This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE INFOCOM 2008 proceedings.

Drive-by Localization of Roadside WiFi Networks

Anand Prabhu Subramanian Pralhad Deshpande Jie Gao Samir R. Das
Computer Science Department, Stony Brook University, Stony Brook, NY 11794-4400, USA
Email: {anandps, pralhad, jgao, samir}@cs.sunysb.edu

***Abstract*—We use a steerable beam directional antenna mounted on a moving vehicle to localize roadside WiFi access points (APs), located outdoors or inside buildings. Localizing APs is an important step towards understanding the topologies and network characteristics of large scale WiFi networks that are deployed in a chaotic fashion in urban areas. The idea is to estimate the angle of arrival of frames transmitted from the AP using signal strength information on different directional beams of the antenna – as the beam continuously rotates while the vehicle is moving. This information together with the GPS locations of the vehicle are used in a triangulation approach to localize the APs. We show how this method must be extended using a clustering approach to account for multi-path reflections in cluttered environments. Our technique is completely passive requiring minimum effort beyond driving the vehicle around in the neighborhood where the APs need to be localized, and is able to improve the localization accuracy by an order of magnitude compared with trilateration approaches using omni-directional antennas, and by a factor of two relative to other known techniques using directional antennas.**

I. Introduction

Localization of the nodes in a WiFi (802.11) network using radio-based information – such as signal strength, angle of arrival (AoA) etc. – is an important problem. The simple reason for this is that devices with WiFi interfaces are ubiquitous and localization of wireless/mobile devices enables interesting applications. GPS receivers are not available with most wireless clients making radio-based localization the only viable option. Needless to mention that GPS may not always work in indoors and in urban canyons.

In the most common approach for radio-based localization for WiFi, location-tagged RF fingerprinting is used for signals from *infrastructure nodes* (i.e., access points or APs). This location information must be independently determined. RADAR [1] and many followup papers [2]–[4] have used this basic method for indoor localization. On the other hand, Intel's Place Lab work [5] used a similar approach for outdoor localization. In a different approach called VORBA [6], rotating directional antennas are used in APs and a combination of signal strength and AoA information is used to localize clients in indoor environments. VORBA does not require RF fingerprints, but needs multiple APs with rotating directional antennas. All these approaches have been primarily used for localizing wireless client nodes, and not the infrastructure.

In this work, we consider the opposite problem – localization of *infrastructure nodes (APs)*. Our goal is to localize them in a passive fashion, i.e., without their direct participations in the localization process. There are tremendous application for such localization. WiFi networks are growing in a viral manner. Many urban regions have a high density of WiFi APs – deployed in a "chaotic" fashion [7] in homes and businesses, in campuses and hotspots, or as a part of a metro or municipal WiFi effort [8]. There is little knowledge about the nature of these networks, e.g., density, connectivity, interference properties, etc. The first step in understanding their nature is estimating locations of the APs. We expect that at the minimum our effort will provide researchers significant datasets for simulations and modeling purposes. We hope that this will eventually lead to significant research as in understanding Internet topology [9], [10] in wired networking context. Other than datasets for research use, learning locations of APs may reveal interesting social aspects. We will show later the accuracy of our localization is good enough to localize APs within the boundary of a typical house or even an apartment. This can lead to interesting data sets for social science – correlating census data (say, level of education or home price) to Internet usage.[1]

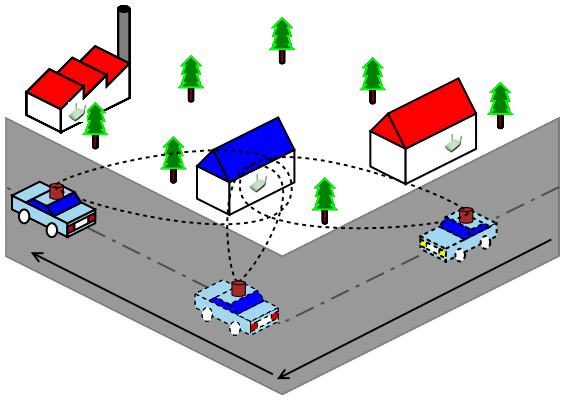


Fig. 1. Drive-by localization of roadside APs.

While several war-driving databases are in existence [12] [13], the location information therein is very primitive. The database simply contains the locations where the APs are heard with a sniffer. As we will show later, even with the most sophisticated techniques this information can only provide very rough location estimates, with errors in hundreds of meters. Our goal here is to be able to improve such location estimates by *an order of magnitude*.

In our approach, we exploit the MobiSteer architecture

[1]We are assuming that having an WiFi AP at home means that residents have broadband connection and use the Internet heavily. Note that we are ignoring privacy aspects. Our technique simply sniffs WiFi frames from streets and public places. This is no different than Google's Street View [11] that takes pictures from streets.

based on our recent work [14]. MOBISTEER uses a steerable beam directional antenna with a WiFi (802.11b/g) client node mounted on a moving car. The antenna coupled with appropriate protocols has been used to improve vehicular connectivity to roadside APs [14]. Here, we use MOBISTEER to gather frames originating at roadside APs on different directional beams to estimate the Angle of Arrival (AoA) of the frames. For robustness reasons, our strategy requires that many samples of AoA information be collected from different locations. Thus a moving car is indeed needed so that many such samples can be collected with relatively little effort. The general idea is driving the car in the neighborhood where APs need to be localized, and collecting GPS-tagged signal strength information on different directional antenna beams for the frames transmitted (e.g., beacons) by the APs. See Figure 1. The approach is purely passive and based on "sniffing" alone; APs are unaware of the localization effort. Because of the use of a car, our work naturally targets outdoor use. However, the APs can be anywhere – either indoor or outdoor. In fact, in most of our experiments they are indeed indoor.

The rest of the paper is organized as follows. In Section II we describe our experimental platform and data collection methods. In Section III we describe our localization approach. The performance results are presented in Section IV. We follow it up with related work and conclusions.

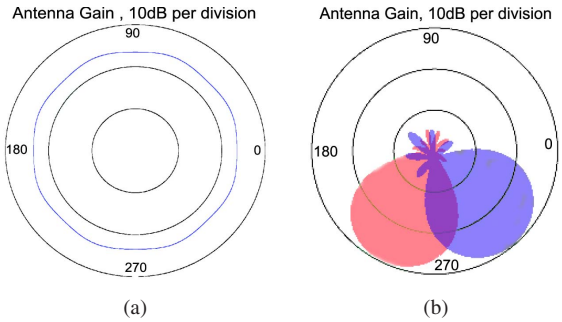
II. EXPERIMENTAL PLATFORM AND SCENARIOS

A. Hardware Setup

Our directional antenna setup uses electronically steerable Phocus Array antennas from Fidelity Comtech [15] for the 2.4 GHz band used in IEEE 802.11b/g. The Phocus Array antenna system consists of eight element *phased arrays* driven by eight individual T/R (transmit-receive) boards that receive radio signals from the wireless card via an eight way RF splitter. The phased arrays combine radio waves by introducing different phase differences and gains in the eight arrays [16] [17]. A T/R board is essentially a vector modulator with bi-directional amplifier controlled by software. Various beam patterns are possible by setting the phases and gains in different boards differently.

The software control on the antenna to produce different beam patterns is achieved via serial-line commands from an embedded computer (a Soekris net4511 board [18]). The beam steering latency has been optimized to 250μs [14]. On the Soekris, we use a 802.11 a/b/g miniPCI card based on Atheros [19] chipset with an external antenna interface. The Soekris computer runs pebble Linux [20] with the Linux 2.4.26 kernel and the widely used `madwifi` [21] device driver for the 802.11 interface.

While many beam patterns are possible using the phased array, the manufacturer ships the antenna with 17 precomputed patterns – one omnidirectional beam and 16 directional beams, each with an approximately 45° half-power beam-width and low sidelobes. The directional gain is about 15dBi. Each directional beam is overlapping with the next beam and is rotated by 22.5° with respect to the next, thus covering the 360° circle with 16 beam patterns. Figure 2 shows the manufacturer provided beam patterns. We refer to the omni-directional beam with beam index 0 and the 16 directional beams we use with beam indices 1 to 16. Adjacent beams are numbered successively. We use a USB-based Garmin [22] GPS receiver inside the car that is connected to the embedded computer. Our experiments with this GPS receiver show a median position accuracy of about 5 meters. The entire hardware setup is called a MOBISTEER node.



(a) (b)

Fig. 2. The beam patterns for phocus array antenna: (a) omni-directional; (b) two directional beams.

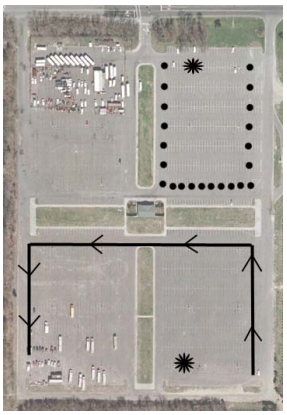
B. Software Setup

The `madwifi` driver allows creation of additional raw *virtual* interface (`ath0raw`) for a physical wireless interface. The virtual interface allows reception of all 802.11 frames (control, management, data) as if in the monitor mode, while the main interface can still operate in the ad hoc or infrastructure mode. We modified Kismet [23], a popular wireless packet sniffer software to *optionally* capture all packets received on the raw virtual interface. Kismet communicates with the GPS server, running as a daemon (`gpsd`), and stamps the current time and GPS coordinates with each received frame from any AP.

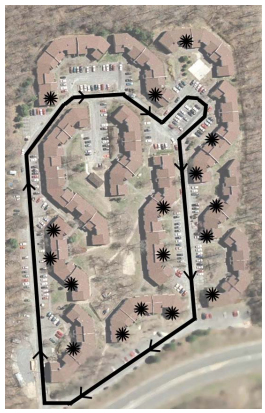
Each received frame (from APs) is also annotated with an index for the current beam pattern on the antenna, orientation of the car (more on this later) and the SNR (Signal to Noise Ratio). The SNR for each received frame is obtained from the radio-tap header appended by the `madwifi` driver for each received frame. The AP's identity (MAC address) is already in the received frame. The tuple *<AP, location, orientation, beam, SNR>* is logged onto the flash memory of the Soekris computer. We also refer to this as a *measurement sample*. SNR is represented in dB, given by $10\log(S/N)$, where S and N are signal power received and noise floor respectively. In Atheros cards, the noise floor N is set at -95 dBm.

C. Experimental Scenarios

Figure 3 shows three representative environments in which we did our experiments. It also shows the actual location of the APs and the driving path of the MOBISTEER node. Figure 3(a) is a large open empty parking lot with no surrounding building.This scenario is used to demonstrate the performance of our localization approach in an uncluttered environment and to create a best case scenario. This is representative of relatively rural or empty environment.



(a) Parking lot (2 APs).



(b) Apartment complex (17 APs).

(c) Office building (2 APs).

Fig. 3. Three experimental scenarios along with the AP locations and the driving trajectories. The black stars are the locations of the APs. The black curve is the trajectory followed by the MOBISTEER Node.

Figure 3(b) and 3(c) are complex environments where there are several two-storied buildings and large trees in close proximity. They are the graduate student apartment complex and the computer science department building in our university, respectively. The APs here are kept indoors as would be normal in such environments. These are relatively challenging scenarios for localization, as there are lots of possibilities of reflections and shadowing. They are representatives of urban homes and offices. The measurements reported here include data collected from APs deployed in 21 different locations in the three scenarios shown in Figure 3.

D. Data Collection

Ideally, we would like to have measurement samples for each AP on all beams at as many points on the roadway around the AP as possible. Samples on all beams let us estimate the AoA – by comparing the received SNRs on all beams. In free space or relatively uncluttered environment, the beam with the highest SNR would be the one pointing towards the AP. A little variation of this also works well in cluttered environments with reflections, which we will describe in the next section.

The complexity in data collection comes from the fact that we have a single radio and single antenna system in the MOBISTEER node. The radio operates on a given channel and the antenna uses a given beam at a time. Thus, all channels and beams need to be scanned. However, for each channel and beam combination the system must hold for certain duration T, where T should be long enough to receive a frame from all APs in range and also long enough so that channel and beam switching latencies can be ignored. We have fixed T to be 100 ms. This time is the default beacon period in most APs and also much longer than the switching latencies. Since T is not insignificant, it is not possible for a moving car to gather measurement samples on all beams exactly at the same location. Our experience has shown that so long as the measurement samples on all beams are within 'close proximity', the errors introduced are not significant in our technique. We have defined 'close proximity' as 5 m, which is similar to the GPS error bound for the GPS receiver we have used. We will discuss more about GPS errors in Section IV.

Still, the car must be driven very slowly. To see this, consider that an entire scan on 16 beams take 1.6 s. Thus, the car should drive maximum 5 m in 1.6 s, i.e., 11.25 km/hr. Such slow driving may not always be practical. We propose to achieve the same effect by driving multiple times on the same route. This approach was also used in [14] to build up an RF signature database, though the context and use of the database were different. One simple approach to reduce the number of runs in the data collection process is to use multiple fixed directional antennas oriented along different directions connected to multiple radios on the moving vehicle. This enables the vehicle to receive frames from APs in all directions throughout the drive.

To summarize, the idea is to drive the car at normal speed appropriate for the roadway used. The drive is simply repeated multiple times such that enough samples are collected. Samples are then clustered such that samples taken within 5 m from one another are assumed to be taken at the same point P. P is assumed to be the centroid of the locations of these samples that are in the cluster. For convenience we will refer to an instance of this point P as the *measurement point*. A larger number of samples naturally provide many such measurement points along the roadway and provides better accuracy for localizing roadside APs. Average SNR is used in the computation when there are multiple samples clustered on the same measurement point for the same beam. Here also, larger number of samples provide better immunity from outliers due to fading. Later in Section IV, we will provide an analysis of sensitivity of our method to the number of samples as number of samples is directly related to the measurement effort.

To simplify the data collection process in the experimental results reported here, we have used a cart pushed at slow walking speed in some instances (scenarios (a) and (c) in Figure 3), instead of using a real vehicle and multiple drives. This enabled us to collect sufficient number of samples per AP on a single "walk". This also enabled us to utilize walking paths and open areas in the university campus where driving is not allowed. For the experiments in scenario (b) in Figure 3, we drove at a very slow speed (approx 10 mph) and repeated the runs multiple times. About 40-60 measurement points are used to localize each AP in the experiments reported here. We used our own APs for the experiments and made them broadcast UDP packets at 250 packets/sec. This let us "speed up" the experiments as we could get many samples on the

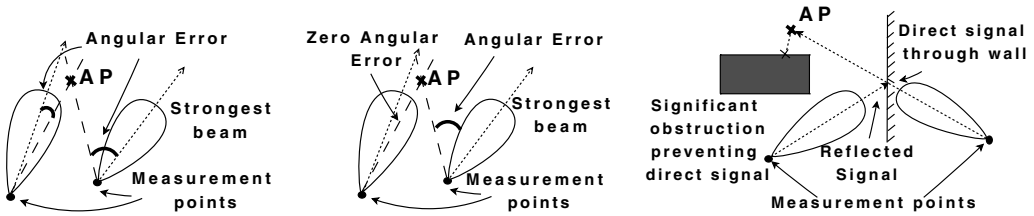


Fig. 4. (a) Angular error using center of directional beam. (b) Angular error considering beamwidth. (c) A scenario demonstrating reflection.

same beam for the same measurement point and used the average SNR for each <*beam,AP*> combination. We used the same channel for all the APs. This speeds up the data collection process further as multiple channels need not be scanned. Note that use of carts or UDP broadcast packets from AP are only used to reduce measurement effort and does not have any fundamental impact on the results.

E. Determining Orientation

Since the antenna is mounted on the car in a fixed fashion, the orientation of the car (with respect to some absolute direction, say magnetic North) also provides an orientation of the antenna setup. This ensures that the absolute direction of the directional beams can be determined by knowing the car's orientation. Orientation of the car can be determined from the *heading* computed from the GPS locations – a method commonly used in navigation systems. This, however, may not provide enough accuracy for quick turns in small spaces. For better accuracy, a digital compass such as [24] could be used.

In the experiments, we indeed used a compass – not digital, but a regular compass with a magnetic needle. During the walks for data collection, we manually ensured that the antenna setup is always oriented in the same direction. In the car experiments, GPS headings were used to compute the orientation. We expect that use of digital compass will simplify the data collection process much and will likely provide better accuracy.

III. Localization Algorithm

A. Preliminaries

Our approach hinges on estimating the AoA of frames from a given AP at each measurement point. The AoA is estimated by noting the average SNR for the frames from a given AP on each directional beam for the same measurement point. The directional beam providing the strongest average SNR (we will call this the *strongest beam*) is expected to point directly to the AP discounting reflections. Thus, the AoA can be estimated by determining the strongest beam and then using the orientation information (Section II-E) to determine the absolute direction of the strongest beam. (Unless mentioned otherwise, the beam direction corresponds to the center of the beam.) Let $\alpha(i)$ denote the absolute direction of the strongest beam at each measurement point (x_i, y_i) along the drive.

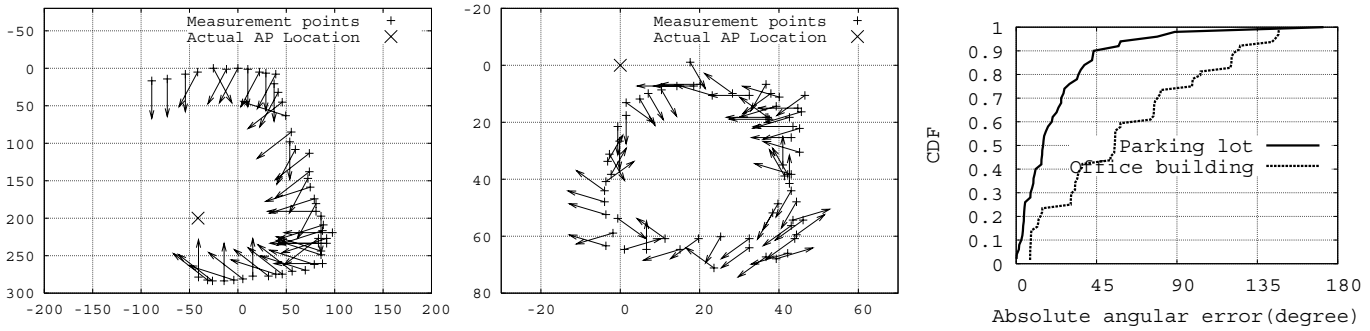
Let us first assume that signal reflections are not present. In this case, the AP can be localized at a point in the 2D plane for which the *sum-square of the angular error* from all the strongest beam directions is minimized similar to the approach used in [6]. See Figure 4(a). Two artifacts, however, complicate this scenario. They are described below.

- *Non-zero beamwidth:* The beams have non-zero width (about 45° between the half-power points on the main lobe). Using the center of the beam for AoA calculations may incur significant error; the non-zero beamwidth must be accounted for. One way to account for this would be to use angular error from the beam sides, i.e., the 'half-power' directions, and pick the minimum of these two errors. In case the direction to the localized point is contained within the beam width, then the error is assumed zero. See Figure 4(b).
- *Reflections:* In most realistic scenarios, radio obstructions and reflections would be present causing the strongest beam point away from the AP. See Figure 4(c). This phenomenon was quite evident in the measurements we did in [14] in cluttered environments (same as Figure 3(b)).

B. Understanding and Modeling Reflections

To understand the impact of reflections, we use Figures 5(a) and (b) to show the measurement points ('+' symbols) and the direction of the strongest beam from each measurement point (arrows). For simplicity we are ignoring the beamwidth issue for now and using the center of the beam for direction. The actual location of the access point is also shown ('$\times$' symbols). These figures correspond to measurements for the empty parking lot (Figure 3(a)) and office building (Figure 3(c)), as described in Section II-C. To complement these figures, we also show the CDF of angular error of the strongest beam from the actual direction towards the AP from each measurement point (Figures 5 (c)).

Note that in the parking lot scenario most arrows are pointing roughly towards the AP as expected. However, the behavior is quite different in the office building scenario. Most of the arrows are pointing in a different direction, presumably due to reflections. Figure 5 (c) qualitatively demonstrates this. Note that for the parking lot scenario, the error CDF is rising sharply, and for the office building scenario, the rise is quite gradual denoting significant errors. The median angular error for the parking lot is about 15° (small) and for the office building is about 55° (unacceptably large). Note also that the 90-percentile error for parking lot is 43° — similar to the beamwidth of the antenna, while the 90-percentile error for office building is again very large (125°).



(a) Parking lot. (b) Office building. (c) CDF of the angular error.

Fig. 5. (a), (b): AP location and directions of the strongest beams in two scenarios demonstrating the impact of reflections. (c): CDF of the angular error of the strongest beam from the actual direction towards the AP in two scenarios.

This study indicates that straight-forward approaches to minimize the sum-square of angular errors [6] using the estimated AoA values can give rise to significant errors in cluttered environments where the strongest signal reaches the MOBISTEER node after one or more reflections. Since this situation will be the common case for roadside APs, modeling reflections is important.

An interesting observation in Figure 5(b) is that while arrows are all pointing to different directions, there seems to be a clustering effect. The arrows are not pointing to random directions but towards one of a handful of possible directions. This is as if each arrow is pointing towards either the real AP or one of its images arising out of presence of reflections. The complexity of modeling reflections now is that there is no knowledge of the number and locations of reflective surfaces that give rise to these images. Thus, there is no knowledge of number of images to look for, and the real AP is indistinguishable from any of its image.

We will approach this problem in the following fashion.

1) Use the well-known *k-means algorithm* [25] to group the measurement points into k clusters such that each group of measurement points have the strongest beam pointing (approximately) towards the same location. These k locations include the real location and the images of the AP. Since the number of reflective surfaces are not known *a priori*, we use the Anderson-Darling normality test [26] to learn the value of k while clustering (more on this in section III-E).
2) Determine which one of these k images is the real AP. We show that it is impossible to determine in a general case, but heuristics can be used quite successfully.

We describe the details in the following subsection.

C. Modeling Reflections by k-Means Clustering

Given the set of tuples $< x_i, y_i, \alpha(i) >$ for each measurement point along the drive, the main idea is to cluster the measurement points into k clusters and find k locations which minimize the *aggregate of intra-cluster sum-square of angular errors*. The angular error is the error considering the beamwidth and is determined as described in Section III-A

In addition, we use weights when minimizing the sum-square of angular errors. The intuition for this is as follows. Since the strongest beam is the one that is important for estimating AoA, we should make distinctions between measurement points with high SNR on the strongest beam compared to those with low SNR on the strongest beam. We use the average SNR (in dB) of packets received in the strongest beam as the weighting function. This intutively puts emphasis on measurement points close to the actual AP location or on those received on a direct beam.

More formally, let w_i be the weights assigned to each measurement point. Let $\alpha_l(i)$ and $\alpha_r(i)$ denote the left and right half-power directions of a beam with absolute angle $\alpha(i)$. We seek to find a k-clustering of the measurement points into k clusters $(S_1, ..., S_k)$ and obtain k locations $L_i = (X_i, Y_i), 1 \leq i \leq k$, that minimize the following objective function:

$$\sum_{i=1}^{k} \sum_{j \in S_i} w_j \cdot \min\{[\alpha_l(j) - \arctan(Y_i - y_j, X_i - x_j)]^2, [\alpha_r(j) - \arctan(Y_i - y_j, X_i - x_j)]^2\}$$

Note that the quantity within $\{.\}$ is the angular error using the non-zero beamwidth idea. Since the angular error can be between 0 and π, if the quantity within $\{.\}$ is larger than π, it is subtracted from 2π.

The k-means clustering algorithm works in the following way. For any given value of k, assume $L_1, \ldots, L_k$ are the k locations of the AP (i.e., real and the images) to be determined. Initially, L_i's are chosen randomly within the 'feasible region'.[2] Each measurement point is mapped to some L_i that provides the minimum angular error for this measurement point. Thus, the measurement points are now clustered into k clusters. The algorithm then repeats the following two steps until convergence.

- Compute a point for each cluster, denoted by C_i, in the feasible region that minimizes the weighted intra-cluster sum-square of the angular errors within each cluster.

[2]In our experiments, we defined the feasible region as a square of side 600m around the measurement points. We assume the transmission range of the AP is not more than 300m and thus the possible location of the AP should be within a region of radius 300m from the measurement points.

- C_i's now become new L_i's. Re-cluster by mapping each measurement point to the L_i that provides the minimum angular error as before.

Convergence is obtained when the clustering does not change. Now we have k locations, $L_1, \ldots, L_k$, which represent the images of the AP including the real location. The next step is to choose one of these k points as the estimated location of the AP.

D. Choosing Real AP Location from k Images

The k-means clustering gives a set of possible locations of the AP, with one of them being the true location and the rest of them reflected images. However, it is hard to distinguish the true location from the image. A simple example is shown as in Figure 6. In the left figure, two measurement points A, B receive signal from the AP. The signal from AP to B is direct but the signal to A was reflected once. Thus at location A the strongest beam points to the image of the AP. Notice that in this figure everything is symmetric, thus we can swap the AP and its image and have another feasible configuration (right). In other words, given a set of k possible locations including the AP and its images, and the way the strongest beam at each measurement point points to these locations, it is impossible to tell the true location of the AP apart from the images, as the same information may admit two (or more) feasible configurations.

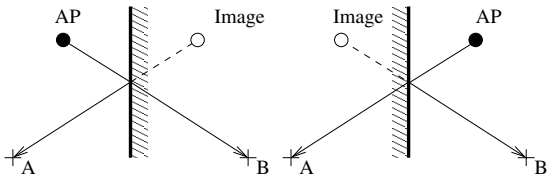


Fig. 6. It is impossible to tell the true location of the AP apart from the images, as the same information may admit two (or more) feasible configurations.

With this difficulty in mind, we use a heuristic to choose the true AP location. Notice that if the strongest beam at a point P points to an image, then the image is *always* farther away from P than the true location. Thus from a particular measurement point's view, the true location must be closer than any of the images. Based on this observation, we propose a simple heuristic that works very well in practice. Each measurement point ranks the k images based on their distances to itself. The nearest image is ranked 1st and the next is ranked 2nd and so on. We compute the weighted (weights being w_j's defined before) sum of the ranks for each image and choose the image with the least value of the weighted sum. In our experiments this always gives the location closest to the true AP location.

E. Learning k for Clustering

One remaining issue is to determine the right value of k to be used in the k-means clustering algorithm. We need some statistical means to estimate k from the measurement data as there is no *a priori* knowledge of number of images. Intuitively, we would like the measurement points to be clustered nicely such that in each cluster the strongest beams all point to the same location L_i. In other words, the angular error within one cluster should be *uni-modal*. Thus, we use the idea from the G-means algorithm [27] and learn the number of clusters, k, by checking whether the angular error values in each cluster follows a Gaussian distribution.

We start with the value of k=1 and successively increment k, performing a k-means clustering in each step as described before. After clustering, we check whether the error values in each cluster satisfy a statistical test for normality. If they do, we stop the procedure; otherwise, we increment k and repeat. We have used the standard Anderson-Darling normality test [26] [28] with a significance level[3] of 1% to test for normality.

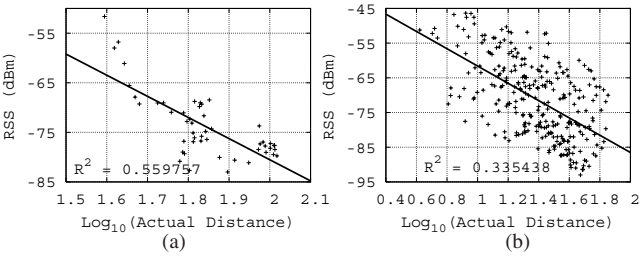


Fig. 7. Relationship between distance and RSS in (a) parking lot (b) apartment complex.

IV. Performance Evaluation

In this section, we present a detailed performance evaluation of our Drive-by Localization approach (DrivebyLoc) using measurements from the 21 APs in three environments as shown in Figure 3. The main comparison points are (i) a trilateration approach using distance information [29] estimated using from signal strength information (SNR) using omni-directional antenna,[4] and (ii) VORBA [6], a localization approach using directional antenna and AoA information. We also study and quantify the effect of several factors that impact the performance of our system.

A. Benefit of Using Directional Antennas and AOA

In Figure 9 (a), we show the CDF of the localization errors for DrivebyLoc and the trilateration approach for the 21 cases studied. The trilateration approach fundamentally depends on deriving distance estimates from the received signal strength (RSS). RSS is derived from SNR assuming constant noise. A simple propagation path loss modeling approach was used to infer distance from RSS following the method used in [30]. The idea is to assume exponential decay of RSS with distance. Thus, RSS (in dBm) should have a linear relationship with the log of distance. <RSS, distance> tuples are collected in the data collection phase *assuming* that the AP locations are

[3]Significance level is the chance of incorrectly judging a set of values to be not Gaussian.

[4]The same antenna with omni-directional beam is used for this study.

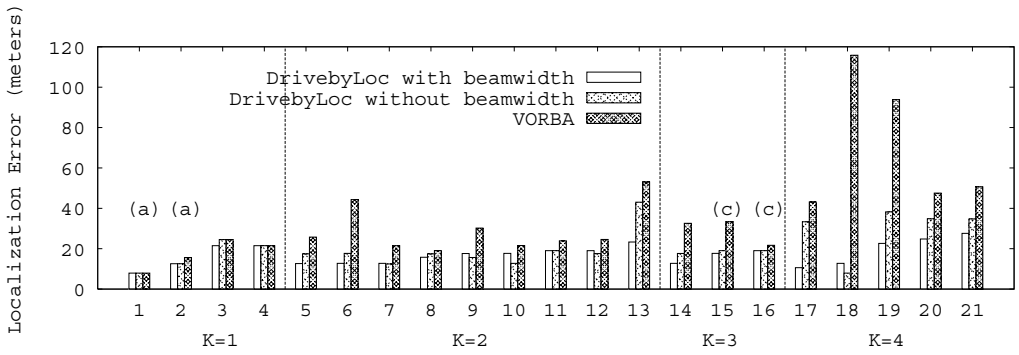


Fig. 8. Comparison of DrivebyLoc and VORBA.

known. This provides the scatterplot in Figure 7. A linear regression (shown) provides the necessary path loss model to be used to estimate distance from RSS. Note that (i) the R^2 value of the regression is not high; and (ii) the parking lot scenario provides a relatively more accurate modeling (higher R^2). The latter is likely due to lesser reflections and shadowing problems.

Figure 9 (a) clearly shows that DrivebyLoc is about an *order of magnitude* better than trilateration. The median localization error in DrivebyLoc is about 15 m, while it is about 128 m in trilateration. In fact, the maximum error in DrivebyLoc is less than 30 m. This shows that it can localize APs within the accuracy of individual homes even in a very cluttered environment.

B. Benefit of Modeling Reflection Using Clustering

We now study the benefit of modeling reflection using the k-means clustering idea and also of modeling non-zero beamwidth. For this purpose, we compare our results with those of VORBA [6]. In VORBA a similar approach is taken, except that a) signal reflections and b) non-zero beamwidth are not modeled and c) no weighting using SNR is used.

Figure 8 shows the localization errors obtained using three approaches in each of our 21 experiments categorized by the value of k learned using the Anderson-Darling normality test. For DrivebyLoc we show the performance with and without modeling of non-zero beamwidth as discussed in Section III-A. Note that the same results for DrivebyLoc with beamwidth modeling was shown before in Figure 9(a) in CDF form.

Note that all three schemes perform almost similarly when $K = 1$. VORBA's performance worsens for larger K. Some very large errors are observed for VORBA $K = 4$ relative to to DrivebyLoc. This indicates the modeling reflections is critical for accurate localization. The impact of modeling beamwidth is usually small and in general it is beneficial, though a few outliers are indeed noticed. But overall it should be recommended that DrivebyLoc be used with modeling beamwidth. Note also that DrivebyLoc's performance with increasing K is relatively stable compared with VORBA. Overall it can be concluded that out of the three differences (see above) between DrivebyLoc and VORBA, modeling reflections has the most impact. Also, overall with the entire data set the median error for DrivebyLoc is about 15 m, while for VORBA it is about 30 m.

For the curious reader, we make a note here which scenarios these 21 sets of experiments correspond to. Note the labels (a) and (c) on top of some of the columns. These correspond to scenarios Figure 3(a) and (c). The rest correspond to the scenario in Figure 3(b). Thus, as expected all schemes get the best results in the parking lot.

C. Impact of GPS Accuracy

As we take measurements while moving in a car, the location of a measurement point is obtained through a GPS unit. GPS devices are known to have errors [31]. In this section, we study the impact of GPS error in our localization approach. In order to quantify the effect of GPS error, we did an experiment in a open parking lot as shown in Figure 3(a) by manually measuring the distance between pairs of measurement points and finding *absolute coordinates* for each point. For this purpose, we used the cart setup as mentioned in Section II-D instead of a car. We moved the cart and stopped at fixed measured distances and took measurements on all directional beams. The dotted lines in Figure 3(a) show the points where the measurements were taken. We also noted the GPS coordinates at each measurement point. The localization error for this particular experiment using the manually measured absolute coordinates is around 12.6 m while the localization error using the GPS coordinates is around 16.4 m. This shows that the GPS errors indeed worsen the performance of our localization approach. To get a better insight about the GPS accuracy, we show the CDF of GPS errors for this particular experiment in Figure 9 (b). The GPS error is computed by finding the difference between actual distance between two points and the distance computed using the GPS coordinates of the two points. The median GPS error is about 1.25 m and this causes an increase in our localization error by about 4 m.

The GPS error could be higher in a cluttered environment with buildings and trees blocking the GPS signals from the satellites. To understand this better, we repeated similar GPS error measurements for the office building scenario. As expected, the errors were higher (median error about 5.5 m) Figure 9 (b), shows the CDF of GPS errors in these two scenarios. This error is quite comparable to the errors DrivebyLoc has in the same scenario (about 20 m). It remains unclear how

the GPS errors are contributing to our localization errors. But given the parking lot experience above, the results presented in Figure 8 likely underestimates DrivebyLoc's performance.

D. Impact of Car Speed

As mentioned in Section II-D, the number of measurement samples could impact the accuracy of DrivebyLoc. The number of samples is inversely proportional to the car speed and directly proportional to the number of runs. Recall that in our data collection approach we either used a car with slow drive (10 mph) or used a cart to walk. Also, in order to 'speed up' the experiments we had the APs broadcast UDP packets at 250 packets/sec which might happen in practise. We are now interested in evaluating what would happen if the car is driven at a more normal speed and APs do not transmit any frame other than beacons at 100ms intervals. We use our existing data to 'simulate' data collection at different speeds and then repeat it to simulate multiple runs. To do this simulation experiment, the existing data is laid out on an imaginary 2D map as points, with each point as if 'lit up' at 100 ms intervals to simulate the corresponding AP beacons. An imaginary car is driven on the same roadway with slightly randomly varying speed (to simulate reality) about an average. Any point lit up within 5 m from the current car location is counted as measurement at the car location. This can be repeated to simulate multiple drives.

In Figure 9 (c), we show the mean localization errors for the 17 experiments in the apartment complex scenario along with the 95% confidence intervals. Two different speeds and multiple drives are shown. As expected, more runs and slower speeds provide better mean localization error and smaller confidence intervals. As discussed in Section II-D, use of multiple directional antennas and multiple radios on the moving vehicle would help to reduce the number of runs significantly.

V. Related Work

RADAR [1] is one of the first systems to do indoor localization of WiFi clients. The key idea is to do an RF fingerprinting *a priori* to collect signal strength values from different APs tagged with location information. When the wireless client needs to be localized, it uses the current signal strength values it receives from different APs and do a lookup on the RF fingerprints. There are several follow up works similar to RADAR [2]–[4]. In particular, Ladd et al [2] improve the accuracy of indoor localization from about 10 m by RADAR to within 1 m by using probabilistic inference of positions from noisy signal information. Using a similar idea, Intel's Place lab work [5] localize wireless clients in outdoor settings. All these ideas suffer from the problem of carefully conducting RF fingerprinting. Also in these works, the emphasis is on localizing WiFi clients unlike our approach.

Use of directional antennas for localization is not new. VORBA [6], one of the significant work in AoA based indoor WiFi localization uses WiFi APs equipped with a rotating directional antenna and estimates AoA information from packets transmitted from clients and uses a simple triangulation approach to find the position of the wireless clients. Our approach compared to VORBA is more robust to AoA information suffering from reflections in cluttered environments that are representative of most urban WiFi deployments.

There have been approaches for indoor localization using other mediums such as ultra sounds, infrared, optical waves ets. Active Badge [32] is an indoor localization system that employs infrared medium. Each user is given an infrared badge and can be localized by IR stations that read the badges. Active Bat [33] uses a similar idea but employs ultrasound medium and has extremely high accuracy in the order of centimeters. The Cricket [34] system from MIT is another indoor positioning system that uses ultrasound combined with RF. It uses several beacons that transmit ultrasound waves deployed in the ceiling of each room in the building. The mobile nodes receiving these waves infer the range and localize themselves. In [35], Nasipuri and El Najjar propose an angle based indoor localization system employing optical waves. They use three rotating optical beacon signal generators that generate regular beacons and wireless sensor nodes equipped with photo sensors determine their locations from the estimated angular separations between the optical sources. This idea is somewhat related our approach in the sense that they also use directional beams and angle information.

Finally, localization in multihop adhoc and sensor networks has been studied in a number of works [29], [36]–[38] and they differ from each other depending on the type of information used for localization such as angles, ranges and connectivity. The idea is to come up with a consistent embedding of the multihop network in either 2D or 3D plane.

VI. Conclusions

In this paper, we have proposed and tested with abundant experiments a system with directional antennas for localizing roadside WiFi access points, by simply driving through the neighborhood where the APs need to be localized. The power of the technique is its complete passive nature. It also does not depend on any prior collection of RF fingerprinting data. While similar approaches have been investigated before [6], our major contribution is identifying that signal reflections can cause significant localization errors and then developing a clustering method to solve this problem. The idea is to recognize *a priori* that there could be images of the AP, and the real AP might be indistinguishable from the images. Thus, we localize these – possibly multiple – images and then use a heuristic to identify the real one among the set of images. The method has demonstrated very satisfactory localization accuracies even in complex environments, compared with existing approaches with omni-directional or directional antennas. In spite of using the toughest scenarios for localization, the localization errors are roughly between 10-30 m in spite of the fact many of our APs were indoors. This is very good compared to the localization error observed in Intel Place Lab outdoor localization effort [5], where the *median* error is between 13-40 m in spite of very careful radio fingerprinting of the environment. This is an impressive performance given

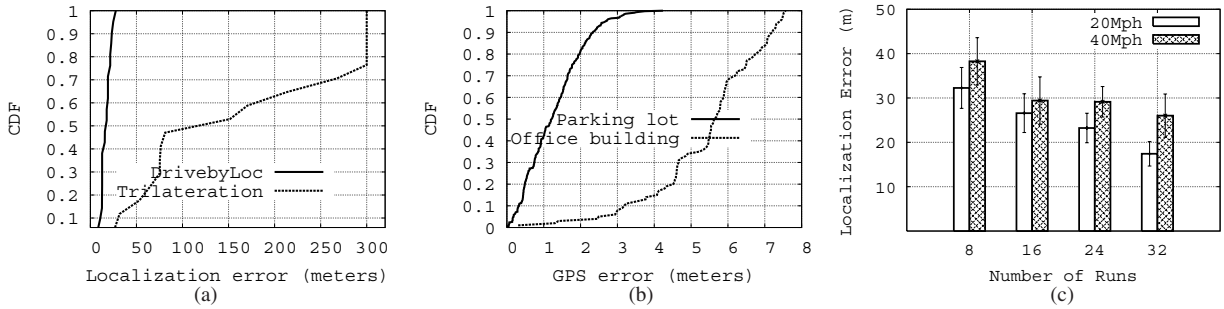


Fig. 9. (a) Localization errors for DrivebyLoc and trilateration.(b) CDF of the GPS errors in two scenarios. (c) Impact of speed and number of runs on localization error.

that we observed up to about 7.5 m of GPS error in similar environments.

While the work so far is limited to 2D, this can be extended to 3D using a combination of antennas so that directivity is on both horizontal and vertical axes can be obtained. We will consider this in the future. We expect that an important fallout of our work will be the creation of a very accurate 'WiFi map' of urban APs with a minimum effort, and eventually motivate novel applications.

ACKNOWLEDGMENT

Anand Prabhu Subramanian, Pralhad Deshpande and Samir Das's research has been partially supported by the NSF grants CNS-0519734 and CNS-0721455. Jie Gao would like to acknowledge the support of NSF CAREER Award CNS-0643687.

REFERENCES

[1] P. Bahl and V. N. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," in *IEEE INFOCOM*, 2000, pp. 775–784.
[2] A. M. Ladd, K. E. Bekris, A. Rudys, G. Marceau, L. E. Kavraki, and D. S. Wallach, "Robotics-based location sensing using wireless ethernet," in *ACM MOBICOM*, Atlanta, GA, September 2002.
[3] P. Krishnan, A. S. Krishnakumar, W.-H. Ju, C. Mallows, and S. Ganu, "A system for LEASE: System for location estimation assisted by stationary emitters for indoor RF wireless networks," in *IEEE Infocom*, Hongkong, March 2004.
[4] M. Youssef, A. Agrawala, and U. Shankar, "WLAN location determination via clustering and probability distributions," University of Maryland, College Park, MD, Tech. Rep., March 2003.
[5] Y.-C. Cheng, Y. Chawathe, A. LaMarca, and J. Krumm, "Accuracy characterization for metropolitan-scale wi-fi localization," in *MOBISYS*, 2005, pp. 233–245.
[6] D. Niculescu and B. Nath, "VOR Base Stations for Indoor 802.11 Positioning," in *ACM MOBICOM*, 2004.
[7] A. Akella, G. Judd, P. Steenkiste, and S. Seshan, "Self Management in Chaotic Wireless Deployments," in *ACM MOBICOM*, Cologne, Germany, August 2005.
[8] "Chaska wireless solutions," http://www.chaska.net/.
[9] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the Internet topology," *Proc. Sigcomm*, pp. 251–262, 1999.
[10] D. Alderson, L. Li, W. Willinger, and J. Doyle, "Understanding Internet topology: principles, models, and validation," *IEEE/ACM Transactions on Networking*, vol. 13, no. 6, pp. 1205–1218, 2005.
[11] "Google Street View," http://maps.google.com/help/maps/streetview/.
[12] "WiFi Maps," http://wifimaps.zhrodague.net/.
[13] "Wireless Geographic Logging Engine (WIGLE)," http://www.wigle.net/.
[14] V. Navda, A. P. Subramanian, K. Dhanasekaran, A. Timm-Giel, and S. R. Das, "MobiSteer: Using Steerable Beam Directional Antenna for Vehicular Network Access," in *MOBISYS*, San Juan, Puerto Rico, June 2007.
[15] "802.11 Phocus Array Antenna System by Fidelity Comtech," http://www.fidelity-comtech.com/.
[16] J. H. Winters and M. J. Gans, "Versus Phased Arrays in Mobile Radio Systems," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 2, pp. 353–362, 1999.
[17] Phocus Array Antenna System: Manual, Fidelity Comtech Inc. 2006.
[18] "Soekris Engineering," http://www.soekris.com/.
[19] "Atheros Communications," http://www.atheros.com.
[20] "NYCwireless Pebble Linux," http://www.nycwireless.net/pebble.
[21] "MADWiFi Project," http://sourceforge.net/projects/madwifi/.
[22] "Garmin GPS System," http://www.garmin.com/.
[23] "Kismet Wireless," http://www.kismetwireless.net/.
[24] "Honeywell HMC6352 Digital Compass Solution," http://www.ssec.honeywell.com/magnetic/datasheets/HMC6352.pdf.
[25] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations," in *5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley*, University of California Press, 1967.
[26] T. Anderson and D. Darling, "Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes," *Annals of Mathematical Statistics*, vol. 23, pp. 193–212, 1952.
[27] G. Hamerly and C. Elkan, "Learning the K in K-mean," *Advances in Neural Information Processing Systems (MIT Press)*, vol. 16, 2004.
[28] M. Stephens, "EDF Statistics for Goodness of Fit and Some Comparisons," *Journal of the American Statistical Association*, vol. 69, pp. 730–737, 1974.
[29] A. Savvides, C.-C. Han, and M. B. Srivastava, "Dynamic fine-grained localization in Ad-Hoc networks of sensors," in *ACM MOBICOM*, 2001, pp. 166–179.
[30] J. Camp, J. Robinson, C. Steger, and E. W. Knightly, "Measurement Driven Deployment of a Two-Tier Urban Mesh Access Network," in *MOBISYS*, Uppasala, Sweden, June 2006.
[31] GPS Errors and Estimating Receiver Accuracy, http://edu-observatory.org/gps/gpsaccuracy.html.
[32] R. Want, A. Hopper, V. Falcao, and J. Gibbons, "The active badge location system," *ACM Transactions on Information Systems*, vol. 10, pp. 91–102, January 1992.
[33] A. Harter, A. Hopper, P. Steggles, A. Ward, and P. Webster, "The anatomy of a context-aware application," in *ACM MOBICOM*, Seattle, Washington, August 1999.
[34] N. Priyantha, A. Miu, H. Balakrishnan, and S. Teller, "The cricket compass for context-aware mobile applications," in *ACM MOBICOM*, Rome, Italy, July 2001.
[35] A. Nasipuri and R. E. Najjar, "Experimental Evaluation of an Angle Based Indoor Localization System," in *WiNMee*, Boston, MA, April 2006.
[36] D. Niculescu and B. Nath, "Ad hoc positioning system (APS)," in *IEEE Globecom*, November 2001.
[37] J. Bruck, J. Gao, and A. Jiang, "Localization and Routing in Sensor Networks by Local Angle Information," in *ACM MOBIHOC*, May 2005.
[38] L. Girod and D. Estrin, "Robust range estimation using acoustic and multimodal sensing," in *International Conference on Intelligent Robots and Systems*, October 2001.