood. Similarly, all the missing RSS values are padded with $minRSS$. In the evaluation section, the performance using three different $apset$ selection scenarios is studied: a) the O-apset—the originally observed, b) the R-apset—the original AP set added with some redundancy, and c) the A-apset—the complete AP set covering the whole area.

However, even though we add redundant AP information to compute likelihood, the ambiguity caused by symmetry remains. In Fig. 2, although the likelihood of each grid point is recalculated over R-apset, there are still multiple light areas surrounding the AP in the top left corner. This is because when we take the error from the propagation model into consideration, the APs with larger modeling error are becoming less determinate. As a result, there are more possible regions appearing in the heat map. Therefore, we utilize a clustering method to further increase the localization accuracy.

b) Clustering: Given the likelihood for each grid point, an intuitive decision can be easily made by picking the one with the highest likelihood. However, this simple solution does not consider the likelihood of its neighbors. Since the probabilities of grid points in a small area are usually very similar, it is not reasonable to consider only the maximum grid candidate when its value is not significantly greater than that of its neighbors. Therefore, we use clustering to increase the localization accuracy. More specifically, we first find the candidate grid points with the top-k likelihood value. Then we classify these candidate grid points into several clusters where each cluster has a maximum radius r, we also define the signature likelihood of a cluster as the average likelihood of all grid points in it. Finally, we select the cluster with the largest signature likelihood, and the returning coordinates are defined as a weighted average of the coordinates of all grid points in the cluster:

$$(x, y) = \sum_{g}^{g \in max\ sig\ cluster} (P^*(\boldsymbol{RSS_{measure}}|\boldsymbol{RSS_g})(x_g, y_g)) \tag{10}$$

where $P^*(\boldsymbol{RSS_{measure}}|\boldsymbol{RSS_g})$ represents the probability normalized to 1 over the grid points in the same cluster, and (x_g, y_g) is the coordinates of grid g. The value of r is crucial—if r is too small, then this approach will degrade to picking a single grid point with the maximum likelihood, but if r is too large, candidate grid points could fall into the same cluster. If these candidate grids are symmetric around a certain point (for example, the light areas are symmetric around the AP in the top-left corner in Fig. 2), the weighted average of these coordinates will be located at such center point, causing unnecessary error. In our experiment, r is empirically set at 6 meters as a constant parameter, the influence of r is further discussed in the evaluation section.

IV. Experimental Results

A. Experiment Setup

We conducted two sets of experiments on the 3rd floor of Davis Hall at SUNY at Buffalo using WiFi APs and Bluetooth beacons. Fig. 3 depicts the overall floor plan of size $76m \times 85m$, including the location of all 14 WiFi APs denoted by triangles. The 802.11 network is officially deployed by UB Computing and Information Technology: no configuration changes have been made to these APs. Although the signal coverage of this WiFi network is fairly dense, one can receive RSS readings from 3 or more APs only at 80% locations. In order to verify the effectiveness of our approach even with

a sparse coverage, we also deployed a Bluetooth network of 12 Bluetooth beacons configured as IBeacon from Gimbal [3] denoted by blue dots. Due to their low transmission power, only about 45% locations are covered by 3 or more beacons.

RSS values of 389 locations are collected as ground truth data and they are depicted as green nodes in Fig. 3. They cover the entire floor except the restroom and server rooms. All RSS values are collected with a Samsung Nexus 5 smartphone running Android version 4.4 ("KitKat"). At each location, WiFi network RSS readings are collected by scanning the channel for two seconds, while a total of four seconds is used to receive Bluetooth RSS measurements due to a longer beacon interval. If multiple RSS samples are received during each scan, the mean value of the samples is used. At each location, the 2D coordinates (x, y) relative to the origin (depicted in Fig. 3) are also collected by a measuring tape. Note that in each experiment, we use only a small subset of the entire location set for training, saving the rest for validation. When establishing the RSS map, we set the grid size to be 0.5 meters, a granularity that provides enough precision for localization.

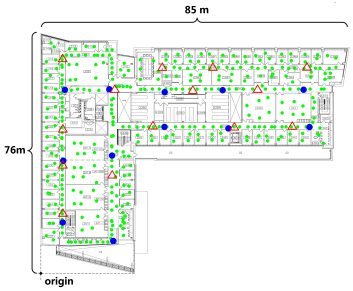

Fig. 3: Floor plan of Davis 3rd floor

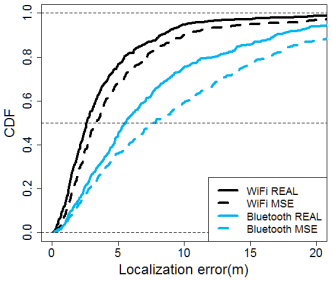

Fig. 4: Localization error CDFs of MSE vs. REAL for WiFi and Bluetooth

B. Performance Evaluation

In this subsection, we will first evaluate the effectiveness of our probabilistic approach by comparing it with MSE [15], keeping the selection of $apset$ and cluster radius the same. MSE can be regarded as a special case of our probabilistic approach, where it assumes that ε_i for any AP i is the same and has value 1. However, in our approach ε_i should be different if the AP is not the same, and this value can be computed from the propagation model. In this experiment, we randomly pick 10% of the location points as a training set, setting aside the remaining 90% for validation. Fig. 4 presents the CDF of the localization errors for these two approaches: for the 802.11 network (black lines), we observe that our approach is slightly better than the LMSE metric, reducing the mean error from $4.94m$ to $3.88m$ and the median error from $3.35m$ to $2.62m$. For the Bluetooth network (blue lines), the improvement is more significant: the median and mean error are reduced by $2.18m$ and $2.5m$ respectively. This is because the Bluetooth signal range is shorter and less stable relative to WiFi, and our probabilistic approach considers modeling errors and thus provides a more robust and accurate comparison metric.

We also study the influence of $apset$ by comparing three different scenarios: (a) the original observable APs (O-apset); (b) a combination of the original AP set with serveral redundant APs (R-apset), and (c) all the APs discovered in the entire floor map (A-apset). We plot the CDF of localization error for WiFi and Bluetooth respectively as shown in Fig. 5 and Fig. 6. Adding redundant APs improves localization accuracy for both networks; however, the performance of second (R-apset) and third (A-apset) scenarios is quite close, which means that utilizing distant APs does not increase localization accuracy. Fig. 7 shows the detailed statistics of error reduction from O-apset to R-apset for both networks, classified by the number of observed APs. The bar height and the number on the top of each bar represent the percentage of validation set and the mean error reduction from each category respectively. We observe that the overall mean error reduction is $1.78m$ for WiFi, while the Bluetooth network sees a much higher improvement of $5.39m$. For Bluetooth, improvement occurs primarily on RSS measurements when fewer than three APs are observed. Even when only one beacon is observed, our approach effectively reduces error by $9.81m$. This result is in accordance with our observation that the probability of seeing fewer than three APs is higher for the Bluetooth network.

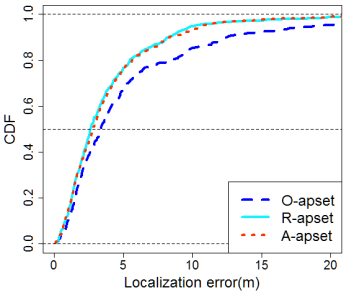

Fig. 5: Comparison of using different AP selection scenario in WiFi network

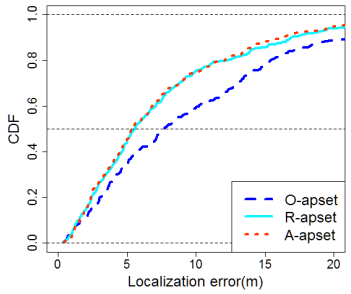

Fig. 6: Comparison of using different AP selection scenario in Bluetooth network

Fig. 8 shows the box plot of localization error for different cluster sizes when 10% locations are used for training. The line inside the box represents the median error, and the top and bottom edge represent the upper and lower quartiles respectively. The whisker shows the maximum and minimum error with a 1.5 interquartile range between the upper and lower quartiles. In this experiment, we vary the cluster radius from 0 to 10 meters, where 0 means clustering is not employed at all. We find that the localization error is reduced when clustering is used ($r > 0$). The median error and confidence interval decrease at first but then increase when r grows larger. This confirms that a large r will produce random errors but a small r cannot capture the full effectiveness of clustering. According to the figure, we set r to be $6m$ for best performance.

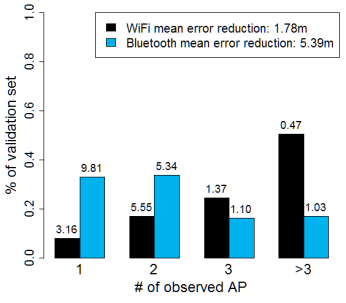

Fig. 7: Error reduction from O-apset to R-apset for WiFi and Bluetooth networks

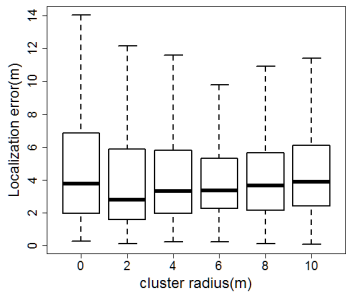

Fig. 8: The influence of cluster radius

C. Cost-Effectiveness of Off-Line Learning

In this subsection, we evaluate the cost-effectiveness of the off-line training phase by varying the size of the training set. We compare our system with support vector regression (SVR)

because SVR has the best empirical performance among other techniques. We use a regression model with Gaussian kernel to relate location coordinates to its RSS measurement:

$$(x, y) \sim SVR(\boldsymbol{RSS}) \tag{11}$$

$\boldsymbol{RSS}$ and (x, y) are the RSS measurement vector and 2D coordinates at a location. To make the RSS vector the same size, we perform the same trick utilized for adding redundant AP information—we pad RSS value of invisible APs with $minRSS$. In order to avoid using overlapped locations for training, we randomly divide the original location set into ten equal parts. For each run we pick out one part, add it to the training set, and leave the rest for validation. The same training set is also used to train the SVR model and predict the (x, y) pair from the RSS vector. Fig. 9 is the box plot of localization errors when we vary the training set from 10% to 90% of 389 locations. We find that when only 10% of the data are used for training, our approach outperforms SVR when 90% of data is used. For SVR, localization accuracy improves significantly when more locations are used for training, however our approach improves very little (mean error from $3.84m$ to $3.53m$) when we increase the training set. In other words, REAL does not rely on a large training set, since the probabilistic approach and clustering method produce robust location searching results although the model is not perfect.

We further reduce the size of training set to determine its lower bound. In order to get every AP trained, we pick the locations which are uniformly distributed in the area. Specifically, we evenly divide the map into equal blocks, then we randomly select one location from each block and add it into training set. For each size we repeat the experiment for 10 times to reduce the influence from randomness. Fig. 10 shows the box plot of the averaged localization error. We observe that even with only 3 training locations, REAL is still able to achieve a moderate localization median error of $6.25m$. The median error reduces to $4.78m$ when the training size is 5, but after that the improvements become smaller, reaching $3.41m$ when 15 locations are used. It shows that REAL can function well with $1 \sim 2$ training locations per $1000m^2$.

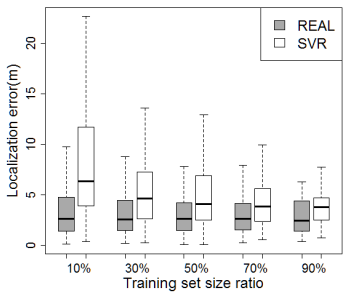

Fig. 9: Comparison of REAL vs. SVR using different size of training set

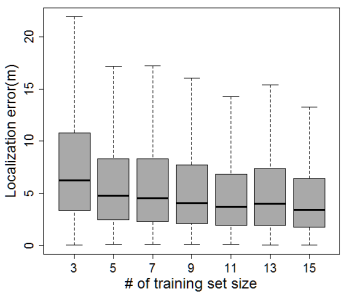

Fig. 10: Lower bound of the training set size for REAL

In this subsection we compare REAL with other typical systems in terms of the training cost (T-cost) and localization median error, where the T-cost is defined as the number of training locations (T-locs) needed for every $1000m^2$. As can be seen from Table I, Horus achieves the minimum localization error at a huge training cost. RADAR has a lower T-cost and uses fewer APs than Horus but results in a larger error. EZ demonstrated that a decent accuracy can be obtained for a large indoor area with a very low T-cost (but note that it also requires boundary GPS information as mentioned earlier) and a sparse coverage with only a few APs. Compared to EZ, REAL achieves a smaller localization error of $4.78m$ with an even lower T-cost of only 5 locations. In addition, the second row shows that even with a sparser AP coverage (i.e., when there are only 9 APs), REAL can also keep the error under 5 meters with a very low T-cost. Hence REAL is a scalable and practical indoor localization system for large indoor areas.

TABLE I: Comparison with other systems

System	T-locs	# AP	Area(m^2)	T-cost	Error(m)
REAL	5	14	3750	1.3	4.78
REAL	14	9	3750	3.7	4.87
Horus	688	21	1768	389	0.46
RADAR	70	3	979	71.5	2.13
EZ	101[2]	12	11466	8.8	5.5

V. Conclusion

In this paper, we have proposed a robust, cost-effective and scalable localization system. More specifically, we have developed a probabilistic approach considering the inaccuracy of the propagation model to calculate the likelihood of seeing an RSS measurement in the RSS map. We have also developed a robust location searching algorithm by utilizing both redundant AP information and clustering method. Our evaluation results in both 802.11 and Bluetooth networks show that the probabilistic approach and our location searching technique improve the localization accuracy. Experiments show that our approach achieves very low localization error even with a small size of only 5 training locations over an area of $3750m^2$. These results clearly demonstrate that REAL is practical for deployment in a large indoor environment.

References

[1] Poovendran, R., Wang, C., Sumit R. *Secure Localization and Time Synchronization for Wireless Sensor and Ad Hoc Networks.* Springer, 2006.

[2] UbiSense Company. http://www.ubisense.net

[3] Gimbal Company. http://www.gimbal.com/

[4] B. Alavi. *Modeling of the TOA-based Distance Measurement Error Using UWB Indoor Radio Measurements.* IEEE Communication Letters 2006. pp 275-277.

[5] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan. *The Cricket Location-Support System.* In Mobicom, 2000.

[6] A. Ward, A. Jones, and A. Hopper. *A New Location Technique for the Active Office.* IEEE PerComm.

[7] EIRIS System. http://www.elcomel.com.ar/english/eiris.htm

[8] R. Want and et al. *The Active Badge Location System.* ACM Transactions on Information Systems, Jan 1992

[9] P. Bahl and V. N. Padmanabhan *RADAR: An Inbuilding RF-based User Location and Tracking System* In INFOCOM, 2000.

[10] M. Youssef and A. Agrawala *The Horus WLAN Location Determination System.* In MobiSys, 2005

[11] Y. Gwon and R. Jain. *Error Characteristics and Calibration-Free Techniques for Wireless LAN-based Location Estimation.* In Mobiwac, 2004

[12] Y. Ji, S. Biaz, S. Pandey, and P. Agrawal. *ARIADNE: A Dynamic Indoor Signal Map Construction and Localization System.* In MobiSys, 2006.

[13] K. Chintalapudi, A. Padmanabha Iyer, and V. N. Padmanabhan *Indoor Localization Without the Pain* In MobiCom, 2010.

[14] S. Y. Seidel, and T. S. Rapport, *914 MHz path loss prediction Model for Indoor Wireless Communications in Multi-floored buildings.* IEEE Trans. on Antennas and Propagation, Feb. 1992

[15] S. Pandey, B. Kim, F. Anjum, and P. Agrawal. *Client assisted location data acquisition scheme for secure enterprise wireless networks.* In IEEE WCNC 2005.

[16] Shanno, David F.; Kettler, Paul C. *Optimal conditioning of quasi-Newton methods.* Math. Comput. 24 (111): 657C664, July 1970

[2]These locations are used for training and localization at the same time