IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 19, 2019, accepted February 28, 2019, date of publication March 13, 2019, date of current version March 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903236

Non-Line-of-Sight Identification Based on Unsupervised Machine Learning in Ultra Wideband Systems

JIANCUN FAN AND AHSAN SALEEM AWAN
School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Corresponding author: Jiancun Fan (fanjc0114@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671367, in part by the Key Research and Development Plan of Shaanxi Province under Grant 2018GY-003, in part by the Project funded by the China Postdoctoral Science Foundation under Grant 2014M560780 and Grant 2015T81031, in part by the Postdoctoral Science Foundation of Shaanxi Province, and in part by the Fundamental Research Funds for the Central Universities.

ABSTRACT Identification of line-of-sight (LOS) and non-line-of-sight (NLOS) propagation conditions is very useful in ultra wideband localization systems. In the identification, supervised machine learning is often used, but it requires exorbitant efforts to maintain and label the LOS and NLOS database. In this paper, we apply unsupervised machine learning approach called ‘‘expectation maximization for Gaussian mixture models’’ to classify LOS and NLOS components. The key advantage of applying unsupervised machine learning is that it does not require any rigorous and explicit labeling of the database at a certain location. The simulation results demonstrate that by using the proposed algorithm, LOS and NLOS signals can be classified with 86.50% correct rate, 12.70% false negative, and 0.8% false positive rate. We also compare the proposed algorithm with the existing cutting-edge supervised machine learning algorithms in terms of computational complexity and signals' classification performance.

INDEX TERMS Expectation maximization, Gaussian mixture models, unsupervised machine learning, ultra wideband systems, non-line-of-sight identification.

I. INTRODUCTION

Localization information is crucial in commercial and military applications [1]–[3]. To improve the localization accuracy, a lot of effective approaches have been proposed [4], [5]. Due to high spatial and temporal resolution, *ultra wideband* (UWB) technology has been used in number of various centimeter level wireless localization applications [6], [7]. Especially, since *global positioning system* (GPS) signals are severely attenuated in harsh indoor environments, UWB localization is considered as an adequate substitute for GPS in indoor environments. However, the localization accuracy in an UWB system is severely affected by the *non line of sight* (NLOS) conditions.

In an indoor environment, signals between the transmitter and receiver are often obstructed by various abundant objects such as people, walls, furniture and doors. If a signal propagates directly between transmitter and receiver it is called as *line of sight* (LOS) condition, on the other hand if there is no direct path between the transmitter and receiver it is known as NLOS condition. In case of NLOS signals the distance between transmitter and receiver will be longer which will result in positive bias in the position estimation. To deal with the effect of NLOS, many research efforts have been devoted to finding better approaches [8]–[15]. In [8], several different localization techniques have been analyzed to approach the *Cramer-Rao lower bound* (CRLB) in an NLOS environment with a single-path prorogation assumption. Furthermore, hybrid *received signal strength* (RSS) and *time-of-arrival* (TOA) based localization method has been proposed in [9] to simultaneously mitigate the effect of NLOS and multipath. A recent study of Mazuelas *et al.* [10] has shown that machine learning techniques can be used to achieve the performance approaching the *Cramer-Rao lower bound* (CRLB) in harsh wireless condition. In [11], the multi-path channel state information is further used to built fingerprint database to implement the localization in outdoor environment. Generally, minimum two LOS anchor

The associate editor coordinating the review of this manuscript and approving it for publication was Oussama Habachi.

nodes are required for *direction-of-arrival* (DOA) localization technique and three LOS anchor nodes are required for TOA localization technique [12]. In this case, multipath and NLOS propagation conditions will cause unreliability of estimated parameters for positioning. Thus identification of LOS and NLOS signals can greatly enhance the localization accuracy [13]. Besides the localization accuracy, NLOS identification can also facilitate the resource allocation, power allocation, and nodes routing in wireless sensor networks [14], [15].

To achieve the objective of NLOS identification in UWB system, several techniques have been investigated in the literature [16]–[21]. Most frequently used methods for channel classification are statistical hypothesis testing [20], [21] and supervised machine learning techniques [16]–[19]. Because of very large bandwidth, UWB signals can be readily discriminated at the receiver. Marano *et al.* [16] have extracted the features, such as received signal energy, maximum amplitude of the received signal, rise time, mean excess delay, *root mean square* (RMS) delay spread and kurtosis for LOS and NLOS conditions from an extremely high resolution *channel impulse response* (CIR) to train the machine learning algorithm. Miao *et al.* [19] utilized only LOS signals to train the one class classification algorithm which is more economical and during the testing phase the algorithm can capture the differences between both LOS and NLOS conditions. However in the supervised machine learning approach a significant investment of time and labor is required to label the channel conditions along with the different features which are extracted from the received waveform. In the supervised machine learning approach, the algorithm must first learn from the labeled data and then it is deployed to discriminate the unlabeled data. Furthermore supervised machine learning techniques need to update the training database when indoor conditions are changed for example furniture in the room is moved from one location to another. Fortunately, unsupervised machine learning approach proposed in this paper obviates the need for explicit and prior labeling of channel conditions thus reducing lot of efforts and time.

The objective of this paper is to study the performance of an unsupervised machine learning algorithm *expectation maximization* (EM) for *Gaussian mixture model* (GMM) to discriminate between LOS and NLOS conditions in indoor environment based on some essential and auxiliary features which are extracted from the waveform at the UWB receiver. The EM algorithm can be used in finding the parameters of GMM components that maximizes the log likelihood whether a signal belongs to LOS or NLOS distribution. To the best of our knowledge, this paper is the first attempt to exploit the unsupervised machine learning algorithm, EM-GMM, for NLOS identification. Advantage of using this algorithm is soft clustering which provides the LOS and NLOS probability for each signal.

The rest of the paper is organized as follows. Channel characteristics, data collection and features extraction are explained in section II. NLOS identification using unsupervised machine leaning algorithm is expressed in section III. Section IV describes the performance evaluation criteria for the proposed EM-GMM approach, also computational complexity and performance results are compared with existing state-of-the-art supervised machine learning algorithms. Finally the paper is concluded in section V.

II. UWB CHANNEL MODEL AND FEATURES EXTRACTION

In this section, we briefly discuss the SG3a UWB channel which includes two typical channel models, CM1 and CM2 [22]. These channel models contain the measurements over the distance of 0-4 meter for both the LOS and NLOS situations separately. SG3a UWB channel model is based on *Saleh-Valenzuela* (S-V) indoor channel modeling which was established in 1987 [23]. In UWB channel model, multipath components arrive at the receiver side in the form of clusters and within each cluster there are several subsequent arrivals which are called rays. Therefore, the channel impulse response of an UWB system can be expressed as

$$h(t) = \sum_{m=0}^{M} \sum_{r=0}^{R} \alpha_{m,r} \exp(j\theta_{m,r})\delta(t - T_m - \tau_{m,r}) \tag{1}$$

where $\alpha_{m,r}$ and $\theta_{m,r}$ are the channel gain and phase of the r^{th} ray in the m^{th} cluster, respectively. T_m is the time of arrival of the first path of the m^{th} cluster and $\tau_{m,r}$ is the delay of the r^{th} ray in the m^{th} cluster. Due to a very large bandwidth, the time-domain transmission signals of UWB is similar as a pulse. Therefore, the received waveform $r(t) = \int_{-\infty}^{\infty} h(\tau)s(t-\tau)d\tau$ has similar characteristic as the channel impulse response, where $s(t)$ is the transmission waveform of the UWB signal. In this study, we discriminate the LOS and NLOS components by exploiting the statistics of the received multipath components. We select three features which are extracted from the received waveform $r(t)$ to identify the NLOS components and define the features set vector P as

$$P = [N_P, \tau_{MED}, \tau_{RMSD}], \tag{2}$$

where

- N_P denotes the number of paths which contain more than 85 percent of the total energy and the energy of the received signal can be obtained by [16]

$$\mathcal{E}_r = \int_{-\infty}^{+\infty} |r(t)|^2 dt. \tag{3}$$

- τ_{MED} denotes *mean excess delay* (MED). For the NLOS components, MED is greater than the LOS components. It can be calculated using the following formula [16]

$$\tau_{MED} = \int_{-\infty}^{+\infty} t\,\Psi(t)\,dt \tag{4}$$

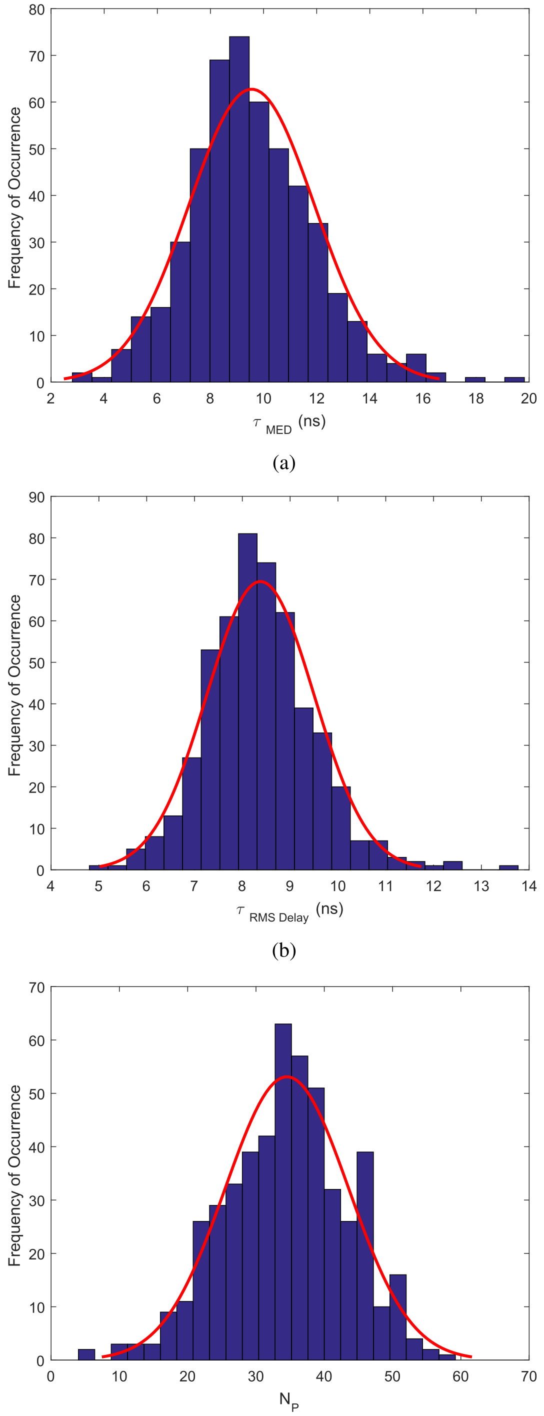
and $\Psi(t) = |r(t)|^2/\mathcal{E}_r$.

- τ_{RMSD} is value of RMS delay spread and depicts the temporal attenuation of the signal energy which is similar to MED. Generally, the RMS delay for NLOS components is greater as compared to the LOS components. RMS delay spread can be expressed as [16]

$$\tau_{RMSD} = \int_{-\infty}^{+\infty} (t - \tau_{MED})^2 \Psi(t)\, dt. \tag{5}$$

Significance of using these three features $[N_P, \tau_{MED}, \tau_{RMSD}]$ for channel classification is that these features are expected to possess substantial differences between LOS and NLOS components and reduce the complexity of NLOS identification since its computational cost is crucially dependent on the dimension of the feature vector P.

To acquire the data set, we have simulated a large number of waveforms in MATLAB program using UWB indoor channel model according to the specifications mentioned in [24] with half of the waveforms in LOS and other half being in NLOS environments.



(a)

(b)

(c)

FIGURE 1. Histograms of NLOS samples of τ_{MED}, $\tau_{RMS\ Delay}$ and N_P. (a) Distribution of τ_{MED}. (b) Distribution of τ_{RMSD}. (c) Distribution of N_P.

III. NLOS IDENTIFICATION

In this section, we first describe the Gaussian mixture model consisting of LOS and NLOS components and then propose a NLOS identification scheme based on the EM algorithm.

A. GAUSSIAN MIXTURE MODELS

The LOS and NLOS channel components tend to follow a certain probability distribution, therefore the mixture of their probability distributions can be used for channel classification by assigning unlabeled data points (unidentified signals) to specific probability distribution which is either LOS or NLOS. Figs. 1 and 2 show that *probability density functions* (PDFs) of selected features comprising of τ_{MED}, $\tau_{RMS\ Delay}$ and N_P for the NLOS components and LOS components can be modeled as Gaussian distribution, respectively. In this case, their mixture model is also the Gaussian model and can be used to classify the LOS and NLOS components. Therefore, once the parameters determining this Gaussian mixture model are obtained, the NLOS identification will be achieved.

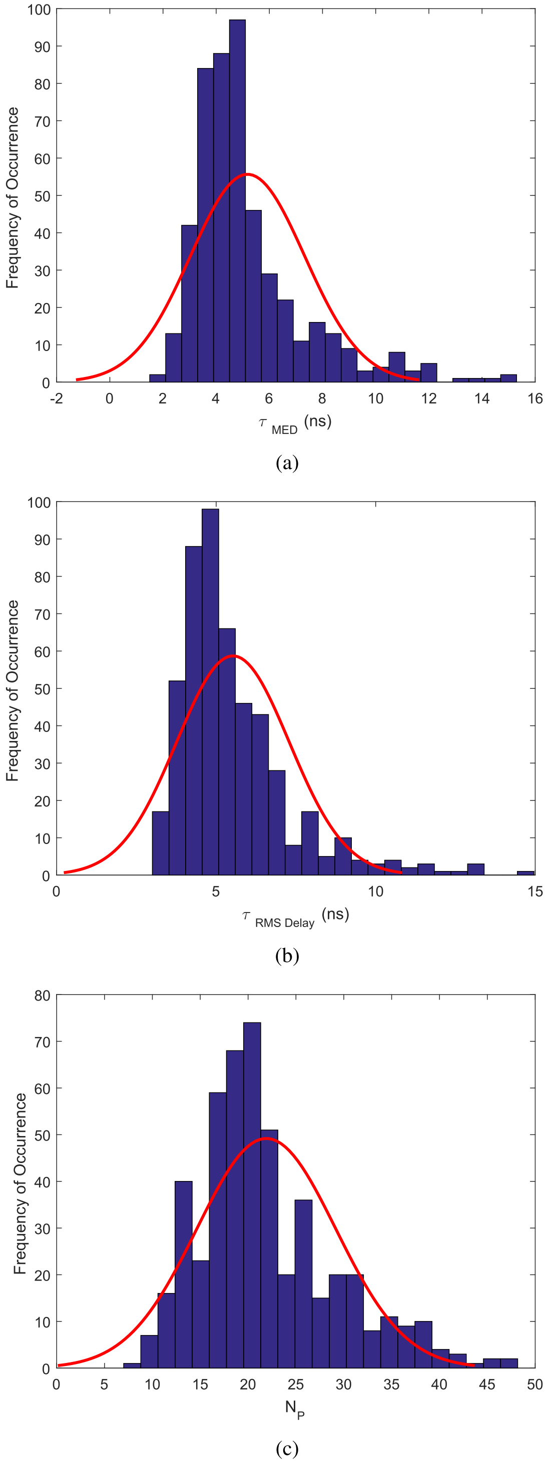
Parameters of the probability distributions are commonly determined by using EM algorithm since gradient based optimization techniques are hard to compute for the mixture of probability densities. As mentioned before, the LOS and NLOS features can be combined into a Gaussian mixture model. Therefore, this Gaussian mixture model with $k = 1$ for LOS components and $k = 2$ for NLOS components can be expressed as following,

$$p(\mathbf{x}_n) = \sum_{k=1}^{2} \omega_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k), \tag{6}$$

where

- ω_k is the mixing coefficient or also known as the weight for each Gaussian distribution. Mixing coefficients always satisfy the below conditions,

$$0 \leq \omega_k \leq 1,$$
$$\sum_{k=1}^{2} \omega_k = 1,$$



(a)

(b)

(c)

FIGURE 2. **Histograms of LOS samples of τ_{MED}, $\tau_{RMS\ Delay}$ and N_P. (a) Distribution of τ_{MED}. (b) Distribution of τ_{RMSD}. (c) Distribution of N_P.**

- $\mathbf{x}_n = (x_{n1}, x_{n2}, x_{n3} \ldots . x_{nD}), 1 \leq n \leq N$, n is the index of data sample, N is the total number of data samples, and D is the number of features. As mentioned before, each data point in this paper can be expressed as $\mathbf{x}_n = [N_P, \tau_{MED}, \tau_{RMSD}]^T$ which can be considered as a point in three dimensional space.

- $\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Theta}_k)$ is a Gaussian probability density which is governed by mean vector μ_k and covariance matrix Θ_k. Multivariate Gaussian distribution can be mathematically expressed as

$$\mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Theta}_k\right) = \frac{\exp\left\{-\frac{1}{2}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Theta}_k^{-1}\left(\mathbf{x}_n - \boldsymbol{\mu}_k\right)\right\}}{(2\pi)^{\frac{D}{2}}\, |\boldsymbol{\Theta}|_k^{\frac{1}{2}}}, \tag{7}$$

  where the Gaussian probability, $\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Theta}_k)$, for $k = 1$ is referred to as the LOS and for $k = 2$ is referred to as NLOS components of the GMM.

In short, a GMM is controlled by a set of three parameters: mean vectors $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$, covariance matrices $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$ and mixing coefficients $\{\omega_1, \omega_2\}$. To discriminate the LOS and NLOS components, we need to estimate the mean vectors $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$ and covariance matrices $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$ for both the distributions. The expectation maximization algorithm estimates the parameters of the latent variables in iterative way and is one of the most elegant techniques for parametric estimation in machine learning [25]. Therefore, an expectation maximization algorithm based on GMM is proposed in the following section.

B. EXPECTATION MAXIMIZATION ALGORITHM FOR GMMS

Expectation maximization algorithm is used to find the maximum likelihood solution of each received signal for the LOS and NLOS Gaussian mixture models. Given a mixture of LOS and NLOS densities, our aim is to maximize the likelihood of each data point $\mathbf{x}_n$ with respect to the parameters. First we initialize the parameters comprising of the means, covariances and weights arbitrarily to maximize the log likelihood function and then update the parameters by switching between following two steps, expectation and maximization iteratively.

1) EXPECTATION STEP (E-STEP)

With the help of Baye's rule we can calculate the posterior probability $\gamma\,(r_{nk})$ for each data point $\mathbf{x}_n$ corresponding to LOS with $k = 1$ and NLOS with $k = 2$ distribution by using the current parameters means, covariances and weights. The posterior probability is defined and calculated by the following mathematical expression [26]:

$$\gamma\,(r_{nk}) = \frac{\omega_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Theta}_k)}{\sum_{j=1}^{2} \omega_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Theta}_j)} \tag{8}$$

where j is the index from 1 to 2, for both the LOS and NLOS distributions.

2) MAXIMIZATION STEP (M-STEP)

In the maximization step, we re-estimate the parameters by making use of the posterior probabilities computed in the E-step such that maximized log likelihood is achieved.

Correspondingly, the parameter update procedures can be expressed as

$$\omega_k^{New} = \frac{\bar{N}_k}{N} \tag{9}$$

$$\boldsymbol{\mu}_k^{New} = \frac{\sum_{n=1}^{N} \gamma\left(r_{nk}\right) \mathbf{x}_n}{\bar{N}_k} \tag{10}$$

$$\boldsymbol{\Theta}_k^{New} = \frac{\sum_{n=1}^{N} \gamma\left(r_{nk}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{New}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{New}\right)^T}{\bar{N}_k} \tag{11}$$

where $\bar{N}_k = \sum_{n=1}^{N} \gamma(r_{nk})$. It should be noted that $\bar{N}_k$ is not an integer, but it can be viewed as the effective number of points allocated to each LOS and NLOS cluster in a physical sense.

Once the above parameters are obtained, we can compute the log likelihood of LOS and NLOS distribution by

$$\ln P\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Theta}, \omega\right) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{2} \omega_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Theta}_k) \right\} \tag{12}$$

In Algorithm 1, we demonstrate the EM algorithm for the Gaussian mixture models. Line 2 to line 11 relates to expectation step and line 12 to line 28 relates to maximization step. Expectation and maximization steps continue until convergence is achieved that is when none of the parameters or log likelihood function is further updated. Each data point x_n has been assigned to either LOS ($k = 1$) or NLOS ($k = 2$) based on the maximum value of its responsibility $\gamma\left(r_{nk}\right)$ for each distribution. This can be expressed as

$$k = \begin{cases} 1, & \text{if } \gamma\left(r_{n1}\right) > \gamma\left(r_{n2}\right), \\ 2, & \text{Otherwise.} \end{cases} \tag{13}$$

After classifying the Gaussian mixture distributions into two components, mixing components with lowest mean value of the selected features $[N_P, \tau_{MED}, \tau_{RMSD}]$ are considered as LOS. In Fig. 3 one thousand data points with their corresponding values of two features τ_{MED} and τ_{RMSD} are depicted in green and blue according to the LOS and NLOS distributions (latent variables). In Fig. 4 data samples are plotted by ignoring the values of data labels. In Fig. 5 data points with colors illustrating the maximum value of responsibility $\gamma\left(r_{nk}\right)$ for LOS ($k = 1$) and NLOS ($k = 2$) components, mean of each distribution is highlighted in black.

IV. PERFORMANCE EVALUATION

To measure the quality of the unsupervised machine learning algorithm EM-GMM, we have adopted the method of external evaluation in which we already have the original labels of each signal. We ignore the data labels and classify the signals into LOS and NLOS and later compare the classification results with the original data labels this gives us the external evaluation of the EM-GMM algorithm. To evaluate the performance, a *confusion matrix* is computed and performance is measured in terms of *Correct rate*, *Error rate*, *LOS detection rate* and *NLOS detection rate*. *Confusion matrix* consists of the following entities:

Algorithm 1 EM-GMM Algorithm

Output: Classification of all the data points into LOS and NLOS probability distributions.
Input: One thousand unlabeled data points.

```
1:  while Mean vectors $\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\}$, covariance matrices $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$ and mixing coefficients $\{\omega_1, \omega_2\}$ are not further updated. do
2:      for $n \leftarrow 1$ to N do
3:          $s \leftarrow 0$
4:          for $k \leftarrow 1$ to K do
5:              $\mathcal{N}_k \leftarrow \omega_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Theta}_k)$
6:              $s \leftarrow s + \mathcal{N}_k$
7:          end for
8:      end for
9:      for $k \leftarrow 1$ to K do
10:         $\gamma\left(r_{nk}\right) \leftarrow \frac{\mathcal{N}_k}{s}$
11:     end for
12:     $\bar{N}_k \leftarrow 0, \omega \leftarrow 0, \mu \leftarrow 0, \sigma \leftarrow 0$
13:     for all $n \in 1 \ldots N, k \in 1 \ldots K$ do
14:         $\bar{N}_k \leftarrow \bar{N}_k + \gamma(r_{nk})$
15:     end for
16:     for all $k \in 1 \ldots K$ do
17:         $\omega_k \leftarrow \frac{\bar{N}_k}{N}$
18:     end for
19:     for all $n \in 1 \ldots N, k \in 1 \ldots K, d \in 1 \ldots D$ do
20:         $\mu_{kd} \leftarrow \mu_{kd} + \frac{r_{nk} x_{nd}}{N_k}$
21:     end for
22:     for all $n \in 1 \ldots N, k \in 1 \ldots K, d \in 1 \ldots D$ do
23:         $\sigma_{kd}^2 \leftarrow \sigma_{kd}^2 + \frac{r_{nk}(x_{nd} - \mu_{kd})^2}{N_k}$
24:     end for
25:     for all $k \in 1 \ldots K$ do
26:         $\boldsymbol{\Theta}_k \leftarrow diag(\sigma_k^2)$
27:     end for
28: end while
```

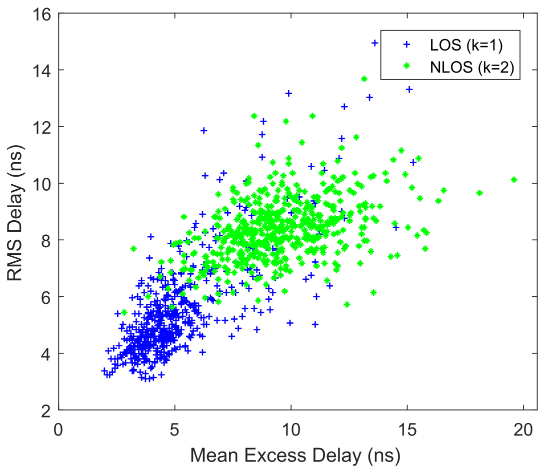


FIGURE 3. LOS and NLOS mixture of probability distributions with actual label of 1000 data points.

TABLE 1. Performance statistics of EM-GMM algorithm.

Feature Set	Correct Rate	Error Rate	TP	FN	TN	FP	LOS Detection Accuracy	NLOS Detection Accuracy
$[N_P]$	0.795	0.205	380	120	415	85	0.760	0.830
$[\tau_{RMSD}]$	0.820	0.180	321	179	499	1	0.642	0.998
$[N_P, \tau_{RMSD}]$	0.830	0.170	333	167	497	3	0.666	0.994
$[\tau_{MED}]$	0.845	0.155	365	135	480	20	0.730	0.930
$[\tau_{RMSD}, \tau_{MED}]$	0.851	0.149	356	144	495	5	0.712	0.990
$[N_P, \tau_{MED}]$	0.852	0.148	370	130	482	18	0.740	0.964
$[N_P, \tau_{MED}, \tau_{RMSD}]$	0.865	0.135	373	127	492	8	0.746	0.984

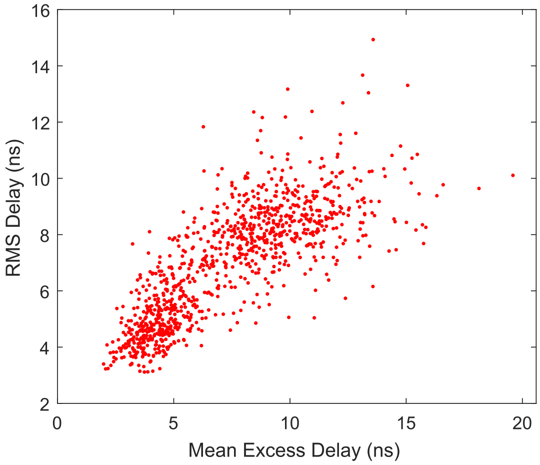


FIGURE 4. Unlabeled data points of LOS and NLOS distributions with unknown parameters mean (μ_1, μ_2) and covariance (Θ_1, Θ_2).

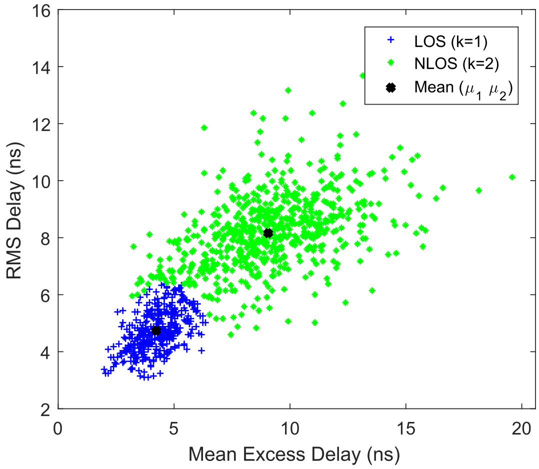


FIGURE 5. Channel classification by EM-GMM algorithm along with mean of LOS (μ_1) and NLOS (μ_2) distributions.

- TP (True Positive): true positive are the signals for which the actual label is LOS and maximum responsibility $\gamma(r_{nk})$ has correctly predicted these signals as LOS.
- FP (False Positive): false positive are the signals for which the actual label is NLOS and maximum responsibility $\gamma(r_{nk})$ has incorrectly predicted these signals as LOS.
- TN (True Negative): true negative are the signals for which the actual label is NLOS and maximum responsibility $\gamma(r_{nk})$ has correctly predicted these signals as NLOS.
- FN (False Negative): false negative are the signals for which the actual label is LOS and maximum responsibility $\gamma(r_{nk})$ has incorrectly predicted these signals as NLOS.

Correspondingly, *Correct Rate, Error Rate, LOS Detection rate* and *NLOS detection rate* can be calculated by the following expressions,

$$Correct\ Rate = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN} \tag{15}$$

$$LOS\ Detection\ Rate = \frac{TP}{TP + FN} \tag{16}$$

$$NLOS\ Detection\ Rate = \frac{TN}{TN + FP} \tag{17}$$

Table 1. represents the performance statistics of EM-GMM algorithm for different set of features. It can be examined that set consisting of all three features wins out in terms of correct classification with 0.8650 correct rate. False negative and false positive rates for this best performing set of feature are calculated as 12.70 and 0.8 percent respectively. It is also observed that numbers of NLOS signals which are incorrectly classified as LOS (false positive samples) are very few in all combination of features as compared to false negative samples. Thus high NLOS identification rate could improve the localization accuracy significantly. Fig. 6 graphically illustrates the *LOS detection rate*, *NLOS detection rate* and *Correct rate* for different set of features.

Table 2 shows computational complexity and performance comparison of EM-GMM algorithm with existing supervised machine learning algorithms, *K-Nearest Neighbor* (KNN), *Naive Bayes* (NB), *Decision Trees* (DT) and *least square support vector machine* (LS-SVM). For simplicity, we use running times to measure the computation complexity. In the simulations, 500 additional LOS and 500 NLOS waveforms are generated to train the supervised machine learning algorithms and MATLAB 2016a is used to perform the comparison. From this table,

TABLE 2. Computational complexity and performance comparison with supervised machine learning algorithms using $\left[N_P, \tau_{MED}, \tau_{RMS\ Delay}\right]$ feature set.

Algorithm	Running Time	LOS Detection Accuracy	NLOS Detection Accuracy	Correct Rate	Error Rate	TP	FN	TN	FP
KNN	0.0371s	0.892	0.956	0.924	0.076	446	54	478	22
NB	0.0394s	0.856	0.880	0.868	0.132	428	72	440	60
DT	0.0559s	0.974	0.944	0.959	0.041	487	13	472	28
LS-SVM	0.1388s	0.988	0.994	0.991	0.009	494	6	497	3
EM-GMM	0.0606s	0.746	0.984	0.865	0.135	373	127	492	8

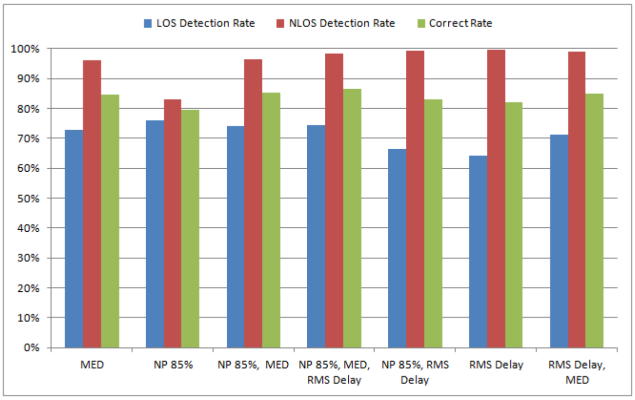


FIGURE 6. Performance graph using different set of features combination.

- Unsupervised machine learning algorithm EM-GMM has a certain performance difference as compared to the supervised machine learning algorithms, but it doesn't require the training data since it does not have the training phase.
- EM-GMM algorithm only takes up to 44% of the running time required by the state-of-the-art supervised machine learning LS-SVM, but it achieves almost the same NLOS Detection Accuracy as the LS-SVM algorithm.

In addition, it is also shown that it is effective to select three features τ_{MED}, τ_{RMSD} and N_P to identify NLOS components in EM-GMM algorithm. Once the NLOS signals are correctly identified, they can be excluded from the localization algorithm to enhance the position accuracy.

V. CONCLUSION

In this paper a novel method called EM-GMM algorithm based on unsupervised machine learning is proposed to identify the NLOS signals. We can conclude that the unsupervised machine learning technique we bring forward is effective for NLOS channel identification and can be used to enhance the localization accuracy. By utilizing this probabilistic method we can get the soft clustering of the data points into LOS and NLOS, especially data points which are not confined but rather loosely attached to a distribution will have an indicator of level of uncertainty over the classification. However the data points which cannot be modeled as Gaussian distribution and also are not linearly separable, such data points might result in incorrect classification. In this study we have used a batch version of unsupervised machine learning algorithm in which all the data points are considered at once and values of latent parameters are discrete. Therefore future works should focus on the need to know the good starting point to initialize the algorithm in actual real time scenarios when signals are received continuously. Verification of the EM-GMM algorithm based on experimental data is also a subject of future work.

REFERENCES

[1] Y. Shen, S. Mazuelas, and M. Z. Win, "Network navigation: Theory and interpretation," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 9, pp. 1823–1834, Oct. 2012.

[2] N. Patwari, J. N. Ash, S. Kyperountas, A. O. Hero, R. L. Moses, and N. S. Correal, "Locating the nodes: Cooperative localization in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 54–69, Jul. 2005.

[3] M. Z. Win *et al.*, "Network localization and navigation via cooperation," *IEEE Commun. Mag.*, vol. 49, no. 5, pp. 56–62, May 2011.

[4] I. Guvenc and C.-C. Chong, "A survey on TOA based wireless localization and NLOS mitigation techniques," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 3, pp. 107–124, Aug. 2009.

[5] C. Laoudias, A. Moreira, S. Kim, S. Lee, L. Wirola, and C. Fischione, "A survey of enabling technologies for network localization, tracking, and navigation," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 3607–3644, 4th Quart., 2018.

[6] D. Dardari, A. Conti, U. Ferner, A. Giorgetti, and M. Z. Win, "Ranging with ultrawide bandwidth signals in multipath environments," *Proc. IEEE*, vol. 97, no. 2, pp. 404–426, Feb. 2009.

[7] L. Yang and G. B. Giannakis, "Ultra-wideband communications: An idea whose time has come," *IEEE Signal Process. Mag.*, vol. 21, no. 6, pp. 26–54, Nov. 2004.

[8] Y. Qi, H. Kobayashi, and H. Suda, "Analysis of wireless geolocation in a non-line-of-sight environment," *IEEE Trans. Wireless Commun.*, vol. 5, no. 3, pp. 672–681, Mar. 2006.

[9] J. Prieto, S. Mazuelas, A. Bahillo, P. Fernandez, R. M. Lorenzo, and E. J. Abril, "Adaptive data fusion for wireless localization in harsh environments," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 1585–1596, Apr. 2012.

[10] S. Mazuelas, A. Conti, J. C. Allen, and M. Z. Win, "Soft range information for network localization," *IEEE Trans. Signal Process.*, vol. 66, no. 12, pp. 3155–3168, Jun. 2018.

[11] S. Chen, J. Fan, X. Luo, and Y. Zhang, "Multipath-based CSI fingerprinting localization with a machine learning approach," in *Proc. Wireless Adv. (WiAd)*, London, U.K., Jun. 2018, pp. 1–5.

[12] R. Zekavat and R. M. Buehrer, *Handbook of Position Location: Theory, Practice and Advances*. Hoboken, NJ, USA: Wiley, 2011.

[13] S. J. Ingram, D. Harmer, and M. Quinlan, "UltraWideBand indoor positioning systems and their use in emergencies," in *Proc. Position Location Navigat. Symp. (PLANS)* , Apr. 2004, pp. 706–715.

[14] V. Sumathy, P. Narayanasmy, K. Baskaran, and T. Purusothaman, "GLS with secure routing in ad-hoc networks," in *Proc. Conf. Convergent Technol. Asia–Pacific Region (TENCON)*, vol. 3, Oct. 2003, pp. 1072–1076.

[15] M. Rahman, M. Mambo, A. Inomata, and E. Okamoto, "An anonymous on-demand position-based routing in mobile ad hoc networks," in *Proc. Int. Symp. Appl. Internet (SAINT)*, Phoenix, AZ, USA, Jan. 2006, pp. 7 and 306.

[16] S. Marano *et al.*, "NLOS identification and mitigation for localization based on UWB experimental data," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 7, pp. 1026–1035, Sep. 2010.

[17] H. Wymeersch, S. Marano, W. M. Gifford, and M. Z. Win, "A machine learning approach to ranging error mitigation for UWB localization," *IEEE Trans. Commun.*, vol. 60, no. 6, pp. 1719–1728, Jun. 2012.

[18] T. Van Nguyen, Y. Jeong, H. Shin, and M. Z. Win, "Machine learning for wideband localization," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 7, pp. 1357–1380, Jul. 2015.

[19] Z.-M. Miao, L.-W. Zhao, W.-W. Yuan, and F.-L. Jin, "Application of one-class classification in NLOS identification of UWB positioning," in *Proc. Int. Conf. Inf. Syst. Artif. Intell. (ISAI)*, Hong Kong, Jun. 2016, pp. 318–322.

[20] S. Venkatesh and R. M. Buehrer, "Non-line-of-sight identification in ultra-wideband systems based on received signal statistics," *IET Microw., Antennas Propag.*, vol. 1, no. 6, pp. 1120–1130, 2007.

[21] K. Yu and Y. J. Guo, "Statistical NLOS identification based on AOA, TOA, and signal strength," *IEEE Trans. Veh. Technol.*, vol. 58, no. 1, pp. 274–286, Jan. 2009.

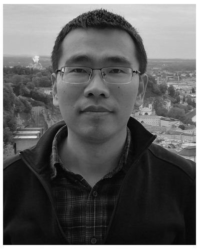
[22] J. Foerster, "Channel modeling sub-committee report final," Tech. Rep. IEEE 802.15–02/490 IEEE P802.15SG3a, Feb. 2003.

[23] A. A. M. Saleh and R. A. Valenzuela, "A statistical model for indoor multipath propagation," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 2, pp. 128–137, Feb. 1987.

[24] Y. S. Cho, J. Kim, W. Y. Yang, and C.-G. Kang, *MIMO-OFDM Wireless Communications with MATLAB*. Hoboken, NJ, USA: Wiley, 2010, pp. 35–40.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). New York, NY, USA: Springer-Verlag, 2006.

[26] C. Guo, H. Fu, and W. Luk, "A fully-pipelined expectation-maximization engine for Gaussian mixture models," in *Proc. Int. Conf. Field-Program. Technol.*, Seoul, South Korea, Dec. 2012, pp. 182–189.



JIANCUN FAN received the B.S. and Ph.D. degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2012, respectively, where he is currently an Associate Professor with the Department of Information and Communications Engineering. From 2009 to 2011, he was a Visiting Scholar with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA. In 2017, he was a Visiting Scholar with the Dresden University of Technology, Dresden, Germany, for three months. His general research interests include signal processing and wireless communications, with an emphasis on MIMO communication, cross-layer optimization for spectral- and energy-efficient networks, wireless localization, practical issues in LTE and 5G systems, and machine learning. He was a recipient of the Best Paper Award at the 20th International Symposium on Wireless Personal Multimedia Communications, in 2017.

AHSAN SALEEM AWAN received the bachelor's degree from the National University of Sciences and Technology, Pakistan, in 2010, and the master's degree from Xi'an Jiaotong University, in 2018. From 2011 to 2016, he was an RF Engineer with Huawei Technologies, Pakistan and Saudi Arabia. He is currently a 5G Research Engineer with the 5G R&D Center, Foxconn, Shenzhen, China. His research interests include mmWave, massive MIMO, wireless localization, and machine learning.