Performance Enhancement of Wi-Fi Fingerprinting-based IPS by Accurate Parameter Estimation of Censored and Dropped Data

Trung Kien VU [1], Manh Kha HOANG [1], Hung Lan LE [2]

[1] Faculty of Electronics Engineering, Hanoi University of Industry, Hanoi 100000,Vietnam
[2] National Center for Technological Progress, Hanoi 129000, Vietnam

{kien.vu, khahoang}@haui.edu.vn, lanlh1960@nacentech.vn

Submitted April 19, 2019 / Accepted September 3, 2019

Abstract. *In complex indoor environments, the censoring, dropping, and multi-component problems may present in the observable data. This is due to the attenuation of signals, the unexpected operation of equipments, and the changing surrounding environment. Censoring refers to the fact that sensors on portable devices are unable to measure Received Signal Strength Index (RSSI) values below a certain threshold, for example, −100 dBm with typical smart phones. Dropping means that, occasionally, RSSI measurements of Wi-Fi access points are not available, although their value is clearly above the censoring threshold. The multi-component problem occurs when the measured data varies due to obstacles as well as user directions; doors closed or open; and so forth. Taking these problems into consideration, this paper proposes a novel approach to enhance the performance of the Wi-Fi Fingerprinting based Indoor Positioning System (WF-IPS). The proposed method is verified through simulated data and real field data. The experimental results show that our proposal outperforms the other state-of-the-art WF-IPS approaches both in positioning accuracy and computational cost.*

Keywords

IPS, EM algorithm, censored and dropped data, Gaussian Mixture Model (GMM), Bayesian Information Criterion (BIC)

1. Introduction

Modeling RSSI distribution in the WF-IPS: Nowadays, indoor positioning, based on Wi-Fi fingerprinting, has attracted significant interest due to its potential to obtain high accuracy at low costs [1], [2]. This method can be well-formulated as pattern recognition system, which operates in two phases: training phase and online localization/ classification phase [3]. In the training phase, measured data at the reference points (RP) from available Wi-Fi access points (AP) are collected to build the database. In the classification phase, the online measured data is compared to the training, and the target position is determined according to the similarity between online data and training data. There are two common methods to be used in the classification phase: the deterministic approach [4], [5], and the probabilistic approach [2], [6–9]. As reported in previous studies, the probabilistic approach seems to outperform the deterministic approach. In the probabilistic approach, the parametric model [6], [8] and nonparametric model [7] can be used to represent the training RSSI distributions. As reported in [10], systems utilizing the parametric model perform better. The reason is the parametric model can take into account the missing signal strength values during the training phase (due to a finite number of measurements) to smooth the distribution shape. This helps to avoid zero probability at those signal strength points. Some studies showed that the majority of RSSI histograms fitted very well with the Gaussian distribution if sufficient samples have been collected [6], [11], [12] while others proposed to model RSSI distribution by the Gaussian Mixture Model (GMM) [8], [11], [13]. It has noted that the GMM extended the single Gaussian process with ability of modeling multi-modal data. Therefore, the GMM is the most feasible parametric model for modeling Wi-Fi RSSI data.

The characteristics of the measured Wi-Fi RSSI data: In this work, the characteristics of the real field Wi-Fi RSSI data have been investigated. In [6], [9], [14], authors have recognized the censoring and dropping problem in the observable data. The Gaussian distribution was chosen as the model for data throughout [6], [9]. In [8], [11], [13], the multi-component problem was noticed. In [11], authors showed that human behaviors in the measurement environment (absence, sitting/standing still, moving randomly and moving specifically) led to the bi-modal phenomena in the experimental data. In this case, using the Gaussian distribution to model the RSSI histogram is not appropriate. In [8], [13], the authors used the GMM to model data owing to the changes in the surrounding environment which would obviously change the measured signal

strength. However, in [8], [11], [13] the censoring and dropping problems have not been considered. With respect to the above mentioned issues, this paper proposes to utilize the GMM including censored and dropped observations (CD-GMM) [15] to model the distribution of the Wi-Fi RSSI data.

Parameter estimation and model selection: For estimating parameters of a probabilistic model in the presence of missing data, the EM algorithm [16], [17] is one of the most feasible estimators among available novel approaches. The results in [15] showed the effectiveness of the EM algorithm for the CD-GMM. However, this approach can only be used for parameter estimation of the GMM with known number of components. In WF-IPS, since the real training data collected at RPs from APs have different distributions, it is necessary to develop an appropriate method to estimate the number of components of the CD-GMM instead of fixing it to a specific number. In [8], the Akaike Information Criterion (AIC) is used to determine the best number of components for the GMM. Authors in [18] proposed a penalized likelihood method for model selection of finite multivariate Gaussian mixture models. This method involves a light computational load and is attractive when there are many possible candidate models. A model selection criterion, based on the sum of weighted real parts of all log-characteristic functions (SWRLCF), was introduced in [19]. This method is suited for large sample applications. The approaches introduced in [8], [18], [19] can select the number of components of GMMs consistently when data are complete. However, in the complex indoor environment, the collected Wi-Fi data are often incomplete due to the censoring and dropping [6], [9], [14].

Scalability of WF-IPS: It is reported that WF-IPS is low cost and easy to deploy [20]. However, once deploying in a large scale, due to a huge amount of RPs, the execution time to produce the positioning results of the IPS must be considered carefully [21]. Therefore, reducing execution time, while maintaining the positioning accuracy, is a challenge when developing a WF-IPS.

The target of this research is to enhance the performance of a WF-IPS including positioning accuracy and computational cost. For the above reasons, this paper proposes a novel method to precisely estimate the number of components as well as their parameters of a GMM in the presence of censored and dropped data. The proposed approach is the combination of Bayesian Information Criterion (BIC) to determine the best number of components of the CD-GMM and EM algorithm to deal with censored and dropped data.

In the following, the characteristics of Wi-Fi RSSI data collected in the indoor environment are investigated; and based on these, we propose a new parameter estimation and model selection algorithm in which the censoring, dropping and multi-component problems are considered (Sec. 2). Section 3 is the evaluation of the effectiveness of proposed approaches in the WF-IPS. The paper is concluded in Sec. 4.

2. Proposed Method

2.1 Parameter Estimation and Model Selection Based on the Characteristics of Real Field Collected Wi-Fi RSSI Data

The characteristics of Wi-Fi RSSI data: According to our data investigation, we have detected three problems with the Wi-Fi RSSI data, namely censoring, dropping and multi-component (Fig. 1) which strongly affect the accuracy of parameter estimates and, consequently, the positioning results.

In Fig. 1(a), (b) and (c), the RP where RSSI measurements were taken is close enough to the AP, hence, all the RSSI values are above the limited sensitivity of the Wi-Fi chipset (in our data set, it is −100 dBm). The distribution of RSSI shown in Fig. 1(b), (c) seems to be drawn from more than one Gaussian component. The reason is the measurements were gathered in varying states such as door opening/closing, the direction of the person who handled collecting equipment (smart phone) had been changed. In these three cases, the distribution of data can be modeled by the standard GMM with one, two and three Gaussian components, respectively.

However, a large number of readings belong to one of the three latter cases which are shown in Fig. 1(d), (e), (f)

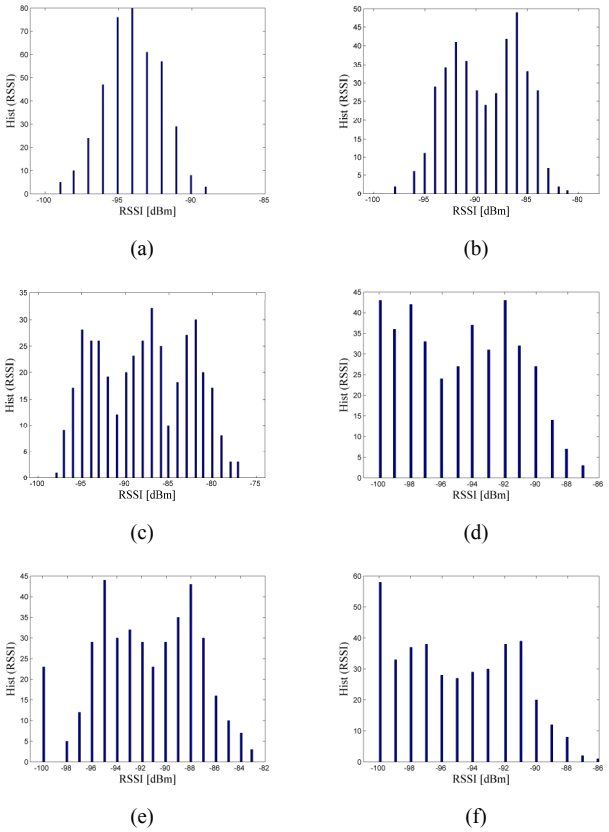

Fig. 1. Histogram of real field data collected from an AP at a RP: (a), (b), (c) complete data; (d) censored data; (e) dropped data; (f) censored and dropped data.

where the unobservable data are presented in the collected data. In Fig. 1(d), training data were taken at a RP far away from an AP, therefore, a certain number of samples were unobservable due to the censoring problem reported, as can be seen by the histogram bar, at −100 dBm. Figure 1(e) shows the presence of dropped data owing to the temporary switching off state of an AP for the energy-saving purpose. In Fig. 1(f), one part of the histogram is missing, which is similar to Fig. 1(d), but the amount of unobservable measurements seems to be larger than the missing part in Fig. 1(e) because the Wi-Fi data might experience both censoring and dropping.

Parameter estimation: Considering all the above phenomena presented in Wi-Fi RSSI data and assuming that data collected from different APs are independent, we propose to model the distribution of data gathered at a RP from each AP by the CD-GMM and estimate its parameters by utilizing the EM algorithm as follows:

E-Step:

Let $\mathbf{y} = [y_1,\ldots,y_N]$ be the set of unobservable, non-censored, non-dropped data (complete data) representing the Wi-Fi RSSIs collected at a RP from an AP, $y_n \in \mathbb{R}$, $n = 1,\ldots, N$, N is the number of elements in $\mathbf{y}$; c is the specific threshold at which a portable device (e.g., smart phone) does not report the signal strength; $\mathbf{x} = [x_1,\ldots,x_N]$ is the set of observable data, censored, possibly dropped data (incomplete data), $x_n = y_n$ only if $y_n > c$ and the dropping problem does not occur, $x_n = c$ means that $y_n \le c$ or the dropping problem occurs.

Since the observations can be considered as incomplete data, instead of computing the likelihood directly, the expected value of log-likelihood of complete data given the observations and old estimated parameters are calculated:

$$\mathrm{Q}\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k)}\right) = E\left\{\ln\left[\mathrm{p}\left(\mathbf{y};\boldsymbol{\Theta}\right)\right]\middle|\mathbf{x};\boldsymbol{\Theta}^{(k)}\right\}. \tag{1}$$

For calculating $\mathrm{Q}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k)})$ in (1), three cases are considered: Data are observable, data are unobservable due to the censoring problem and data are unobservable due to the dropping problem. Finally, the detailed calculation of $\mathrm{Q}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k)})$ are given in (2) with some definitions are as follows:

$v_n\,(n = 1,\ldots, N)$ are hidden binary variables indicating whether y_n is unobservable ($v_n = 1 \approx x_n = c$) or observable ($v_n = 0 \approx x_n = y_n$);

$\boldsymbol{\Theta} = \{[w_1,\ldots, w_J]; [\mu_1,\ldots, \mu_J]; [\sigma_1,\ldots, \sigma_J]\}$ is the set of parameters of the GMM; J is the number of Gaussian components; $w_j\,(j = 1,\ldots, J)$ are positive mixing weights which sum up to 1; $\theta_j = [\mu_j, \sigma_j]$ is the set of parameters of the j^{th} Gaussian component; ψ is the dropped rate; k is the iteration index; $\mathcal{N}(\ldots;\theta)$ is the Gaussian distribution parameterized by θ.

The other terms in (2) are given in (3)–(6).

M-Step:

Re-estimated parameters at the $(k+1)^{\text{th}}$ iteration are obtained by computing the partial derivatives of $\mathrm{Q}(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k)})$ in (2) w.r.t. the elements of μ_j, σ_j, w_j, ψ and setting them to zero, then we arrived at formulae given in (7)–(10).

In (7) and (8), the notations $\mathrm{I}_1(\theta_j^{(k)})$ and $\mathrm{I}_2(\theta_j^{(k)})$ are given in (11), (12).

As can be seen in (2), (7)–(10), collected data, including observable, censored and dropped samples are contributed to the estimate, simultaneously. This means the proposed EM algorithm can deal with all the mentioned phenomena presented in collected data.

$$\begin{aligned}&\mathrm{Q}\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(k)}\right)\\ &= \sum_{n=1}^{N}\sum_{j=1}^{J}\left(1-v_n\right)\Upsilon\left(x_n;\Theta_j^{(k)}\right)\left[\ln\left(w_j\right)+\ln\left(1-\psi\right)+\ln\left(\mathcal{N}\left(x_n;\theta_j\right)\right)\right]\\ &+\sum_{n=1}^{N}\sum_{j=1}^{J}v_n\beta\left(\Theta_j^{(k)}\right)\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\int_{-\infty}^{c}\left[\ln\left[w_j\left(1-\psi\right)\right]+\ln\left(\mathcal{N}\left(y_n;\theta_j\right)\right)\right]\frac{\mathcal{N}\left(y_n;\theta_j^{(k)}\right)}{\mathrm{I}_0\left(\theta_j^{(k)}\right)}\mathrm{d}y_n+\sum_{n=1}^{N}\sum_{j=1}^{J}v_n w_j^{(k)}\left[1-\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\right]\ln\left(\psi\right),\end{aligned} \tag{2}$$

$$\Upsilon\left(x_n;\Theta_j^{(k)}\right)=\frac{w_j^{(k)}\mathcal{N}\left(x_n;\theta_j^{(k)}\right)}{\sum_{j=1}^{J}w_j^{(k)}\mathcal{N}\left(x_n;\theta_j^{(k)}\right)}, \tag{3}$$

$$\mathrm{I}_0\left(\theta_j^{(k)}\right)=\int_{-\infty}^{c}\mathcal{N}\left(y_n;\theta_j^{(k)}\right)\mathrm{d}y_n=\frac{1}{2}\mathrm{erfc}\left(-\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right), \tag{4}$$

$$\beta\left(\Theta_j^{(k)}\right)=\frac{w_j^{(k)}\,\mathrm{I}_0\left(\theta_j^{(k)}\right)}{\sum_{j=1}^{J}w_j^{(k)}\,\mathrm{I}_0\left(\theta_j^{(k)}\right)}, \tag{5}$$

$$\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)=\frac{\left(1-\psi^{(k)}\right)\sum_{j=1}^{J}w_j^{(k)}\,\mathrm{I}_0\left(\theta_j^{(k)}\right)}{\left(1-\psi^{(k)}\right)\sum_{j=1}^{J}w_j^{(k)}\,\mathrm{I}_0\left(\theta_j^{(k)}\right)+\psi^{(k)}}, \tag{6}$$

$$\mu_j^{(k+1)}=\frac{\sum_{n=1}^{N}\left(1-v_n\right)\Upsilon\left(x_n;\Theta_j^{(k)}\right)x_n+\beta\left(\Theta_j^{(k)}\right)\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\frac{\mathrm{I}_1\left(\theta_j^{(k)}\right)}{\mathrm{I}_0\left(\theta_j^{(k)}\right)}\sum_{n=1}^{N}v_n}{\sum_{n=1}^{N}\left(1-v_n\right)\Upsilon\left(x_n;\Theta_j^{(k)}\right)+\beta\left(\Theta_j^{(k)}\right)\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\sum_{n=1}^{N}v_n}, \tag{7}$$

$$\left(\sigma_j^2\right)^{(k+1)}=\frac{\sum_{n=1}^{N}\left(1-v_n\right)\Upsilon\left(x_n;\Theta_j^{(k)}\right)\left(x_n-\mu_j^{(k)}\right)^2+\beta\left(\Theta_j^{(k)}\right)\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\left[\frac{\mathrm{I}_2\left(\theta_j^{(k)}\right)}{\mathrm{I}_0\left(\theta_j^{(k)}\right)}-\frac{2\mu_j^{(k)}\mathrm{I}_1\left(\theta_j^{(k)}\right)}{\mathrm{I}_0\left(\theta_j^{(k)}\right)}+\left(\mu_j^{(k)}\right)^2\right]\sum_{n=1}^{N}v_n}{\sum_{n=1}^{N}\left(1-v_n\right)\Upsilon\left(x_n;\Theta_j^{(k)}\right)+\beta\left(\Theta_j^{(k)}\right)\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\sum_{n=1}^{N}v_n}, \tag{8}$$

$$w_j^{(k+1)}=\frac{\sum_{n=1}^{N}\left(1-v_n\right)\Upsilon\left(x_n;\Theta_j^{(k)}\right)+\beta\left(\Theta_j^{(k)}\right)\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\sum_{n=1}^{N}v_n+\left[1-\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\right]\sum_{n=1}^{N}v_n}{N}, \tag{9}$$

$$\psi^{(k+1)}=\frac{\left[1-\alpha\left(\boldsymbol{\Theta}^{(k)},\psi^{(k)}\right)\right]\sum_{n=1}^{N}v_n}{N}, \tag{10}$$

$$\mathrm{I}_1\left(\theta_j^{(k)}\right)=\int_{-\infty}^{c}y_n\mathcal{N}\left(y_n;\theta_j^{(k)}\right)\mathrm{d}y_n=\mu_j^{(k)}\mathrm{I}_0\left(\theta_j^{(k)}\right)-\frac{1}{\sqrt{2\pi}}\sigma_j^{(k)}\exp\left[-\left(\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right)^2\right], \tag{11}$$

$$\mathrm{I}_2\left(\theta_j^{(k)}\right)=\int_{-\infty}^{c}y_n^2\mathcal{N}\left(y_n;\theta_j^{(k)}\right)\mathrm{d}y_n=\left[\left(\mu_j^{(k)}\right)^2+\left(\sigma_j^{(k)}\right)^2\right]\mathrm{I}_0\left(\theta_j^{(k)}\right)-\frac{1}{\sqrt{2\pi}}\sigma_j^{(k)}\left(c+\mu_j^{(k)}\right)\exp\left[-\left(\frac{c-\mu_j^{(k)}}{\sqrt{2}\sigma_j^{(k)}}\right)^2\right]. \tag{12}$$

Model selection: As mentioned in the first part of this sub-section, the distribution of collected RSSIs might be drawn from one, two, three or several Gaussian components while the presented EM algorithm must use an assumption of the number of Gaussian components (J). For this reason, an extended BIC was developed to estimate the number of components in the CD-GMM as follows:

The penalty function (PF) of BIC for the GMM is

$$\mathrm{PF}_{\mathrm{BIC}}\left(\boldsymbol{\Theta}^J\right)=-2\ln\left[\mathcal{L}\left(\boldsymbol{\Theta}^J\mid\mathbf{x}\right)\right]+N_{\mathrm{Ps}}\ln\left(N\right). \tag{13}$$

In (13), $\boldsymbol{\Theta}^J$ is the set of parameters of GMM with J Gaussian components; $\mathcal{L}\left(\boldsymbol{\Theta}^J\mid\mathbf{x}\right)$ is the likelihood; N_{Ps} is the total number of parameters in the GMM; N is the number of measurements.

In the GMM, since $w_1=1-\sum_{j=2}^{J}w_j$, total number of parameters is $N_{\mathrm{Ps}}=3J-1$; in the CD-GMM, one additional parameter (ψ) is used to model RSSI distribution and then, equation (13) becomes:

$$\mathrm{PF}_{\mathrm{BIC-CD}}\left(\boldsymbol{\Theta}^J,\psi\right)=-2\ln\left[\mathcal{L}\left(\boldsymbol{\Theta}^J,\psi\mid\mathbf{x}\right)\right]+3J\ln\left(N\right). \tag{14}$$

Here, $\mathrm{PF}_{\mathrm{BIC-CD}}(\boldsymbol{\Theta}^J,\psi)$ is the PF of extended BIC, in which both observable and unobservable data are considered. The term $\ln\left[\mathcal{L}\left(\boldsymbol{\Theta}^J,\psi\mid\mathbf{x}\right)\right]$ in (14) can be calculated as follows:

$$\ln\left[\mathcal{L}\left(\boldsymbol{\Theta}^J,\psi\mid\mathbf{x}\right)\right]=\sum_{n=1}^{N}\ln\left[\sum_{j=1}^{J}w_j\,\mathrm{p}\left(x_n;\theta_j,\psi\right)\right]. \tag{15}$$

In (15), the term $\mathrm{p}(x_n;\theta_j,\psi)$ is the continuous probability density function parameterized by θ_j, ψ.

Let d_n ($n=1,\ldots,N$) be hidden binary variables indicating whether an observation (y_n) is dropped ($d_n=1$) or not ($d_n=0$), for calculating $\mathrm{p}(x_n;\theta_j,\psi)$, three cases are considered:

- Data that are observable $\left(\begin{cases}d_n=0\\y_n>c\end{cases}\approx\begin{cases}v_n=0\\x_n=y_n\end{cases}\right)$:

$$\begin{aligned}&\mathrm{p}\left(x_n;\theta_j,\psi\right)\\&=\sum_{d_n=0}^{d_n=1}\int_{-\infty}^{+\infty}\mathrm{p}\left(x_n\mid y_n,d_n;\theta_j\right)\mathrm{p}\left(y_n\mid d_n;\theta_j\right)\mathrm{P}\left(d_n;\theta_j\right)\mathrm{d}y_n\\&=\int_{c}^{+\infty}\delta\left(x_n-y_n\right)\mathrm{p}\left(x_n;\theta_j,\psi\right)\mathrm{P}\left(d_n=0;\theta_j\right)\mathrm{d}y_n\\&=\left(1-\psi\right)\mathcal{N}\left(x_n;\theta_j\right).\end{aligned} \tag{16}$$

- Data that are unobservable owing to the censoring problem $\left(\begin{cases}d_n=0\\y_n\le c\end{cases}\approx\begin{cases}v_n=1\\x_n=c\end{cases}\right)$:

$$\begin{aligned}&\mathrm{p}\left(x_n;\theta_j,\psi\right)\\&=\int_{-\infty}^{c}\delta\left(x_n-c\right)\mathrm{p}\left(y_n;\theta_j,\psi\right)\mathrm{P}\left(d_n=0;\theta_j\right)\mathrm{d}y_n\\&=\left(1-\psi\right)\mathrm{I}_0\left(\theta_j\right).\end{aligned} \tag{17}$$

- Data that are unobservable due to the dropping problem ($d_n=1\approx v_n=1$):

$$\mathrm{p}\left(x_n;\theta_j,\psi\right)=\int_{-\infty}^{c}\delta\left(y_n-c\right)\mathrm{P}\left(d_n=1;\theta_j\right)\mathrm{d}y_n=\psi. \tag{18}$$

Using equations (15)–(18), equation (14) ends up with

$$\begin{aligned}&\ln\left[\mathcal{L}\left(\Theta^J,\psi\mid x\right)\right]\\&=\sum_{n=1}^{N}\left(1-v_n\right)\ln\left[\left(1-\psi\right)\sum_{j=1}^{J}w_j\mathcal{N}\left(x_n;\theta_j\right)\right]\\&+\sum_{n=1}^{N}v_n\left\{\ln\left[\left(1-\psi\right)\sum_{j=1}^{J}w_jI_0\left(\theta_j\right)\right]+\ln\left(\psi\right)\right\}.\end{aligned} \tag{19}$$

Then, the PF of the extended BIC in (14) arrives at

$$\begin{aligned}&\mathrm{PF}_{\mathrm{BIC-CD}}(\boldsymbol{\Theta}^{J},\psi)\\&=-2\sum_{n=1}^{N}(1-v_n)\ln\left[(1-\psi)\sum_{j=1}^{J}w_j\mathcal{N}\left(x_n;\theta_j\right)\right]\\&-2\sum_{n=1}^{N}v_n\left\{\ln\left[(1-\psi)\sum_{j=1}^{J}w_j\,\mathrm{I}_0\left(\theta_j\right)\right]+\ln(\psi)\right\}\\&+3J\ln(N).\end{aligned}\tag{20}$$

Using the EM algorithm and the PF of the extended BIC calculated in (20), the proposed algorithm for parameter estimation and model selection is as follows:

Input: A set of incomplete data ($\mathbf{x}$), convergence threshold of the EM algorithm for CD-GMM (ε) and the maximum number of Gaussian components ($J_{\max}$) for calculating PFs.

Output: The estimated number of Gaussian components ($\hat{J}$) and estimated parameters ($\hat{\boldsymbol{\Theta}}^{\hat{J}},\hat{\psi}$) in the CD-GMM using to model the distribution of $\mathbf{x}$.

The details of the algorithm are shown in Fig. 2.

2.2 Positioning/Classification

The parameter estimation and model selection algorithm mentioned in Sec. 2.1 is done for all RPs and it is done for the measurements of each AP separately. Let Q and N_{AP} are the total number of RPs and APs, respectively, the final estimated parameters of the q^{th} RP (q = 1,…, Q) and the i^{th} AP (i = 1,…, N_{AP}) are denoted by $[\hat{\boldsymbol{\Theta}}_{q,i},\hat{\psi}_{q,i}]$, where $\hat{\boldsymbol{\Theta}}_{q,i}=\{[\hat{w}_{q,i,1},\hat{\theta}_{q,i,1}],...,[\hat{w}_{q,i,\hat{J}_{q,i}},\hat{\theta}_{q,i,\hat{J}_{q,i}}]\}$, $\hat{J}_{q,i}$ is the number of Gaussian components in the CD-GMM using to model RSSI distribution collected at the q^{th} RP from the i^{th} AP, estimated by applying the method proposed in Sec. 2.1.

Indoor localization can be formulated as a classification problem, where the classes are RPs. During online classification/positioning phase, to estimate the target's position, a MAP (maximum a posteriori) based classification rule is developed as follows.

First, the posterior is calculated:

$$\mathrm{p}\left(\ell_q\mid\mathbf{x}^{\mathrm{on}}\right)=\frac{\prod_{i=1}^{N_{\mathrm{AP}}}\mathrm{p}\left(x_i^{\mathrm{on}}\mid\ell_q\right)\mathrm{P}\left(\ell_q\right)}{\sum_{q'=1}^{Q}\prod_{i=1}^{N_{\mathrm{AP}}}\mathrm{p}\left(x_i^{\mathrm{on}}\mid\ell_{q'}\right)\mathrm{P}\left(\ell_{q'}\right)}.\tag{21}$$

In (21), ℓ_q is the position of the q^{th} RP; $\mathbf{x}^{\mathrm{on}}$ is a set of online measurements, x_i^{on} is the RSSI value measured from the i^{th} AP (i = 1,…, N_{AP}). We considered that RSSI measurements of different APs are independent, and the prior $\mathrm{P}(\ell_q)$ is equal for all locations.

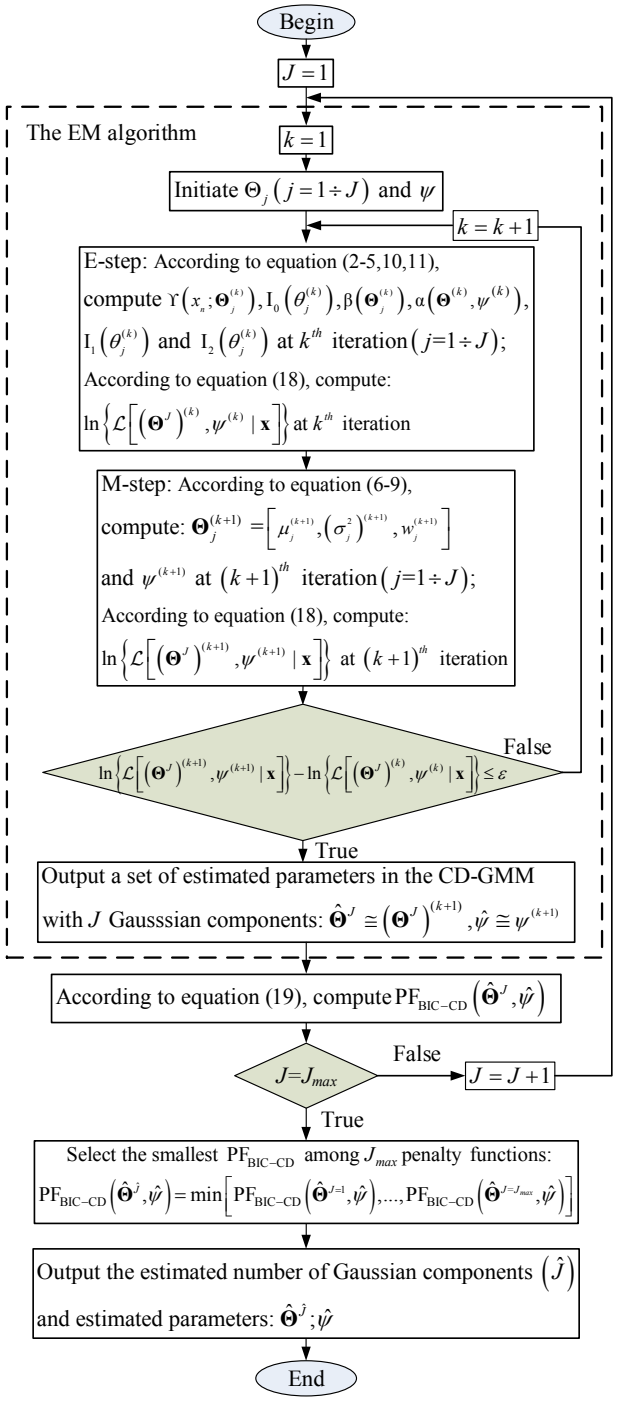

Fig. 2. The parameter estimation and model selection algorithm.

The online measurements might be suffered from censoring or dropping or both of them. Therefore, the likelihood $\mathrm{p}(x_i^{\mathrm{on}}|\ell_q)$ in (21) can be calculated for the censored and dropped Gaussian mixture data as follows:

$$\mathrm{p}\left(x_i^{\mathrm{on}}\mid\ell_q\right)=\begin{cases}\left(1-\hat{\psi}_{q,i}\right)\sum_{j=1}^{\hat{J}_{q,i}}\hat{w}_{q,i,j}\mathcal{N}\left(x_i^{\mathrm{on}};\hat{\theta}_{q,i,j}\right) & \text{if } x_i>c\\ \left(1-\hat{\psi}_{q,i}\right)\sum_{j=1}^{\hat{J}_{q,i}}\hat{w}_{q,i,j}\,\mathrm{I}_0\left(\hat{\theta}_{q,i,j}\right)+\hat{\psi}_{q,i} & \text{if } x_i=c\end{cases}\tag{22}$$

Let K_{NN} be the number of the nearest neighbors chosen among RPs by taking those with the largest posteriors. The estimated position of the mobile object is obtained by:

$$\hat{\ell}\left(\mathbf{x}^{\mathrm{on}}\right)=\frac{\sum_{q\in K_{\mathrm{NN}}}\ell_q\,\mathrm{p}\left(\ell_q\,|\,\mathbf{x}^{\mathrm{on}}\right)}{\sum_{q\in K_{\mathrm{NN}}}\mathrm{p}\left(\ell_q\,|\,\mathbf{x}^{\mathrm{on}}\right)}. \tag{23}$$

3. Results

3.1 Estimating the Number of Components in the GMM

In this simulation, our proposal mentioned in Sec. 2.1 and three other approaches [8], [18], [19] were applied to estimate the number of components in the GMM with artificial data as follows: First, a random integer number which refers to the true number of components of artificial mixture data (J) was generated within a range of 1÷4. Next, according to the value of J, one of four sets of parameters defined in Tab. 1 was selected to generate 1000 data samples (complete data). The incomplete data ($\mathbf{x}$) are the censored, possibly dropped data, gathered by changing the limited sensitivity of Wi-Fi sensors (c) and dropped rate (ψ). After 1000 experiments, different levels between the true number (J) and estimated number ($\hat{J}$) of Gaussian components were recorded in Tab. 2.

As can be seen in Tab. 2, our proposed method introduced far better results than other approaches, especially when data are suffered from censoring or dropping or both of them. This can be explained as follows: our proposed method utilized the extended version of the EM algorithm in which both observable data ($x_n = y_n$) and unobservable data ($x_n = c$) are contributed to the estimates. When data are unobservable owing to the censoring and dropping problems, this algorithm produces a lot better results compared to the standard EM algorithm introduced in [8], [18], [19]. Moreover, in the PF of AIC in [8], the PF of BIC in [18] and SWRLCF in [19], unobservable data had almost no practical contribution while they really contributed to the likelihood in PF of our proposal, as mentioned in Sec. 2.1 (equation (20)).

3.2 Positioning Accuracy

In order to evaluate the positioning accuracy of the proposed method, compared to the three state-of-art approaches [7–9] on real data, we have used the Wi-Fi RSSI data measured at 25 RPs (black dots) of an office building as shown in Fig. 3.

In the training phase, RSSI values were taken at 25 RPs (25 free positions, without wall, furniture), roughly evenly distributed, resulting in an average distance of 2.7 m between two locations. At each RP, 400 measurements were collected from each available AP. Training

Set 1 ($J=1$)	$\mu=-90;\sigma=2;w=1$
Set 2 ($J=2$)	$\mu=[-90,-84];\sigma=[2,2];w=[0.5,0.5]$
Set 3 ($J=3$)	$\mu=[-90,-84,-78];\sigma=[2,2,2];w=[0.33,0.33,0.34]$
Set 4 ($J=4$)	$\mu=[-90,-84,-78,-72];\sigma=[2,2,2,2];$ $w=[0.25,0.25,0.25,0.25]$

Tab. 1. The four sets of true parameters using to generate artificial data.

$c=$	Methods	Probability	$\psi=$ 0	0.1	0.2
−94 [dBm]	Standard EM for GMM and AIC[8]	$\mathrm{P}(J=\hat{J})$	0.28	0.01	0.01
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.21	0.48	0.45
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.51	0.51	0.54
	Standard EM for GMM and BIC[18]	$\mathrm{P}(J=\hat{J})$	0.82	0.05	0.02
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.15	0.6	0.62
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.03	0.36	0.36
	Standard EM for GMM and SWRLCF[19]	$\mathrm{P}(J=\hat{J})$	0.53	0.02	0.01
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.27	0.9	0.88
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.2	0.08	0.11
	Proposed	$\mathrm{P}(J=\hat{J})$	0.86	0.82	0.81
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.13	0.15	0.16
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.01	0.03	0.03
−92 [dBm]	Standard EM for GMM and AIC[8]	$\mathrm{P}(J=\hat{J})$	0.01	0.01	0.01
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.31	0.27	0.22
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.68	0.72	0.78
	Standard EM for GMM and BIC[18]	$\mathrm{P}(J=\hat{J})$	0.01	0.01	0.01
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.39	0.37	0.3
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.6	0.62	0.69
	Standard EM for GMM and SWRLCF[19]	$\mathrm{P}(J=\hat{J})$	0.52	0.02	0.01
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.39	0.78	0.77
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.09	0.2	0.22
	Proposed	$\mathrm{P}(J=\hat{J})$	0.82	0.8	0.79
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.16	0.18	0.2
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.02	0.02	0.01
−90 [dBm]	Standard EM for GMM and AIC[8]	$\mathrm{P}(J=\hat{J})$	0.01	0.01	0.01
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.3	0.27	0.2
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.69	0.72	0.79
	Standard EM for GMM and BIC[18]	$\mathrm{P}(J=\hat{J})$	0.01	0.01	0.01
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.38	0.36	0.28
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.61	0.63	0.71
	Standard EM for GMM and SWRLCF[19]	$\mathrm{P}(J=\hat{J})$	0.02	0.01	0.02
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.75	0.71	0.67
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.23	0.28	0.31
	Proposed	$\mathrm{P}(J=\hat{J})$	0.78	0.77	0.76
		$\mathrm{P}(\lvert J-\hat{J}\rvert=1)$	0.21	0.22	0.22
		$\mathrm{P}(\lvert J-\hat{J}\rvert\geq 2)$	0.01	0.01	0.02

Tab. 2. Different levels between J and $\hat{J}$ of four approaches

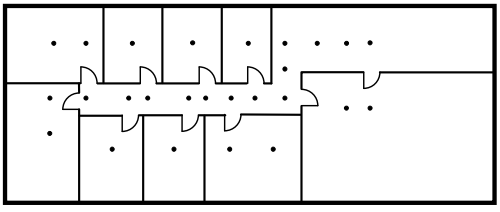

Fig. 3. The layout of area where the Wi-Fi RSSIs had been conducted.

measurements were gathered at four different times of the day including morning, noon, afternoon and evening (100 samples per section). The direction of the collecting equipment (smart phone) was also changed during the measurement collection. In each direction of 0°, 90°, 180° and 270°, 25 measurements were collected. There are 8 APs which are available at all positions of 25 RPs among 26 APs detected in collected training data. The strongest AP selection strategy [10] was applied to select 4 APs which have the largest mean of RSSI values and use them to build the radio-map by utilizing the algorithm introduced in Sec. 2.1. The convergence threshold of the EM algorithm was set to 10^{-6} ($\varepsilon = 10^{-6}$) and the maximum number of Gaussian components for calculating PFs was set to 6 ($J_{\max} = 6$).

In the online phase, 100 sets ($\mathbf{x}^{\text{on}}$) of Wi-Fi RSSI measurements were gathered at the positions of 25 RPs (4 sets per RP) in the same scenarios with the training data. The MAP method presented in Sec. 2.2 was applied to estimate the target's position. The number of nearest neighbors K_{NN} is 3 ($K_{\text{NN}} = 3$). After 100 experiments, positioning results were calculated and reported in Fig. 4 and Tab. 3.

Figure 4 shows the Cumulative Distribution Function (CDF) of positioning error as a function of the distance for four methods. The CDF is defined as the probability that the positioning error (e) is lower than a certain distance (d):

$$\text{CDF}_e(d) = \text{P}(e < d),\ d \geq 0\,. \tag{24}$$

Furthermore, mean (μ_{DE}) and variance (σ^2_{DE}) of distance error of four approaches were recorded in Tab. 3.

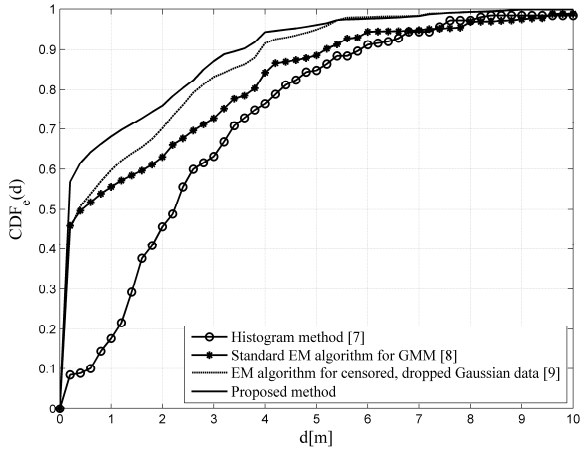

Fig. 4. Comparison of positioning error of four approaches with real data.

Approach	μ_{DE} [m]	σ^2_{DE} [m]
Histogram [7]	2.7807	5.0517
Standard EM algorithm for GMM [8]	2.1843	4.9812
EM algorithm for censored, dropped Gaussian data [9]	1.5607	4.3945
Proposed method	1.0686	2.9862

Tab. 3. Mean and variance of distance error of four approaches.

It can be seen in Fig. 4 and Tab. 3, among four probabilistic approaches, the histogram method [7] which utilized non-parametric model produced the lowest positioning accuracy. The three remaining parametric model approaches showed different levels of positioning accuracy. These can be expounded as follows:

In [8], the standard EM algorithm for GMM was applied to build the radio-map in the training phase. This approach can deal with the multi-component problem. However, as mentioned in [15], when censored and dropped data presented in collected data, the standard EM algorithm produced biased estimates and, hence, it leads to high positioning error.

In [9], the censoring and dropping problems were considered but, in this work, the distribution of collected Wi-Fi RSSIs was assumed to be a Gaussian distribution while it might be drawn from two or more Gaussian components, according to our data investigation. Therefore, this proposal produced a lower positioning accuracy compared to our proposed method.

By utilizing the novel method presented in Sec. 2.1, the censoring, dropping and the multi-component issues have been solved, simultaneously. For this reason, although the test area is limited, the experimental results in Fig. 4 and Tab. 3 indicate that the proposed approach is significantly better than the other approaches.

3.3 The Computational Cost

Beside the positioning accuracy, the computational cost of localization procedure is one of the most important metrics in the WF-IPSs. Systems which have a lower computational cost will spend less time on resulting the target's position. According to (21)–(23), excepting the number of RPs (Q) and number of Aps (N_{AP}), the computational cost highly depends on the estimated number of Gaussian components ($\hat{J}_{q,i}$). In order to evaluate the positioning accuracy and the computational cost, we performed four experiments with the same collected data as mentioned in Sec. 3.2, but different numbers of Gaussian components were selected. In the first experiment, the number of components and parameters in the CD-GMM were estimated by applying the algorithm introduced in Fig. 2 ($J = \hat{J}$). In the experiment 2, 3 and 4, the number of components was fixed by 2, 3 and 4 ($J = 2$, 3 and 4), respectively; parameters were estimated by using the EM algorithm for CD-GMM mentioned in Sec. 2.1. After 100 experiments, the mean time spent

Experiment	**1** ($J = \hat{J}$)	**2** ($J = 2$)	**3** ($J = 3$)	**4** ($J = 4$)
Mean of t_{ETP} [ms]	257.577	340.931	369.604	400.335
μ_{DE}[m]	1.0686	1.1401	1.0372	1.0058
σ^2_{DE}[m]	2.9862	3.0388	2.9685	2.9527

Tab. 4. Mean of t_{ETP}, mean and variance of distance error of four experiments.

on estimating the target's position (t_{ETP}), the mean and variance of distance error were recorded in Tab. 4.

As can be seen in Tab. 4, the four experiments introduced about the same positioning accuracy, but very different t_{ETP}. When the proposed PF of extended BIC was applied, the t_{ETP} reduced by 25%, 30% and 36% compared to fixing the number of components by 2, 3 and 4, respectively. This demonstrates that our proposed method not only improved positioning accuracy, but also introduced the least computational cost.

4. Conclusion

The performance of the WF-IPS is of particular interest. In this paper, novel approaches are introduced to take into account the phenomena presented in real field data. When the censoring, dropping and multi-component problems occurred simultaneously, by utilizing our proposed method, the positioning results of the WF-IPS improved considerably. The experiment in the complex indoor environment showed that the mean distance error is at least 0.4921 m lower than available fingerprinting based probabilistic approaches. On the other hand, by applying our proposed PF of extended BIC, both the number of components and parameters in the CD-GMM are accurately estimated, which leads to better performance in both positioning accuracy and computational cost.

The computational cost of the WF-IPS is proportion to number of RPs and parameters of each distribution to be stored in the database. While this proposal has solved the latter, the former still remains. Once the deployment of the WF-IPS is in a large scale, the searching domain becomes large, too. Therefore, in the future work, we will find a solution to reduce the searching domain which helps to further reduce the execution time in the localization phase. On the other hand, there were still some high positioning errors (5% position estimates had errors which are higher than 4 m). These errors can be explained as follows: In the complex indoor environment, some unexpected reasons (for example: the moving of people, the unexpected operation of APs) might cause the unusual fluctuation of Wi-Fi signal strength that have not been captured during the training phase. As a consequence, some unwonted samples might present in collected data in the online positioning phase. This led to some outliers reflected by some position estimates had high errors. For solving this problem, our approach and the dead reckoning will be combined in the IPS in the future work.

References

[1] BRENA, R. F., GARCÍA-VÁZQUEZ, J. P., GALVÁN TEJADA, C. E., et al. Evolution of indoor positioning technologies: A survey. *Journal of Sensors,* 2017, vol. 2017, p. 1–21. DOI: 10.1155/2017/2630413

[2] STELLA, M., RUSSO, M., BEGUSIĆ, D. RF localization in indoor environment. *Radioengineering*, 2012, vol. 21, no. 2, p. 557–567. ISSN: 1210-2512

[3] HE, S., GARRY CHAN, S.-H. Wi-Fi fingerprint-based indoor positioning: Recent advances and comparisons. *IEEE Communications Surveys and Tutorials*, 2016, vol. 18, no. 1, p. 466–490. DOI: 10.1109/COMST.2015.2464084

[4] BAHL, P., PADMANABHAN, V. N. RADAR: An in-building RF-based user location and tracking system. In *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies.* Tel Aviv (Israel), 2000, p. 775–784. DOI: 10.1109/INFCOM.2000.832252

[5] DING, X., WANG, B., WANG, Z. Dynamic threshold location algorithm based on fingerprinting method. *Wiley ETRI Journal*, 2018, vol. 40, no. 4, p. 531–536. DOI: 10.4218/etrij.2017-0155

[6] HOANG, M. K., HAEB-UMBACH, R. Parameter estimation and classification of censored Gaussian data with application to Wi-Fi indoor positioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver (BC, Canada), 2013, p. 3721–3725. DOI: 10.1109/ICASSP.2013.6638353

[7] DORTZ, N. L., GAIN, F., ZETTERBERG, P. Wi-Fi fingerprint indoor positioning system using probability distribution comparison. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Kyoto (Japan), 2012, p. 2301–2304. DOI: 10.1109/ICASSP.2012.6288374

[8] TSENG, C. H., YEN, J. Enhanced Gaussian mixture model for indoor positioning accuracy. In *International Computer Symposium (ICS)*. Chiayi (Taiwan), December 2016, p. 462–466. DOI: 10.1109/ICS.2016.0099

[9] HOANG, M. K., SCHMALENSTROEER, J., HAEB-UMBACH, R. Aligning training models with smartphone properties in Wi-Fi fingerprinting based indoor localization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Brisbane (QLD, Australia), April 2015, p. 1981–1985. DOI: 10.1109/ICASSP.2015.7178317

[10] YOUSSEF, M., AGRAWALA, A. The Horus WLAN location determination system. In *International Conference on Mobile Systems, Applications, and Services (MobiSys)*. Washington (USA), June 2005, p. 201–218. DOI: 10.1145/1067170.1067193

[11] LUO, J., ZHAN, X. Characterization of smart phone received signal strength indication for WLAN indoor positioning accuracy improvement. *Journal of Networks*, 2014, vol. 9, no. 3. DOI: 10.4304/jnw.9.3.739-746

[12] KAEMARUNGSI, K., KRISHNAMURTHY, P. Properties of indoor received signal strength for WLAN location fingerprinting. In *The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services*. Boston (USA), August 2004, p. 14–23. DOI: 10.1109/MOBIQ.2004.1331706

[13] ALFAKIH, M., KECHE, M., BENOUDNINE, H. Gaussian mixture modeling for indoor positioning WI-FI systems. In *3rd International Conference on Control, Engineering & Information Technology (CEIT)*. Tlemcen (Algeria), May 2015, p. 1–5. DOI: 10.1109/CEIT.2015.7233072

[14] BELLER, S. Model adaptation to improve fingerprinting based indoor navigation. (in German: Modelladaption zur Verbesserung von Fingerprinting basierter Indoornavigation.) *Master Thesis.* University of Paderborn, 2014.

[15] VU, T. K., HOANG, M. K., LE, H. L. An EM algorithm for GMM parameter estimation in the presence of censored and dropped data with potential application for indoor positioning. *ICT Express*, 2019, vol. 5, no. 2, p. 120–123. DOI: 10.1016/j.icte.2018.08.001

[16] DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, vol. 39, no. 1, p. 1–38.

[17] LEE, G., SCOTT, C. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics and Data Analysis*, 2012, vol. 56, no. 9, p. 2816–2829. DOI: 10.1016/j.csda.2012.03.003

[18] HUANG, T., PENG, H., ZHANG, K. Model selection for Gaussian mixture models. *Statistica Sinica*, 2013, vol. 27, p. 147–169. DOI: 10.5705/ss.2014.105

[19] XIE, C. H., CHANG, J. Y., LIU, Y. J. Estimating the number of components in Gaussian mixture models adaptively for medical image. *Optik*, 2013, vol. 124, no. 23, p. 6216–6221. DOI: 10.1016/j.ijleo.2013.05.028

[20] HUSSIEN, H. M., SHIFERAW, Y. N., TESHALE, N. B. Survey on indoor positioning techniques and systems. In *International Conference on Information and Communication Technology for Develoment for Africa*. Bahir Dar (Ethiopia), September 2017, p. 46–55. DOI: 10.1007/978-3-319-95153-9_5

[21] XIA, S., LIU, Y., YUAN, G., et al. Indoor fingerprint positioning based on Wi-Fi: An overview. *International Journal of Geo-Information*, 2017, vol. 6, no. 5, p. 135–149. DOI: 10.3390/ijgi6050135

About the Authors

Trung Kien VU (corresponding author) was born in 1977. He received his M.Sc. from the Hanoi University of Science and Technology in 2006. His research interests include positioning/navigation, signal processing, wireless technology.

Manh Kha HOANG was born in 1979. He received his Ph.D. from the University of Paderborn in 2016. His research interests include positioning/navigation, signal processing, cognitive radio, smart antennas.

Hung Lan LE was born in 1960. He was appointed professor in 2013. His research interests cover smart control, swarm robotics, artificial neural network and fuzzy logic. He was a member of technical program committee of several international conferences.