This paper was presented as part of the main technical program at IEEE INFOCOM 2011

Implications of Device Diversity for Organic Localization

Jun-geun Park
MIT CS & AI Lab
Cambridge, MA, USA
jgpark@csail.mit.edu

Dorothy Curtis
MIT CS & AI Lab
Cambridge, MA, USA
dcurtis@csail.mit.edu

Seth Teller
MIT CS & AI Lab
Cambridge, MA, USA
teller@csail.mit.edu

Jonathan Ledlie
Nokia Research Center
Cambridge, MA, USA
jonathan.ledlie@nokia.com

***Abstract*—Many indoor localization methods are based on the association of 802.11 wireless RF signals from wireless access points (WAPs) with location labels. An "organic" RF positioning system relies on regular users, not dedicated surveyors, to build the map of RF fingerprints to location labels. However, signal variation due to device heterogeneity may degrade localization performance.**

We analyze the diversity of those signal characteristics pertinent to indoor localization — signal strength and AP detection — as measured by a variety of 802.11 devices. We first analyze signal strength diversity, and show that pairwise linear transformation alone does not solve the problem. We propose kernel estimation with a wide kernel width to reduce the difference in probability estimates. We also investigate diversity in access point detection. We demonstrate that localization performance may degrade significantly when AP detection rate is used as a feature for localization, and correlate the loss of performance to a device dissimilarity measure captured by Kullback-Leibler divergence. Based on this analysis, we show that using only signal strength, without incorporating negative evidence, achieves good localization performance when devices are heterogeneous.

I. Introduction

As more people carry devices that can determine their location, numerous applications have emerged that rely on accurate and continuous location information. Typical examples include route finding, photo and video geotagging, friend finding, targeted advertising and coupons, suggesting local points of interest, and inferring a user's context. GPS [1] is the primary source of location estimates for many applications, which function as expected only when location is available. This failure typically occurs indoors, but may happen outdoors as well, in "urban canyons" or in bad weather where the satellite signals required for GPS are obscured. In order to improve the availability of these applications, researchers have worked for two decades on extensions and alternatives to GPS to provide location information ubiquitously [2]–[5].

Early work on alternatives to GPS typically explored instrumenting a lab space with beacons and mobile devices with receivers. For example, ActiveBadge [2] used infra-red beacons; Cricket [3] used RF/ultrasound beacons. However, as IEEE 802.11 access point coverage has become nearly universal in most urban environments [6], localization methods that rely on WiFi beaconing have become the primary alternative to GPS, particularly in indoor environments. Beginning with work by Bahl and Padmanabhan [7], these methods achieve accuracy comparable to that of GPS — typically a few meters [8] — and, therefore, can often support location-based applications at no additional infrastructure cost.

The basic survey/use model for WiFi-based localization is well-understood [4], [7]–[11]. First, an expert surveyor walks around a building, collecting a set of scans, or a *fingerprint*, for each room. In aggregate, the fingerprints for a building make up its *signal map*. Second, to estimate its location, a user's device observes its own fingerprint and finds the closest match in the signal map, returning the corresponding coordinate or semantic name.

While this curated model is practical for highly managed and popular spaces, such as airports and city streets, it does not effectively extend positioning applications into schools, shops, offices, and homes — the long tail of spaces where people spend most of their time. This is chiefly because expert surveying is time-consuming, intrusive, requires maintenance over time, and expensive: a commercial survey of a single multi-story building can cost more than $10,000.

In order to overcome these shortcomings and extend WiFi-based localization and its applications into this much broader range of environments, several recent localization systems have adopted user-collected fingerprints [12]–[15]. These systems follow a Wikipedia-style crowdsourcing model with respect to populating and maintaining the fingerprint database: a small fraction of local users contribute survey information about their locations, while the vast majority of users simply benefit from the database. Figure 1 illustrates this process for an office environment. In contrast to survey-driven WiFi collection and maintenance, these crowd-sourced, or "organic," localization systems must determine when to prompt user/surveyors for input and must exclude invalid fingerprints automatically [15].

A key problem for organic indoor localization is that device diversity introduces a new, complicating variable. Where expert surveys may use an expensive RF-scanner whose properties, *e.g.* dynamic range, have been rigorously calibrated, this level of standardization and equipment cannot be expected in organic surveying with typical consumer-grade laptops and cellphones. The "device heterogeneity" problem occurs when a user/surveyor's and a standard user's devices are different, which is the common case in organic location systems. This paper addresses device heterogeneity, significantly expanding

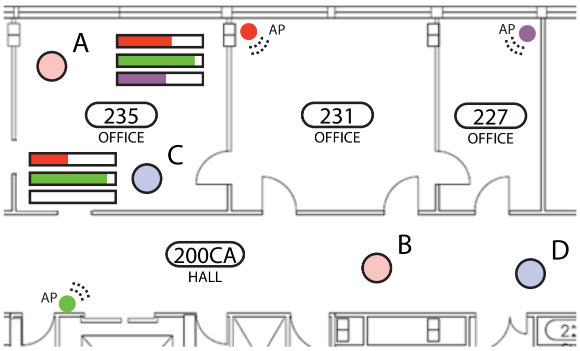


Fig. 1. In an "organic" indoor location system, a small fraction of users contribute RF signatures for each space, while most user's devices simply use the resulting database for positioning. For example, users A and B could be contributors to the office environment shown, whereas C and D are non-contributing users. The bars next to A and C represent the set of signal strengths (the fingerprint) seen by their devices when they are in room 235. If A and C have different devices, the fingerprints observed may differ, even if they are in the same room at the same time. This paper examines how to achieve accurate indoor localization in this situation, which is becoming more common as more location applications rely on RF signatures in GPS-denied environments.

the potential of organic localization systems for real-world deployment.

Previous work has acknowledged the problem of sharing fingerprints between heterogeneous devices and suggested it could be solved by a simple linear transformation between received signal strength indicator (RSSI) values [8]. We demonstrate that linear transformation of signal strengths is insufficient for cross-device localization. Instead, we find that a wide kernel applied to the signal strength distribution provides significantly better end-to-end accuracy than linear transformation, because the former captures the primary difference across devices: signal strength dispersion. A second complication afflicts fingerprint sharing: the set of access points detected by each device can be different. We show experimentally that the number and the identity of APs detected by each device can vary widely. As a consequence, a common alternative to RSSI-based localization — relying on access point presence or absence — fails when this type of fingerprint is shared across heterogeneous device types. We show through an information-theory argument that augmenting RSSI-based localization methods with presence/absence will actually degrade performance.

This paper makes the following contributions:

- We analyze the problem of localizing with RF fingerprints surveyed from and shared by different types of devices: *i.e.* , the use of heterogeneous devices for organic indoor localization.
- We show that using kernel estimation with wide kernel widths to transform the received signal strengths from one device to another is beneficial for cross-device localization.
- We show why using AP presence/absence information yields poor performance for cross-device localization, and relate it to the device dissimilarity captured by Kullback-Leibler divergence.
- We contribute a public RF-scan dataset, collected from six heterogeneous devices, for other researchers to analyze.

II. Background

We describe our client-server model for collecting and sharing organic fingerprints (§II-A), review Bayesian localization (§II-B), then cover the three features most commonly used for Bayesian localization (§II-B1,§II-B2,§II-B3).

A. Model

Our model for signature collection, validation, sharing, and use is based on our previous work deploying an Organic Indoor Location (OIL) system [15]. OIL follows a client-server design: devices collect scans, users associate scans with a location label to form a fingerprint, and devices then send this fingerprint to a server; fingerprints are shared across devices via the server; localization itself is performed on each mobile device. This design is similar to other organic localization systems, such as RedPin [13] and ActiveCampus [12], all of which must grapple with device heterogeneity.

B. Bayesian Localization

For localization, we use the naïve Bayes classifier, which has a number of properties that make it suitable for user-contribution-based localization systems for mobile devices: low overhead of model update, simple computation on the client, and good performance despite its simplicity.

Given a set of fingerprinted locations (*training data*) and a WiFi scan observation (*test data*), the Bayesian localization method infers the most likely location $\hat{l}$ of the mobile device using Bayes' rule. Let L and O denote the random variables for location and observation respectively. Given a WiFi scan observation $o \in O$, the posterior probability of being in location $l \in L$ is given by the Bayes' rule:

$$p_{L|O}(l|o) = \frac{p_{O|L}(o|l)\ p_L(l)}{p_O(o)}. \tag{1}$$

Note that the observation likelihood $p_O(o)$ is fixed and can be ignored in what follows. If we assume that the prior probability, $p_L(l)$, is uniform, the *maximum a posteriori* (MAP) estimate, with which posterior probability is maximized, is given as follows:

$$\hat{l} = l_{MAP} = \underset{l \in L}{\operatorname{argmax}} \left[p_{O|L}(o|l) \right]. \tag{2}$$

The MAP estimate depends on the class-conditional probability $p_{O|L}(o|l)$, which is described by features from WiFi measurement characteristics. The common features used for WiFi localization are signal strength and AP presence/absence. For all models, we assume that each feature is conditionally independent of every other feature given a location, yielding the naïve Bayes classifier.

In the following sections, we describe how to model the class-conditional probability for each feature, along with a hybrid model that uses both features.

1) Signal Strength: In signal-strength-based localization, each observation consists of a vector of signal strengths $O = (S_1, S_2, \ldots, S_k)$ for k access points. Suppose, at the positioning phase, only $m \leq k$ access points are observed with a signal strength value. Let M denote the index of the observed access points. Then, the decision rule (Equation 2) becomes:

$$\hat{l} = \underset{l \in L}{\operatorname{argmax}} \left[\prod_{i \in M} p_{S_i|L}(s_i|l) \right]. \tag{3}$$

The class-conditional probability $p_{S_i|L}(\cdot|\cdot)$ can be estimated from training data in different ways, by modeling it as a categorical distribution (histogram), a Gaussian distribution (with maximum-likelihood parameter estimates), or a kernel density estimator (Parzen window estimator). The latter accounts for the variance of each sample.

The kernel density estimator $\hat{p}^k_X(\cdot)$ [16] estimates the probability density function $p_X(\cdot)$ as follows:

$$\hat{p}^k_X(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4}$$

where x_i is an observed sample of random variable X, h is a kernel width, and $K(\cdot)$ is a kernel function. A Gaussian kernel is often used for the kernel function. The kernel width determines the degree of sample "smoothing" effected by the kernel.

The estimated probability density function is discretized and m-estimate smoothing is applied to avoid the zero-probability problem in Bayesian classification as well as to provide a regularization effect [17]. Given a discretized probability mass function $\hat{p}_X(x)$ computed from either histogram, Gaussian, or kernel estimator, our final probability estimate $p_X(x)$ using the m-estimator is:

$$p_X(x) = \frac{N\hat{p}_X(x) + \Phi\bar{p}_X(x)}{N + \Phi} \tag{5}$$

where N is the weight of the observed histogram, $\bar{p}_X(x)$ is a "prior" probability, and Φ determines how much weight we attribute to the prior $\bar{p}_X(x)$. We use the uniform prior, and $\Phi = N/10$ for the experiments.

2) AP Detection: Another feature vector can be constructed to reflect the presence or absence of access points. If we model the presence/absence of the signal from a certain AP as a Bernoulli process, the observation follows a multivariate Bernoulli model in which $O = (J_1, J_2, \ldots, J_k)$ for k APs, where J_i is a binary variable with 1 indicating presence of signal from AP i and 0 indicating absence. In this framework, the decision rule (Equation 2) becomes:

$$\hat{l} = \underset{l \in L}{\operatorname{argmax}} \left[\prod_{1 \leq i \leq k} \left\{p_{J_i|L}(1|l)\right\}^{J_i} \left\{1 - p_{J_i|L}(1|l)\right\}^{1-J_i} \right] \tag{6}$$

where $p_{J_i|L}(1|l)$ is the probability that AP i is detected in a WiFi measurement at location l.

This localization algorithm requires only presence/absence information of access points, which can be easily obtained from any WiFi device, and from which constructed fingerprints are compact compared to those of signal-strength-based localization. Because of these merits, presence/absence information is particularly well-suited to large-scale, coarse-grained localization.

However, this formulation explicitly considers negative evidence — absence — of a signal from a certain access point; we show later that this becomes problematic when different devices detect partially disjoint sets of APs.

3) Hybrid: Signal Strength and AP Detection: The detection probability $p_{J_i|L}(1|l)$ can be used together with the signal strength as feature variables. In this formulation, the observation variable becomes $O = ((J_1, S_1), (J_2, S_2), \ldots, (J_k, S_k))$. The signal strength variable S_i is conditioned on the detection variable J_i, and must be marginalized if it is not observed ($J_i = 0$). Based on this, we derive the following classification rule from Equation 2 as follows:

$$\hat{l} = \underset{l \in L}{\operatorname{argmax}} \left[\prod_{1 \leq i \leq k} \left\{p_{J_i|L}(1|l)\ p_{S_i|J_i,L}(s_i|1,l)\right\}^{J_i} \left\{1 - p_{J_i|L}(1|l)\right\}^{\alpha(1-J_i)} \right] \tag{7}$$

where α, $0 \leq \alpha \leq 1$, is a discounting factor which determines how much to discount negative evidence. This prevents localization from being dominated by negative evidence if many access points are present but each WiFi scan captures only a small fraction of them. When $\alpha = 1$, the formula is identical to the one used in [9]; when $\alpha = 0$, it is equivalent to [18]. We set $\alpha = 1$ for our experiments.

III. Experimental Setup

In order to examine the effect of device diversity on indoor positioning, we collected WiFi scans from six different devices at 18 locations in one building. We used two different commodity laptops, a netbook, a mobile phone, and two tablet computers. The tablets were the same model, illustrating the homogeneous organic localization case. Table I summarizes the devices we compared for the experiment.

The six devices were placed on a rolling cart, enabling simultaneous data collection, with all logging to local storage

Device	WiFi Chipset	OS	Kernel
Clevo D901C laptop	Intel 5300AGN (802.11a/b/g/n)	Linux Ubuntu 10.04	Linux 2.6.32
Asus EEE900A netbook	Atheros AR5001 (802.11b/g)	Linux Ubuntu 10.04	Linux 2.6.32
Lenovo Thinkpad X61 laptop	Intel 4965AGN (802.11a/b/g/n)	Linux Ubuntu 10.04	Linux 2.6.32
Nokia N810 tablet (x2)	Conexant CX3110X (802.11b/g)	Maemo OS2008	Linux 2.6.21
Nokia N95 cellphone	TI OMAP2420 (802.11b/g)	Symbian S60 FP1	EKA2

TABLE I
DEVICES USED FOR DATA COLLECTION.

to avoid using their radios for data transmission while data collection was in progress. The device radios performed no activity other than scanning.

In each location, each device recorded WiFi scans for seven minutes. Scans were taken near each device's peak rate, with a one-second gap between scans. Because of the variation in the time each device takes to complete a scan, this resulted in a maximum of 552 and a minimum of 61 scans collected at any location; the EEE900A laptop, for example, often took seven seconds to complete a scan request. While this difference would affect time-to-update performance for moving users, our analysis ignores this factor because it is not relevant for instantaneous localization. To remove the effect of this factor from our results, in each experiment we selected 60 scans at random per device from each seven-minute period. (See `http://rvsn.csail.mit.edu/location` for the raw data, and the samples used for each experiment.)

IV. HETEROGENEOUS WIFI SIGNAL STRENGTHS

This section shows that signal strength scans from different devices exhibit not only a linear shift in signal strength but also a difference in dispersion. This suggests that sharing fingerprints between different devices would be more effective with "smoothed" signal strength values, *e.g.* a wide kernel function. We then show that using a wide kernel to share signatures does indeed lead to a significant improvement in accuracy.

A. Analysis of Pairwise Device Calibration

Previous work suggested that inter-device calibration can be achieved by applying a linear transformation of signal strength values from one device to the other [8], [19]. Such a linear function can be estimated from WiFi scan data taken from both devices at the same time and place.

We compare signal strength measurements of the six devices in our dataset, showing pairwise scatter plots in Figure 2. We observe a strong correlation in the *mean value* of signal strength measurements between every pair of devices. Therefore, as suggested by previous work, we first attempt the following simple procedure for cross-device localization.

1) **Pairwise device calibration.** For every pair of devices A and B, the coefficients for linear transformation from device A to device B are computed as:

$$\bar{S}^B_{i,l} = \beta^B_A \bar{S}^A_{i,l} + \alpha^B_A, \quad (8)$$

where $\bar{S}^A_{i,l}$ ($\bar{S}^B_{i,l}$) denotes the mean signal strength value of device A (device B) for WAP i at location l, and α^B_A and β^B_A denote linear coefficients for the transformation from device A to B.

2) **Positioning.** If device A is used for training and B is used for positioning:
   a) Linear transformation from B to A is applied to test scans of device B.
   b) Device B is then localized using device A's training data.

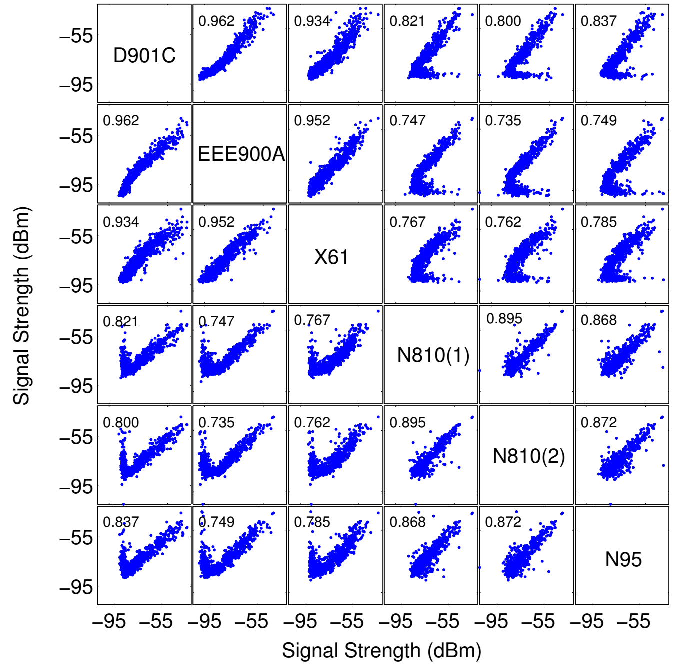


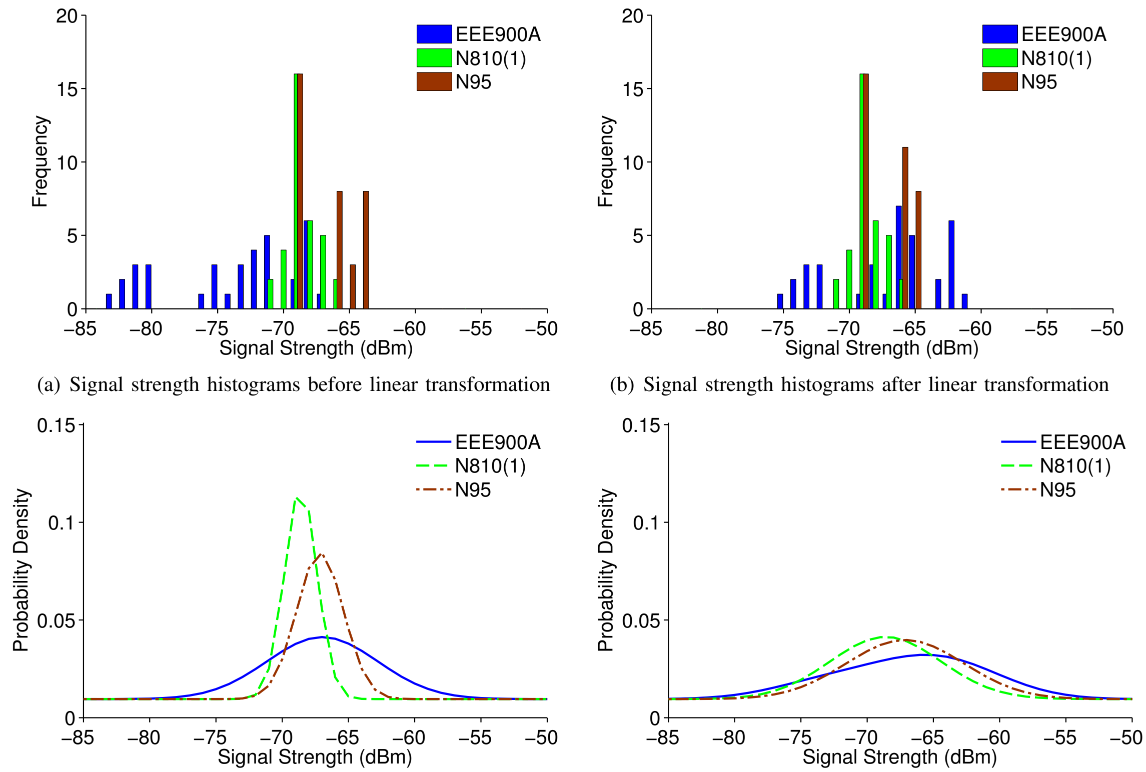
Fig. 2. Each point in each scatter-plot represents a pair of mean RSSI values from the same access point. The values were observed by pairs of devices placed in the same location for the same time interval. For example, the scatter-plot in row 3, column 4, compares the scans of the `N810(1)` with the Thinkpad `X61`. Most pairs show a strong linear correlation, but some devices, *e.g.* N810s, show noisy values at low signal strength ranges. The Pearson correlation coefficient for each dataset is given in the corresponding scatter-plot.

The linear transformation is computed by linear least squares, with sample pairs differing by more than 20 dB excluded from fitting.

Figure 3 shows the resulting localization error in meters when WiFi measurements from each device are tested against training data from device `N810(2)`. Other combinations of devices showed similar characteristics. For baseline evaluation, we used the Gaussian distribution for class-conditional probabilities.

As Figure 3 shows, linear transformation with Gaussian class-conditional probability improved localization performance significantly only for `EEE900A`, while it did not provide significant improvement for other devices. Among every combination of training and test device, linear transformation improved performance significantly only when `EEE900A` was used for either the training or test device. For the 10 device pairs including `EEE900A`, the improvements in spot-on accuracy and error distance were 29.8% and 5.47 meters respectively, while for the other 20 device pairs, improvements were 3.29% and 0.418 meters, which are not significant.

This observation led us to investigate the net effect of linear transformation on actual WiFi measurements of each device. Table II shows the dynamic range of each device. Only `EEE900A` was significantly different from other devices with respect to the dynamic range of signal strengths. As a result, the linear transformations shift signal strengths only



(a) Signal strength histograms before linear transformation

(b) Signal strength histograms after linear transformation

(c) Gaussian estimation

(d) Kernel estimation

Fig. 4. Kernel density estimation vs. Gaussian estimation. In 4(a), the raw data show the histogram of RSSI values from a single AP that three devices observed in the same room during the same seven-minute window. Only the signal strengths for `EEE900A` are considerably different, while the difference between `N810(1)` and `N95` is smaller. In Figure 4(b), the linear transformation is effective for `EEE900A`, while its effect is minimal for `N95`. Even after transformation, the dispersion and shape of signal strength values for each device differ significantly. Therefore, the Gaussian probability estimates (and histogram) for these devices differ significantly (Figure 4(c)), and this difference adversely affects localization. For example, a signal strength of -64 dBm is observed often for `N95`, but has near-zero probability if Gaussian-fitted training data for `N810(1)` is used. Estimation with a wide kernel significantly reduces the difference between different devices (Figure 4(d)).

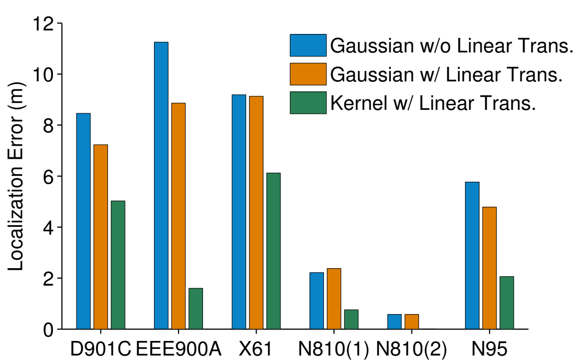


Fig. 3. Localization accuracy, as measured in mean physical distance away from the correct room, when the `N810(2)` acted as the organic surveyor and the other devices act as users.

slightly for most devices, except for `EEE900A`. However, the mean deviation (regression residual) of signal strength values, excluding outliers, from the linear transformation lines was 3.5 dB. This means that the amount of deviation of each signal strength value is comparable to the amount of global shift by linear transformation, except for `EEE900A`. Thus linear transformations are ineffective for other devices. Figure 4(a)–4(c) illustrate the details.

B. Kernel Density Estimation

This observation implies that the major characteristics of signal strength diversity lie not only in the linear difference between devices, but also in the different local deviation and shape of individual signal strength distributions. While global linear transformation may be able to adjust for large differences in dynamic range, it fails to adjust for local differences that are specific for a certain location and an AP. In order to reduce such differences in signal strength distributions across devices, we consider kernel density estimation (Eq. 4) in computing individual class-conditional probabilities.

Kernel density estimation takes the noisiness of individual

(dBm)

%	D901C	EEE900A	X61	N810(1)	N810(2)	N95
0	-92	-106	-93	-92	-110	-90
25	-86	-98	-87	-81	-83	-81
50	-81	-90	-81	-76	-77	-75
75	-72	-79	-69	-69	-70	-68
100	-25	-41	-29	-35	-39	-35

TABLE II
DYNAMIC RANGE OF THE TEST DEVICES (IN PERCENTILES)

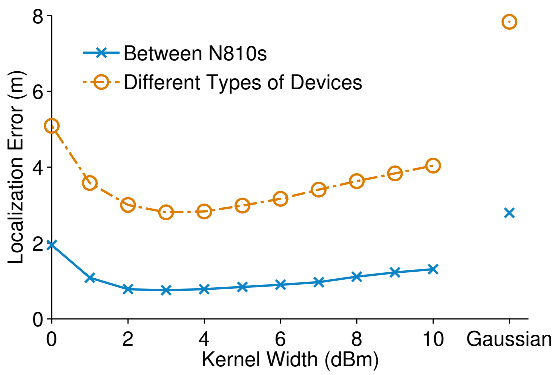


Fig. 5. Use of kernels significantly improves cross-device localization. Localization performance with varying kernel widths for the same type of devices (N810) and for the heterogeneous pairs of devices (all excluding N810 pairs) are shown. The results with Gaussian density estimation are also provided for comparison. As the kernel compensates for the difference between signal strength distributions across devices, localization accuracy improves significantly. Improvement is greater for different device types than for homogeneous devices. However, if too wide of a kernel is used, localization performance starts to degrade as RSS differences arising from true changes in location are masked.

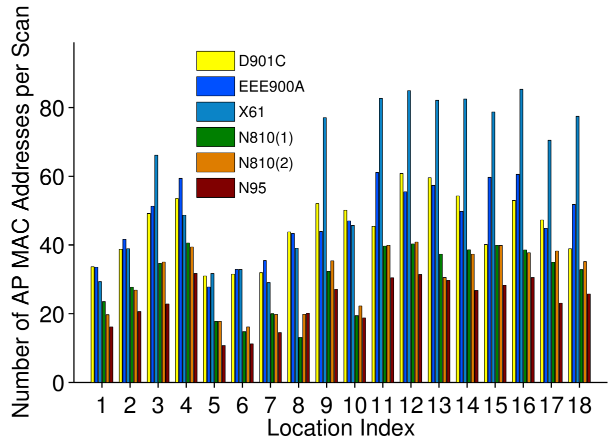


Fig. 6. Number of distinct AP MAC addresses observed per WiFi scan for each device. The number of APs differs by more than a factor of two. In general, devices with larger form factors (laptops) observe more APs than smaller devices (tablets and cellphones).

samples into account. Here, we use kernel density estimation to compensate for the difference between signal strength distributions across different devices. We evaluated a Gaussian kernel with widths varying from 1 dBm to 10 dBm. An example of kernel estimate with width 4 dBm is shown in Figure 4(d). Figure 3 also shows enhanced localization accuracy when kernel estimation with the same width is used, if `N810(2)` acts as a training device.

Figure 5 shows the effect of kernel width on cross-device localization, and compares kernel estimation to the histogram method and Gaussian density estimation. We show localization error between the same type of device (*i.e.* between `N810(1)` and `N810(2)`) and the error between different types of devices separately. A kernel width of 3 dBm provided the best localization performance with our dataset. Not surprisingly, the effect of kernel estimation is more significant for different device types, as their deviation was greater. Neither raw histogram estimation (kernel width $\rightarrow$ 0 dBm) nor Gaussian density estimation perform well, particularly for localization between different device types.

The standard deviation computed from the signal strength samples taken from one device for two minutes was approximately 2 dBm. Compared to this value, the best kernel width of 3 – 4 dBm for cross-device localization is somewhat higher than the smaller-scale variation of a specific device type. The reason is that, as each device shows a different dispersion and shape of its signal strength distribution, a strategy of doing more "smoothing" than that required for single device localization is more effective.

V. Feature Design of Localization Algorithms for Organic Indoor Localization

This section first analyzes another characteristic of the wireless scan signal — visibility, or *detection* of access points by diverse devices. The detection feature has been used exclusively, or augmented with the signal strength feature in various contexts of RF-based localization [9], [18], [20], [21]. We compare its use to the signal strength feature, and discuss feature design of localization algorithms for heterogeneous devices.

A. Analysis of Detection Rate for Diverse Devices

In any practical, large-scale localization system based on wireless networks, a mobile device captures only a subset of all access points "visible" at a given location because some access points are only intermittently detected by the mobile device. Factors may be both environmental, such as multipath fading, and transient, such as occlusion by humans or other objects. In addition, the OS or driver may allow only a limited time for collecting scan information, so APs may be missed if the beacon and driver are off-cycle.

A localization algorithm can use the probability of observing an individual access point, or detection rate, to give a different weight for the features it uses, *i.e.* signal strength likelihood. The rationale behind using AP presence/absence is that more frequently observed access points may be more informative for distinguishing locations. Alternatively, presence/absence can be used as the exclusive factor, providing coarser precision at a lower information cost, as no RSSI values are used; this may be beneficial when the (binary) signal maps for physically large areas are stored on low-memory mobile devices.

However, different devices not only detect WiFi signal strengths differently, as we saw in Section IV, but also differ in the sets of APs that they observe. One device may detect a nearby AP consistently, while another may not detect it at all (Figures 6 and 7). This occurs because of differences in frequency band (2.4 GHz and/or 5 GHz), radio/antenna sensitivity, firmware/driver implementations, and other factors.

Consequently, localization performance can degrade if the detection probability for each access point is used as a feature for localization. To illustrate this problem, we consider three

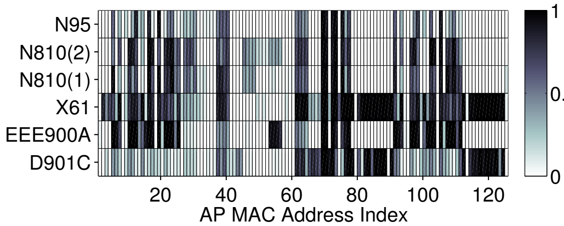


Fig. 7. Presence/absence pattern across different devices. While the primary difference is due to frequency (*i.e.* , `D901C` and `X61` operate at both 2.4 GHz and 5 GHz bands while others use only the 2.4 GHz band), there is still considerable difference across same-band devices. This difference reduces the efficacy of localizers that rely on AP presence/absence as a feature.

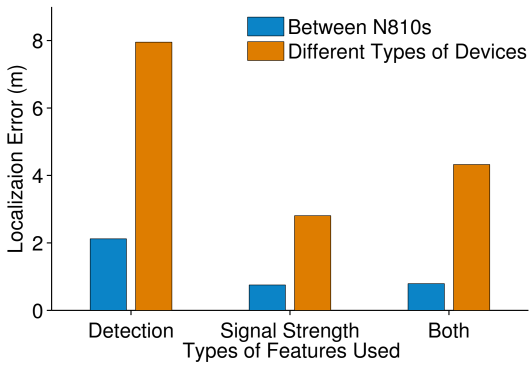


Fig. 8. Comparison of localization error using three feature types: AP Detection, Signal Strength, and Both. Because different devices observe partially disjoint sets of APs, it is better not to include AP presence/absence in cross-device localization. Degradation in localization error due to use of the detection rate feature is most prominent between different device types.

different types of Bayesian localization algorithms according to the degree of detection information used in the localization process (Section II-B): the localization rule in Equation 3 is least dependent on the presence/absence information, while Equation 6 uses the detection rate exclusively.

We compare the performance of each Bayesian localization feature choice in Figure 8. The algorithm that is exclusively detection-based exhibits the worst performance; the gap between the same type of device (N810) and the different types of devices is also the largest. The signal-strength-based algorithm performs best among the three, as it does not rely explicitly on presence/absence information. The localization algorithm that uses both signal strength and presence/absence information shows comparable localization accuracy for same-type-of-device localization, but accuracy is significantly degraded for cross-device localization.

B. Effect of Dissimilarity in AP Detection on Localization

In order to better understand the effect of AP detection on the performance of each localization algorithm, we consider Kullback-Leibler divergence (KLD) [22], an information-theoretic measure that captures the asymmetric dissimilarity between two probability distributions. As in Section II-B2, we model the presence/absence of each access point as a Bernoulli process. For access point i, $1 \leq i \leq k$, let J_i^S and J_i^T denote the binary random variables associated with

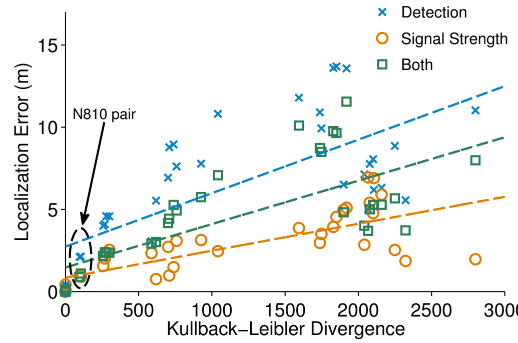


Fig. 9. Relation between KLD and localization error for three localization algorithms. Each point represents detection dissimilarity captured by KLD and localization error between a pair of devices. When two devices differ more in WAP detection, localization based on detection degrades more significantly. Cross-device localization using devices of the same type (`N810 pair`) does not show such degradation.

the Bernoulli processes for test device S and training device T respectively. Then, given location l, the Kullback-Leibler divergence of J_i^S over J_i^T for access point i is given by:

$$D_{KL}(J_i^S || J_i^T; l) = \sum_{x \in \{0,1\}} p^S_{J_i|L}(x|l) \log \frac{p^S_{J_i|L}(x|l)}{p^T_{J_i|L}(x|l)}. \quad (9)$$

where $p^S_{J_i|L}(x|l)$ (or $p^T_{J_i|L}(x|l)$) is the probability that access point i is detected by device S (or T) in location l for each WiFi measurement.

To compute the overall divergence of device S over device T, Equation 9 is summed over all k access points and all locations in the data:

$$D_{KL}(S||T) = \sum_{l \in L} \sum_{1 \leq i \leq k} \sum_{x \in \{0,1\}} p^S_{J_i|L}(x|l) \log \frac{p^S_{J_i|L}(x|l)}{p^T_{J_i|L}(x|l)}. \quad (10)$$

The Kullback-Leibler divergence of S over T can be considered a measure of how much extra information is required to encode the detection process for device S when using the detection process for device T. Thus, it naturally captures the "divergence" of testing device S when training data from device T is used.

We compared KLD from each training-test pair of devices with the location error between them. Figure 9 shows the correlation between KLD and the localization error for each localization algorithm. The effect of device difference captured by KLD is more correlated with detection-based and hybrid localization errors, than with errors observed using a signal-strength-based localization algorithm.

C. Feature Design of Localization Algorithms for Heterogeneous Devices

The previous two sections (§V-A,§V-B)) showed that AP visibility varies considerably across heterogeneous devices, and that dissimilarity in AP detection adversely affects cross-localization. In general, we found that using only the signal strength feature, without incorporating negative evidence, showed the best localization performance for heterogeneous devices.

As shown in Figure 8, the detection rate feature augmented in Equation 7 does not give much extra information for distinguishing locations over signal strength, even between same-type devices (the N810s). Similarly, the localization algorithm based solely on the signal strength feature is also affected by dissimilarity of AP detection (§ V-B). This is because the use of m-estimate smoothing applied to the class-conditional probability (§II-B1) for each AP implicitly encodes the detection information of that AP. For example, if there is no observation in the training data for a certain AP in a certain location, but the AP is detected during localization, the corresponding class-conditional probability is initialized as a uniform distribution, which encodes the least amount of information possible from that new observation. Consequently, the signal-strength-based localization algorithm will assign a minimal score to that location according to Equation 5, in which $N = 0$. However, as more readings are observed, the probability distribution will converge to the empirical maximum-likelihood estimate.

Therefore, the localization algorithm using only signal strengths as features is also weakly affected by dissimilarity in AP detection. Given location l, if the test device observes a new access point i which was not observed by the training device in the same location, Equation 3 assigns a minimal score determined from the m-estimator for class-conditional probability to location l for feature i. If the same access point i is present in another location l' instead, this may bias the localization decision to l' over l.

However, the effect of dissimilarity in AP detection is less significant than with algorithms that explicitly use detection rate, because negative evidence — failure to observe access point i — will not be directly incorporated into the localization score in Equation 3. In this sense, incorporation of evidence is asymmetric, and the effect of a mismatch in AP detection is less severe than in the algorithms that incorporate AP absence information directly (Equation 6 and 7).

Even for algorithms that do not explicitly use detection rate, we expect that the presence or absence of a certain access point will implicitly affect localization results. For example, many instance-based classification algorithms, such as k-nearest-neighbor or support vector machines, require choosing a value for each missing entry in each instance. A typical value used for WiFi localization is -100 dBm, encoding prior information that non-detected APs are expected to be far away, and that if they were detected it would be with low signal strength. However, this effect is symmetric in contrast to the signal-strength-based Bayesian algorithm presented in this paper.

VI. Related Work

Relatively few researchers have addressed the problem of using heterogeneous devices for localization. For GSM localization, Chen *et al.* tested cross-device localization using three different devices, showing that the heterogeneity of training and test devices considerably degrades the accuracy of their fingerprinting method [23]. Kaemarungsi compared RSSI values from different devices, but did not evaluate their effect on localization [24].

Researchers have proposed several methods for compensating for differences in signal strengths or RSSI values. Linear transformation from one device to another has been computed either manually or on-line using an expectation-maximization algorithm [8], [19], [25]. Dong *et al.* suggested using the difference between signal strengths across access points, rather than the absolute signal strength vector, as a localization feature [26]. While the difference between signal strength values is a major factor in localization using heterogeneous devices, we showed that the algorithm must be designed to compensate for the different shape and dispersion of signal strength values among devices.

Detection rate, or response rate, of access points has also been used for RF localization. Bargh and Groote used the inquiry response rate of Bluetooth devices for indoor localization, as signal strength for Bluetooth devices is not readily available without connection establishment [20]. In contrast, 802.11 devices can scan access points without establishing connections. For WiFi localization, Cheng *et al.* considered response rate as an alternative set of features for localization and showed that its performance is comparable to that of signal strength based localization [21]. Our results show that while it is possible to use response rate as a feature, doing so will not increase the accuracy of cross-device localization.

VII. Conclusion

This paper analyzed device diversity and its effect on localization. We reported simultaneous collection of data from six 802.11 devices in 18 indoor locations. While there is a clear linear correlation of signal strengths across devices, linear transformation alone is not enough for cross-localization: we find that local variations are on the same order of magnitude as the compensation provided by linear transformation. Instead, wide smoothing can accommodate the different shapes of signal strength distributions across devices, and proves effective for cross-localization. We also found that access point detection rates vary widely across client devices. As a result, incorporating access point presence and absence — in particular, relying solely on this factor to reduce storage costs and simplify positioning — provides poor localization performance when fingerprints are shared across different devices. To better understand this issue, we used Kullback-Leibler divergence to capture device differences with respect to AP detection, and showed that a correlation exists between detection similarity and localization accuracy.

Acknowledgment

We thank the Nokia Research Center Cambridge (US) for their financial support and technical engagement, and those in the MIT Stata Center who contributed to our data collection efforts and provided valuable feedback on early versions of our prototype.

REFERENCES

[1] B. Hofmann-Wellenhof, H. Lichtenegger, and J. Collins, *GPS Theory and Practice*. Springer, 1997.

[2] R. Want, V. Falcao, and J. Gibbons, "The Active Badge Location System," *ACM Transactions on Information Systems*, vol. 10, pp. 91–102, 1992.

[3] N. Priyantha, A. Chakraborty, and H. Balakrishnan, "The Cricket Location-Support System," in *Proc. Annual International Conference on Mobile Computing and Networking (MobiCom)*, Boston, MA, Aug. 2000, pp. 32–43.

[4] A. LaMarca, Y. Chawathe, S. Consolvo, J. Hightower, I. E. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, J. Tabert, P. Powledge, G. Borriello, and B. N. Schilit, "Place Lab: Device Positioning Using Radio Beacons in the Wild," in *Proc. International Conference on Pervasive Computing (Pervasive)*, Munich, Germany, May 2005, pp. 116–133.

[5] M. Azizyan, I. Constandache, and R. R. Choudhury, "SurroundSense: Mobile Phone Localization via Ambience Fingerprinting," in *Proc. Annual International Conference on Mobile Computing and Networking (MobiCom)*, Beijing, China, Sep. 2009, pp. 261–272.

[6] V. Bychkovsky, B. Hull, A. Miu, H. Balakrishnan, and S. Madden, "A Measurement Study of Vehicular Internet Access Using In-Situ 802.11 Networks," in *Proc. Annual International Conference on Mobile Computing and Networking (MobiCom)*, Los Angeles, CA, Sep. 2006.

[7] P. Bahl and V. N. Padmanabhan, "RADAR: An In-Building RF-Based User Location and Tracking System," in *Proc. IEEE International Conference on Computer Communications (INFOCOM)*, Tel Aviv, Israel, Mar. 2000, pp. 775–784.

[8] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach, and L. E. Kavraki, "Practical Robust Localization over Large-Scale 802.11 Wireless Networks," in *Proc. Annual International Conference on Mobile Computing and Networking (MobiCom)*, Philadelphia, PA, Sep. 2004, pp. 70–84.

[9] J. Krumm and E. Horvitz, "LOCADIO: Inferring Motion and Location from Wi-Fi Signal Strengths," in *Proc. International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous)*, Boston, MA, Aug. 2004, pp. 4–14.

[10] "Ekahau Positioning Engine," http://www.ekahau.com.

[11] J. Krumm and J. Platt, "Minimizing Calibration Effort for an Indoor 802.11 Device Location Measurement System," Microsoft Research, Tech. Rep. MSR-TR-03-82, Nov. 2003.

[12] E. S. Bhasker, S. W. Brown, and W. G. Griswold, "Employing User Feedback for Fast, Accurate, Low-Maintenance Geolocationing," in *Proc. IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Orlando, FL, Mar. 2004, pp. 111–120.

[13] P. Bolliger, "Redpin – Adaptive, Zero-Configuration Indoor Localization through User Collaboration," in *Proc. International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT)*, San Francisco, CA, Sep. 2008, pp. 55–60.

[14] A. Barry, B. Fischer, and M. Chang, "A Long-Duration Study of User-Trained 802.11 Localization," in *Proc. International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT)*, Orlando, FL, Sep. 2009, pp. 197–212.

[15] J. Park, B. Charrow, D. Curtis, J. Battat, E. Minkov, J. Hicks, S. Teller, and J. Ledlie, "Growing an Organic Indoor Location System," in *Proc. International Conference on Mobile Systems, Applications, and Services (MobiSys)*, San Francisco, CA, Jun. 2010, pp. 271–284.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classfication*, 2nd ed. Wiley-Interscience, 2000.

[17] J. Cussens, "Bayes and Pseudo-Bayes Estimates of Conditional Probabilities and Their Reliability," in *Proc. European Conference on Machine Learning (ECML)*, Vienna, Austria, Apr. 1993, pp. 136–152.

[18] P. Bolliger, K. Partridge, M. Chu, and M. Langheinrich, "Improving Location Fingerprinting through Motion Detection and Asynchronous Interval Labeling," in *Proc. International Symposium on Location and Context Awareness (LoCA)*, Tokyo, Japan, May 2009, pp. 37–51.

[19] A. W. Tsui, Y.-H. Chuang, and M. Hao-Hua Chu, "Unsupervised Learning for Solving RSS Hardware Variance Problem in WiFi Localization," *Mobile Networks and Applications*, vol. 14, no. 5, pp. 677–691, 2009.

[20] M. S. Bargh and R. de Groote, "Indoor Localization Based on Response Rate of Bluetooth Inquiries," in *Proc. International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT)*. San Francisco, CA: ACM, Sep. 2008, pp. 49–54.

[21] Y.-C. Cheng, Y. Chawathe, A. LaMarca, and J. Krumm, "Accuracy Characterization for Metropolitan-scale Wi-Fi Localization," in *Proc. International Conference on Mobile Systems, Applications, and Services (MobiSys)*, Seattle, WA, Jun. 2005, pp. 233–245.

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley and Sons, 2006.

[23] M. Y. Chen, T. Sohn, D. Chmelev, D. Haehnel, J. Hightower, J. Hughes, A. LaMarca, F. Potter, I. Smith, and A. Varshavsky, "Practical Metropolitan-Scale Positioning for GSM Phones," in *Proc. International Conference on Ubiquitous Computing (UbiComp)*, Orange County, CA, Sep. 2006, pp. 225–242.

[24] K. Kaemarungsi, "Distribution of WLAN Received Signal Strength Indication for Indoor Location Determination," in *Proc. International Symposium on Wireless Pervasive Computing (ISWPC)*, Phuket, Thailand, Jan. 2006.

[25] M. B. Kjærgaard, "Automatic Mitigation of Sensor Variations for Signal Strength Based Location Systems," in *Proc. International Workshop on Location and Context Awareness (LoCA)*, Dublin, Ireland, May 2006, pp. 30–47.

[26] F. Dong, Y. Chen, J. Liu, Q. Ning, and S. Piao, "A Calibration-Free Localization Solution for Handling Signal Strength Variance," in *Proc. International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments (MELT)*, Orlando, FL, Sep. 2009, pp. 79–90.