JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.7 No.1 January 2025: 63-70
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i1.12581

Indoor Positioning System using Gaussian Mixture Model on BLE Fingerprint

Maximilianus Maria Kolbe Lie[1*], Bakti Amirul Jabar[2]

[1,2] Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
lie.kolbe@binus.edu; bakti.jabar@binus.ac.id

*Correspondence: lie.kolbe@binus.edu

Abstract: ***After the release of Bluetooth Low Energy (BLE), people have been trying to use Bluetooth as an alternative source to solve indoor positioning. Unfortunately, due to the nature of Bluetooth about proximity, the propagated signal is very fluctuating. This decreases the accuracy considerably and has become one of the main problems in using Bluetooth. To combat the signal fluctuations, we propose a fingerprinting-based concept of using received signal strength (RSS) frequency distribution values as the data in the radio map, which is termed Frequency Distribution Radio Map (FDRM). We also propose a probabilistic fingerprinting-based algorithm utilizing FDRM using Gaussian Mixture Model (GMM) as the probability distribution function (PDF). In the offline phase, we compare 2 methods: k-Means only, and k-Means with Expectation-Maximization (EM); to learn the FDRM. This resulting a probability distribution function (PDF) of the RSS in each reference points for each BLEs. In the online phase, k-Nearest Neighbour (KNN) and weighted average are used to estimate the receiver's location. The proposed method is validated over 3 different datasets taken from a 4 m x 6 m classroom equipped with chairs and tables. The experiment shows that the proposed fingerprint and model are better in capturing the environment, including the signal fluctuation. By using only k-Means in obtaining the GMM, it achieved mean error of 98.18 cm and standard deviation of 56.11 cm. By adding EM, there will be a trade-off between mean error with better standard deviation and lower computing time. It achieved standard deviation of 47.99 cm and mean error of 112.24 cm.***

Keywords: ***Bluetooth Low Energy; Frequency Distribution Radio Map; Probabilistic Fingerprinting; Gaussian Mixture Model; k-Means; Expectation-Maximization.***

I. INTRODUCTION

Global Positioning System (GPS) has an important role in nowadays life, providing location-based service. However, GPS depends on GPS signal which can be weakened by obstacles such as building blocks. Therefore, GPS is unreliable when the receiver is inside a building. To address this problem, researchers have been working on an Indoor Positioning System (IPS). A reliable IPS has been very crucial these days in a lot of Internet of Things (IoT) projects and smart home projects (Kim, Jeong, & Park, 2013) (Ke, Wu, Chan, & Lu, 2018). IPS also has an important role in some indoor public places. Other than indoor positioning or navigation (Ramani & Tank., 2014) (Ruggiero, Charith, Song, & Lucia, 2018) (U.S. Patent No. 8,866,673, 2014) (Yang, Wang, & Zhang, 2015), IPS is required for a more sophisticated management system such as surveillance in hospital (Fisher, 2006) and vehicle tracking in construction sites under a tunnel (Woo, et al., 2011).

Scientists have been working on a wireless IPS by using various sources such as infrared (Lee, 2004), ultrasound (Medina, Segura, & Torre, 2013), audible sound (Mandal, et al., 2005), sensor (Haque, 2014), and radio frequency (RF). However, object obstructions and signal reflections will affect the received signal strength (RSS). A more sophisticated estimation algorithm is required to handle this problem. Generally, fingerprinting-based algorithm will provide a higher accuracy as it represents the environment. The environment is captured in the offline phase by picking RSS vector in several reference points that is called radio map. This radio map will be used as a reference to estimate the receiver's location in the online phase. The disadvantage of using fingerprinting-based algorithm is the offline phase that is tedious and time-consuming as the size of the room increases (Hossain & Soh, 2015).

RF signals such as Wi-Fi, RFID (Papapostolou & Chaouchi, 2011), and Bluetooth has become the most reliable solution due to its low-cost and reasonable accuracy. Wi-Fi has become the most mature research topic amongst other RF signals, exploiting the channel state information (CSI) on advanced Wi-Fi network interface card (NIC) (Wang, Gao, & Mao, CSI Phase Fingerprinting for Indoor Localization, 2017) (Wang, Gao, Mao, & Pandey, DeepFi: Deep Learning for Indoor Fingerprinting Using Channel State Information, 2015). With a high accuracy, researchers have started to work on IPS with human intervention (Yang, Wu, & Liu, 2012). However, Bluetooth using Bluetooth Lower Energy (BLE) provides lower energy consumption, lower network latency, cheaper price, and ideal for single-hop communication (Gomez, Oller, & Paradells, 2012). It is possible to use an array of BLEs to cover larger area while keeping the accuracy high in a reasonable cost.

The downside of using Bluetooth is that the signals fluctuates more than other RF signals. This is a big problem in using Bluetooth as the RSS in the online phase might differ by a large margin compared to the radio map. With the radio map is less related to the RSS, the algorithm is also less accurate to estimate the receiver's location.

Typically, a single RSS vectors in radio map is obtained by averaging several RSS values over time to increase its

reliability. However, in our case, the average value barely represents the signal propagation as presented in Fig 1. The RSS value varies in range of ±6 dBm from its average. Later we suspect that the frequency distribution of the RSS values is more informative than its average value.

This paper is a proof-of-concept on using RSS frequency distribution values that we called Frequency Distribution Radio Map (FDRM), instead of RSS average values. We also propose a probabilistic fingerprinting FDRM-based method

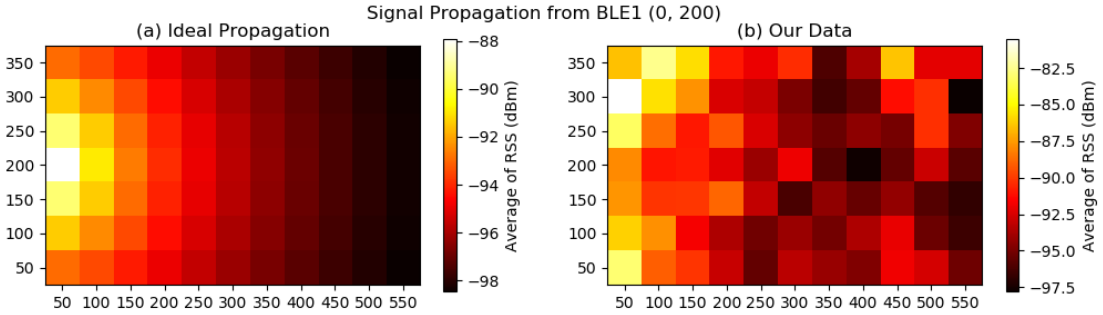


Fig. 1. RSS does not follow the propagation loss model due to fluctuation.

using Gaussian Mixture Model (GMM) as the probability distribution function (PDF). The experiment was held in a 4 m x 6 m room using 4 BLEs. Three different datasets are used to evaluate our proposed method.

There are several positioning algorithms, which are usually classified into 2 types: fingerprinting and non-fingerprinting. Fingerprinting uses reference points collected in the offline phase, and the data will be used to estimate the location in the online phase. Non-fingerprinting algorithm does not use any reference points; therefore, it needs other data to estimate. Measurement such as angle of arrival (AoA) (Wong, Klukas, & Messier, 2008), time of arrival (ToA), time difference of arrival (TDoA) (Han, Lu, & Lan, 2010) and RSS can be use instead of reference points. Therefore, non-fingerprinting algorithm does not require offline phase, which is an advantage by itself. Unfortunately, Bluetooth provides less precision of time synchronization, and it is hard to measure angle (Wang, Yang, Zhao, Liu, & Cuthbert, 2013). Typically, non-fingerprinting method on BLE will rely on RSS.

Since there is only online phase, estimation should be done by using only RSS vectors. The only way to estimate location is to convert RSS into distance using some kind of propagation-loss model. The distance between location and each BLE can be used to find the location by geometrical algorithm such as Trilateration/Triangulation (Paterna, Auge, Aspas, & Bullones, 2017), Heron-Bilateration (Chung-Hao Huang, 2015) and Least Square (Li, 2014) algorithm. However, the accuracy of the estimation will heavily rely on the propagation-loss model. Since Bluetooth has a lot of fluctuation, therefore there will be a lot of false distance value that will lower the accuracy. Researchers has been trying to improve the propagation-loss model specifically for BLE (Onofre, Silvestre, Pimentão, & Sousa, 2016). However, different environment such as object obstructions will also resulting a different propagation model. Some research tried to use machine learning to model the propagation loss (Chandel, Ahmed, Arora, & Ghose, 2016).

In 2000, Bahl and Padmanabhas (Bahl & Padmanabhan, 2000) as they recorded radio signals that will be used to validate and estimate the location. Later this method is termed Fingerprinting, with the recorded radio signals termed as radio map. Fingerprinting consists of 2 steps: offline phase to collect the radio map from reference points, and online phase to estimate the asked location. Generally fingerprinting has a better accuracy because the reference points will represent the environment. There are 2 types of fingerprinting algorithm: deterministic fingerprinting and probabilistic fingerprinting. One method in deterministic fingerprinting is to use KNN with certain RSS distance metric as a degree of similarity between RSS vectors. Yu-Chi and Pei-Chun in their research shows that Chebyshev distance provides higher accuracy compared to Euclidean Distance (Pu & You, 2018). A method named Enhanced weighted KNN on Wi-Fi fingerprint was proposed (Shin, Lee, Lee, & Kim, 2012) with the ability to dynamically change the k value to enhance the accuracy even further.

Probabilistic fingerprinting method use a posteriori probability of location, given the RSS vector received in the online phase to estimate in the online phase. A posteriori probability functions of RSS vector given location of reference point are also defined in the offline phase. These 2 probabilities are related to each other through Bayes' Theorem. Typically, researchers are working on the different probability function used in the reference points. Xuyu et al. proposed DeepFi method, using Autoencoder and Radial Basis Function (RBF) to calculate the probability of input given location (Wang, Gao, Mao, & Pandey, DeepFi: Deep Learning for Indoor Fingerprinting Using Channel State Information, 2015), but they used Wi-Fi CSI instead of the regular RSS. Their method achieved mean error of 94.25 cm which one of the best algorithms in IPS using CSI.

A similar work to our method has been researched before. Alfakih used GMM to approximate the PDF of RSS in Wi-Fi fingerprints (Alfakih, Keche, & Benoudnine, 2015) resulting an improvement in the overall performance. Abhishek et al. also mentioned in their research that GMM using EM will have less intensive training process compared to other fingerprinting method (Goswami, Ortiz, & Das, 2011). GMM can also handle time varying issues since the PDF is able to learn the environment including the time variation.

Other than a positioning algorithm, GMM can also be used as a signal strength prediction in a wireless network (Prashant, M, Shreyas, Chaithanya, & Kuttaiah, 2009). Therefore, when fluctuations are happening or the device is disconnected from the network, the device can use the predicted signal strength instead of the RSS. The signal strength can be predicted from the RSS history using Markov

Chain. This might increase the accuracy of a tracking algorithm, however the possibility of using GMM signal strength prediction on BLE has not yet been explored.

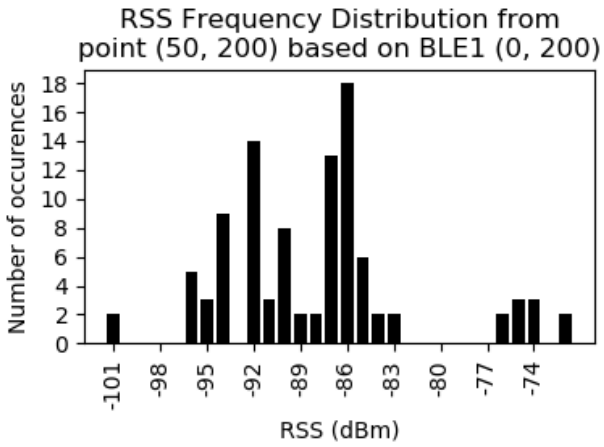


Fig. 2. RSS Frequency Distribution is more informative compared to the mean value.

II. METHODS

In theory, reference points should be able to capture the environment, and this includes the object obstructions and signal fluctuations. However, as mentioned in Fig 1, using mean RSS as the RSS vectors barely represent the propagation model. By looking at Fig 2, RSS frequency distribution is more informative and representative than a single mean value. Therefore, we propose a method utilizing the frequency distribution as the RSS vectors.

Suppose that there are N reference points and M BLEs numbered from 1 to M. Define RP as the reference point and TP as the testing point that we want to estimate. The RSS vector in the reference point is called $\vec{r}$ and the RSS vector in the testing point is called $\vec{t}$. The vector $\vec{r}$ consists of M different RSS frequency distribution taken from each BLEs. Meanwhile the vector $\vec{t}$ consists of M numbers representing the RSS value received at the online phase.

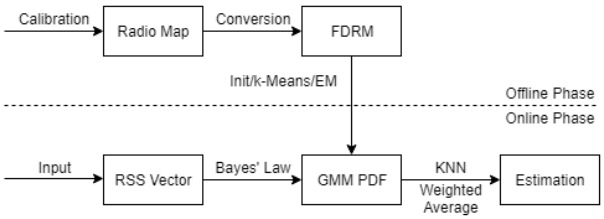


Fig. 3. Flowchart of the proposed algorithm.

A. Offline Phase: Gaussian Mixture Model

The posterior probability of receiving certain RSS value from the i-th BLE given its location is represented as the PDF of GMM. The GMM is calculated as a sum of several Gaussian PDFs that is weighted by its mixture coefficient. Suppose that there are g different Gaussians numbered from 1 to g, therefore

$$P(rss_i|RP) = \sum_{j=1}^{g} w_j \, G(rss_i|\mu_j, \sigma_j^2)\,; \quad (1)$$

The GMM will vary based 4 parameters: the number of Gaussians, the centre of Gaussians, the covariance of Gaussians, and the mixture coefficients. These values can be obtained through 3 steps: initialization, k-Means, and Expectation-Maximization (EM).

First, the GMM parameters will be initialized based on the frequency distribution obtained from the calibration process. The number of Gaussians will be the number of distinct RSS values, with the centre being the RSS value itself and the covariance is by default set to 1. The mixture coefficient will be set to the ratio between its frequency and the total frequency. These initial parameters themselves are already forming a GMM and ready to be tested. However, the next step is optional, with the intention to improve the result even further.

The next step is k-Means that will update the GMM parameters except the covariances. First, we can choose k random Gaussians as long as $k \leq g$, and this will be the initial value for our k-Means. This step is completed after iterating the k-Means until it converges. The result of this step is k Gaussians with new center, with the mixture coefficient being sum of frequencies that is in the same cluster with the center, divided by the total frequency.

Since the result of k-Means is also a GMM, the final step which is EM is also optional. By using EM, we are starting to change the covariances of the Gaussians. This algorithm also might change the number of Gaussians, since there is a possibility that the mixture coefficient will converge to 0. When the mixture coefficient is 0, we can neglect that Gaussian.

First, the initial value for the EM is the parameters gained from k-Means, with the initial value of the covariance. Every iteration consists of 2 steps: E-step and M-step. In E-step, the likelihood of every data is calculated based on the current parameter. Meanwhile in M-step, the parameter will be updated in order to maximize the likelihood function. By iterating EM until it converges, we finally get the final parameters for our GMM, which represents the maximum likelihood of the data.

B. Online Phase: K-Nearest Neighbour & Weighted Sum

In probabilistic fingerprinting, posterior probability is assigned to each RPs as the probability of them becoming the estimation given RSS vector received in that point. By assuming each BLEs are conditionally independent between each other, the posterior probability of a RP is described as follows:

$$P(RP|\vec{t}) = \prod_{i=1}^{m} P(RP|t_i) \quad (2)$$

with $P(EP = RP|t_i)$ is the return value by inserting the value of $tRSS_i$ to the PDF at point RP that is taken from the i-th BLE. Then we can use KNN to choose k RPs with highest probability returned by the PDF. Suppose that the k RPs is numbered from 1 to k, then using Bayes' theorem we can calculate the value of:

$$P(RP|t_i) = \frac{P(RP)P(t_i|RP)}{\sum_{j=1}^{k} P(RP_j)P(t_j|RP_j)} \quad (3)$$

with $P(t_i|RP)$ being a Gaussian Mixture Model PDF given location RP using the i-th BLE. By assuming an equal distribution of probability between each RPs, the value of $P(RP|t_i)$ can be simplified into:

$$P(RP|t_i) = \frac{P(t_i|RP)}{\sum_{j=1}^{k} P(t_i|RP_j)} \quad (4)$$

This a posteriori probability can be used as the weight of the k chosen RPs in weighted average to calculate the estimated location.

$$\overrightarrow{TP} = \sum_{i=1}^{k} P(RP_i|\vec{t})\overrightarrow{RP_i}\,;\ \sum_{i=1}^{k} P(RP_i|\vec{t}) = 1 \quad (5)$$

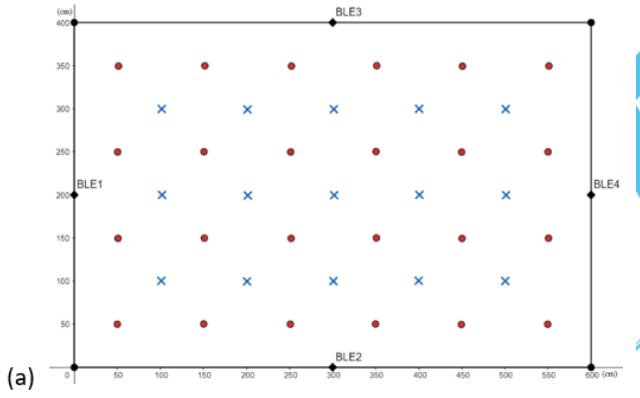
In our method, KNN has the role of eliminating RPs that are unlikely to become the estimation point. If such points are included in the weighted average, it will increase the

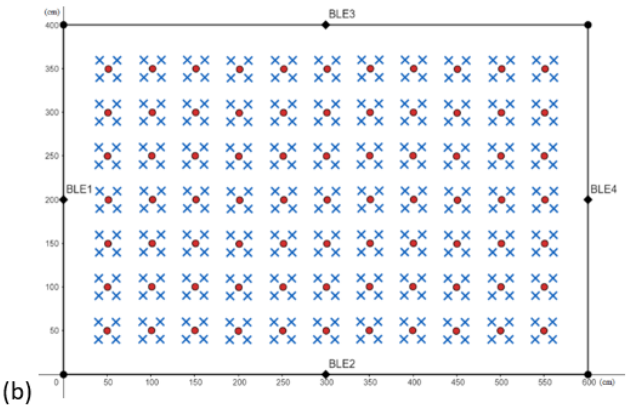
distance error. If the *k* value in KNN is too small, then it will eliminate too much information, resulting an unreliable estimation. If the *k* value in KNN is too high, then the KNN will barely give any impact to the overall estimation. Therefore the *k* value is included as the parameter explored in this experiment.

III. RESULTS AND DISCUSSION

A. Experimental Design

In this experiment, we tested 2 different method to construct GMM: k-Means only, and k-Means with EM. As a benchmark, we use KNN with Chebyshev Distance as the distance metric (Pu & You, 2018). The experiment was conducted in a 4 m x 6 m classroom equipped with chairs and tables. We also prepared 3 datasets gathered from the same room, with different configuration of RPs and TPs as shown in Fig 4. The data collection process follows the configuration in dataset 2, as the other datasets can be obtained from dataset 2. In total, there will be 4 BLEs used in this experiment, each has the configuration of TX Power 3 with broadcast interval 500 ms. The BLE signal was received through a smartphone at a height of 125 cm, using our BLE RSS Application that was created in Android Studio.





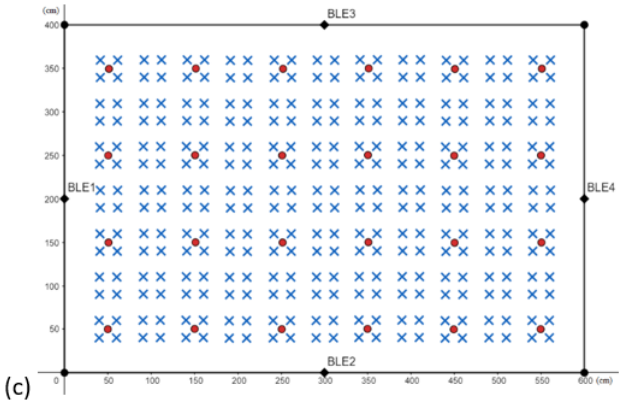


Fig. 4. RPs (red dot) and TPs (blue cross) from dataset 1 (a), 2 (b), and 3 (c).

For each RPs, 100 signal data will be received from each BLE. These data will be converted into frequency distribution for the FDRM. Meanwhile for each TPs, 10 signal data will be received from each BLE. However, since the RSS vector for the TP only contains RSS values, therefore we took the mean value from the 10 RSS values.

Between 3 datasets, our original dataset was dataset 2. Each RP forms a grid with the distance between adjacent points is 50 cm. In total, there are 77 RPs. The TPs are ±10 cm from the x-value and y-value of every RP. Therefore, each RPs will have 4 TPs surrounding it, and in total there are 308 TPs.

However, the dataset that is used in some research is similar to the dataset 1. Therefore, we included this dataset as a benchmark. The distance between adjacent RPs is 100 cm, and in total there are 24 RPs. The TPs are the middle of every grid, resulting a total of 15 TPs. Notice that both RPs and TPs in dataset 1 are subset of the RPs in dataset 2. Therefore, there is no additional data collection to gather this dataset. Dataset 1 should be easier to estimate compared to dataset 2, since the formation is much simpler.

Finally, dataset 3 is the combination of dataset 1 and 2. The RPs used are the 24 points from dataset 1, meanwhile the TPs used are the 308 points from dataset 2. The reason why this dataset is used is to evaluate the relation between the number of RPs with the overall performance. Having more RPs usually resulting in better accuracy. However, since the calibration process is time-consuming, it is better to have an enough RPs while maintaining similar performance.

Notice that there are points in dataset 2 and 3 that are outside the reach of the RPs. Since the sum of weights is always 1, the resulting estimation will be always towards some RPs. Therefore, every point that the x value is not between 50 and 550, or the y value is not between 50 and 350, is unreachable by using weighted average no matter what method is used. This is one of the disadvantages on using weighted average in fingerprinting-based algorithm as it can only estimates location between RPs. However, our dataset is intended to contain such points since this case will happen in real scenario. Another solution is to have a better configuration of RPs. However, we also want to test out if the proposed method can minimize the distance error.

Table- I. Summary of differences between 3 datasets.

Dataset	#RP	#Samplings per RP	RP Grid Size (cm)	#TP	#Samplings per TP
1	24	100	100	15	100
2	77	100	50	308	10
3	24	100	100	308	10

In order to achieve the best result from each method, all possible parameters from each method were explored. The parameter for KNN and k-means is the *k* value, meanwhile EM relies on the initial value taken from the k-means. We also evaluate the performance from each iteration of EM to understand what impact EM will give to the GMM. All algorithm used in this experiment was implemented and tested in C++11. The indicators used to evaluate performance of the methods are mean distance error and standard deviation. The distance error between the actual location and the estimation is calculated using Euclidean Distance.

In addition, there is a small chance in probabilistic fingerprinting of getting 0 probability from each RPs, resulting estimation cannot be calculated. In this case, the estimation is considered failed and the distance error will be set to the maximum distance between TPs to compensate this issue. In dataset 1, the maximum distance between TPs is 447.21 cm. Meanwhile in dataset 2 and 3, the maximum distance between TPs is 610.57 cm.

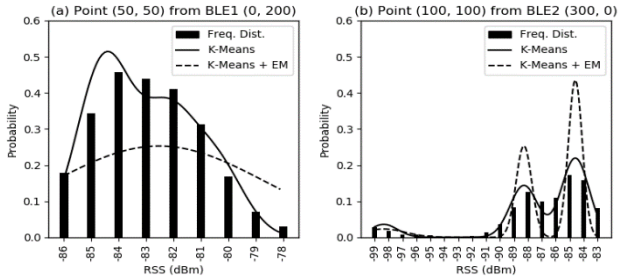


Fig. 5. Behaviour of K-Means and K-Means + EM on the frequency distribution.

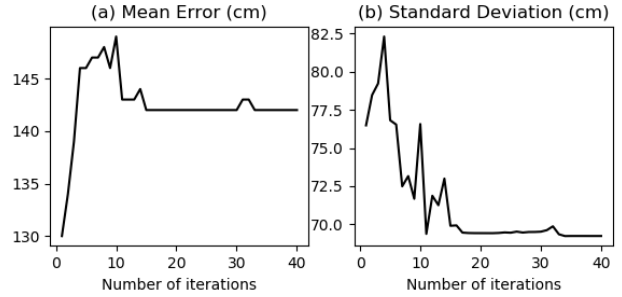


Fig. 6. The effect of EM towards the positioning result over its iteration.

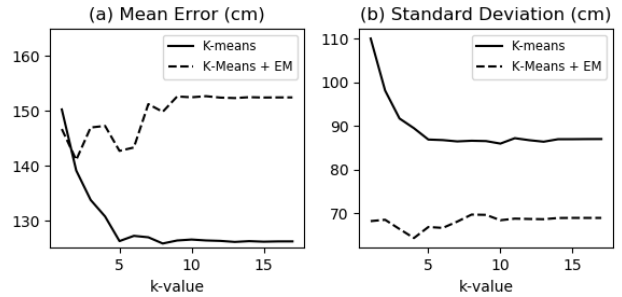


Fig. 7. The effect of K-Means towards the positioning result over the value of k.

B. Experimental Result

In each RPs, the number of distinct RSS values will vary from 6 to 19 values. When the frequency is distributed equally, this will give a bad PDF that might lower the accuracy. In Fig 5, the left chart shows a PDF that is distributed quite equally. The GMM is also less representative compared to the right chart. The reason is because corner points are far from BLEs, therefore the RSS is less accurate than points in the middle of the room. By applying EM, it changes the PDF to a simpler PDF with smaller value. Meanwhile in a better PDF, applying EM reduces the covariance of the Gaussians. Furthermore, the GMM from k-Means seems fitter to the frequency distribution compared to the k-Means + EM. This can be caused by either the GMM from k-Means is still overfit, or the GMM from k-Means + EM is already underfit.

Even though EM is not mandatory, Fig 7 shows us the role of EM in this method by comparing the estimation over the iteration. At the beginning, the mean error is very low, and the standard deviation is acceptable. Then, in the first couple iteration, the GMM performs worse than the initial value. Then, the mean error is started to converge, meanwhile the standard deviation keeps getting lower and lower. Finally, the GMM is started to converge between 100-200 iteration (157 iteration for Fig 7), resulting the lowest standard deviation. This concludes that it is necessary to iterate EM until the GMM converges to achieve best result.

In our experiment, the maximum number of distinct RSS values amongst all RPs is 19. When $k = 1$, the GMM is already converged, therefore the EM immediately stops, resulting the same GMM from k-Means. When $k = all$, all the distinct RSS values are treated as different Gaussians, therefore the k-Means immediately stops, resulting a GMM from the initial FDRM.

Table- II. Statistical result from the best performance of each method in 3 datasets.

Indicator	No GMM	K-Means	K-Means + EM
Dataset 1			
Mean error (cm)	125.84	98.18	112.24
Std. dev. (cm)	74.92	56.11	47.99
Min error (cm)	14.71	12.040	10.79
Max error (cm)	284.97	199.39	180.81
90th percentile (cm)	197.56	161.60	165.65
Error < 100 cm (%)	40	53.33	46.67
Error > 200cm (%)	13.33	0	0
Dataset 2			
Mean error (cm)	141.23	125.83	141.06
Std. dev. (cm)	71.26	86.64	68.57
Min error (cm)	11.20	2.58	8.75
Max error (cm)	450.92	3 points not found	400.25
90th percentile (cm)	229.09	211.01	230.94
Error < 100 cm (%)	30.84	43.51	30.19
Error > 200cm (%)	20.45	12.99	17.53
Dataset 3			
Mean error (cm)	148.84	136.23	140.06
Std. dev. (cm)	70.35	105.79	69.13
Min error (cm)	18.60	6.108	11.48
Max error (cm)	443.68	7 points not found	366.13
90th percentile (cm)	244.25	244.11	232.75
Error < 100 cm (%)	27.92	42.86	29.87
Error > 200cm (%)	23.70	16.56	18.51

The lowest mean error that k-Means achieved is 125.83 cm when $k = 9$. When the value of k is too small, the GMM will underfit. The same thing happens when the value of k is too big, the GMM will overfit. Therefore when $k = 9$, it is enough to preserve the best GMM without removing too much Gaussians. However, this does not happen with the result from EM. The mean error seems unstable with small value of k, but at the same time approaching worse mean error as the value of k increases. For all values of k, while the mean error of k-Means is consistently lower than EM, the standard deviation is consistently higher. This proves our hypothesis that the role of EM in this method is to achieve lower standard deviation. The lowest standard deviation achieved by EM is 61.59 cm when $k = 6$. Unfortunately, when $k = 6$, the mean error of EM is 151.90 cm which is considerably bad. Other value of k that has low standard deviation with reasonable mean error is $k = 3$. The mean error is 141.06 cm, while the standard deviation is 68.57 cm.

Finally, by comparing the result between each method in our 3 datasets, this experiment proves our concept in using FDRM instead of the regular radio map. Our proposed method also outperforms the regular KNN in all datasets. This concludes that GMM captures the environment better, therefore resulting a better estimation. The method using only k-Means in constructing the GMM serves the overall best result between 3 methods. Our hypothesis is shown to be correct consistently through all datasets. EM holds the lowest standard deviation amongst 3 method, while holding better accuracy than the regular KNN at the same time.

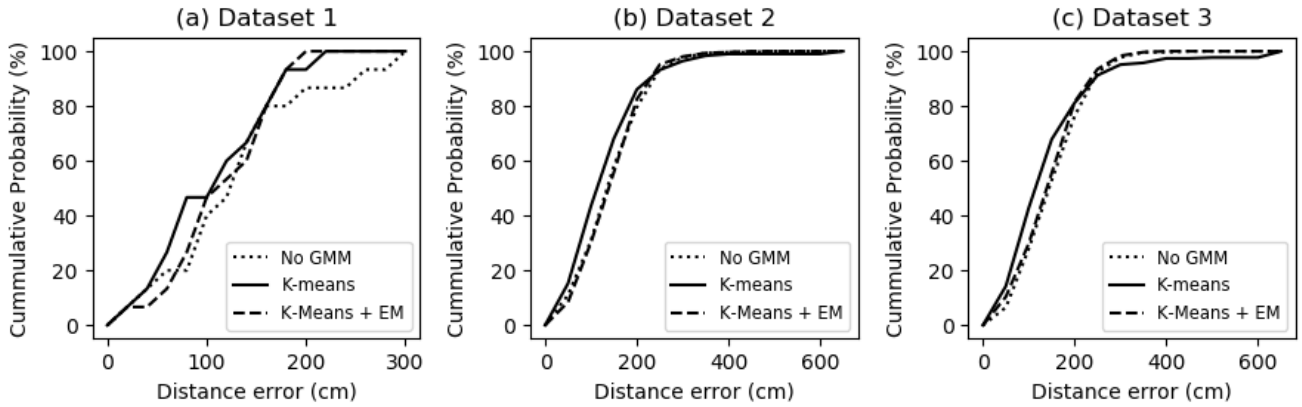


Fig 8. Cumulative Distribution Function (CDF) of all methods between dataset 1 (a), 2 (b), and (c).

Table- III. Parameters used between methods that achieved the result in Table III.

Dataset	No GMM	K-Means		K-Means + EM	
	KNN	*K-Means*	*KNN*	*K-Means*	*KNN*
1	6	4	6	8	7
2	14	9	15	3	14
3	5	12	11	8	5

By comparing the result in dataset 2 and dataset 3, we also found out that k-Means GMM will perform better when using more RPs. Even though the performance can be increased by using more RPs, it will require more work in the calibration process. The number of RPs in dataset 2 is more than 3 times the number of RPs in dataset 3, however the mean error is only 11 cm better. Therefore, less RPs can be used if the system does not require a precise estimation.

Contrary to the k-Means GMM and the benchmark method, k-Means + EM performs slightly better with less RPs. This is beyond our expectation since generally fingerprinting-based method will perform better with more RPs. This contradicts Abhishek's research on WiGEM that the accuracy of GMM will be better as the finer the grid size is (Goswami, Ortiz, & Das, 2011). The maximum distance error in dataset 3 is significantly lower compared to dataset 2, and this lower the mean error by 1.00 cm. Also, the mean error is only 3.82 cm higher compared to k-Means GMM, but the standard deviation is 36.66 cm lower. This concludes that k-Means + EM is preferable if there is a smaller number of RPs.

By analysing Fig 8, k-Means GMM always performs better at 80% of the data compared to the other methods. However, the remaining 20% are quite bad, and this lowers the overall performance of the method. It turns out that 7.14% in dataset 2 and 8.12% in dataset 3 are caused by the points outside the convex hull of the RPs, which is the main limitation of weighted sum. This shows that k-Means GMM is worse in reducing the distance error on such points. Meanwhile k-Means + EM GMM performs consistently better compared to No GMM, especially in dataset 2 and dataset 3.

One of the disadvantages of using our method has been mentioned previously, that there is a small chance of getting no estimation due to having 0 probability in each RPs. By looking at the k-Means only method, there are: 0 points not found in dataset 1, 3 points not found in dataset 2, and 7 points not found in dataset 3. This issue will happen if the GMM has not captured all possibilities in the environment. As result, it causes the method using k-Means GMM has the highest standard deviation compared to the other method, even though it provides the lowest mean error.

By analysing Table 2., there are several solutions to handle this issue:

- Better calibration process. This can be done by increasing the number of RPs, as shown from the result from dataset 2 and dataset 3 in Table 2. More signal sampling can be done to make sure the GMM has captured the whole environment. Unfortunately, fixing the issue from this step is time consuming.
- Better GMM. As shown in Table 2, GMM constructed by k-Means and EM does not have this issue throughout all

3 datasets. However, since the mean error from this method is higher, this is not necessarily the best solution either.

In summary, both k-Means GMM and k-Means + EM GMM have shown an overall improvement of mean error and standard deviation compared to the benchmark method, which is the KNN with Chebyshev distance as the distance metric. Unfortunately, due to some fail estimation in dataset 2 and dataset 3, the k-Means GMM has the worst standard deviation. Therefore,

- Use k-Means if: the system requires the best performance overall and have considered to collect more RPs in the calibration process.
- Use k-Means + EM if: the system requires lowest standard deviation and small computing time, but still have a low mean error with small number of RPs.

IV. CONCLUSION

In this paper, we proposed a concept on using frequency distribution as the RSS vector used in the radio map, named FDRM. We also proposed a probabilistic fingerprinting-based algorithm using GMM as the PDF. In the offline phase, The GMM is obtained from the FDRM, then it can be improved by using either k-Means only or k-Means with EM. In the online phase, the GMM will be used as the distance metric in KNN and the weight in weighted average. This method is validated using 3 different datasets, taken from a 4 m x 6 m classroom. Our experiment shows that FDRM is more informative compared to the mean, and our proposed method giving better performance in all indicators. K-Means provides a more complex GMM, resulting the lowest mean error in all datasets. K-Means + EM provides a simpler GMM, resulting the lowest standard deviation but a higher mean error in all datasets. Overall, our experiment shows that the proposed fingerprint and method can learn BLE signal fluctuation better than the benchmark method.

In the future, we want to explore the possibilities of FDRM even further. The probability distribution of the RSS is not necessarily a Gaussian, therefore a more sophisticated algorithm such as neural networks can be used instead. The frequency distribution can also be used in deterministic fingerprinting by calculating degree of similarity between the input and the FDRM.

REFERENCES

Alfakih, M., Keche, M., & Benoudnine, H. (2015). Gaussian Mixture Modeling for Indoor Positioning WIFI System. *2015 3rd International Conference on Contro, Engineering & Information Technology (CEIT)*. Tlemcen.

Bahl, P., & Padmanabhan, V. N. (2000). RADAR: An In-Building RF-based User Location and Tracking System. *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*. Tel Aviv, Israel.

Chandel, V., Ahmed, N., Arora, S., & Ghose, A. (2016). InLoc: An end-to-end robust indoor localization and routing solution using mobile phones and BLE beacons. *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. Alcala de Henares.

Chung-Hao Huang, L.-H. L.-L.-H. (2015). Real-Time RFID Indoor Positioning System Based on Kalman-Filter Drift Removal and Heron-Bilateration Location Estimation. *IEEE Trans. Instrumentation and Measurement, 64*(3), 728-739.

Fisher, J. A. (2006). Indoor positioning and digital management: Emerging surveillance regimes in hospitals. In *Surveillance and Security: Technological Politics and Power in Everyday Life* (pp. 89-100). Abingdon: Routledge.

Gomez, C., Oller, J., & Paradells, J. (2012). Overview and Evaluation of Bluetooth Low Energy: An Emerging Low-Power Wireless Technology. *Sensors, 12*(9), 11734-11753.

Goswami, A., Ortiz, L. E., & Das, S. R. (2011). WiGEM: a learning-based approach for indoor localization. *CoNEXT '11 Proceedings of the Seventh COnference on emerging Networking EXperiments and Technologies*. Tokyo.

Han, T., Lu, X., & Lan, Q. (2010). Pattern recognition based Kalman filter for indoor localization using TDOA algorithm. *Applied Mathematical Modelling, 34*(10), 2893-2900.

Haque, I. T. (2014). A sensor based indoor localization through fingerprinting. *Journal of Network and Computer Applications, 44*, 220-229.

Hossain, A. K., & Soh, W.-S. (2015). A Survey of Calibration-free Indoor Positioning Systems. *Computer Communications, 66*, 1-13.

Ke, C., Wu, M., Chan, Y., & Lu, K. (2018). Developing a BLE Beacon-Based Location System Using Location Fingerprint Positioning for Smart Home Power Management. *Energies, 11*(12), 3464.

Kim, S.-C., Jeong, Y.-S., & Park, S.-O. (2013). RFID-based indoor location tracking to ensure the safety of the elderly in smart home environments. *Personal and ubiquitous computing, 17*(8), 1699-1707.

Lee, C. (2004). Indoor positioning system based on incident angles of infrared emitters. *30th Annual Conference of IEEE Industrial Electronics Society, 2004. IECON 2004*. Busan, South Korea.

Li, H. (2014). Low-Cost 3D Bluetooth Indoor Positioning with Least Square. *Wireless Personal Communications, 78*(2), 1331-1344.

Mandal, A., Lopes, C. V., Givargis, T., Haghighat, A., Jurdak, R., & Baldi, P. (2005). Beep: 3D Indoor Positioning Using Audible Sound. *IEEE Consumer Communications and Networking Conference (CCNC'05)*. Las Vegas.

Medina, C., Segura, J. C., & Torre, Á. l. (2013). Ultrasound Indoor Positioning System Based on a Low-Power Wireless Sensor Network Providing Sub-Centimeter Accuracy. *Sensors, 13*(3), 3501-3526.

Mendelson, E. (2014, October 21). *U.S. Patent No. 8,866,673.*

Onofre, S., Silvestre, P. M., Pimentão, J. P., & Sousa, P. (2016). Surpassing Bluetooth Low Energy Limitations on Distance Determination. *2016 IEEE International Power Electronics and Motion Control Conference (PEMC).* Varna.

Papapostolou, A., & Chaouchi, H. (2011). RFID-assisted indoor localization and the impact of interference on its performance. *Journal of Network and Computer Applications, 34*(3), 902-913.

Paterna, V. C., Auge, C. A., Aspas, J. P., & Bullones, M. A. (2017). A Bluetooth Low Energy Indoor Positioning System with Channel Diversity, Weighted Trilateration and Kalman Filtering. *Sensors, 17*(12), 2927.

Prashant, K., M, N. A., Shreyas, N., Chaithanya, N., & Kuttaiah, P. (2009). Gaussian Mixture Model-Expectation Maximization based Signal Strength Prediction for Seamless Connectivity in Hybrid Wireless Networks. *MoMM '09 Proceedings of the 7th International Conferencee on Advances in Mobile Computing and Multimedia.* Kuala Lumpur, Malaysia.

Pu, Y., & You, P. (2018). Indoor Positioning System Based on BLE Location Fingerprinting with classification approach. *Applied Mathematical Modelling, 62*, 654-663.

Ramani, S. V., & Tank., Y. N. (2014). Indoor Navigation On Google Maps And Localization Using RSS Fingerprinting. *International Journal of Engineering Trends and Technology, 11*(4), 171-173.

Ruggiero, L., Charith, D., Song, X., & Lucia, B. (2018). Investigating pedestrian navigation in indoor open space environments using big data. *Applied Mathematical Modelling, 62*, 499-509.

Shin, B., Lee, J. H., Lee, T., & Kim, H. S. (2012). Enhanced weighted K-nearest neighbor algorithm for indoor Wi-Fi positioning systems. *2012 8th International Conference on Computing Technology and Information Management (NCM and ICNIT).* Seoul.

Wang, X., Gao, L., & Mao, S. (2017). CSI Phase Fingerprinting for Indoor Localization. *IEEE Internet of Things Journal, 3*(6), 1113-1123.

Wang, X., Gao, L., Mao, S., & Pandey, S. (2015). DeepFi: Deep Learning for Indoor Fingerprinting Using Channel State Information. *2015 IEEE wireless communications and networking conference (WCNC).* New Orleans.

Wang, Y., Yang, X., Zhao, Y., Liu, Y., & Cuthbert, L. (2013). Bluetooth Positioning using RSSI and Triangulation Methods. *2013 IEEE 10th Consumer Communications and Networking Conference (CCNC).* Las Vegas.

Wong, C., Klukas, R., & Messier, G. G. (2008). Using WLAN infrastructure for angle-of-arrival indoor user location. *2008 IEEE 68th Vehicular Technology Conference.* Calgary.

Woo, S., Jeong, S., Mok, E., Xia, L., Choi, C., Pyeon, M., & Heo, J. (2011). Application of Wifi-based Indoor Positioning System For Labor Tracking At Construction Sites A Case Study In Guangzhou MTR. *Automation in Construction, 20*(1), 3-13.

Yang, J., Wang, Z., & Zhang, X. (2015). An iBeacon-based Indoor Positioning Systems for Hospitals. *International Journal of Smart Home, 9*(7), 161-168.

Yang, Z., Wu, C., & Liu, Y. (2012). Locating in Fingerprint Space: Wireless Indoor Localization with Little Human Intervention. *Proceedings of the 18th annual international conference on Mobile computing and networking.* Istanbul.