Error Analysis for Fingerprint-based Localization

Yunye Jin, Wee-Seng Soh, and Wai-Choong Wong
Department of Electrical and Computer Engineering
National University of Singapore
Email: {g0700214, elesohws, elewwcl}@nus.edu.sg

***Abstract*—In this paper, we derive the theoretical error Probability Density Function (PDF) and Region of Confidence (RoC) conditioned on the on-line signal parameter vector, for a generalized fingerprint-based localization system. As the computations of these terms require the exact expression of the joint PDF for both the device location and the on-line signal parameter vector, which is often not available practically, we propose to approximate this joint PDF by Nonparametric Kernel Density Estimation techniques using the training fingerprints.**

***Index Terms*—Fingerprint-based localization, error analysis, nonparametric density estimation, Region of Confidence.**

I. Introduction

Because of easy implementation and cost-effectiveness, fingerprint-based methods have gained more popularity among practical indoor localization systems, compared to trilateration-based methods, which cost more in terms of RF bandwidth, hardware needs, and computational overhead [1].

In a typical fingerprint-based localization system, several Access Points (APs) are broadcasting beacon frames periodically. During an off-line "training phase", vectors of location-dependent signal parameter, most commonly Received Signal Strength (RSS), are collected at a number of "training locations" as location fingerprints. Various algorithms can be used to estimate user location when an on-line RSS vector is captured. The K-Nearest-Neighbor (KNN) scheme [2] returns the location estimate as the average of the coordinates of the K training locations whose fingerprint vectors have the shortest Euclidean distances to the on-line RSS vector. A special case of KNN is the Nearest Neighbor in Signal Space (NNSS) [2], in which $K = 1$. The probabilistic approach, [3], [4], uses the training data to construct the Probability Density Function (PDF) for the device location conditioned on the observed on-line RSS vector. The conditional expectation of the location is returned as the estimate. It has been reported that KNN and the probabilistic approach have similar performance [3], [4].

In practice, error PDF and Region of Confidence (RoC) conditioned on the on-line RSS vector not only conveniently indicate the reliability of the current location estimate, but also facilitate the fusion of multiple sensors [5]. Due to the presence of multipath propagation, noise, and interference, there can be significant temporal and spatial variations in the on-line RSS vectors. As illustrated in Fig. 1, different samples of on-line RSS vectors can result in different estimated locations and radii of RoCs, even if they were collected at the same actual location.

There are very few works which study the theoretical error analysis of fingerprint-based localization systems, while taking the current on-line RSS vector into account. The analyses in [6] and [7] are only applicable to NNSS. On-line error analysis for more advanced schemes such as KNN and probabilistic

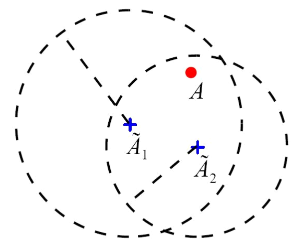

Fig. 1. Estimated locations and RoCs based on two different on-line RSS vectors collected at the same actual location.

approach have not been explored. [8] formulates RoC geometrically to filter outliers in localization results. However, the formulation is only validated empirically, without any theoretical justifications. On the other hand, [9] proposed to use the covariance of the estimates as a confidence measure.

In this paper, we will derive the exact theoretical expression of both the error PDF and RoC, conditioned on the observed on-line RSS vector, for a fingerprint-based localization system. As the computations of the relevant terms require exact knowledge of the joint PDF for the location and the on-line RSS vector, which is practically not available, we approximate this joint PDF by Nonparametric Kernel Density Estimation (NKDE) techniques using the training fingerprints.

Theoretically, our proposed scheme is also applicable to different location fingerprints other than RSS. However, in order to easily verify our proposed scheme experimentally, we focus our discussion on the RSS fingerprint in this paper.

II. Nonparametric Kernel Density Estimation

Assume that we have collected N data samples as the training data set, $(\mathbf{c}_n, \mathbf{s}_n), n = 1, 2, ..., N$, where $\mathbf{s}_n = [s_{n,1}, s_{n,2}, ..., s_{n,M}]^T$ is the M-dimensional fingerprint vector of the n^{th} training sample, and $\mathbf{c}_n = [x_n, y_n]^T$ is the 2-D coordinates of the training location at which the n^{th} sample is taken. Note that, different RSS vectors taken at the same training location are treated as different training samples. Also, let the vectors, $\mathbf{s} = [s_1, s_2, ..., s_M]^T$ and $\mathbf{c} = [x, y]^T$, denote the on-line RSS vector and the actual device location coordinates, respectively. For the convenience of discussions, let the column vector $\mathbf{u} = [x, y, s_1, s_2, ..., s_M]^T$, and $\mathbf{u}_n = [x_n, y_n, s_{n,1}, s_{n,2}, ..., s_{n,M}]^T$. The dimension of the vectors $\mathbf{u}$ and $\mathbf{u}_n$ is therefore $D = M + 2$.

In order to approximate the joint PDF of $\mathbf{c}$ and $\mathbf{s}$, $f_{\mathbf{c},\mathbf{s}}(\mathbf{u})$, which is now the multivariate PDF of $\mathbf{u}$, a kernel function, $K_{\mathbf{H}_n}(\mathbf{u} - \mathbf{u}_n) = \frac{1}{|\mathbf{H}_n|} K(\mathbf{H}_n^{-1} \cdot (\mathbf{u} - \mathbf{u}_n))$, can be placed at each training sample $\mathbf{u}_n$, where the choice of $K(\mathbf{z})$ determines the functional form of the kernel, and the "bandwidth matrix",

$\mathbf{H}_n$, controls the spread and orientation of the kernel function. Therefore, the multivariate density estimation of $f_{\mathbf{c},\mathbf{s}}(\mathbf{u})$ is,

$$\hat{f}_{\mathbf{c},\mathbf{s}}(\mathbf{u}) = \frac{1}{N}\sum_{n=1}^{N} K_{\mathbf{H}_n}(\mathbf{u}-\mathbf{u}_n). \tag{1}$$

In this paper, we have adopted the popular Gaussian kernel,

$$K(\mathbf{z}) = \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}\mathbf{z}^T \cdot \mathbf{z}\right). \tag{2}$$

The choice of the bandwidth matrix, $\mathbf{H}_n$, is critical to the density estimation. In this paper, we have adopted the local adaptive bandwidth selection method introduced in [10], which is briefly described below.

First, we compute a fixed "pilot bandwidth matrix", $\mathbf{H}' = b \cdot \mathbf{R}^{1/2}$. The scalar, $b = (\frac{4}{2D+1})^{1/(D+4)} \cdot N^{-1/(D+4)}$, is the optimal plug-in bandwidth when the distribution of the underlying D-dimensional vector elements are independent and Normal [10]. The matrix, $\mathbf{R}$, is the sample covariance matrix of the vectors, $\mathbf{u}_n, n = 1, 2, ..., N$. It is used to reduce the inter-dependence between elements in the vector $\mathbf{u}$. Using $\mathbf{H}'$ and (1), a "pilot density" value, $\hat{f}'(\mathbf{u}_n)$, for each training data point $\mathbf{u}_n$ can be computed. Let $g = \sqrt[N]{\prod_{n=1}^{N} \hat{f}'(\mathbf{u}_n)}$ be the geometric mean of the pilot density values. The local adaptive bandwidth for the training data sample, $\mathbf{u}_n$, is then, $\mathbf{H}_n = (\frac{\hat{f}'(\mathbf{u}_n)}{g})^{-1/2} \cdot \mathbf{H}'$ [10].

Similar to (1), $f_{\mathbf{s}}(\mathbf{s})$ can be estimated as,

$$\hat{f}_{\mathbf{s}}(\mathbf{s}) = \frac{1}{N}\sum_{n=1}^{N} K_{\mathbf{H}_n^s}(\mathbf{s}-\mathbf{s}_n). \tag{3}$$

Obtaining $\mathbf{H}_n^s$ is easy given that $\mathbf{H}_n$ is already computed. This is because, theoretically,

$$\begin{aligned} \hat{f}_{\mathbf{s}}(\mathbf{s}) &= \int \hat{f}_{\mathbf{c},\mathbf{s}}(\mathbf{u})\ d\mathbf{c} \\ &= \frac{1}{N}\sum_{n=1}^{N} \int K_{\mathbf{H}_n}(\mathbf{u}-\mathbf{u}_n)\ d\mathbf{c}, \end{aligned} \tag{4}$$

where each $K_{\mathbf{H}_n}(\mathbf{u}-\mathbf{u}_n)$ is equivalently a multivariate Gaussian PDF characterized by mean vector $\mathbf{u}_n$ and covariance matrix $\mathbf{H}_n\mathbf{H}_n^T$. Therefore, each integration in (4) results in a marginal Gaussian PDF characterized by mean vector $\mathbf{s}_n$, and the $M \times M$ covariance matrix $\mathbf{\Phi}_n^{\mathbf{ss}}$, which is a sub-matrix of $\mathbf{H}_n\mathbf{H}_n^T$ corresponding to the auto-covariance of $\mathbf{s}$, i.e.,

$$\mathbf{H}_n\mathbf{H}_n^T = \begin{bmatrix} \mathbf{\Phi}_n^{\mathbf{cc}} & \mathbf{\Phi}_n^{\mathbf{cs}} \\ \mathbf{\Phi}_n^{\mathbf{sc}} & \mathbf{\Phi}_n^{\mathbf{ss}} \end{bmatrix}. \tag{5}$$

Therefore, $\mathbf{H}_n^s = [\mathbf{\Phi}_n^{\mathbf{ss}}]^{1/2}$.

III. Theoretical Error Performance Analysis

Recall that $\mathbf{c} = [x, y]^T$ is the actual on-line device location, which is unknown. Let $\tilde{\mathbf{c}} = [\tilde{x}, \tilde{y}]^T$ be the location estimate provided by any one of the aforementioned fingerprint-based algorithms. The error vector from the location estimate to the actual device location is defined as,

$$\mathbf{e} = [x-\tilde{x},\ y-\tilde{y}]^T. \tag{6}$$

Let ρ and θ be the length and angle of the error vector $\mathbf{e}$, respectively. We have,

$$[\rho\cos\theta,\ \rho\sin\theta]^T = [x-\tilde{x},\ y-\tilde{y}]^T. \tag{7}$$

And hence,

$$[x,\ y]^T = [\tilde{x}+\rho\cos\theta,\ \tilde{y}+\rho\sin\theta]^T. \tag{8}$$

The Jacobian matrix $\mathbf{J}$ is,

$$\mathbf{J} = \begin{bmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\rho\sin\theta \\ \sin\theta & \rho\cos\theta \end{bmatrix}. \tag{9}$$

Therefore, the determinant of $\mathbf{J}$ is simply ρ.

If $f_{\mathbf{c},\mathbf{s}}(\mathbf{c}, \mathbf{s})$ is the joint PDF of the on-line RSS vector $\mathbf{s}$ and the actual device location $\mathbf{c}$, we can perform the transformation from $[x,\ y]^T$ to $[\rho,\ \theta]^T$ as follows.

$$f_{[\rho,\ \theta]^T,\mathbf{s}}([\rho,\ \theta]^T, \mathbf{s}) = f_{\mathbf{c},\mathbf{s}}([\tilde{x}+\rho\cos\theta,\ \tilde{y}+\rho\sin\theta]^T, \mathbf{s}) \cdot \rho. \tag{10}$$

Integrating over θ, we have,

$$f_{\rho,\mathbf{s}}(\rho, \mathbf{s}) = \int_0^{2\pi} f_{[\rho,\ \theta]^T,\mathbf{s}}([\rho,\ \theta]^T, \mathbf{s})\ d\theta. \tag{11}$$

Once the joint PDF of ρ and $\mathbf{s}$ is obtained, the PDF of localization error ρ conditioned on the on-line RSS vector $\mathbf{s}$ is simply,

$$\begin{aligned} f_{\rho|\mathbf{s}}(\rho|\mathbf{s}) &= \frac{f_{\rho,\mathbf{s}}(\rho, \mathbf{s})}{f_{\mathbf{s}}(\mathbf{s})} \\ &= \frac{\int_0^{2\pi} f_{\mathbf{c},\mathbf{s}}([\tilde{x}+\rho\cos\theta,\ \tilde{y}+\rho\sin\theta]^T, \mathbf{s}) \cdot \rho\ d\theta}{f_{\mathbf{s}}(\mathbf{s})}. \end{aligned} \tag{12}$$

Substituting (1) and (3) into (12), we have,

$$\hat{f}_{\rho|\mathbf{s}}(\rho|\mathbf{s}) = \left[\sum_{n=1}^{N} K_{\mathbf{H}_n^s}(\mathbf{s}-\mathbf{s}_n)\right]^{-1} \cdot \left[\int_0^{2\pi} \sum_{n=1}^{N} K_{\mathbf{H}_n}\left(\begin{bmatrix} \tilde{x}+\rho\cos\theta \\ \tilde{y}+\rho\sin\theta \\ \mathbf{s} \end{bmatrix} - \mathbf{u}_n\right) \cdot \rho\ d\theta\right]. \tag{13}$$

The conditional probability that the localization error is less than a given distance, r_0, can then be estimated as,

$$P_0 = \hat{P}(\rho \le r_0|\mathbf{s}) = \int_{\rho=0}^{r_0} \hat{f}_{\rho|\mathbf{s}}(\rho|\mathbf{s})\ d\rho. \tag{14}$$

For a given P_0, (14) could be used to obtain the radius of the corresponding RoC (e.g. 90% RoC), which is commonly shown as a circle centering the estimated location on a map.

IV. Experimental Verifications and Discussions

A. *Testbed Setup and Experimental Equipments*

We have set up the experimental testbed in our lab, as shown in Fig. 2. Three Linksys-WRT54G wireless routers are deployed in the testbed as APs, broadcasting beacon frames periodically in channel 1, 6, and 11. A Fujitsu S6410 notebook equipped with an Intel WiFi 4965AGN adapter, is used for RSS measurements. The Linux packet sniffer, tcpdump, is used to monitor the beacon frames transmitted by the APs. The MAC addresses of APs, timestamps, and the Received Signal Strength Indicator (RSSI) values are retrieved from the radiotap header of the captured packet. Note that, although the beacon frames from the APs arrive asynchronously, we can still use the timestamps of the arriving packets to align the reported RSSI values and form RSS vectors.

The size of the testbed is approximately 130 m^2. Within the accessible area of the testbed, 125 training locations and 126 testing locations are uniformly selected, such that the spacing

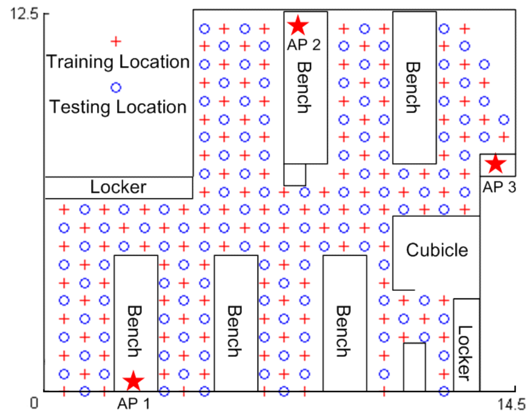

Fig. 2. Layout of the experimental testbed.

between adjacent training locations is 0.85 m and the spacing between a training location and its nearest testing location is 0.6 m. At each training location, 50 training RSS vectors are collected. At each testing location, 5 testing RSS vectors are collected, resulting in 630 testing cases in our experiment.

B. Statistical Verification

Let us denote the l^{th} testing data sample as, $(\mathbf{c}^{(l)}, \mathbf{s}^{(l)})$, and the corresponding location estimate as $\hat{\mathbf{c}}^{(l)}$, for $l = 1, 2, ..., L$, where $L = 630$. From the testing samples, we can compute the error PDF, $\hat{f}_{\rho|\mathbf{s}^{(l)}}(\rho|\mathbf{s}^{(l)})$, for each testing on-line RSS vector, $\mathbf{s}^{(l)}$. However, it is not possible to verify its correctness individually, since the pair of estimated location and ground truth corresponding to $\mathbf{s}^{(l)}$ only gives us a single error distance value for ρ. Therefore, rather than verifying each error PDF individually, we derive the overall error PDF conditioned on the entire testing set, S_{test}, i.e.,

$$\hat{f}(\rho|S_{\text{test}}) = \frac{\sum_{l=1}^{L} \hat{f}_{\rho|\mathbf{s}^{(l)}}(\rho|\mathbf{s}^{(l)}) \cdot \hat{f}_{\mathbf{s}^{(l)}}(\mathbf{s}^{(l)})}{\sum_{l=1}^{L} \hat{f}_{\mathbf{s}^{(l)}}(\mathbf{s}^{(l)})}, \qquad (15)$$

where $\hat{f}_{\rho|\mathbf{s}^{(l)}}(\rho|\mathbf{s}^{(l)})$ and $\hat{f}_{\mathbf{s}^{(l)}}(\mathbf{s}^{(l)})$ can be obtained from (13) and (3). From here, we can estimate the overall error Cumulative Density Function (CDF) and compare it with the empirical error CDF to indirectly verify the correctness of our approach. In order to predict the error CDF, we compute $\hat{f}(\rho|S_{\text{test}})$ for ρ ranging from 0 m to 10.5 m (experimentally determined), with a step size of 0.5 m. Simple rectangle-rule-based numerical integration is then applied to give the discrete error CDF prediction. For the empirical error CDF, we apply the Kaplan-Meier algorithm implemented in the MATLAB "ecdf()" function, on the actual error distances.

We have chosen the two most widely adopted fingerprint-based localization methods, namely, KNN and probabilistic approach, for our study. As shown in Fig. 3, despite the testbed difference, the experimental results are comparable with that in [3] and [4]. In both cases, the predicted error CDFs computed by our proposed scheme track the empirical error CDFs closely. In particular, comparison of the mean error distance, and error distances corresponding to 0.25, 0.50, and 0.75 overall cumulative error probabilities (CEP) between the predicted and the empirical data, are presented in Table I.

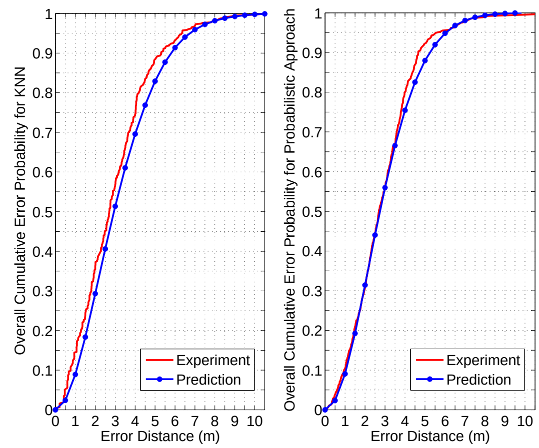

Fig. 3. Comparison of predicted and empirical error CDFs.

TABLE I
COMPARISON BETWEEN EMPIRICAL AND PREDICTED ERROR (IN METERS)

	KNN		Probabilistic	
	Empirical	Predicted	Empirical	Predicted
CEP = 0.25	1.51	1.81	1.74	1.73
CEP = 0.50	2.69	2.93	2.69	2.75
CEP = 0.75	4.02	4.37	3.79	3.97
Mean Error	2.94	3.24	2.88	2.99

V. CONCLUSION AND FUTURE WORK

In this paper, we have derived the theoretical error PDF and RoC for a generalized fingerprint-based localization system conditioned on the on-line RSS vector. We also propose to utilize the multivariate NKDE techniques in order to facilitate the computations in practical cases. The effectiveness of our proposed scheme has been verified in a realistic experimental testbed. We point out two future directions. First, we aim to study the effect of training data size on the performance of the proposed scheme. Second, the robustness of the proposed method in testbeds with more extensive area and variations of building structures should be verified.

REFERENCES

[1] C. K. Seow, S. Y. Tan, "Non-Line-of-Sight localization in multipath environments," *IEEE Trans. Mobile Computing*, vol. 7, no. 5, pp. 647-660, May 2008.

[2] P. Bahl, V. Padmanabhan, "RADAR: An in-building RF-based user location and tracking system," *IEEE INFOCOM*, vol. 2, pp. 775-784, Mar. 2000.

[3] P. Myllymäki, T. Roos, H. Tirri, P. Misikangas, J. Sievänen, "A probabilistic approach to WLAN user location estimation," *International Journal of Wireless Information Networks*, vol. 9, pp. 155-164, Jul. 2002.

[4] A. Kushki, K. N. Plataniotis, A. N. Venetsanopoulos, "Kernel-based positioning in wireless local area networks," *IEEE Trans. Mobile Computing*, vol. 6, no. 6, pp. 689-705, Jun. 2007.

[5] H. Wu, M. Siegel, R. Stiefelhagen, J. Yang, "Sensor fusion using Dempster-Shafer theory," *IEEE IMTC*, vol. 1, pp. 7-12, 2002.

[6] N. Swangmuang, P. Krishnamurthy, "Location fingerprint analyses toward efficient indoor positioning," *IEEE PerCom.*, pp. 100-109, Mar. 2008.

[7] J. Yang, Y. Chen, "A theoretical analysis of wireless localization using RF-based fingerprint matching," *IEEE IPDPS*, pp. 1-6, Apr. 2008.

[8] Y. Gwon, R. Jain, T. Kawahara, "Robust indoor location estimation of stationary and mobile users," *IEEE INFOCOM 2004*, vol. 2, pp. 1032-1043, Mar. 2004.

[9] A. Kushki, K. N. Plataniotis, A. N. Venetsanopoulos, "Intelligent dynamic radio tracking in indoor wireless local area networks," *IEEE Trans. Mobile Computing*, vol. 9, no. 3, pp. 405-419, Mar. 2010.

[10] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, 1986.