In Proceedings of the 3rd Annual IEEE International Conference on Pervasive Computing and Communications (IEEE PerCom 2005), Hawaii USA, March, 2005. Pages 95--104.

Reducing the Calibration Effort for Location Estimation Using Unlabeled Samples

Xiaoyong Chai and Qiang Yang
Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong, China
{carnamel, qyang}@cs.ust.hk

Abstract

WLAN location estimation based on 802.11 signal strength is becoming increasingly prevalent in today's pervasive computing applications. As alternative to the well-established deterministic approaches, probabilistic location determination techniques show good performance and become more and more popular. However, in order for these techniques to achieve a high level of accuracy, adequate training samples should be collected offline for calibration. As a result, a great amount of manual effort is incurred. In this paper, we aim to solve the problem by reducing both the sampling time and the number of locations sampled in constructing the radio map. A learning algorithm is proposed to build location estimation systems based on a small fraction of the calibration data traditional techniques require and a collection of user traces that can be cheaply obtained. Our experiments show that unlabeled user traces can be used to compensate the effects of reducing calibration effort and even improve the system performance. Consequently, manual effort can be significantly reduced while a high level of accuracy is still achieved.

1. Introduction

With the recent development in mobile computing devices and wireless techniques, location-aware systems are of a growing interest and are becoming increasingly popular as well as practical. In building such systems, a fundamental issue is to know the locations of mobile devices in a wireless environment, where an important goal is to increase the accuracy of location estimation. In the indoor settings, radio frequency (RF) based techniques are particularly effective among the existing solutions because they provide ubiquitous coverage and use the inexpensive wireless LAN (WLAN) as the fundamental infrastructure. In recent years, a variety of systems have emerged [1] [2] [6] [9] [13] [16].

Most RF-based systems estimate locations by measuring the strength of the signals propagated from the access points (AP's) in the environment. They usually work in two phases [16]: an *offline* training phase and an *online* location estimation phase. In the offline phase, a so-called *radio map* is built. In the online phase, the strength of received signals is used to lookup the radio map to estimate the location. A radio map is a table of signal strength values received at selected locations from the AP's in the area of interest. RF signals provide rich information on locations since the signal strength varies noticeably with the distance between the AP's and the physical locations where a wireless device is located. Location estimation is a challenging problem because of the non-trivial ways in which signals propagate. As a result, a large number of samples should be collected offline for calibration in order to make the radio map robust to the noisy signals. To obtain the signals, a calibration process is very labor intensive. Let N_s denote the sampling time at each location and N_l the number of selected locations. The amount of calibration effort can thus be quantitatively measured as $N_s \times N_l$. Suppose in a small environment with 100 locations ($N_l = 100$) and 100 samples are collected at each location, one sample per second ($N_s = 100$). Typically several hours should be spent to collect such an amount of calibration data, let alone the tedious labelling process. The problem is more serious when the area of concern, such as a shopping mall, is very large and where spatially high-density calibration is needed. In this paper, we focus on how to significantly reduce the offline calibration effort while still achieving high accuracy in location estimation through machine learning techniques.

One way to reduce the manual effort is through reducing both N_s and N_l. However, simply reducing N_s and N_l brings a side effect of lowering the accuracy. Experiments show that 26% of accuracy is lost when N_s and N_l

are both reduced by two-thirds. To make up for the loss of accuracy, in this paper, we propose a novel EM-based algorithm that makes use of user traces. While calibration data can be viewed as labeled samples since the true positions from which these samples are taken are known (labeled), user traces are sequences of signal strength recording user's movement in the environment. These are unlabeled samples because the signal strength received during the movement is recorded without any position label. The most attractive property of user traces is that without the labelling process, sequences of samples can be collected easily and inexpensively. Using a Hidden Markov Model to model user traces, our method provide a way to build probabilistic estimation systems that require only a small fraction of the calibration data. Trained from a limited number of labeled samples, the system can gradually improve its performance when more and more user traces are obtained. Experiments show that when all the calibration data are used, an accuracy of 85% with three meters is obtained using a Bayesian estimation method. By using 60 unlabeled traces, the same accuracy can be achieved which only requires as 1/6 calibration data as before. Moreover, by using 100 traces, the accuracy reaches 86% when only 1/9 calibration data is needed. Therefore, the manual effort can be significantly reduced while even higher accuracy can be achieved.

Our contributions are as follows. First, we empirically study the influence of reducing the calibration effort on the accuracy of location estimation. Both the factors N_s and N_l are considered. Second, we propose a learning algorithm that makes use of the unlabeled trace data to supplement a limited number of labeled samples. The resulting system can be initialized from a limited number of sampled data and gradually improve its performance by using more and more unlabeled traces. As a consequence, much offline manual effort can be significantly reduced. Finally, we evaluate our algorithm by conducting experiments in a real-world indoor environment.

2. Location estimation based on 802.11 signal strength

2.1. Overview of previous work

In general, location estimation can be classified into two categories: *deterministic* techniques and *probabilistic* techniques. Deterministic techniques [1][2][14] use deterministic inference methods to estimate a user's location, such as Triangulation and K-nearest neighbor averaging (KNN). The RADAR system [1][2], one of the pioneering and most comprehensive work using signal strength measurements, is based on KNN to infer a user's location. It maintains a radio map with which each online signal strength measurement is then compared. The coordinates of the best K location matches are averaged to give an estimation.

Probabilistic techniques [6][9][15][16] form the second category. They are also called distribution-based techniques since they store the signal strength distributions from the AP's as the information for the radio map. In contrast to the first category, in the second category, probabilistic inference methods are used to estimate a user's location. In [16], locations in the area are pre-clustered into groups so as to reduce the computational cost of searching the radio map. In [15], correlation among consecutive samples from the AP's is introduced to enhance the system performance. Furthermore, in [9] and [6], spatial and motion constraints are utilized in a postprocessing step to refine the estimation. The core to all these techniques is the use of Bayesian inference to compute the posterior probabilities over locations.

However, compared with the various techniques on location estimation, in previous literature little attention has been paid to the issue of reducing calibration effort. To the best of our knowledge, [8] is among the only work that explicitly considers minimizing calibration effort for indoor 802.11 location estimation system. In their work, they observed that it is unnecessary to spend much time at each location. Formulating the problem as one of interpolation, they showed that a significant fraction of calibration locations can also be skipped. In our work, we not only consider how to significantly reduce the manual effort but also consider how to use information extracted from user traces to supplement the reduced amount of calibration data. The similar idea of using unlabeled trace data to improve localization accuracy was also explored in [3]. By assuming piecewise linear Gaussian distribution over locations, they employed a version of Monte Carlo localization algorithm for tracking people. Unlabeled traces are used to tune a motion model so as to adapt it to individual persons, exploiting regularities when a person navigates the environment. However, directly refining the radio map was not considered in their work.

2.2. Noisy characteristics of wireless channel

The IEEE 802.11b standard works over the radio frequencies in the 2.4 GHz band. The standard is widespread since the band is license-free at most places around the world. It is also attractive because the RF-based techniques are popular and inexpensive, providing much ubiquitous coverage and requiring little overhead.

A WLAN and a wireless device held by a user have different functionality: AP's in the WLAN broadcast signals and the wireless device acts as a sensor which senses the location by analyzing the signals received. Although signal strength varies noticeably with the distance between AP's and the wireless device, accurate location estimation us-

ing measurements of signal strength is still a difficult task due to the *noisy characteristics* of signal propagation. Subject to reflection, refraction, diffraction and absorption by structures and even human bodies, signal propagation suffers from severe multi-path fading effects in an indoor environment [7]. As a result, a transmitted signal can reach the device through different paths, each having its own amplitude and phase. These different components combine and reproduce a distorted version of the original signal. Moreover, even changes in the environmental conditions, such as temperature or humidity, also affect the signals to a large extent.

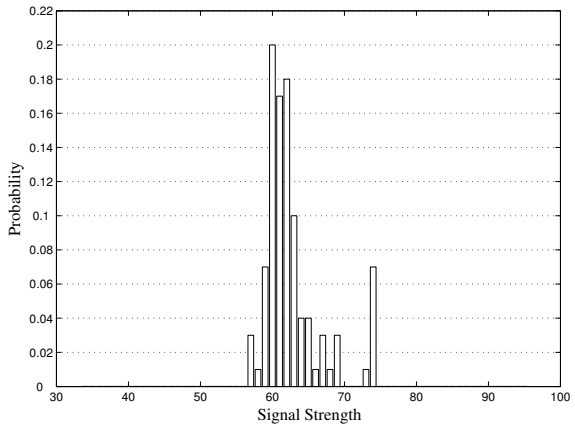

Figure 1. Signal strength distribution at a fixed location

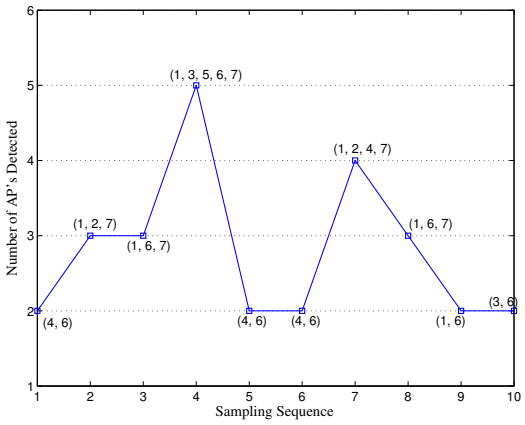

Figure 2. Variation of AP coverage over a fixed location

Figure 1 gives a typical example of a normalized histogram of signal strength received from an AP at a fixed location. Several hundreds of measurements were sampled to construct the histogram. It is clear from the figure that even at a fixed location, the signal strength received from the same AP varies with time. Furthermore, the number of AP's covering a location also varies with time. As shown in Figure 2, not only the number of AP's changes over time, the group of AP's detectable at the location also changes as well, as indicated by the numbers beside each point. For example, the fourth sample in Figure 2 contains signals from AP's 1, 3, 5, 6 and 7, while the fifth sample contains signals from AP's 4 and 6.

2.3. Probability-based location estimation

Since our work lies in the category of probabilistic techniques, in this section we introduce the Bayesian framework of location estimation methods. In general, an estimation is represented as a probability distribution over all the locations in the area of interest. The Bayesian inference method is used to compute a distribution conditioning on the observed signal strength. Finally, the estimated location is the one with the maximum probability in the resulting distribution.

Formally, we model the physical area of interest as a finite location-state space $\mathbb{L} = \{l_1, \ldots, l_n\}$. The state space $\mathbb{L}$ is defined as a set of physical locations with x- and y-coordinates:

$$\mathbb{L} = \{l_1 = (x_1, y_1), \ldots, l_n = (x_n, y_n)\}.$$

Each location l represents a grid cell on the hallways in the environment.

All possible signal strength values are modelled as a finite observation space $\mathbb{O} = \{o_1, \ldots, o_m\}$. An observation o in the observation space $\mathbb{O}$ consists of a set of signal strength measurements received from k access points. k is the number of AP's which have the most strongest signals. Normally, in an environment, signals from many AP's are detectable somewhere, either located within the area of concern, such as a hallway, or located outside. A subset of k AP's is selected so as to reduce the computational cost. Thus, each observation o is represented as a vector of k pairs as follows:

$$o =< (b_1, ss_1), \ldots, (b_k, ss_k) >$$

where b_i represents the ith AP scanned and ss_i is the signal strength received from b_i. Since signals are noisy and a single scan may probably miss some AP's, we take the signal strength measurement ss_i average over every second.

In the offline training phase, calibration data are collected at each location l_i. That is, signal strength measurements are recorded at each location as observations. After the data are collected, we build a histogram of observation for each AP b_j at each location l_i. This is done by constructing the conditional probability $Pr(ss_j|b_j, l_i)$, which is the

probability that AP b_j has the signal strength measurement ss_j at location l_i. By making an independence assumption among signals from different AP's, we multiply all these probabilities to obtain the conditional probability of receiving a particular observation o at location l_i as follows:

$$Pr(o|l_i) = \prod_{j=1}^{k} Pr(ss_j|b_j, l_i),$$

which is exactly the content of a radio map introduced before. In the online phase, a posterior distribution over all the locations is computed using Bayes rule:

$$Pr(l_i|o^*) = \frac{Pr(o^*|l_i)Pr(l_i)}{\sum_{i=1}^{n} Pr(o^*|l_i)Pr(l_i)}$$

where o^* is a new observation obtained. $Pr(l_i)$ encodes the prior knowledge about where a user may probably be. $Pr(l_i)$ can be set as the uniform distribution, assuming every position is equally likely. The estimated location l^* is the one which obtains the maximum value of the posterior probability: $l^* = \arg\max Pr(l_i|o^*)$.

3. Reducing offline calibration effort

As discussed before, the amount of calibration effort is determined by two factors: N_s and N_l. N_s is the sampling time spent at each location to collect signal samples and N_l is the number of locations to sample from. Therefore, we consider reducing offline calibration effort by two methods: reducing N_s (M_1) and reducing N_l (M_2). An illustration is shown in Table 1.

	Reduce N_l	
Reduce N_s	No	Yes
No	—	M_2
Yes	M_1	$M_1 + M_2$

Table 1. Different methods to reduce calibration effort

3.1. M_1 : Reducing the sampling time at each location

One method to reduce calibration effort is to reduce the sampling time N_s. We call this method M_1. Although it is not necessary to spend much time at each location during calibration, it normally requires tens or even hundreds of samples to stabilize signal strength distributions and reduce the influence of noisy wireless channel. When the calibration data are scarce, only five or ten samples are available at each location, the limited samples are not representative enough, because the resulting distributions are easily biased. As a consequence, when a measurement is obtained online, it can be easily rejected as an outlier only because it does not appear in the training data. Although there are techniques to smooth the distributions [11], they are still far from satisfactory when the training data are insufficient. Our experiments reveal that only 3% of accuracy is lost when N_s is reduced from 60 to 30 at each location, which is a good tradeoff since half of the effort can be saved. However, accuracy decreases by 12% when N_s is further reduced to 10. Therefore, reducing N_s has its limitation in achieving significant reduction in manual effort.

3.2. M_2 : Reducing the number of locations sampled

The other method to reduce the calibration effort is to reduce the number of locations N_l on which we collect samples offline. We call this method M_2. This is done as follows. Instead of sampling at each location in $\mathbb{L}$, we collect samples at a subset of locations $\mathbb{L}_1 \in \mathbb{L}$ and skip the rest of locations $\mathbb{L}_2$ ($\mathbb{L}_2 = \mathbb{L} - \mathbb{L}_1$). Signal strength distributions at locations in $\mathbb{L}_1$ can be built in the same way as given in Section 2.3. However, the resulting radio map is incomplete since the distributions at those skipped locations are missing. To solve the problem, we use an interpolation method to make up for the missing distributions. The idea is to interpolate the missing distributions from the available ones. An illustration is shown in Figure 3, where l_a, l_b ($\in \mathbb{L}_1$) are the locations directly sampled and l_c ($\in \mathbb{L}_2$) is one of the locations skipped between l_a and l_b. d_1 and d_2 are the distances from l_c to l_a and l_c to l_b, respectively. We interpolate the signal strength distribution at l_c from those at l_a and l_b as follows (Equation (1)):

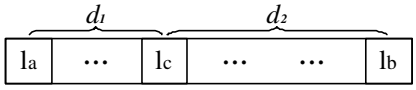

Figure 3. An illustration of interpolation, where $Pr(o|l_c)$ is interpolated from $Pr(o|l_a)$ and $Pr(o|l_b)$

$$Pr(o_j|l_c) = \frac{d_2}{d} Pr(o_j|l_a) + \frac{d_1}{d} Pr(o_j|l_b), \qquad o_j \in \mathbb{O} \tag{1}$$

where $d = d_1 + d_2$. The idea is that the similarity between $Pr(o_j|l_c)$ and $Pr(o_j|l_a)$, the signal strength distribution at the interpolated location l_c and that at the sampled location l_a, depends on the distance between l_c and l_a. This is intuitive since the closer two locations are, the more similar the signals received at them. In Equation (1), such dependence

is assumed to be linear and the coefficients $\frac{d_1}{d}$ and $\frac{d_2}{d}$ are used to normalize the resulting distribution. More complex nonlinear relations can also be assumed, although we do not investigate them in this work.

Using interpolation, we can obtain a complete radio map which only requires as $\frac{||\mathbb{L}_1||}{||\mathbb{L}||}$ amount of the manual effort as before. However, since only a fraction of locations are directly sampled, M_2's performance is inferior to that of the radio map built with all the locations sampled. As will be shown in Section 5.3, reducing two-thirds of N_l incurs a loss of 16% in accuracy.

Manual effort can be significantly reduced if we use both methods ($M_1 + M_2$) simultaneously. However, such reduction is obtained at the loss of high accuracy. Therefore, we look for other sources to supplement the reduced calibration data and this is achieved by utilizing unlabeled user traces.

4. M^*: Using unlabeled traces to improve the radio map

User traces are sequences of signal strength measurements recording user's movement in the environment. The main difference between calibration data and user traces lies in whether the true position where an observation is taken is known or not. Each sample of the calibration data has its location label, and therefore it is recorded as a pair (o, l), where l is the location at which o is taken. On the other hand, a user trace has no location label assigned when recorded. It appears as a sequence of observed samples $< o^1, o^2, \ldots, o^m >$ and therefore cannot be used directly for training as calibration data. While labelling signal samples with the correct locations is time-consuming, collecting them is relatively easy. This is especially true when samples are collected consecutively when a user is walking around in the environment. Therefore, an interesting question is how to extract useful information contained in user traces to improve a radio map built from a limited amount of calibration data. In this paper, we propose a method in which we use a Hidden Markov Model to model user traces and apply EM algorithm to improve the radio map. We call this method M^*.

4.1. Modelling user traces using Hidden Markov Model (HMM)

We use an HMM to model user traces. HMM is a well-known technique in pattern recognition and has a wide range of applications [4][12]. In pervasive computing, HMM and its variation have been successfully used in tracking and recognizing human activities [10]. An HMM is a stochastic finite state machine which models a Markov process with parameters. It is termed "hidden" since the internal states of the process are viewed as hidden and only the outputs of the states are observable. In modelling user traces, the underlying process is a user's sequential changes in location, where the user's locations are the hidden internal states and the signal strength measurements are the observations.

For our purpose, we define an HMM on the location-state space $\mathbb{L}$ and the observation space $\mathbb{O}$, both of which are given in Section 2.3. The HMM consists of a radio map $\lambda = \{Pr(o_j|l_i)\}$, a location-state transition matrix $A = \{Pr(l_j|l_i)\}$, and an initial state distribution $\pi = \{Pr(l_i)\}$. Both λ and π are also given in Section 2.3. The radio map λ is a set of conditional probabilities which give the likelihood of obtaining signal strength measurement $o \in \mathbb{O}$ at the location $l \in \mathbb{L}$. The transition matrix A indicates how a user travels through the state space. While a user can freely navigate the environment, his movement subjects to certain constraints imposed by the environment. For example, he can only walk on hallways but cannot walk across rooms. Also, the user has limited mobility. That is, he does not move too quickly in an indoor environment, only moving to the locations nearby at consecutive time steps. All this prior information can be encoded into A by setting proper $Pr(l_j|l_i)$. Given an observation sequence, which is an unlabeled trace of signal strength measurements, the well-known Viterbi algorithm [12] can be used to infer the most probable hidden state sequence in HMM, which is a sequence of user's location changes.

4.2. Improve the radio map λ using EM algorithm

When the calibration data are insufficient, a radio map built from a small number of labeled samples is easily biased. To reduce the calibration effort and still achieve good performance, we apply the EM algorithm [5] to improve the radio map using unlabeled traces. Let λ^0 denote the initial radio map which is built from a limited amount of labeled calibration data. In the case that interpolation is used, λ^0 is the interpolated radio map resulted. Let A^0 and π^0 be the initial location-state transition matrix and the initial state distribution, both of which are set a priori. Then, an HMM can be initialized by the model parameter $\theta^0 = (\lambda^0, A^0, \pi^0)$. Given a set of unlabeled traces T, EM is used to adjust the model parameter $\theta = (\lambda, A, \pi)$ iteratively to find θ^* such that the likelihood $Pr(T|\theta^*)$ is maximized. Here, $Pr(T|\theta)$ is calculated as follows:

$$\begin{aligned} Pr(T|\theta) &= \prod_{t \in T} Pr(t|\theta) = \prod_{t \in T} \sum_{q} Pr(t|q,\theta) Pr(q|\theta) \\ &= \prod_{t \in T} \sum_{q} \Big(Pr(l^1) Pr(o^1|l^1) \times \\ &\qquad \prod_{k=2}^{n_t} Pr(l^k|l^{k-1}) Pr(o^k|l^k) \Big) \end{aligned} \quad (2)$$

In Equation (2), $t = (o^1, o^2, \ldots, o^{n_t})$ is a trace of length n_t and $q = (l^1, l^2, \ldots, l^{n_t})$ is one possible location sequence of the same length as t. $Pr(T|\theta)$ represents the likelihood with which we obtain the traces of signal strength measurements in T given the model parameter θ. θ^* maximizing the likelihood $Pr(T|\theta^*)$ means that the parameter θ^* best explain the signal sequences in the traces. Therefore, starting from an initially biased radio map λ^0, λ^0 is adjusted to best explain the set of unlabeled traces. Meanwhile, the traces are implicitly used to improve the radio map and the useful information in T is thus extracted.

The EM algorithm is an iterative process through two steps: an Expectation step (E-step) and a Maximization step (M-step). During the iterations, a sequence of model parameters $\theta^0, \theta^1 \ldots \theta^*$ is generated, where θ^0 is the initial parameter and θ^* is the converged parameter obtained when the iteration terminates. The standard method maximizes the so-called Q-function defined as follows:

$$Q(\theta, \theta^k) = \sum_{t \in T} \sum_{q} \log Pr(t, q|\theta) Pr(t, q|\theta^k), \quad (3)$$

where θ^k is the parameter obtained after the kth iteration. In the E-steps, Q-function (3) is calculated, and in the M-steps, maximization is taken over θ:

$$\theta^{k+1} = \arg\max_{\theta} Q(\theta, \theta^k).$$

In particular, the M-step in the $(k+1)$th iteration for the radio map $\lambda^{k+1} = \{Pr(o_j|l_i)^{k+1}\}$ is as follows:

$$Pr(o_j|l_i)^{(k+1)} = \frac{\sum_{t \in T} \sum_{s=1}^{t_n} Pr(t, l^s = i|\theta^k) \delta(o^s, o_j)}{\sum_{t \in T} \sum_{s=1}^{t_n} Pr(t, l^s = i|\theta^k)}, \quad (4)$$

where $\delta(x, y)$ is a function such that $\delta(x, y) = 1$ if $x = y$, otherwise $\delta(x, y) = 0$. The EM algorithm guarantees that $Pr(T|\theta^{k+1}) \geq Pr(T|\theta^k)$ and the parameter will converge to θ^* when the likelihood does not change in consecutive iterations. Interested readers please refer to [5]. When a new radio map λ^* is learned, it can be used to substitute the initial map λ for location estimation. To avoid the bias towards unlabeled traces, we take an additional step. We use λ^* to label the traces and thus obtain a new set of labeled samples. This new set of samples, together with the original calibration data, produces a modified radio map λ', which is used in the online phase.

5. Experimental results

In this section, we evaluate the performance of our proposed algorithm. First, we empirically study the effects of reducing the sampling time N_s and the number of sampled locations N_l on accuracy. Then, we present the results on using unlabeled traces to improve the system performance.

5.1. Experimental setup

Our experimental testbed was set up in the office area of CS Department in the Academic Building of Hong Kong University of Science and Technology. The building is equipped with an IEEE 802.11b wireless network in the 2.4 GHz frequency bandwidth. The layout of the floor is shown in Figure 4. This area has a dimension of 64 meters by 50 meters. Experiments were carried out in the four hallways (HW1$\sim$HW4) and two rooms (Room1 and Room2) as labeled in the figure. The four hallways measure 19.5, 37.5, 46 and 21 in meters, respectively. To form the location-state space, the environment was modelled as a space of 99 locations, each representing a 1.5-by-1.5 meter grid cell. Using the device driver and API we developed, we carried an IBM laptop with a standard wireless Ethernet card to collect calibration data and record user traces. For the calibration data, one hundred samples were collected at each location, one sample per second. Finally, traces were recorded when a user navigated the environment, walking through the hallways.

5.2. Accuracy v.s. sampling time

Experiments were first carried out to study the effect of varying the length of sampling time on accuracy (M_1). We simulated the effect of reducing sampling time by only using the first N_s collected samples for training. That is, a radio map was built using N_s samples at each location. Evaluation was done by testing on the rest of samples. Starting from five samples per location, we increased the number by five at each time to simulate the effect of gradually increasing the calibration effort. The results of the estimation accuracy with the number of training samples ranging from 5 to 60 are shown in Figure 5.

When the training samples are scarce, increasing the amount of calibration data has a great influence on accuracy. As shown in Figure 5, the accuracy increases by 10.2% as the number of training samples increases from five to ten. Enhancement is less significant when more training samples are available. Overall, the discrepancy can be as large as 22.3%, ranging from 62.8% ($N_s = 5$) to 85.1% ($N_s = 60$). As we can see, reducing the sampling time can significantly degrade the system performance.

5.3. Accuracy v.s. number of locations sampled

Another set of experiments was performed to examine the effect of interpolating a radio map (M_2). For this purpose, we simulated the effect of reducing the number of sampled locations N_l as follows. Out of all the 99 locations in modelling the test environment, we selected 31 locations

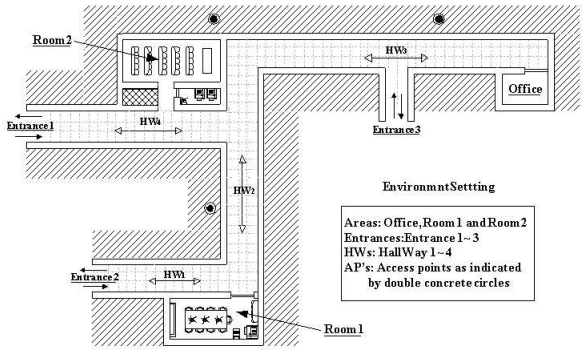

Figure 4. The layout of the office area of CS Department of Hong Kong University of Science and Technology

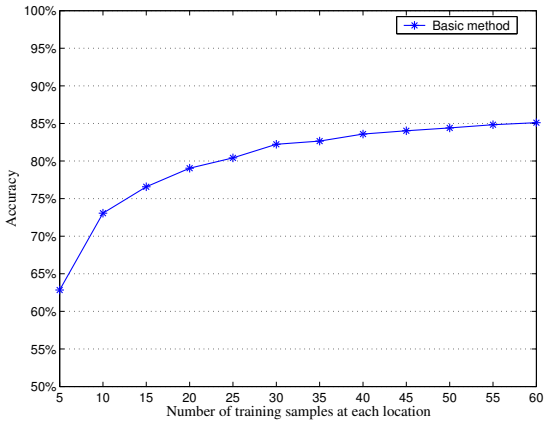

Figure 5. Accuracy v.s. number of training samples at each location (M_1)

by skipping every several locations between them, as illustrated in Figure 6. The 31 locations form the subset $\mathbb{L}_1$ and the other 68 skipped locations form $\mathbb{L}_2$. Calibration data at the locations in $\mathbb{L}_1$ were still used to construct the signal strength distributions while the data at the locations in $\mathbb{L}_2$ were no longer used for training but only for testing. The distributions at those locations in $\mathbb{L}_2$ were built using the interpolation method. After a radio map was interpolated, we then measured how the location estimation accuracy is affected at both the locations sampled and the locations interpolated.

Figure 7 shows the effect of reducing the number of locations sampled. For illustration, the factor of reducing the sampling time N_s was also considered. For a fixed number of training samples, for example $N_s = 20$, three measurements were taken. The first one is the *sampled accuracy* at the locations in $\mathbb{L}_1$, whose signal strength distributions were built directly from the calibration data ($N_s = 20$). The second measurement is the *interpolation accuracy* at the locations in $\mathbb{L}_2$, whose distributions were interpolated from the sampled locations with the calibration data $N_s = 20$. The last one is the *overall accuracy* of the resulting interpolated radio map, which is obtained by taking average over all the locations in $\mathbb{L}$. As we can see from the three curves, both the sampled accuracy on $\mathbb{L}_1$ and the interpolated accuracy on $\mathbb{L}_2$ increase as more calibration data are available. This is intuitive since as more training samples are obtained at the sampled locations, the sampled signal strength distributions and subsequently the interpolated distributions are less biased. However, since the interpolated accuracy is about 20% lower than the sampled accuracy, the overall accuracy is lower than that shown in Figure 5.

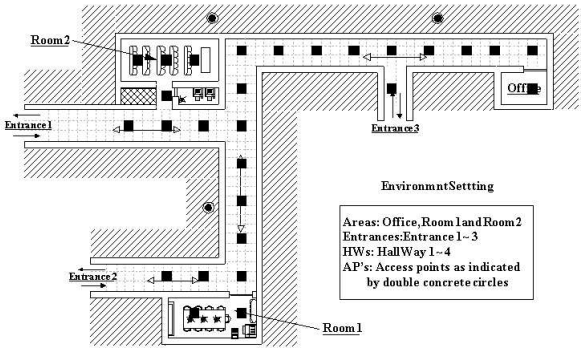

Figure 6. Layout illustration of reducing the number of locations sampled. Dark dots are the sampled locations.

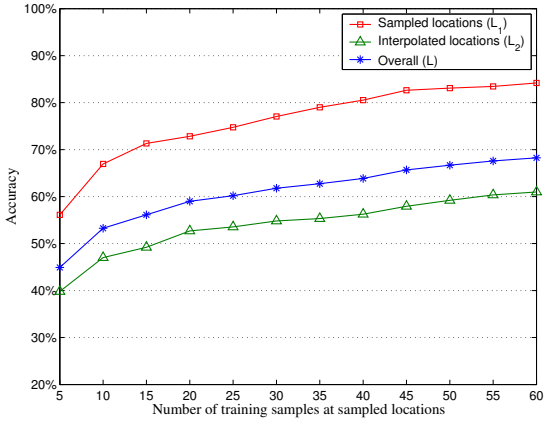

Figure 7. Performance at sampled and interpolated locations with varying sampling time ($M_1 + M_2$)

It is also interesting to compare the two methods, M_1 and M_2, in terms of their effects on accuracy. From Fig-

ures 5 and 7, we can see that reducing the sampling time is more effective than reducing the number of location sampled. The accuracy decreases by 6% when the sampling time is reduced by 2/3 ($N_s = 20$), while the accuracy decreased by 16% when roughly 2/3 locations are skipped ($N_l = 31$).

5.4. Using unlabeled traces to improve the performance

To measure the performance of using unlabeled traces, we first initialized radio maps using both the methods M_1 and M_2, and then used the method M^* to improve them. To investigate the utility of unlabeled traces, we also varied the number of traces.

Figure 8 shows the improvement in accuracy using unlabeled traces, where $N_l = 99$ and the amount of calibration data is fixed at $N_s = 20$. When no learning is performed – the number of traces used is zero, the accuracy is about 79%. The accuracy goes up as the number of traces increases. Improvement is about 4% when 20 traces are used and 9% by using 100 traces. At this point, the radio map tends to be stabilized as the influence of using more traces is lessened.

Figure 9 shows the effect of using unlabeled traces to reduce the sampling time. The dashed curve is the same one as in Figure 5. It is denoted as "Basic (0 Trs)" since only the calibration data are used. The other three curves show the performance of improved radio maps learned by the EM algorithm using 20, 60 and 100 traces. The improvement is significant when the calibration data are extremely scarce. At the point where $N_s = 5$, an increase of 12.8% is achieved using 20 traces and 23.8% using 100 traces. It shows that by using unlabeled traces, we can progressively reduce the sampling time and a high level of accuracy can still be achieved.

Experiments were also conducted to evaluate the learning algorithm when both N_s and N_l are reduced. The results are shown in Figure 10. The overall accuracy on $\mathbb{L}$ from Figure 7 is shown as the dashed curve for comparison. As we can see, the improvement is significant. When $N_s = 5$, we achieve an increase of about 17.2% using 20 traces and about 33.2% using 100 traces. To be more illustrative, the improvement at both the sampled locations in $\mathbb{L}_1$ and the interpolated locations in $\mathbb{L}_2$ is shown separately in Figures 11 and 12. Unlabeled traces are particularly effective in adjusting the distributions at the interpolated locations.

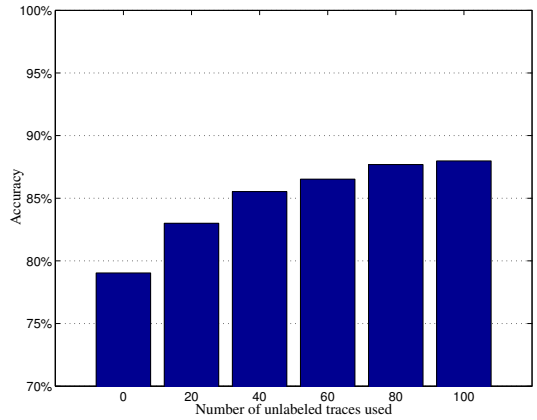

Figure 8. Improvement achieved through using a increasing number of traces (M^*)

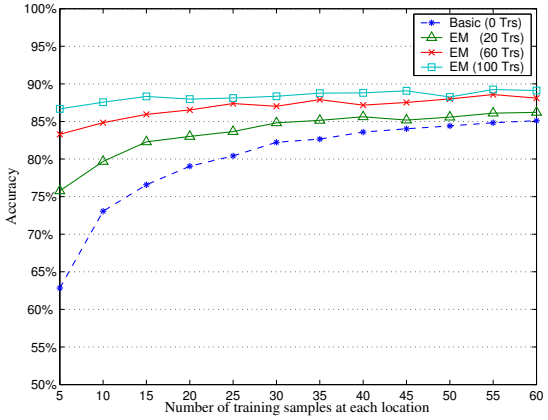

Figure 9. Effect of varying the number of traces on reducing the sampling times ($M_1 + M^*$)

6. Conclusion and Future Work

In this work, we empirically study the effect of reducing the calibration effort on estimation accuracy by reducing both the sampling time and the number of locations sampled. A learning algorithm is proposed to use unlabeled traces to improve system performance. Experiments show that unlabeled traces can be used to compensate the effects of reducing calibration data. As a result, manual effort can be significantly reduced while high accuracy can still be achieved.

In the future, we plan to extend our work in several directions. First, we will examine the effects of varying the amount of interpolation on accuracy. It is interesting to see how the accuracy will change when more locations are interpolated from less sampled locations. Second, we plan to take complex environment dynamics into consideration. For example, in building a location sensing system, the radio map of daytime can be much different from that of the

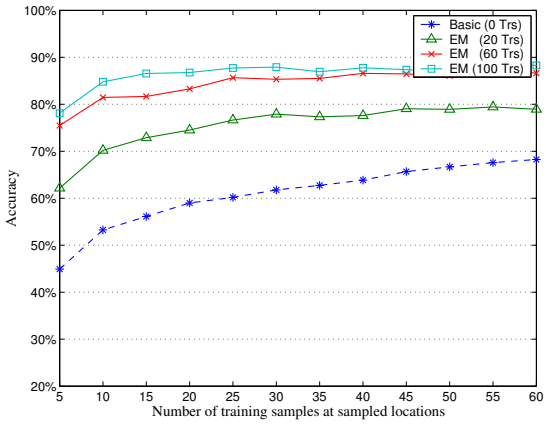

Figure 10. Accuracy improvement over all the locations in $\mathbb{L}$ using $M_1 + M_2 + M^*$

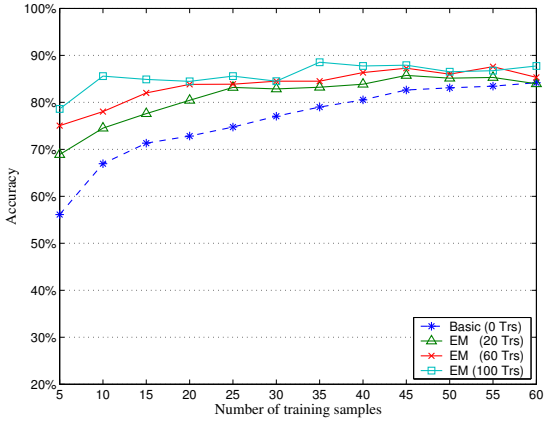

Figure 11. Accuracy improvement over sampled locations in $\mathbb{L}_1$ using $M_1 + M_2 + M^*$

nighttime. Instead of building radio maps for different periods of time, we are investigating methods to find a mapping between them and thus eliminate the need for tedious and repeated manual effort to update the radio maps.

7 Acknowledgments

This work is supported by a grant from Hong Kong Research Grant Committee RGC HKUST 6180/02E.

References

[1] P. Bahl, A. Balachandran, and V. Padmanabhan. Enhancements to the RADAR user location and tracking system. Technical report, Microsoft Research, February 2000.

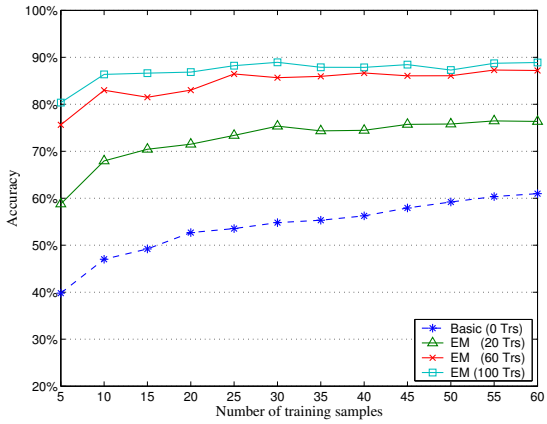

Figure 12. Accuracy improvement over interpolated locations in $\mathbb{L}_2$ using $M_1 + M_2 + M^*$

[2] P. Bahl and V. N. Padmanabhan. RADAR: An in-building RF-based user location and tracking system. In *Proceedings of IEEE INFOCOM2000*, pages 775–784, 2000.

[3] M. Berna, B. Lisien, B. Sellner, G. Gordon, F. Pfenning, and S. Thrun. A learning algorithm for localizing people based on wireless signal strength that uses labeled and unlabeled data. In *IJCAI'03*, Acapulco, Mexico, August 2003.

[4] H. Bunke and T. Caelli. *Hidden Markov Models - Applications in Computer Vision*. World Scientific Publishing Co, 2001.

[5] A. P. Dempster, N. M. Laird, and D. B.Rubin. Maximum likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38, 1977.

[6] C. Gentile and L. K. Berndt. Robust location using system dynamics and motion constraints. In *IEEE Conference on Communications*, June 2004.

[7] H. Hashemi. The indoor radio propagation channel. In *Proceedings of the IEEE*, volume 81, pages 943–968, 1993.

[8] J. Krumm and J. C. Platt. Minimizing calibration effort for an indoor 802.11 device location measurement system. Technical report, Microsoft Research, 2003.

[9] A. Ladd, K. Bekris, G. Marceau, A. Rudys, L. Kavraki, and D. Wallach. Robotics-based location sensing using wireless ethernet. In *Proceedings of MOBICOM2002*, Atlanta, Georgia, USA, September 2002.

[10] S. Luhr, H. H. Bui, S. Venkatesh, and G. A. West. Recognition of human activity through hierarchical stochastic learning. In *First IEEE International Conference on Pervasive Computing and Communications*, March 2003.

[11] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *Proc. of AAAI/ICML-98 Workshop on Learning for Text Categorization*, pages 41–48, 1998.

[12] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *Proc. IEEE*, volume 77, pages 257–286, 1989.

[13] T. Roos, P. Myllymaki, H. Tirri, P. Misikangas, and J. Sievanen. A probabilistic approach to WLAN user location esti-

mation. *International Journal of Wireless Information Networks*, 9(3):155–164, July 2002.

[14] A. Smailagic, D. P. Siewiorek, J. Anhalt, D. Kogan, and Y. Wang. Location sensing and privacy in a context aware computing environment. *Pervasing Computing*, 2001.

[15] M. Youssef and A. Agrawala. Handling samples correlation in the horus system. In *IEEE INFOCOM 2004*, 2004.

[16] M. Youssef, A. Agrawala, and U. Shankar. WLAN location determination via clustering and probability distributions. In *Proceedings of IEEE PerCom2003*, March 2003.