WKNN indoor Wi-Fi localization method using k-means clustering based radio mapping

Siyang Liu, Raul de Lacerda, Jocelyn Fiorina

▶ **To cite this version:**

Siyang Liu, Raul de Lacerda, Jocelyn Fiorina. WKNN indoor Wi-Fi localization method using k-means clustering based radio mapping. 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), Apr 2021, Helsinki, Finland. 10.1109/VTC2021-Spring51267.2021.9448961 . hal-03697714

HAL Id: hal-03697714
https://hal.science/hal-03697714v1

Submitted on 17 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

WKNN indoor Wi-Fi localization method using k-means clustering based radio mapping

Siyang LIU
Laboratoire des Signaux et Systèmes
Université Paris-Saclay, CNRS, CentraleSupélec
Gif-sur-Yvette, France
siyang.liu@centralesupelec.fr

Raul DE LACERDA
Laboratoire des Signaux et Systèmes
Université Paris-Saclay, CNRS, CentraleSupélec
Gif-sur-Yvette, France
raul.delacerda@centralesupelec.fr

Jocelyn FIORINA
Laboratoire des Signaux et Systèmes
Université Paris-Saclay, CNRS, CentraleSupélec
Gif-sur-Yvette, France
jocelyn.fiorina@centralesupelec.fr

***Abstract*—Wifi fingerprinting using received signal strength has been widely studied for indoor localization. Classic similarity-based methods like weighted K-nearest neighbor (WKNN) localize targets by searching the best matching fingerprint in the dataset. Performance of these methods suffers from RSS variance and they are slow under a large size of fingerprint dataset. In this paper, we propose a WKNN localization strategy using k-means clustering radio mapping to improve localization precision while mitigating computational complexity.**

***Index Terms*—Indoor localization, WiFi fingerprinting, RSS, k-means clustering, WKNN**

I. Introduction

With rapid development of Internet of Things (IoT), the need of location based services such as asset management, routing, and equipment positioning has also grown overtime. Global Positioning System (GPS), which supports a large number of location based applications is not adapted for indoor environments due to signal attenuation and scattering caused by walls and other obstacles[1]. Therefore, indoor localization system becomes an area of interest.

Localization can be performed using fingerprinting approach which exploits the mapping between measurements and positions[1], [2]. The process of fingerprinting includes an offline training phase and an online localization phase.

Received signal strength (RSS) is a commonly used feature for fingerprinting. However, it is easily affected by environment factors and it fluctuates even in the same position [3]. One way to address RSS variance problem is to use probabilistic approach. A. Haeberlen et al. [4] proposed a Gaussian fit sensor model, which is robust against untrained fluctuations. Another way to compensate is to sample multiple times in the same reference point (RP) over a long period of time. With redundant measurements, M. Adriano et al. [5] proposed to only use the minimum feature distance between test sample and all RSS samples at the same RP for WKNN which ensures that each RP will only contribute with the best similarity value.

Even though sampling RSS multiple times can provide more information for localization, a large fingerprint database not only results in high power consumption but also long processing time during online phase especially for similarity based methods like WKNN [6]. Improving accuracy and reducing computational complexity are two important directions of current localization studies [7].

Many methods are proposed to reduce the dataset used for similarity comparison mainly by reducing the number of fingerprints or the dimension of each sample. Radio mapping methods compress radio database during the offline phase. A. Arya et al. [8] used hierarchical clustering on the entire dataset taking into account both the location and the radio components of samples. S. G. Lee et al. used k-means clustering to reduce redundant measurements on each RP [9]. Clustering methods divide the entire radio map into clusters during offline phase and localize the TP inside the most relevant cluster [7]. Reduction of dataset can also take place during the online phase by filtering the radio map for relevant RPs [10]. AP selection and data quantisation use prior knowledge to reduce sample dimension [10], [11]. Feature selection methods like Principle Component Analysis (PCA) technique recover the low-rank matrix from a noisy RSS matrix [12]. Deep neural network such as auto-encoder can also be used to reduce the dimension of feature vectors [13].

In this paper, a strategy combining radio mapping based on k-means clustering and WKNN algorithm is proposed aiming to improve localization efficiency while maintaining accuracy. To further reduce fingerprint data size, a variant of the proposed method is also introduced, adding an extra spatial filtering step. The remainder of the paper is organized as follows: Section II presents notations and system model. The proposed improved WiFi fingerprinting method and its variant are presented in Section III. Section IV presents numerical results for performance evaluation, and Section V concludes the paper.

II. System Model

We consider an Area of Interest (AoI) with R fixed reference points $\{RP_r\}_{r=1}^{R}$ with known coordinates (x_r, y_r) and M fixed access points $\{AP_m\}_{m=1}^{M}$ whose positions are not specified. A RSS sample of an arbitrary point on the AOI would be a vector of size $M \times 1$ constituted by the Received Signal Strength Indicator (RSSI) of all M APs. Let's assume that at each RP several measurements were taken, where the number

This work is financially supported by China Scholarship Council from the Ministry of Education of P.R. China.

of measurements are represented by $\{N_r\}_{r=1}^{R}$ and we define the total number of RP measurements as $N = \sum N_r$.

At test point q, RSS sample is taken forming a vector of size M. Sample distance between samples i and j is given by:

$$ds_{i,j} = \sqrt{\sum_{m=1}^{M} (RSSI_j^m - RSSI_i^m)^2}. \quad (1)$$

Based on the fact that for each RP_r has N_r RSS samples, similarity is employed to identify the smallest distance between all N_r RSS samples and the given test sample:

$$s_{r,q} = \min\{ds_{i,q}\}_{i=1}^{N_r}. \quad (2)$$

Similarity-based localization algorithms search the best matching fingerprint on the training set. For each test sample, K RPs with highest similarity are selected whose coordinates denotes as (x_k, y_k) $(k = 1, ..., K)$. WKNN algorithm estimates TP coordinates as a weighted average of these selected RPs. The weight is chosen as a decreasing function of similarity assuming that RP with higher similarity is closer to the TP. In this paper, the weight of RP_k is chosen as:

$$w_k = \frac{1}{s_{k,q}^2} \quad (k = 1, ..., K). \quad (3)$$

The estimated TP coordinates are obtained as:

$$\begin{aligned} \hat{x}_q &= \frac{\sum_{k=1}^{K} w_k x_k}{\sum_{k=1}^{K} w_k} \\ \hat{y}_q &= \frac{\sum_{k=1}^{K} w_k y_k}{\sum_{k=1}^{K} w_k}. \end{aligned} \quad (4)$$

The number of K should be chosen considering the trade-off between accuracy and complexity. Assuming the true coordinates of TP_q is $(x_q, y_q) \in AOI$, then we can obtain the error of localization as:

$$err_q = \sqrt{(\hat{x}_q - x_q)^2 + (\hat{y}_q - y_q)^2}. \quad (5)$$

III. Methodology

In this section, a radio mapping scheme aiming to improve localization efficiency by reducing the number of fingerprint samples and details of the localization strategy are presented. A block diagram of the proposed fingerprinting method and its variant in dashed line is shown as Fig. 1.

First, fingerprint samples measured at the same RP are put into the same group. Then using k-means algorithm each group of samples is divided into a certain number of non-overlapping clusters. For each cluster, one representative sample will be obtained, forming a new fingerprint set that is smaller and more robust to noise and signal fluctuation.

For k-means algorithm, the number of clusters kc is a parameter that should be determined before using. Choosing kc is often a decision based on prior knowledge, assumptions, and practical experience which is difficult when the data has many dimensions[14].

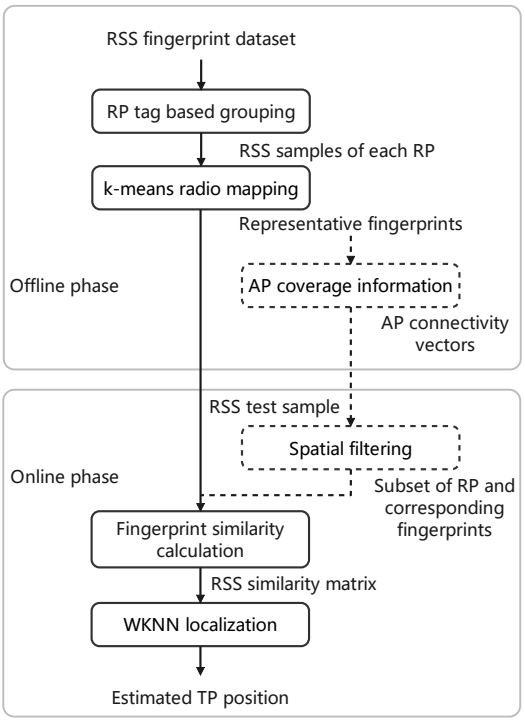

Figure 1. Block diagram for the proposed localization method

Given a certain value of kc, a global silhouette index $S_{r,kc}$ for RP_r is defined to evaluate the quality of clustering [15]. First, we consider a RSS sample i in the p-th cluster C_p with n_p samples. Using sample distance definition (1), $a(i)$ is defined as the mean distance between this sample and other samples in the same cluster shown as:

$$a(i) = \frac{1}{n_p - 1} \sum_{i' \neq i}^{n_p} ds_{i,i'} \quad for \quad i, i' \in C_p. \quad (6)$$

Then we define the mean distance between sample i and another cluster $C_{p'}$ of the same reference point with $n_{p'}$ samples shown as:

$$dc(i, C_{p'}) = \frac{1}{n_{p'}} \sum_{j=1}^{n_{p'}} ds_{i,j} \quad \text{for} \quad i \in C_p \text{ and } j \in C_{p'}. \quad (7)$$

Let us also denote by $b(i)$ the smallest of these distances:

$$b(i) = \min_{p' \neq p} dc(i, C_{p'}) \quad (8)$$

The cluster $C_{p'}$ that presents the lowest distance represents the most probable cluster to consider the RSS sample i other than cluster C_p. A silhouette index is defined to describe relative similarity between sample i and its closest other cluster:

$$S(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \quad (9)$$

A global silhouette index $S_{r,kc}$ for RP_r is defined as the mean value of $S(i)$ considering all RSS samples of this RP. A large $S_{r,kc}$ indicates that samples are well clustered.

To determine the optimal value of cluster number kc, inspired by [9] we perform k-means clustering on each RP with kc as an integer value that varies from 2 to 6 and then the final cluster number of RP_r is chosen as the one that maximizes $S_{r,kc}$. Once cluster number kc for a RP is determined, kc representative samples can be obtained from these clusters as weighted centres and silhouette index $S(i)$ serves as the weight. A weighted average allows well classified samples in the cluster to contribute more to representative samples.

After radio mapping, N' $(2R < N' < 6R)$ number of representative samples are generated. Since N' depends only on the RP number R and the range of kc we set, using this representative sample set with limited size, the complexity of WKNN is also limited.

A step of spatial filtering is added before WKNN localization as a variant of the proposed method to further reduce the size of fingerprint dataset. For each test sample, spatial filtering selects a subset of RPs that are spatially relevant to the TP using AP coverage information. A binary AP coverage vector is generated for each point as $\boldsymbol{I} = [I^1...I^M]$, where $I^m = 1$ if the RSS reading from AP_m is greater than receiver sensitivity for at least 90% of the time[10]. Hamming distance is then computed for AP vector of all RPs and TPs. For a TP, only RPs with distance smaller than a threshold $\alpha M(0 \leq \alpha \leq 1)$ are taken for localization.

IV. Performance Evaluation

In this section, two datasets are used for performance evaluation, a simulated fingerprinting dataset and an open access dataset UjiIndoorLoc. Two other localization methods as well as another radio mapping method are taken for comparison.

A. Datasets for simulation

A simulated dataset is generated based on Friis Radiation Propagation model noted as dataset 1. 9 APs with fixed position are deployed evenly on the surface of $200*200$ meters. 1600 reference points are spread in a grid of 5 meters and for each RP 10 RSS samples are generated. To evaluate the localization method, 500 test points are randomly selected on the surface, each with one RSS sample. A Gaussian white noise with SNR of $100dBm$ is added to both training and test set.

Dataset 2 is a public assess localization database UjiIndoorLoc with 520 APs over 3 buildings [16]. RSSI is represented as negative integer values ranging from -104dBm (extremely poor signal) to 0dBm. In this dataset positive default value 100 is used to denote when an AP was not detected. This value does not reflect the received signal strength or distance to the AP but it will greatly affect the calculation of similarity. A range of other values are considered as replacement[5] and for the proposed method ($k = 3$ for WKNN) the best results are achieved with -95 as shown in Table I. Noted that in this dataset, position information of each point is given as longitude, latitude, floor and building ID. To calculate positioning error, the vertical axis is normalized considering a floor height of 4 meters.

Table I
Impact of default RSS value on the proposed method

def RSS(dbM)	*Building acc(%)*	*Floor acc(%)*	*Position error(m)*
100	99.19	75.52	13.51
-85	98.47	91.18	13.49
-95	**99.46**	**91.27**	**8.62**
-105	99.82	89.83	9.01
-115	99.82	88.12	9.89

B. Comparison to prior work

In this section, performance of the proposed method and its variant is compared to other radio mapping and localization methods. WKNN proposed in [5] is applied to the entire original fingerprint dataset. Method proposed in [9] use k-means clustering method to produce representative samples for each RP then use NN for localization. Average linkage hierarchical clustering is used as comparison for radio mapping which is suggested to outperform k-means clustering under big cluster number [8]. Noted that k-means radio mapping in the proposed method is different from the one compared in [8]. The former one is conducted on samples of the same RP whereas the latter one is used directly on the entire fingerprint dataset. Variant of the proposed method adds a step of spatial filtering after radio mapping to further reduce sample number for localization.

Comparative methods as well as their computation complexity **zuo2008kernel**, [17] are shown in Table II. Note that M is the number of APs while N, N' and N'' ($N > N' > N''$) corresponds to the number of original RSS samples, generated representative samples and filtered samples. K is the parameter for WKNN. k_h is the number of clusters for hierarchical clustering. kc and L is the number of clusters and iterations preset for the k-means.

Table II
Comparative methods

Method	Computational complexity	
	Offline	*Online*
WKNN [5]	-	$\mathcal{O}(MN + KN)$
k-means+NN [9]	$\mathcal{O}(MNLk_c)$	$\mathcal{O}(MN' + N')$
Hierarchical clustering+NN [8]	$\mathcal{O}(MN^2k_h)$	$\mathcal{O}(MN' + N')$
Proposed method	$\mathcal{O}(MNLk_c)$	$\mathcal{O}(MN' + KN')$
Proposed method'	$\mathcal{O}(MNLk_c)$	$\mathcal{O}(N' + MN'' + KN'')$

By taking more neighbor RPs into account, WKNN has slightly higher complexity than NN. From the expression of complexity we can see that reducing data size has a direct impact on WKNN efficiency for the online phase. For the proposed method, given a defined range of cluster number kc, for example, from two to six, the size of representative fingerprint set $2R < N' < 6R$ gives maximum complexity of the WKNN as $\mathcal{O}(M6R + K6R)$. N' depends on RP number R as well as the preset range of kc and it does not grow with

the number samples taken at the same RP making the gain on localization efficiency more significant with more samples taken at each RP. By adding spatial filtering step, complexity for WKNN is further reduced with the cost of some extra operations for filtering during the online phase. The overall complexity for proposed method' is smaller than the proposed method if the size reduction of filtering is big enough.

C. Simulation results

For the proposed method and its variant, parameter k for WKNN and spatial filtering threshold α need to be tuned first to achieve better performance.

Localization performance of WKNN with and without k-means radio mapping under different value of k is shown in Fig. 2 in which average error 1 corresponds to the mean error of all the samples in the validation set whereas average error 2 considers only the error of samples who are correctly classify for buildings and floors. On both datasets, when k is small, the localization error reduces as k increases. On dataset 2, localization error increases after a certain value of k which is different from dataset 1 where the radio map is more ideal. To balance between accuracy and complexity, $k = 3$ is chosen for further simulations. Also, the performance of both methods are quite close if the value of k is well chosen meaning that using k-means clustering can reduce the complexity for WKNN without compromising the performance too much.

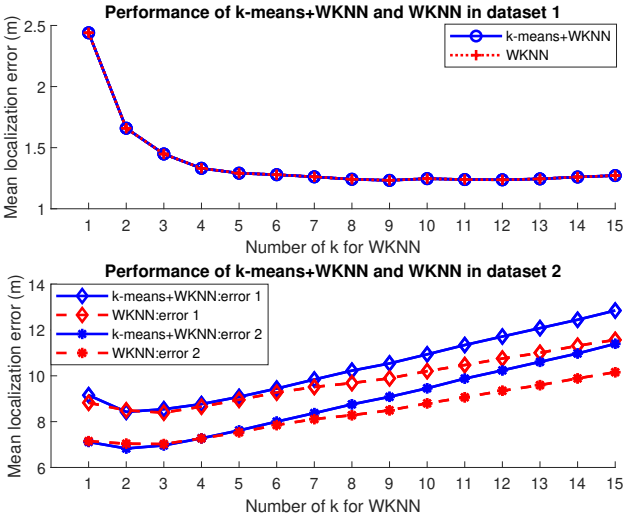

Figure 2. Performance of k-means+WKNN and WKNN

Spatial filtering parameter α affects the number of samples filtered for localization. With a larger value of α, more fingerprints will pass through the filter and $\alpha = 1$ corresponds to no filtering. Since spatial filtering depends on AP coverage information and all samples in dataset 1 have the same AP coverage, proposed method' is only applied on dataset 2.

Table III presents impact of α on localization performance in which the ratio of filtering is defined as the ratio of sample number after and before filtration. As the the value of α increases, more survey points are kept after the filtering step and hence smaller localization error. Since dataset 2 is quite sparse, a small value of α already lets a large percentage of samples pass the filter but if the value is below 0.05, all points are filtered out and it is unable to perform localization. In this section, $\alpha = 0.06$ is chosen for simulations.

Table III
IMPACT OF α ON PERFORMANCE OF THE PROPOSED METHOD'

α	*Mean error(m)*	*Ratio of filtering*(%)
0.05	-	0
0.06	8.73	45.52
0.08	8.54	74.34
0.1	8.54	91.51
0.2	8.54	99.62

Then performance of the proposed and comparative methods is compared on both datasets. Mean error, building and floor classification accuracy as well as data size for localization are shown in Table.IV while Fig. 3 presents the distribution of localization error. For comparison, the maximum cluster number for hierarchical clustering is set to be the same as the number of representative samples generated using k-means clustering.

Table IV
PERFORMANCE OF THE PROPOSED AND COMPARATIVE METHODS

	Mean error(m)	*Building(%)*	*Floor (%)*	*Data size*
	Dataset 1			
NN	*2.44*	-	-	*16000*
WKNN	*1.45*	-	-	*16000*
k-means+NN	*2.44*	-	-	*9557*
HC+NN	*2.44*	-	-	*9557*
Proposed method	*1.45*	-		*9557*
	Dataset 2			
NN	*8.82*	*99.64*	*91.18*	*19861*
WKNN	*8.39*	*99.64*	*91.18*	*19861*
k-means+NN	*9.15*	*99.55*	*91.09*	*4202*
HC+NN	*15.10*	*74.62*	*54.28*	*4202*
Proposed method	*8.54*	*99.55*	*91.09*	*4202*
Proposed method'	*8.73*	*99.64*	*91.18*	*1775*

As we can see, in both datasets, adding a radio mapping step effectively reduces the data size for localization by 40% and 78%, respectively. Besides the gain in efficiency, radio mapping step does not improve localization performance. Adding k-means clustering, mean localization error only increase slightly comparing to only using NN or WKNN. Hierarchical clustering on the other hand, results in a significant decrease in terms of performance and it is also with high complexity as shown in TableII above. Results on both datasets also show an improvement of localization error using WKNN instead of NN. On dataset 1, mean error is reduced by 1 meter using WKNN. Comparing the proposed method to k-means+NN, the use of WKNN for localization leads to a mean error that is 0.6 meter smaller.

By combining k-means radio mapping with WKNN, the proposed method finds a way to balance between efficiency

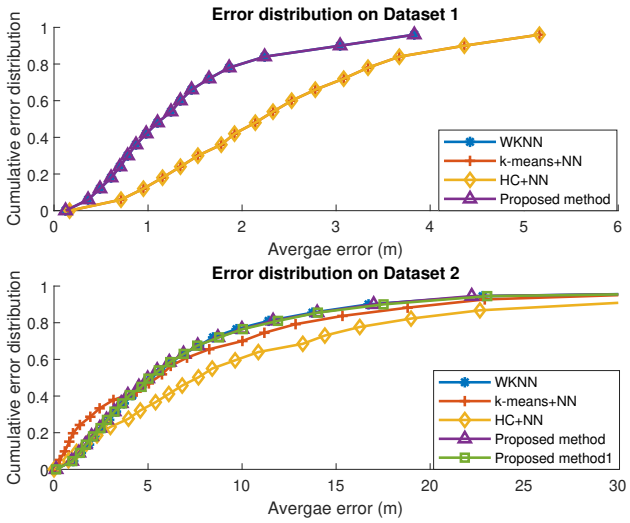

Figure 3. Cumulative average error on two datasets

and accuracy. Adding a spatial filtering step in proposed method' further reduces the average data size for localization by 57% with a small increase in localization error.

V. Conclusion and Future Work

In this paper, we propose a WKNN localization strategy with limited complexity using k-means clustering based radio mapping to improve efficiency of localization without sacrificing performance too much. This mapping scheme reduces the complexity of WKNN by producing a new dataset with fewer and limited number of representative samples. Simulations on two datasets show that this mapping scheme does not improve localization accuracy but with well chosen parameters the increase of mean error is minor and therefore, this strategy finds a balance between precision and efficiency.

As future work, reducing complexity can be investigated for example by using neural network to extract feature with smaller dimension or using clustering methods to further divide the feature space into smaller subareas.

References

[1] X. Wang, L. Gao, S. Mao and S. Pandey, 'Csi-based fingerprinting for indoor localization: A deep learning approach,' *IEEE Transactions on Vehicular Technology*, vol. 66, no. 1, pp. 763–776, 2017.

[2] G. Bhatti, 'Machine learning based localization in large-scale wireless sensor networks,' *Sensors*, vol. 18, no. 12, p. 4179, 2018.

[3] X. Tian, R. Shen, D. Liu, Y. Wen and X. Wang, 'Performance analysis of rss fingerprinting based indoor localization,' *IEEE Transactions on Mobile Computing*, vol. 16, no. 10, pp. 2847–2861, 2017.

[4] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach and L. E. Kavraki, 'Practical robust localization over large-scale 802.11 wireless networks,' in *Proceedings of the 10th annual international conference on Mobile computing and networking*, 2004, pp. 70–84.

[5] A. Moreira, M. J. Nicolau, F. Meneses and A. Costa, 'Wi-fi fingerprinting in the real world - rtls@um at the evaal competition,' in *2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2015, pp. 1–10.

[6] W. Cherif, 'Optimization of k-nn algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis,' *Procedia Computer Science*, vol. 127, pp. 293–299, 2018.

[7] B. Wang, X. Liu, B. Yu, R. Jia and X. Gan, 'An improved wifi positioning method based on fingerprint clustering and signal weighted euclidean distance,' *Sensors*, vol. 19, no. 10, p. 2300, 2019.

[8] A. Arya, P. Godlewski and P. Mellé, 'A hierarchical clustering technique for radio map compression in location fingerprinting systems,' in *2010 IEEE 71st Vehicular Technology Conference*, IEEE, 2010, pp. 1–5.

[9] S. G. Lee and C. Lee, 'Developing an improved fingerprint positioning radio map using the k-means clustering algorithm,' in *2020 International Conference on Information Networking (ICOIN)*, IEEE, 2020, pp. 761–765.

[10] A. Kushki, K. N. Plataniotis and A. N. Venetsanopoulos, 'Kernel-based positioning in wireless local area networks,' *IEEE transactions on mobile computing*, vol. 6, no. 6, pp. 689–705, 2007.

[11] S. Khandker, J. Torres-Sospedra and T. Ristaniemi, 'Analysis of received signal strength quantization in fingerprinting localization,' *Sensors*, vol. 20, no. 11, p. 3203, 2020.

[12] L. Zhang, T. Tan, Y. Gong and W. Yang, 'Fingerprint database reconstruction based on robust pca for indoor localization,' *Sensors*, vol. 19, no. 11, p. 2537, 2019.

[13] C. Xiao, D. Yang, Z. Chen and G. Tan, '3-d ble indoor localization based on denoising autoencoder,' *IEEE Access*, vol. 5, pp. 12 751–12 760, 2017.

[14] G. Hamerly and C. Elkan, 'Learning the k in k-means,' in *Advances in neural information processing systems*, 2004, pp. 281–288.

[15] B. Desgraupes, 'Clustering indices,' *University of Paris Ouest-Lab Modal'X*, vol. 1, p. 34, 2013.

[16] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau and J. Huerta, 'Ujiindoorloc: A new multi-building and multi-floor database for wlan fingerprint-based indoor localization problems,' in *2014 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 2014, pp. 261–270.

[17] D. Xu and Y. Tian, 'A comprehensive survey of clustering algorithms,' *Annals of Data Science*, vol. 2, no. 2, pp. 165–193, 2015.