Doi: 10.31276/VJSTE.61(1).03-08

Gaussian Mixture Modeling for Wi-Fi fingerprinting based indoor positioning in the presence of censored data

Trung Kien Vu[1*], Hung Lan Le[2]

[1]*Faculty of Electronics, Hanoi University of Industry*
[2]*National Center for Technological Progress*

Received 4 September 2018; accepted 6 December 2018

Abstract:

In complex indoor environments, due to the attenuation of the signal and the changing surrounding environment, the censoring and multi-component problems may be present in the observed data. Censoring refers to the fact that sensors on portable devices cannot measure Received Signal Strength Index (RSSI) values below a specific threshold, such as -100 dBm. The multi-component problem occurs when the measured data varies due to obstacles and user directions, whether the door is closed or open, etc. By accounting for these problems, this paper proposes to model the RSSI probability density distributions using the Censoring Gaussian Mixture Model (C-GMM) and develop the Expectation-Maximization (EM) algorithm to estimate the parameters of this model in the offline phase of the Wi-Fi fingerprinting based Indoor Positioning Systems (IPS). The simulation results demonstrate the effectiveness of the proposed method.

Keywords: **censored data, EM algorithm, fingerprinting, Gaussian Mixture Model, IPS.**

Classification numbers: **1.3, 2.3**

Introduction

With the popularity of wireless local area networks (WLAN), Wi-Fi based indoor positioning techniques are widely used for indoor user localization. Most popular Wi-Fi positioning methods use the received signal strength indication (RSSI). Among available approaches, fingerprinting appears to be the most feasible method for positioning in the indoor environment [1]. This method estimates the position of an object and relies on training data from a set of reference points (RP) with known locations. Fingerprinting-based methods consist of two phases, namely the offline phase and the online phase. In the offline phase, the training data (i.e., RSSI) are collected at the RPs and used to build the database, which is often called the radio map. During the online phase, the online measurements are compared against the training data at every RP. The position of the RP whose training data most closely match the online data can be regarded as the estimated position of the object.

To represent the training data in probabilistic approaches, the parametric model and nonparametric model are two basic categories which are commonly used. The systems which utilized the parametric model had more advantages than the nonparametric model [2].

The probability density function (PDF) of the observed data is assumed to be the single Gaussian in the presence of censoring and dropping problems [3, 4]. Censoring occurs due to the limited sensitivity of Wi-Fi sensors or the sensor driver, which does not intentionally report the overly weak observed signal strengths; in other words, the smart phones do not report the signal strength if it is below a specific threshold, e.g., -100 dBm with typical smart phones.

An EM algorithm was proposed to estimate the

*Corresponding author: Email: vutrungkienfee@gmail.com

parameters of censored and dropped single Gaussian data. Experimental results with real field data can demonstrate the effectiveness of this proposal relative to the others, but the multi-component was not considered.

In [5, 6], the multi-component problem has been noted. In [6], the authors illustrated that human behaviors in the measurement environment (absence, sitting or standing still, moving randomly, and moving specifically) result in the bi-modal phenomena in the experimental data. In this case, using a single Gaussian distribution to model the RSSI histogram is not appropriate. In [5], the Gaussian Mixture Model (GMM) was proposed to model the RSSI measurements. Positioning results were improved relative to the single Gaussian model. However, the censoring problem has not been considered in these studies, although they clearly occurred, as discussed in [3, 4].

In [7, 8], the authors introduced EM algorithms for parameter estimation of the grouped, truncated, and censored data. This proposal can solve the bias of parameter estimation, but the censoring and multi-component problems have not been resolved.

This paper accounts for all of the problems discussed and proposes to develop a new extended version of the EM algorithm to enhance the quality of estimated parameters in the offline phase and there by improve the performance of the Wi-Fi fingerprinting-based IPS.

Proposed methods

This section delineates the proposed method, which relies on the characteristics of the collected Wi-Fi RSSI data for enhancing the accuracy of the fingerprinting-based indoor positioning system (Fig. 1). First, a C-GMM is introduced to model the RSSI distribution in the presence of censored mixture data. Second, an extended EM algorithm is developed to estimate the parameters of this model. This algorithm is employed during the offline phase. Third, in the online phase, the localization and classification procedure is based upon the Maximum a Posteriori (MAP) method.

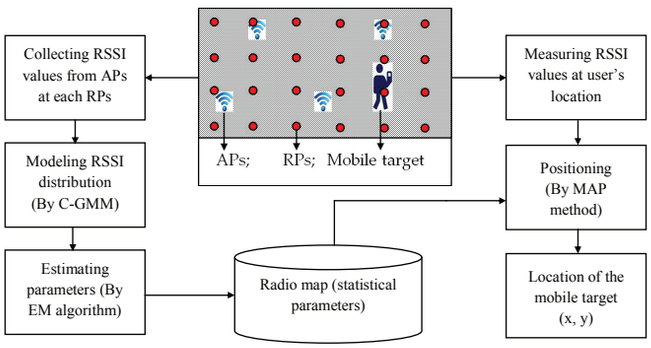

Fig. 1. Block diagram of the proposed Wi-Fi fingerprinting-based IPS.

Modeling RSSI distribution by the C-GMM

Below are several important definitions:

$\vec{y} = [y_1, y_2, \cdots, y_N]$; $y_n \in R$; $n = 1 \div N$ is the set of complete data (non-censored data), y_n are i.i.d. random variables. c is the specific threshold at which a portable device, e.g., smart phone does not report the signal strength. $\vec{x} = [x_1, \cdots, x_N]$ is the set of observable data (censored data), $x_n = \begin{cases} y_n, \text{if } y_n > c \\ c, \text{if } y_n \le c \end{cases}$.

Figure 2 illustrates the measurement model.

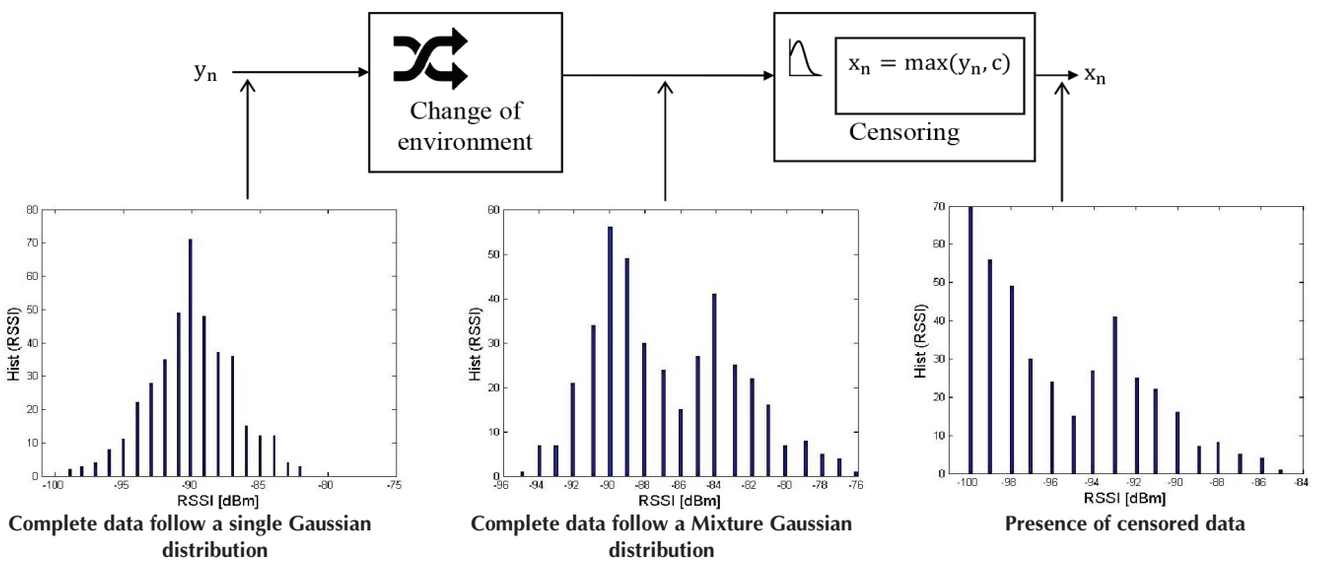

Fig. 2. Proposed measurement model.

Parameter estimation in the offline phase

$\vec{\Theta} = [w_1, \cdots, w_J; \mu_1, \cdots, \mu_J; \sigma_1, \cdots, \sigma_J]$ is the set of parameters of the GMM. The GMM includes J Gaussian components; the j^{th} component (j = 1~J) is parameterized by $\theta_j = [\mu_j, \sigma_j]$. w_j are positive mixing weights which sum up to one.

The likelihood of $\vec{y}$ following a GMM is as follows:

$$p(\vec{y};\vec{\Theta}) = \prod_{n=1}^{N}\sum_{j=1}^{J} w_j p(y_n;\theta_j) \quad (1)$$

To simplify the calculation of the summation in Eq. (1), a set of auxiliary variables $\vec{\Delta} = \begin{bmatrix} \Delta_{11} & \cdots & \Delta_{1J} \\ \vdots & \ddots & \vdots \\ \Delta_{N1} & \cdots & \Delta_{NJ} \end{bmatrix}$ were introduced, where $\Delta_{nj} = 0$ if y_n does not pertain to the j^{th} Gaussian component; otherwise, $\Delta_{nj} = 1$. Then, Eq. (1) becomes the following:

$$p(\vec{y},\vec{\Delta};\vec{\Theta}) = \prod_{n=1}^{N}\prod_{j=1}^{J}[w_j p(y_n;\theta_j)]^{\Delta_{nj}} \quad (2)$$

Then, the log-likelihood is as follows:

$$\ln[p(\vec{y},\vec{\Delta};\vec{\Theta})] = \sum_{n=1}^{N}\sum_{j=1}^{J}\Delta_{nj}\{\ln(w_j) + \ln[p(y_n;\theta_j)]\} \quad (3)$$

E-step:

The expected log-likelihood of the complete data $\vec{y}$ given the observable data $\vec{x}$ is the following:

$$\begin{aligned} Q(\vec{\Theta},\vec{\Theta}^{(k)}) &= E\{\ln[p(\vec{y},\vec{\Delta};\vec{\Theta})]|\vec{x};\vec{\Theta}^{(k)}\} \\ &= \sum_{n=1}^{N}\sum_{j=1}^{J}\int_{-\infty}^{+\infty}\Delta_{nj}\{\ln(w_j)+\ln[p(y_n;\theta_j)]\}\,p(\Delta_{nj},y_n|x_n;\Theta_j^{(k)})dy_n \end{aligned} \quad (4)$$

In Eq. (4), $\vec{\Theta}^{(k)}$ indicates the current estimated parameters, and k is the iteration index. Introducing a set of binary variables $\vec{z} = (z_1, \cdots, z_N)$, where $z_n = 0$ when the n^{th} measurement is observable ($x_n = y_n$) and z_n=1 when the n^{th} measurement is not observable ($x_n = c$), the summand in Eq. (4) can be written as follows:

$$Q(\vec{\Theta},\vec{\Theta}^{(k)}) = \sum_{n=1}^{N}\sum_{j=1}^{J}[(1-z_n)F_1 + z_n F_2] \quad (5)$$

In Eq. (5),

$$F_1 = \gamma_j(x_n;\Theta_j^{(k)})\{\ln(w_j) + \ln[\mathcal{N}(x_n;\theta_j)]\};$$

$$F_2 = \beta_j(\Theta_j^{(k)})\left\{\ln(w_j) + \int_{-\infty}^{c}\ln[\mathcal{N}(y_n;\theta_j)]\frac{\mathcal{N}(y_n;\theta_j^{(k)})}{I_0(\theta_j^{(k)})}dy_n\right\}$$

Here, $\mathcal{N}(x,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is the Gaussian probability density function, and

$$I_0(\theta_j^{(k)}) = \int_{-\infty}^{c}\mathcal{N}(y,\theta_j^{(k)})dy = \frac{1}{2}\mathrm{erfc}\left(-\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right)$$

$$I_1(\theta_j^{(k)}) = \int_{-\infty}^{+\infty}y\mathcal{N}(y,\theta_j^{(k)})dy = \mu_j^{(k)}I_0(\theta_j^{(k)}) - \frac{1}{\sqrt{2\pi}}\sigma_j^{(k)}\exp\left[-\left(\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right)^2\right]$$

$$\begin{aligned} I_2(\theta_j^{(k)}) &= \int_{-\infty}^{c}y^2\mathcal{N}(y,\theta_j^{(k)})dy \\ &= \left[(\sigma_j^2)^{(k)} + (\mu_j^2)^{(k)}\right]I_0(\theta_j^{(k)}) - \frac{1}{\sqrt{2\pi}}\sigma_j^{(k)}\mu_j^{(k)}\exp\left[-\left(\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right)^2\right] \end{aligned}$$

$$\gamma_j(x_n;\theta_j^{(k)}) = \frac{w_j^{(k)}\mathcal{N}(x_n;\theta_j^{(k)})}{\sum_{i=1}^{J}w_i^{(k)}\mathcal{N}(x_n;\theta_i^{(k)})}$$

$$\beta_j(\Theta_j^{(k)}) = \frac{w_j^{(k)}I_0(\theta_j^{(k)})}{\sum_{i=1}^{J}w_i^{(k)}I_0(\theta_i^{(k)})}$$

M-step:

Computing the derivative of the auxiliary function in Eq. (5), the following iterative parameter estimation formulae can be readily derived as follows:

$$\mu_j^{(k+1)} = \frac{\sum_{n=1}^{N}(1-z_n)\gamma_j(x_n;\Theta_j^{(k)})x_n + \beta_j(\Theta_j^{(k)})\frac{I_1(\theta_j^{(k)})}{I_0(\theta_j^{(k)})}\sum_{n=1}^{N}z_n}{\sum_{n=1}^{N}(1-z_n)\gamma_j(x_n;\Theta_j^{(k)}) + \beta_j(\Theta_j^{(k)})\sum_{n=1}^{N}z_n} \quad (6)$$

$$\begin{aligned} (\sigma_j^2)^{(k+1)} &= \frac{\sum_{n=1}^{N}(1-z_n)\gamma_j(x_n;\Theta_j^{(k)})(x_n-\mu_j)^2}{\sum_{n=1}^{N}(1-z_n)\gamma_j(x_n;\Theta_j^{(k)}) + \beta_j(\Theta_j^{(k)})\sum_{n=1}^{N}z_n} \\ &+ \frac{\beta_j(\Theta^{(k)})\left(\frac{I_2(\theta_j^{(k)})}{I_0(\theta_j^{(k)})} - \frac{2\mu_j I_1(\theta_j^{(k)})}{I_0(\theta_j^{(k)})} + \mu_j^2\right)\sum_{n=1}^{N}z_n}{\sum_{n=1}^{N}(1-z_n)\gamma_j(x_n;\Theta_j^{(k)}) + \beta_j(\Theta_j^{(k)})\sum_{n=1}^{N}z_n} \end{aligned} \quad (7)$$

$$w_j^{(k+1)} = \frac{\sum_{n=1}^{N}(1-z_n)\gamma_j(x_n;\Theta_j^{(k)}) + \beta_j(\Theta_j^{(k)})\sum_{n=1}^{N}z_n}{N} \quad (8)$$

The EM algorithm stops when the convergence criterion is satisfied or when the max iteration is achieved. After convergence, the estimated parameters are as follows:

$$\mu_j^{(k+1)} \approx \mu_j^{(k)} := \hat{\mu}_j; \sigma_j^{(k+1)} \approx \sigma_j^{(k)} := \hat{\sigma}_j; w_j^{(k+1)} \approx w_j^{(k)} := \hat{w}_j \quad (9)$$

Given equations (6÷8), both observable and censored mixture data contribute to the estimates. Moreover, if the data

are complete data (c = -∞), then these equations are reduced to the standard EM algorithm for the mixture Gaussian data [5]. On the other hand, if the data have a single Gaussian distribution and suffered from the censoring problem, by setting J = 1, over three formulae become those reported in [3]. This means that the proposal can handle both the censoring and multi-component problems presented in the Wi-Fi RSSI data.

The online classification and positioning phase

This sub-section utilizes the Maximum a Posteriori (MAP) method to perform the classification. For each reference position l_k, the parameters of the C-GMM class conditional density $p_Y(y|l_k)$ of RSSI measurements are estimated using equations (6÷9). During online classification, the MAP is used to estimate the user's location. First, the posterior is calculated as follows:

$$P(\ell_k|\vec{x}) = \frac{\prod_{i=1}^{N_{AP}} p(x_i|\ell_k)P(\ell_k)}{\sum_{k'=1}^{K}\prod_{i=1}^{N_{AP}} p(x_i|\ell_{k'})P(\ell_{k'})} \quad (10)$$

In Eq. (10), K and N_{AP} represent the total number of RPs and APs, respectively. x_i is the online measurement from i^{th} AP, and $\vec{x}$ is the set of x_i ($i=1\div N_{AP}$). It has been considered that the RSSI measurements of different APs are independent, and the prior $P(l_k)$ is equal for all locations.

The likelihood $p(x_i \mid l_k)$ can be calculated as follows:

$$p(x_i|\ell_k) = \begin{cases} \sum_{j=1}^{J} \hat{w}_{k,i,j}\mathcal{N}(x_i;\hat{\theta}_{k,i,j}), & \text{if } x_i > c \\ \sum_{j=1}^{J} \hat{w}_{k,i,j} I_0(c;\hat{\theta}_{k,i,j}), & \text{if } x_i = c \end{cases} \quad (11)$$

In Eq. (11), $\hat{\theta}_{k,i,j}, \hat{w}_{k,i,j}$ are estimated parameters at the k^{th} RP of the i^{th} AP in the offline phase.

The estimated position of the mobile object is obtained by the following:

$$\hat{\ell}(\vec{x}) = \frac{\sum_{k\in P}\ell_k p(\ell_k|\vec{x})}{\sum_{k\in P} p(\ell_k|\vec{x})} \quad (12)$$

Simulation results and discussion

Parameter estimation

To evaluate the effectiveness of the proposed EM algorithm, complete data $\vec{y}$ with the following parameters has been generated (true parameters):

$$N = 1000; J = 2; [w_1, w_2] = [0.5, 0.5]; [\sigma_1, \sigma_2] = [3, 4];$$
$$[\mu_1, \mu_2] = [-90, -80].$$

Observable data $\vec{x}$ are performed censoring as follows: $x_n = \max(y_n, c)$. The censoring threshold c was changed from $\mu_1 - 2\sigma_1$ to $\mu_1 + 2\sigma_1$. Table 1 indicates the mean of Kullback Leibler (KL) divergence [9] between true parameters and estimated parameters after 1,000 experiments.

Table 1. Parameter estimation compared by mean of KL using the Monte Carlo sampling method.

C (dBm)	Standard EM algorithm for GMM [5]	After [3]	Proposed EM algorithm for C-GMM
-96	0.0018	0.0664	0.0016
-93	0.0329	0.0679	0.0031
-90	3.1491	0.0798	0.0092
-87	5.6358	0.0886	0.0124
-84	7.2847	0.0972	0.0473

As is evident, when $c = \mu_1 - 2\sigma_1 = -96$, data nearly do not suffer from censoring (almost complete); the proposal and the standard EM algorithm for GMM produced the same results. However, when c changes from -93 to -84, the proposed EM algorithm introduces improved results.

Positioning accuracy

To evaluate the effectiveness of the proposed approach in the Wi-Fi fingerprinting-based IPS, a floor plan with 100 RPs (small red circles) and 10 APs (green circles) has been generated, as illustrated in Fig. 3. The first experiment was setup as follows: In the offline phase, 400 measurements are collected for each RP. The measured data at 50% of the training positions (RPs) follow the single Gaussians, randomly; the rest follows the GMMs, and the number of components is J = 2, 3, 4, 5, 6, respectively (10% for each model). Measured data at RPs were computed by the log-distance path loss model and by adding a Gaussian with a mean of zero and a standard deviation of two for reflecting the fluctuation of the signal [10]. The limited sensitivity of the Wi-Fi sensor was set to -100 dBm (c = -100). The radio map was developed by employing equations (6÷9) with J=4 and methods, which were proposed in [3, 5]. For the online localization phase, 100 simulations were performed. Each simulation, one online measurement per position was generated in the same scenarios with the training data, and the MAP method was used for computing the final position estimate, as presented in sub-section 2.3.

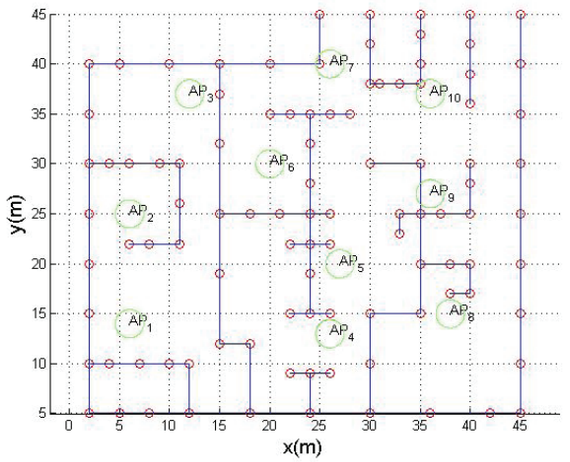

Fig. 3. The computer-generated floor plan.

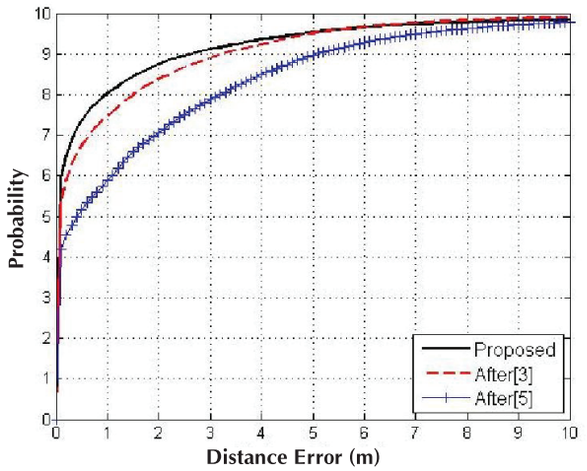

Fig. 4. Comparison of positioning results when the observable training data ratio was 69.77%, the observable online data ratio was 69.82%.

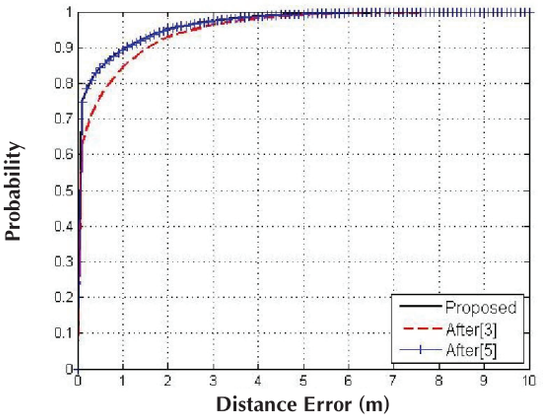

Fig. 5. Comparison of positioning results when the observable training and online data ratio were 100%.

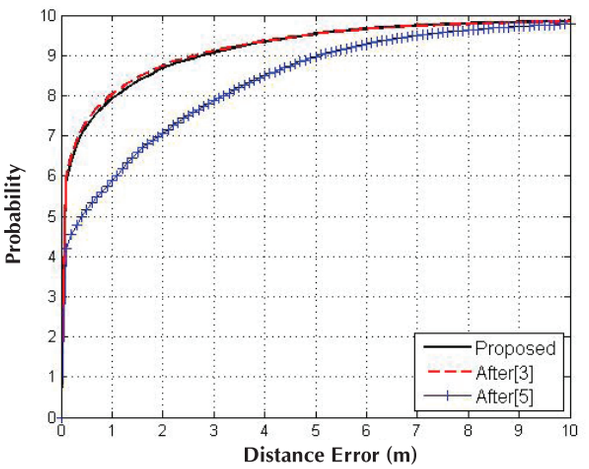

Fig. 6. Comparison of positioning results when the observable training data ratio was 69.24% the observable online data ratio was 69.74%.

Figure 4 illustrates the probability that the positioning error is lower than a specific distance. The estimated position is the specific position at which the mobile target had collected the online measurements. The plots in the figure are computed by averaging the positioning results of 100 simulations. It is evident that the proposed method outperforms the others, particularly when the error distance is smaller than 2 meters. In term of Wi-Fi fingerprinting-based indoor positioning, while the proposal in [3] is unable to solve the multi-component problem in the observed data, authors of [5] have not considered the censoring problem in their research. This simulation result demonstrates that the proposal can cope with the phenomena presented in the measured Wi-Fi RSSI data.

In the second experiment, data were gathered in the same manner as in the first experiment, but the limited sensitivity of the Wi-Fi sensor was changed to a value which is smaller than the smallest value of collected Wi-Fi RSSI. This means that collected data at all RPs of all APs are complete. Fig. 5 validated that the proposal and the standard EM algorithm for GMM [5] presented the same results, which means that the C-GMM is still appropriate to model complete mixture data.

The experiment setup for the results in the Fig. 6 is the same as in the first experiment; however, the measured data all RPs of all APs follow the single Gaussian distribution and exert an influence on censoring. It is apparent that the approach nevertheless works as effectively as the method proposed in [3].

Moreover, Table 2 indicates the properties of the Mean Distance Error (MDE) of the three experiments.

Table 2. MDE (m).

	After [3]	**After** [5]	**Proposed**
Experiment 1	1.3428	1.6321	1.0452
Experiment 2	1.0402	0.4856	0.4863
Experiment 3	0.9920	1.7395	1.0012

Conclusions

This paper has presented and analyzed an EM algorithm for estimating the parameters of the GMM in the presence of censored mixture data. The results have demonstrated that the algorithm delivers less biased and more efficient estimates relative to existing methods. Further, it has illustrated the enhancement of the Wi-Fi fingerprinting-based indoor positioning system when the novel method was employed. Experimental results on artificial data verify that the proposal produces optimal accuracy of positioning among available approaches. Future research will make substantial use of labor work for gathering real data and evaluate the proposed method. In addition, reducing the computational cost in the online phase and using sensors on the portable devices to predict the current position of the moving objects can significantly enhance the real-time performance of the IPS.

The authors declare that there is no conflict of interest regarding the publication of this article.

REFERENCES

[1] L. Mainetti, L. Patrono, and I. Sergi (2014), "A survey on indoor positioning systems", *Proceedings of 22nd Int. Conf. on Software, Telecommunications and Computer Networks* (SoftCOM).

[2] K. Kaemarungsi and P. Krishnamurth (2004), "Modeling of indoor positioning systems based on location fingerprinting", *Proceedings of the INFOCOM*, Hong Kong.

[3] K. Hoang and R. Haeb-Umbach (2013), "Parameter estimation and classication of censored Gaussian data with application to Wi-Fi indoor positioning", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Vancouver.

[4] K. Hoang, J. Schmalenstroeer, and R. Haeb-Umbach (2015), "Aligning training models with smartphone properties in Wi-Fi fingerprinting based indoor localization", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Brisbane.

[5] M. Alfakih, M. Keche, and H. Benoudnine (2015), "Gaussian mixture modeling for indoor positioning Wi-Fi systems", *3rd Int. Conf. on Control, Engineering and Information Technology* (CEIT), Tlemcen, Algeria.

[6] Jiayou Luo and Xingqun Zhan (2014), "Characterization of smart phone received signal strength indication for WLAN indoor positioning accuracy improvement", *Journal of Networks*, **9(3)**, pp.739-746.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, series B (Methodological), pp.1-38.

[8] G. Lee and C. Scott (2012), "EM algorithms for multivariate Gaussian mixture models with truncated and censored data", *Computational Statistics & Data Analysis*, **56(9)**, pp.2816-2829.

[9] J.R. Hershey and P.A. Olsen (2007), "Approximating the Kullback Leibler divergence between Gaussian mixture models", *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Honolulu.

[10] C. Gustafson, T. Abbas, D. Bolin, and F. Tufvesson (2015), "Statistical modeling and estimation of censored pathloss data", *IEEE Wireless Comm. Letters*, **4(5)**, pp.569-572.