Outdoor Location Estimation Using Received Signal Strength Feedback

Kejiong Li, Peng Jiang, Eliane L. Bodanese, and John Bigham

***Abstract*—An improvement to outdoor fingerprint location estimation based on clustering received signal strength (RSS) from base stations (BSs) is presented. The novel features that contribute to the greater accuracy are the use of deviations from the path loss model for each RSS component rather than the raw RSS for clustering; the accurate estimation of the cluster membership probability; the optimized trade-off between cluster size and accuracy of cluster modeling; and clusters are invariant to the BS or relay station (RS) power.**

***Index Terms*—Location estimation, clustering, fingerprint.**

I. Introduction

FINGERPRINTING is the most widely used technique for positioning. It overcomes the limitations of traditional triangulation approaches and performs well in non-line-of-sight circumstances. A typical fingerprint-based localization system only needs to collect the measurements of Received Signal Strength (RSS), or other useful parameters at some known locations to form a location fingerprints database (a.k.a. radio map) during the training phase. Fingerprinting can be broadly categorized into deterministic and probabilistic approaches. This letter uses a deterministic approach. One of the simplest deterministic approaches estimates the location of an observed RSS tuple as the average of the of locations of the K-Nearest-Neighbors (KNN) in RSS space as measured by the Euclidean distance from the observed RSS to the training RSS tuples [3]. [4] and [5] improve the accuracy by using a weighted average of the coordinates of the K nearest training samples. The weight value is taken as the inverse of the Euclidian distance between the observed RSS measurement and the RSS measurements of the K nearest training samples. In this case, this method is referred to as the Weighted K-Nearest Neighbor (WKNN). The results in [5] indicate that the KNN and the WKNN can provide greater accuracy than the Nearest Neighbor (NN) method, particularly for K = 3 and K = 4. However, for high density radio maps, the NN method can perform as well as more complicated methods [6]. Probabilistic approaches use the training RSS tuples to construct conditional probability density functions of the fingerprint tuples given locations, and utilize Bayes' theorem to compute the posterior probabilities of possible locations given a new RSS tuple. However, because of the complexity, the authors in [7], for example, assume the elements of the fingerprint tuple are statistically independent from each other, which does not hold in a real environment. A classification of well known fingerprint approaches using three key features is shown in Table I.

Manuscript received August 26, 2011. The associate editor coordinating the review of this letter and approving it for publication was C. Assi.

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

Digital Object Identifier 10.1109/LCOMM.2012.050912.111805

TABLE I
CLASSIFICATION OF COMMON FINGERPRINTS SCHEMES

Localization Scheme	Area of Deployment	Partition Model Cluster/Global	Estimation Techniques
RADAR [3]	WLAN Indoor	Global	Deterministic
LANDMARC [4]	WLAN Indoor	Global	Deterministic
Ref. [7]	WLAN Indoor	Global	Probabilistic
Horus [8]	WLAN Indoor	Cluster	Probabilistic
Proposed Method	Outdoor	Cluster	Deterministic

Previous fingerprint studies in the literature mainly focus on indoor localization due to the difficulty in acquiring and analyzing the large amounts of data that need to be processed. In this letter, we investigate the use of RSS for outdoor mobile localization that efficiently utilizes the training RSS data. The training RSS data is partitioned into clusters [2] and the mobile location is estimated by which cluster it belongs to and its relative location in that cluster.

The novel features that contribute to the greater accuracy are: a) the use of deviations from the observed path loss model for each RSS component rather than the raw RSS. This also results in the clusters being invariant to the BS/RS power; b) the accurate estimation of the cluster membership probability and the optimal number of clusters to manage the trade-off between cluster size and accuracy of cluster modeling; c) refinements of the intersection approach for improved performance. The datasets used are described in [1] and are available for use.

II. Location Estimation

A. Training Phase

In this phase, RSS and reference location data are collected and models created to represent the data using clustering and regression techniques. For this paper, it can be considered the training is done once and for all. Clusters are used instead of a uniform grid for a better model of complex topography as grid boundaries and topographic features do not necessarily align. (This has been verified in direct comparisons on the data sets described here.) Previous clustering localization research [8] does not pay attention to the cluster stability and scalability issues without losing important correlation information. In our clustering scheme, the Affinity Propagation (AP) method [9] is used for clustering and the Venn Probability Machine (VPM) algorithm [10] is used to predict probability of cluster membership and manage the trade-off between estimation accuracy of cluster identification and the number of clusters to select the best clustering scheme.

Assume that in the training stage, the following data is collected for a set of n MSs: the MS geographic location and the RSS measurements from neighboring transmitters. Let $R^i = (r_1^i, r_2^i, ..., r_q^i)$ represent the set of RSS from MS i from q antennas, i.e. BSs and RSs, in the area of interest. In this letter, the deviations from the q RSS log-distance path loss

models create a tuple $\rho^i = (\rho_1^i, \rho_2^i, ..., \rho_q^i)$. Clustering is based on the Mahalanobis distance. For any two MSs, such as MS i and MS k, the similarity between them can be expressed as:

$$s(i,k) = -\sqrt{(\rho^i - \rho^k)^T M^{-1} (\rho^i - \rho^k)},\ \forall k \neq i \quad (1)$$

Here M is the $q \times q$ covariance matrix in signal space, which describes the mutual dependence of the signal strength received by any two MSs from different BSs. It can be estimated as $M = [m_{j,p}]$, where $m_{j,p} = \frac{1}{n}\sum_{i=1}^{n}(\rho_j^i - \overline{\rho_j})(\rho_p^i - \overline{\rho_p})^T,\ 1 \leq j, p \leq q$, and $\overline{\rho_j}$ is the average RSS deviation value of MSs from BS j, $\overline{\rho_j} = \frac{1}{n}\sum_{i=1}^{n}\rho_j^i$.

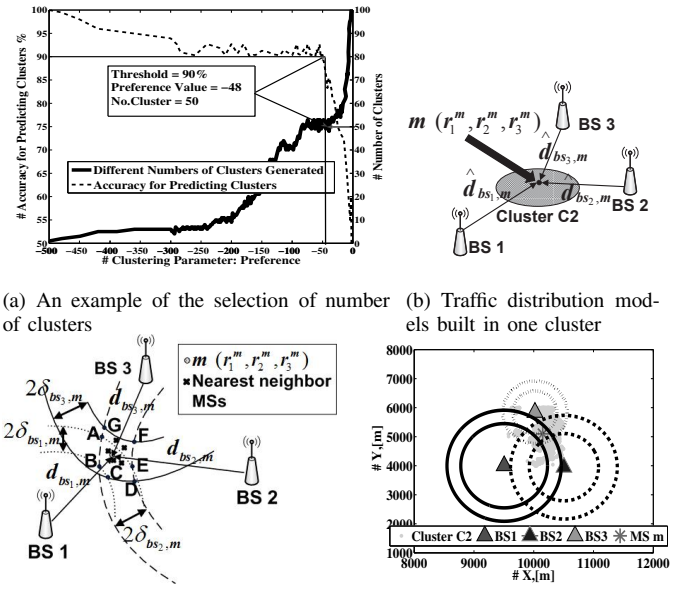
Using deviations means we can eliminate to some extent the effects of distance dependent path loss attenuation, and so better capture the effects of multipath and shadowing, which mainly depend on the topography. Using the raw RSS leads to clusters where the similarity is dominated by the distance path loss, which is approximated anyway by the estimated path loss model. Direct comparisons on different data sets are given in section C later. The deviations from the log-distance path loss model are obtained based on the RSS data during the training stage. P_{rss} at the distance d from each transmitter can be formulated as in the equation below:

$$P_{rss} - P_{tr} = \kappa + \beta \log(d/d_0) \quad (2)$$

Where P_{tr} represents the transmit power of BS, d_0 is reference distance for the antenna area. The values of parameter β and k are heavily dependent on the environment and are estimated by a least squares regression model from the training data according to (2).

An example is given in Fig. 1(a) to illustrate the identification of the optimal number of clusters. The thick blue line represents the relationship between the cluster preference parameter value and the number of clusters generated. The green dashed line depicts the dependence of the cluster prediction accuracy on the number of clusters created. There is a trade-off between the number of generated clusters and location estimation: the greater number of clusters generated in the training period reduces the cluster prediction accuracy, but results in a relatively higher precision in location estimation conditional on the cluster. The objective is to find the right balance between the accuracy of cluster identification and the number of clusters. Here, the optimal number of clusters is taken as the maximum number of clusters that satisfies the accuracy requirement of cluster prediction. The cluster shape is determined by the location of training data in this cluster. Seen from Fig. 1(a), here the threshold accuracy of cluster identification is taken as 90%, the corresponding maximum number of clusters is 50. The training data is collected first and the collection time depends on the size of area required. To take physical environmental changes into account, our method also creates RSS coverage models and tests for discrepancies in the coverage models created from additional data that can be collected periodically.

So in summary in the training phase, the deviations are computed by computing the regression equation (2) for each BS using all the training data points. Then the clustering is performed and for each cluster for each BS, a regression model is built using the training data in that cluster according to (2).



(a) An example of the selection of number of clusters

(b) Traffic distribution models built in one cluster

(c) Uncertainty area of location estimation area

(d) A sample of the uncertainty area

Fig. 1. Cluster selection and intersection estimation

B. Location Estimation Phase

In the location estimation phase, mobile location is estimated based on RSS data for each cluster. Suppose the terrain is divided into a set of clusters $C = \{c_1, c_2, ..., c_N\}$, and N is the total number of clusters. Given a new MS m with observed RSS tuple, (r_1^m, r_2^m, r_3^m), from e.g. three neighboring BSs, the process of estimation of MS m's location is described below in detail:

Step 1: Use the VPM method based on the KNN algorithm to calculate the probability of MS m belonging to every possible cluster ID, and then assign the cluster ID with the highest probability to MS m. Call this cluster c_2.

Step 2: Use the previously computed regression model for each BS in this cluster to estimate the geographical distance from each BS to MS m, $\hat{d}_{bs_i,m} (i = 1, 2, 3)$, as shown in Fig. 1(b).

Step 3: Calculate the distance between the RSS tuples of training data in cluster c_2 and MS m's RSS tuple, and then use KNN algorithm to find MS m's K nearest neighbors in the training data from cluster c_2. Let K_1 represent the number of K nearest neighbors of MS m and $NL = \{nn_1, nn_2, ..., nn_{k_1}\}$ the neighbor list. In order to estimate of the precision for each calculated distance, a distance range for each $\hat{d}_{bs_i,m}$ is computed, viz. $\delta_{bs_i,m} (i = 1, 2, 3)$. To do this, we first work out the MS m's K_1 nearest neighbors individual deviations of the distances from the centroid of the cluster, which can be given by:

$$\delta_{bs_i,nn_t} = \sqrt{|d_{bs_i,nn_t} - \overline{d_{bs_i,nn_{\{nn \in NL\}}}}|} \quad (3)$$

Where $\overline{d_{bs_i,nn_{\{nn \in NL\}}}}$ is the centroid of the K_1 nearest neighbors in training data from cluster c_2 and is defined as:

$$\overline{d_{bs_i,nn_{\{nn \in NL\}}}} = \sum_{t=1}^{K_1} d_{bs_i,nn_t}/K_1 \quad (4)$$

Then compute the weighted mean (μ) and standard deviation (σ) of the deviations $\delta_{bs_i,nn_t}, nn_t \in NL$. Since the RSS distribution is skew (by observation), a two sided confidence

interval is created by applying *Chebyshev's inequality*. Specifically the $\delta_{bs_i,m}$ are chosen so that:

$$P(|\delta_{bs_i,m_{\{m\in NL\}}} - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2} \quad (5)$$

On the right hand side of (5), the value of $\frac{1}{\lambda^2}$ is set to be 0.01, which means that whatever the distribution is, there is always at least 99% of the probability inside the distance band interval. We now take the value of the upper bound of the distance band interval as the uncertainly band value, $\delta_{bs_i,m}$. Thus, the possible distance band $d_{bs_i,m}$ between the BS i and MS m is:

$$d_{bs_i,m} \in \left[\hat{d}_{bs_i,m} - \delta_{bs_i,m}, \hat{d}_{bs_i,m} + \delta_{bs_i,m}\right] \quad (6)$$

As illustrated in Fig. 1(c), the position of MS m can be determined from the overlapping areas formed by circular bands associated with likely distance from each source (now BSs) similarly to [11]. Distinct from [11], we apply this idea within each cluster rather than over the whole area; we estimate a distribution for the distance based on K nearest neighbor density and then compute confidence bands based on (5) (rather than uniform); and we do not assume all the bands overlap (not true in our complex environment data) and hence consider other situations. This is considered in step 4.

Step 4: At this step, different cases need to be considered in order to find the most likely intersection area that MS m might be located in. We use KNN to find MS m's $K_2(K_2 \leq 5)$ nearest neighbors in the training data from cluster c_2, and in each intersection area (area where all intersect, areas where each pair intersect, etc.) find the number of nearest neighbors. For example, in Fig. 1(c), among MS m's 5 nearest neighbors, 3 MSs are within the intersection area of the bands from the three BSs, i.e. ABCDEA in Fig. 1(c). If $B_{bs_i}(i = 1, 2, 3)$ is the distance bands from each BS, the common intersection area ABCDEA can be simplified to $\{B_{bs_1} \cap B_{bs_2} \cap B_{bs_3}\}$. The other 2 MSs are located within AEFGA. Then choose the intersection area from the candidate areas for MS m as the one with the highest frequency, here it is $\{B_{bs_1} \cap B_{bs_2} \cap B_{bs_3}\}$.

Step 5: In $\{B_{bs_1} \cap B_{bs_2} \cap B_{bs_3}\}$, we can find MS m's three nearest neighbors in training data from cluster c_2, and the mean value of the location of these three points is the estimation of MS m's location. Fig.1 (d) provides a sample to display how the method above process with a real training data set.

III. Performance Evaluation

In this section, the localization accuracy is tested with two sets of data. In all cases, the MS were stationary or walking. One is data generated by a network planning tool ASSET 3G for the island of Jersey (primarily rural) and the other data is collected by an Android mobile app for the Queen Mary campus (city). In the simulations, we randomly divided the data into two equal sets. The first is treated as training data and their exact location coordinates are assumed known. The other half is used for testing and uses only RSS tuples. The results are compared with KNN algorithm using different nearest neighbor contexts: one is Global-KNN that uses KNN in the whole area and another one is Cluster-KNN that utilizes KNN based on our optimized clustering scheme.

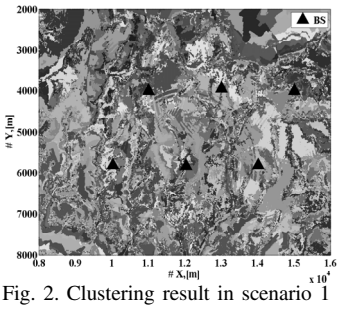


Fig. 2. Clustering result in scenario 1

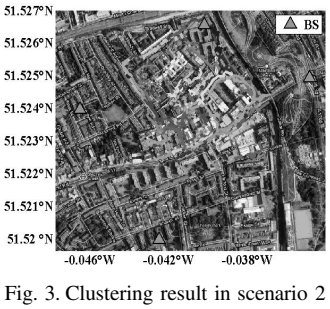


Fig. 3. Clustering result in scenario 2

A. Scenario 1: the island of Jersey area

We choose six BSs in the centre of the island covering an area of 8km x 6km. Fig. 2 shows the optimized clustering in this area. 160 clusters are created. Different colors represent different clusters. The clusters generally represent the topographical features including the contours of highways and roads. Table II summarizes the information in terms of the mean, variance, 50th, 75th and 90th percentile values of the error distance for each method. The intersection method performs better than the other two KNN methods, e.g. 90% of distance errors are within 23.3m, whereas Cluster-KNN and Global-KNN report 42.5m and 26.8m respectively for the same cumulative probability. Global-KNN can achieve relatively high accuracy but takes a longer time to calculate than both other methods. For KNN, a larger value of K leads to *lower* estimation accuracy.

In summary, given the large number of test points and the complexity of the model, it seems feasible to adopt our algorithms to estimate location based on the clustering scheme that can have a mapping to the topography meaningful in a large area.

B. Scenario 2: Queen Mary campus

To test the proposed mechanisms in an urban area, the RSS data of a GSM network was collected by a mobile app on an Android smart phone. The mobile app records the RSSs from nearby BSs and the corresponding location with GPS. The locations of BSs are obtained from the service provided by Sony Ericsson lab. Fig. 3 shows the topographic map of Queen Mary campus that covers a 475m x 365m area. The nearest four BSs are considered in this letter. The colored-line area represents the result of clustering 9277 test points. In this case, the optimal number of clusters is 70. As shown in Table II, the intersection method again outperforms the Global-KNN and (slightly) Cluster-KNN. For the intersection approach, the mean measurement error is around 22.8m with 26.0m^2 variance.

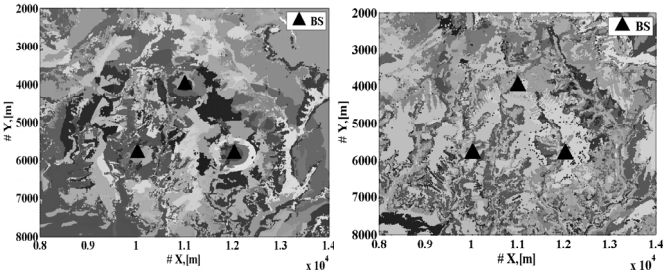
Unlike the rural environment in Scenario 1, the signal in complex urban environment undergoes additional attenuation and fluctuates rapidly due to many obstructions i.e. high buildings. These results indicate that the intersection method provides relatively high location estimation accuracy with a small amount of training data in a small range of area. Since the ground truth is taken as the GPS, and it has inaccuracies the variance is an overestimate.

C. RSS deviations from path loss and raw RSS

Using the deviations helps in two ways: a) Clusters generated by the deviations data are a better reflection of the

TABLE II
COMPARISON OF ESTIMATION ERROR BETWEEN PROPOSED METHOD, CLUSTER-KNN AND GLOBAL-KNN METHODS IN SCENARIO 1,2(IN METERS)

	Scenario 1			Scenario 2		
	Inter-section	Cluster -KNN	Global -KNN	Inter-section	Cluster -KNN	Global -KNN
Mean Error	10.5	17.8	11.5	22.8	27.7	51.6
Variance	15.6	31.3	15.1	26.0	37.7	58.0
50 Percentile	5.3	7.8	6.1	13.4	14.3	31.3
75 Percentile	11.1	17.3	11.8	30.6	32.5	72.6
90 Percentile	23.3	42.5	26.8	50.2	58.2	135.3



(a) Clustering based on raw RSS (b) Clustering based on deviation RSS

Fig. 4. A comparison of clustering the results using raw RSS and deviations of RSS from expected path loss. N.B. Colors are reused, so the same color can represent different clusters.

topography. For the Jersey data, Fig. 4 depicts the comparisons of clustering distribution between using the raw RSS and deviation RSS when the same number of clusters is created. It can easily be seen that using the deviation RSS has achieved significant better result than raw RSS. If we do not use the deviations effectively spurious clusters are generated that reflect the decay in power. Using the raw RSS leads to clusters where the similarity is dominated by the distance path loss when near a BS and this is approximated anyway by the estimated path loss model. For example, if the world was uniform, then ring segments are generated that simply reflect the decay with distance rather than topography. The evidence for this is observed in Fig. 4(a) when we use the raw RSS. On the contrary, Fig. 4(b) illustrates that using the deviations the mapping of the clusters onto the geographical locations shows more scatter than if clustering is performed on the raw data in Fig. 4(a).

b) The location estimation is more accurate on both data sets when using the deviations, e.g. Fig. 5 shows an example of the cumulative distribution function of the error distance for intersection method based on both data sets in the Queen Mary Scenario, under the premise that the same number of clusters is created. For the intersection method the mean values of the distance error based on deviations is 22.8m, whereas based on the raw RSS data the mean is 72.6m. The invariance of the clusters to transmit powers using the deviations is described in [2].

IV. CONCLUSION

Results presented show that the proposed scheme finds more accurate locations and outperforms the Global-KNN and Cluster-KNN approaches for all numbers of NNs tested. Using deviations is seen to give better accuracy in a complex environment. The clustering scheme also has the following advantages: a) it allows selective additional data collection to enhance accuracy, e.g. near cluster boundaries; b) it is not sensitive to transmit power changes; c) it can be coupled with location tracking and other non RSS data to improve accuracy. Whilst the comparisons in this letter are based on stationary and walking speed observations, future work is looking at the use of auxiliary information, such as the direction from the compass, accelerometer readings and route information to enhance accuracy and reduce battery usage when people are moving at faster speeds.

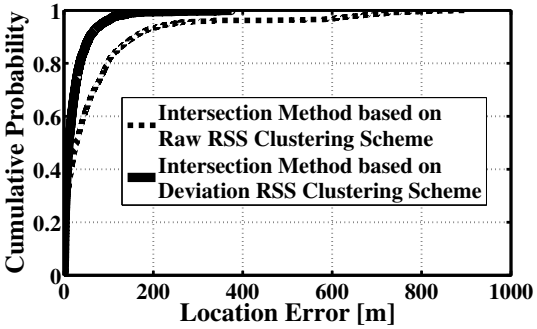


Fig. 5. Location estimation results based on raw RSS and deviation RSS (Queen Mary data set). Probability that the error is less than a chosen value.

ACKNOWLEDGMENT

A special thanks to Aircom International Ltd for providing their network planning tool ASSET 3G and for the data provided on pilot signal strengths for the island of Jersey.

REFERENCES

[1] Open Google Project, http://code.google.com/p/location-estimation-trials/
[2] K. Li, P. Jiang and J. Bigham, "Partitioning the wireless environment for determining radio coverage and traffic distribution with user feedback," *2011 National Conference on Communications*.
[3] P. Bahl and V. N. Padmanabhan, "RADAR: an in-building RF-based user location and tracking system," in *Proc. 2000 IEEE INFOCOM*, vol. 2, pp. 775–784.
[4] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, "LANDMARC: indoor location sensing using active RFID," *Wireless Networks*, vol. 10, pp. 701–710, 2004.
[5] B. Li, J. Salter, A. G. Dempster, and C. Rizos, "Indoor positioning techniques based on wireless LAN," *School of Surveying and Spatial Information Systems*, UNSW, Sydney, Australia, 2006.
[6] V. Honkavirta, T. Perala, S. Ali-Loytty, and R. Piche, "A comparative survey of WLAN location fingerprinting methods," in *Proc. 2009 Workshop on Positioning, Navigation and Communication*, pp. 243–251.
[7] T. Roos, P. Myllymki, H. Tirri, and P. Misikangas, "A probabilistic approach to WLAN user location estimation," *Int'l J. Wireless Inf. Networks*, vol. 9, no. 3, pp. 155–164, July 2002.
[8] M. Youssef, A. Agrawala, and A. U. Shankar, "WLAN location determination via clustering and probability distributions," in *Proc. 2003 IEEE International Conference on Pervasive Computing and Communications*, pp. 143–150.
[9] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.
[10] V. Vovk, G. Shafer, and I. Nouretdinov, "Self-calibrating probability forecasting," *Advances in Neural Information Processing Systems 16*, 2003.
[11] Y. Ji, "Practical precision bound for indoor location determination," in *Proc. 2010 International Conference on Computer and Information Application*, pp. 410–413.