Outdoor Location Estimation in Changeable Environments

Kejiong Li, John Bigham, Eliane L. Bodanese, and Laurissa Tokarchuk

***Abstract*—One approach to location estimation constructs a radio map of received signal strength (RSS) measurements at different known locations. However, location-based systems that depend on RSS alone are susceptible to inaccuracies caused by several factors, such as changes in humidity, temperature, the number of users and the physical environment. In this paper, we present a novel algorithm in which one radio map is generated as a reference map and adjusted for different run-time environmental conditions. A small sample of new RSS samples for the new environment is collected, with locations, and used to build a model to calibrate new measurements to the reference radio map. The calibration is not uniform and depends on the observed RSS. The effectiveness of the proposed method is demonstrated using real GSM data sets collected from a three-day music festival in London Victoria Park. Results are presented with and without applying the correction. A state-of-the-art cluster-based deterministic location estimation algorithm is used throughout.**

***Index Terms*—Location estimation, received signal strength, radio map.**

I. Introduction

THE recent proliferation of location-aware services has necessitated the development of outdoor and indoor positioning applications. Although the global positioning system (GPS) is the most popular positioning system for mobile devices, it is not always the best for widespread commercial use as: a) it relies on special hardware, has high complexity, battery consumption and latency; b) in dense environments such as urban areas with many high buildings, mountainous terrain and indoor areas, the access to GPS signals is often limited. Many localization systems utilize the signal-strength values received from base stations (BSs) or relay stations (RSs) or access points (APs) to estimate the location of mobile user, based on deterministic or probabilistic techniques.

Of particular relevance to this paper are methods based on received signal strength (RSS). These have been widely investigated principally in the context of indoor location estimation. This is because the data required to create the RSS database is readily collected indoors. RSS fingerprint-based localization has the potential to overcome the limitations of traditional triangulation approaches as it performs relatively well for non-line-of-sight circumstances where the alternative of modeling the nonlinear and noisy patterns of realistic radio signals is quite challenging. It requires less battery resources than receiving GPS signals and less run-time computational resources than triangulation calculation. Furthermore, GSM RSS and WiFi RSS data can be integrated to enhance accuracy.

Manuscript received June 20, 2013. The associate editor coordinating the review of this letter and approving it for publication was H. Wymeersch.

The authors are with the School of Electronic Engineering and Computer Science, Queen Mary, University of London, UK. J. Bigham is the corresponding author (e-mail: john.bigham@eecs.qmul.ac.uk).

Digital Object Identifier 10.1109/LCOMM.2013.092813.131427

RSS-based fingerprinting localization typically involves two phases: *training* and *online estimation*. In the training phase, RSS are collected at known locations to form a location fingerprints database (a.k.a. radio map). This generated radio map consists of many location-RSS vector pairs. Every RSS tuple is the location fingerprint of its corresponding location. In the online phase, new RSS observations measured at unknown positions are compared with all the fingerprints in the radio map to estimate their locations based on the preferred algorithm and distance function.

Most previous work assumes the radio map is static. During the training phase, after generating the radio map, location estimation models are built between the RSS and their corresponding location information. These models are applied with the radio map for further location estimation without making any adaptation to the new RSS measurements. However, the observed RSS measurement may significantly deviate from those stored in the radio map due to the changes in humidity, temperature, physical environment and the mobile users' hard-to-predict movements. Consequently, location-based systems that depend on static radio maps have been criticized because of their often substantial inaccuracies.

To take dynamic environmental changes into account, [1] [2] [3] [4] have proposed different approaches utilized inside buildings. By using highly distributed additional hardware, [1] uses a small number of stationary emitters and sniffers to assist location estimation, in order to obtain new RSS to update the radio maps in WLAN networks. [2] adapts a static radio map by calibrating new RSS samples at a few known locations. [3] applies a model-tree-based method, called LEMT, to adapt radio maps by only using a few reference points in an 801.11b wireless network. LEMT requires building a model tree at each location to capture the global relationship between the RSS received at various locations and those received at reference points. It requires additional sensors to keep on recording RSS all the time. [4] proposes an unsupervised learning scheme to automatically solve the hardware variance problem (chipset, antenna differences) in WiFi localization. A transformation function for mapping WiFi signal patterns from an unknown tracking device to the training device under which the radio map is calibrated is learnt using different learning methods. This technique uses a linear mapping relevant for device variations but does not reflect, e.g. RSS variances caused by population density. We include their batch linear regression learning algorithm in our comparisons.

Distinct from our previous work [5], which focuses on accurate and efficient algorithms for positioning based on a static radio map, here we consider how to adjust this map so that it can apply to new environmental conditions. In this paper we present a novel algorithm to allow an existing radio

map, which is built for a specific weather condition and user population density, to be reused by adjusting a small set of real-time RSS data collected in the new environment. This calibration process is not uniform over the area of consideration. We cluster the deviations from the existing radio map and apply a correction based on the cluster a run-time RSS observation lies in. We show that this technique can effectively accommodate the variations of signal strength due to different weather conditions and different population densities, without rebuilding the radio maps for each possible weather condition and user density. We use the cluster-based intersection location estimation algorithm derived from our previous research in the comparisons [5]. This is a deterministic intersection method that partitions the radio map using Affinity Propagation [6] clustering (where the optimal cluster size is determined by a Venn Probability Machine [7]). We use the Mahalanobis distance function to avoid giving too much weight to correlated RSS values in the distance function. The approach does not assume homogeneous transmission as the radio map is based on deviations of the raw RSS from a reference path loss model for each RSS component. In this way, the distributions of the clusters are more related to the topography than the distance from the BS. The estimated reference path loss coefficients depend on the environment and are estimated by the least squares from the training data for the transmitter [5]. Nevertheless, we believe the results of this paper are not specific to the location estimation algorithm used.

II. Location Estimation

A. The Overview of Proposed Localization System

There are three steps: firstly, RSS with corresponding location data are collected at different random locations in a reference environment and the radio map is created, e.g. using clustering and regression techniques as described in [5]. This collection of data is called the *training data* (*TR*). Secondly, under the desired different weather condition or mobile user density, a small set of RSS data and corresponding GPS signals are collected and used to build updating patterns (to be described). This RSS data is called the *secondary training data* (*STR*). Finally, newly measured RSS values are shifted based on the updating patterns, so that they can be regarded as being measured under the reference condition. Hence they can be used for positioning so as to find the best location estimate.

B. Training Phase

In this stage, suppose that there is a set of mobile stations (MSs) collected in the reference environment T_0 in the area of interest: the MS geographic location and corresponding RSS measurements from nearby transmitters (e.g. BSs) are recorded. The collection of this data is taken as *the training data set* (*TR*). For the j-th training data, let $\vec{r}_j(T_0) = (r_{j,1}(T_0), ..., r_{j,q}(T_0))$ represents the signal strength vector received by the MS from q antennas, i.e. BSs and RSs. $\vec{l}_j(T_0)$ represents its corresponding geographic location.

Secondary training data set (*STR*): Let n be the total number of data elements in the *STR* that are measured in environment T_σ. Let $R(T_\sigma) = (\vec{r}_1(T_\sigma), ..., \vec{r}_i(T_\sigma), ..., \vec{r}_n(T_\sigma))$ denote the RSS measurements from nearby transmitters, where $\vec{r}_i(T_\sigma) = (r_{i,1}(T_\sigma), ..., r_{i,q}(T_\sigma))$ is a q dimensional vector of RSS received by *STR* element i (i.e. a MS) from q antennas. $L(T_\sigma) = \left(\vec{l}_1(T_\sigma), ..., \vec{l}_i(T_\sigma), ..., \vec{l}_n(T_\sigma)\right)$ consists of geographic locations. $\vec{l}_i(T_\sigma)$ is the 2-D position coordinates of *STR* i.

For the i-th element of the *STR*, its measured RSS values $\vec{r}_i(T_\sigma)$ are adjusted to create $\vec{r}\,'_i(T_0)$, so that the estimated signal strength values $\vec{r}\,'_i(T_0)$ can be treated as if it were collected in the reference environment T_0.

Step 1: Find the K (e.g. K = 3) nearest neighbors of *STR* i from *TR* in location space (not RSS space), and the IDs of these neighbor *TRs* are recorded in set U_i. So the physical location of the k-th ($1 \le k \le K$) neighbor can be given as $\vec{l}_{U_i(k)}(T_0)$, and $\vec{r}_{U_i(k)}(T_0)$ denotes its corresponding RSS measurements. Therefore, the location distance between *STR* i and its k-th neighbor can be give as

$$d_k = \left\| \vec{l}_i(T_\sigma) - \vec{l}_{U_i(k)}(T_0) \right\| \tag{1}$$

Step 2: Calculate an estimated RSS values for *STR* i that can be regarded as measured in the reference environment T_0, which can be expressed as

$$\vec{r}\,'_i(T_0) = \sum_{k=1}^{K} w_k \vec{r}_{U_i(k)}(T_0) \tag{2}$$

where w_k is a normalized weight for the k-th neighbor:

$$w_k = \frac{1}{d_k \sum_{i=1}^{K} \frac{1}{d_i}} \tag{3}$$

Step 3: Obtain a vector of difference values of *STR* i between its estimated RSS values $\vec{r}\,'_i(T_0)$ and measured RSS values $\vec{r}_i(T_\sigma)$.

$$\vec{\Delta}_i = \vec{r}\,'_i(T_0) - \vec{r}_i(T_\sigma) \tag{4}$$

Step 4: Repeat **Step 1-3** another $(n-1)$ times for all the other *STRs*, thus every *STR* has a vector of difference values.

Step 5: Apply our clustering scheme to cluster the n difference values. Let G_i stand for the cluster that $\vec{\Delta}_i$ belongs to. Assume that G_i contains N_i vectors of difference values including $\vec{\Delta}_i$, so the average of all the difference vectors in cluster G_i can be assigned to *STR* i ($1 \le i \le n$) as:

$$\bar{\Delta}_i = \frac{1}{N_i} \sum_j \vec{\Delta}_j, \left\{ 1 \le j \le n \mid \vec{\Delta}_j \in G_i \right\} \tag{5}$$

C. Online Location Estimation Phase

During the online phase for the environment T_σ, given a new MS ms with observed RSS tuple $\vec{r}_{ms}(T_\sigma)$ from q BSs, the process of estimating ms's location $\hat{l}_{ms}$ is as follows.

Step 1: Find MS ms's K' (e.g. K'=3) nearest neighbors in the *STRs* (using Eq. (6)) and store their IDs in V_{ms}. So $\vec{r}_{V_{ms}(k')}(T_\sigma)$ and $\vec{l}_{V_{ms}(k')}(T_\sigma)$ can denote the RSS sets and locations of the k'-th *STR* neighbor of MS ms, respectively. By using the Mahalanobis distance in signal space, we can obtain the similarity between the MS ms's RSS values and its k'-th neighbor *STR*'s RSS values:

$$s_{k'} = \sqrt{\left(\vec{r}_{ms}(T_\sigma) - \vec{r}_{V_{ms}(k')}(T_\sigma)\right)^T \Sigma^{-1} \left(\vec{r}_{ms}(T_\sigma) - \vec{r}_{V_{ms}(k')}(T_\sigma)\right)} \tag{6}$$

TABLE I
ENVIRONMENT INFORMATION DURING THE THREE DAYS IN LONDON VICTORIA PARK

Day	Temperature	Humidity	Cloud Amount	Precip Amount	User Density
Day 1	20°C	73%	42%	0.3mm	10,000
Day 2	19°C	64%	54%	0.0mm	30,000
Day 3	16°C	77%	84%	1.3mm	9,000

Here Σ is a $q \times q$ covariance matrix in signal space that describes the mutual dependencies of the received signal strength.

Step 2: Based on the similarity in signal space, each of these K' *STR* neighbors can be assigned a weight using:

$$w_{k'}^{'} = \frac{1}{s_{k'} \sum_{j=1}^{K'} \frac{1}{s_j}} \tag{7}$$

Step 3: Calibrate the RSS tuple of MS ms to what it would be as if it were measured in the reference environment T_0 by

$$\vec{r}'_{ms}(T_0) = \vec{r}'_{ms}(T_\sigma) + \sum_{k'=1}^{K'} w_{k'}^{'} \bar{\Delta}_{V_{ms}(k')} \tag{8}$$

Since the above calibration process focuses on eliminating the impact of environmental factors, such as weather condition and mobile population density, the calibrated RSS value $\vec{r}'_{ms}(T_0)$ can be regarded as measured in the same environment T_0 as training data. So the calibrated RSS value can be used for position estimation with the cluster-based intersection approach described in [5].

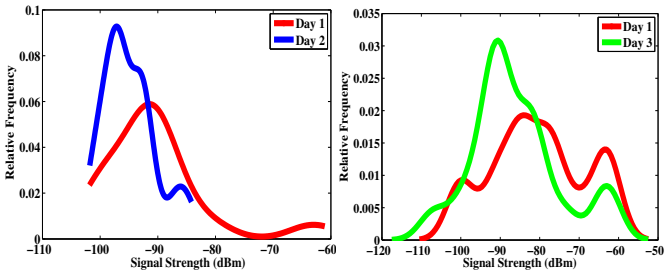
III. PERFORMANCE EVALUATION

We conducted the experiments in a three-day music festival held in London Victoria Park (analogous to a rural setting) that covers a 450m x 240m area. Due to the different activities and venues of the music festival, the walking paths on different days are different. We partition the data sets according to the day collected, that is, Day 1, Day 2 and Day 3. The weather and population density information during these three days is shown in Table I according to [8]. The RSS data of a GSM network is collected by a mobile app on an Android smart phone. The app records the received signal strength from each of the surrounding BSs and GPS latitude and longitude every 1 second as we move around the outdoor venue. The locations of all the nearby BSs are obtained from the Sony Ericsson server. More details about these three scenarios including the downloadable raw data can be found in [9].

2095 RSS samples collected on the first day are used as the training data set to create clusters. In this case, the optimal number of clusters is 52. 2050 and 3424 RSS measurements are also collected along with their location coordinates on the second day and the third day, respectively. The performance of the proposed calibration method is compared with the method introduced in [4], named here as online regression learning method. The comparison results between cluster-based intersection [5] and K-nearest Neighbor (KNN) localization [10] algorithms with and without using these two calibration schemes on two different days are presented.

A. Impact of Environmental Changes

In the graphs below we illustrate the changes in RSS for different conditions on these days of the festival: (1) Similar weather, different population density in Fig. 1(a); (2) Different weather, similar population density in Fig. 1(b).



(a) (b)

Fig. 1. (a) The comparisons of RSS distributions over Day 1 (medium attendance) and Day 2 (large attendance) at fixed locations from a typical BS (b) The comparisons of RSS distributions over Day 1 (dry and sunny) and Day 3 (wet and small attendance) at fixed locations from a typical BS

It can be seen that the signal strength values received from the same BS at a fixed location may vary significantly. From Table I, we can see that Day 1 was sunny and dry and with a moderate number of visitors, while Day 2 had the same weather condition but had a much larger audience. Day 3 was cloudy and wet and with a slight drop in the user numbers compared with Day 1. Because of the different activities and venue layouts of the music festival on different days, there are only a few locations that are measured with RSS and same GPS signals in all the three days. Therefore, we make pairwise comparisons of the RSS distributions between Day 1 and Day 2, and Day 1 and Day 3, to analyze the impact of environmental factors, e.g. weather condition and population density, on RSS measurements in two cases.

Fig. 1(a) and Fig. 1(b) illustrate two comparisons of RSS distributions, both of which are processed with the kernel density estimate method. RSS data in each comparison are measured at the same locations from the same BS. We can see that in each figure the signal strength of the peaks vary from each other, probably because of the different environments in each comparison. We can conclude that the RSS distributions from the same BS vary both with different audience numbers, and weather condition variations during the three days even at the fixed locations. These variations imply that depending on the original radio map generated in the training phase, the position estimation results might be inaccurate when the physical environment changes.

B. The Effect of the Number of Secondary Training Samples

The location estimation accuracy in a new environmental condition depends on the number of secondary training RSS tuples collected in the new environment: a larger number of secondary training data leads to higher accuracy of location estimation but more time for training. This is why it is useful to have a method that is efficient in its use of the sample data. Fig. 2 reports the effect of using different numbers of secondary training samples, which are taken from Day 2 and Day 3 based on different calibration methods using the cluster-based intersection method, and shows the estimation accuracy. The efficiency can be seen from Fig. 2, where, using the clustering approach for example, and only taking 650 secondary training samples, then the proposed calibration

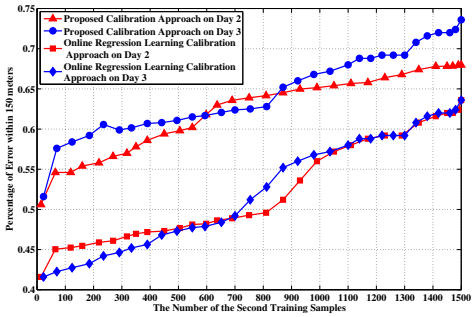
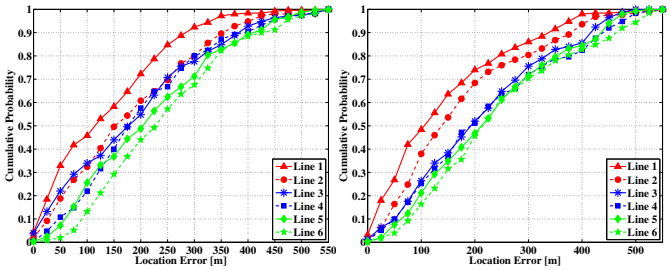


Fig. 2. Percentage of error within 150 meters versus the number of secondary training samples using the cluster-based intersection method with different calibration schemes for Music Festival



(a) Day 2 (b) Day 3

Fig. 3. Cumulative percentage of error for different algorithms: Line 1 (Cluster-based Intersection with Proposed Calibration), Line 2 (KNN with Proposed Calibration), Line 3 (Cluster-based Intersection with Online Regression Learning Calibration), Line 4 (KNN with Online Regression Learning Calibration), Line 5 (Cluster-based Intersection without Calibration), Line 6 (KNN without Calibration)

method already outperforms the online regression learning method, even when it uses 1500 samples. This results in a reduction of costs of site survey and data collection. Hence, from this figure, given a required accuracy, e.g. 60%, the sample sizes from Day 2 and Day 3 need to be chosen to 520 and 360 respectively when using our proposed method.

C. Positioning Performance

We choose 520 and 360 random RSS measurements from Day 2 and Day 3 respectively as the secondary training data set to adapt the built radio map and their location coordinates are assumed known. The remainder of the measurements are used for location estimation with only their RSS values. Their GPS values are only used for the subsequent validation. Fig. 3 depicts the cumulative distribution function (CDF) of the error distance for the cluster-based intersection method and KNN method with and without using the different calibrated schemes on these two days. Comparison of the two figures clearly shows that the data update is effective, especially for our proposed correction scheme and that our cluster-based intersection outperforms the traditional approaches. More specifically, as seen from Fig. 3(a), the percentage of errors less than 150 meters in cluster-based Intersection and KNN methods with online regression learning update scheme are 40.4% and 32.4% respectively whereas those localization methods using our proposed update scheme are 60.0% and 45.6% on Day 2. Similarly, we can observe from Fig. 3(b) that our proposed calibration method can perform better than the online regression learning method on Day 3, e.g. for the cluster-based intersection approach, the mean measurement error by using our proposed correction method is around 151.7m, while using the online regression learning method and without any correction report 203.4m and 236.9m respectively. From the experiments, we can conclude that the proposed method can adapt better to changeable environments compared with conventional static fingerprint-based positioning method. The online regression approach is different, in that it does not require new GPS values. It uses correlation between RSS tuples to match locations. It was developed for inside location estimation. Hence, it is inevitably less accurate than when GPS information is added. When GPS is added, our method remains superior in our experiments.

IV. Conclusion

In this paper, we have described a novel RSS-based outdoor location estimation approach that can adapt to environmental changes. The proposed method only needs one full radio map built for a specific environmental condition or user population density. A small set of data measured in a new environment is compared within the existing radio map and a model that can calibrate the run-time RSS data for the new environment is created. Thus, the calibrated RSS data can be regarded as measured in the same reference environment as the training data. The improvement in location estimation is tested, and the results show that the proposed algorithms achieve a considerable accuracy and efficiency advantages in a real environment. In future work, we will focus on adapting Bayesian methods into a larger outdoor environment over different seasons to further validate these results. We are also incorporating the user's movement trajectories to further improve the accuracy of location estimation.

References

[1] P. Krishnan, A. Krishnakumar, W. H. Ju, C. Mallows, and S. Ganu, "A system for LEASE: location estimation assisted by stationery emitters for indoor RF wireless networks," in *Proc. 2004 IEEE Joint Conf. of the IEEE Comput. and Commun. Soc.*, pp. 1001–1011

[2] A. Haeberlen and A. Rudys, "Practical robust localization over large-scale 802.11 wireless networks," in *Proc. 2004 ACM Int. Conf. on Mobile Computing and Networking*, pp. 70–84.

[3] J. Yin, Q. Yang, and L. M. Ni, "Learning adaptive temporal radio maps for signal-strength-based location estimation," *IEEE Trans. Mobile Comput.*, vol. 7, pp. 869–883, July 2008.

[4] W. Tsui, Y. H. Chuang, and H. H. Chu, "Unsupervised learning for solving RSS hardware variance problem in WiFi localization," *Mobile Networks and Applications*, vol. 14, no. 5, pp. 677–691, 2009

[5] K. Li, P. Jiang, E. L. Bodanese, and J. Bigham, "Outdoor location estimation using received signal strength feedback," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 978–981, July 2012.

[6] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[7] V. Vovk, G. Shafer and I. Nouretdinov, "Self-calibrating probability forecasting," in *Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[8] Weather2. Avaliable: http://www.myweather2.com/

[9] Open Google Project, location-estimation-trials. Available: http://code.google.com/p/location-estimation-trials/.

[10] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *Proc. 2000 IEEE Joint Conf. of the IEEE Comput. and Commun. Soc.*, vol. 2, pp. 775–784