IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 31 March 2024, accepted 21 May 2024, date of publication 24 May 2024, date of current version 3 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3405350

RESEARCH ARTICLE

Enhancing Fingerprint Localization Accuracy With Inverse Weight-Normalized Context Similarity Coefficient-Based Fingerprint Similarity Metric

ABDULMALIK SHEHU YARO[1,2], FILIP MALY[1], PAVEL PRAZAK[1], AND KAREL MALÝ[1]

[1]Department of Informatics and Quantitative Methods, Faculty of Informatics and Management, University of Hradec Králové, 500 03 Hradec Králové, Czech Republic

[2]Department of Electronics and Telecommunications Engineering, Faculty of Engineering, Ahmadu Bello University, Zaria 810106, Nigeria

Corresponding author: Abdulmalik Shehu Yaro (abdulmalik.yaro@uhk.cz)

This work was supported by the Excellence project run at the Faculty of Informatics and Management, University of Hradec Králové, Czech Republic.

ABSTRACT Distance-based metrics are the most common fingerprint similarity metrics used in fingerprint database clustering and localization processes in a fingerprint-based localization system. In this paper, however, a less common but promising pattern-based fingerprint similarity metric is proposed as an alternative to the distance-base metric. The proposed fingerprint similarity metric is based on an inverse weight (IW) normalization of the context similarity coefficient (CSC)-based similarity metric measure. The clustering and localization performance of the fingerprint-based localization system with the proposed IW-CSC-based fingerprint similarity metric is determined and compared to the square Euclidean, Manhattan, and cosine distance-based metrics. The k-means algorithm with a k-means++ cluster initialization process is considered for fingerprint database clustering, while the k-nearest neighbor (k-NN) algorithm is considered for localization. Based on the four fingerprint databases considered, the proposed IW-CSC-based metric has the slowest localization time with moderate clustering performance. However, it has the best localization performance, which is at least 52% higher than the localization performances of the three distance-base metrics considered. The proposed IW-CSC-based metric is recommended as an alternative to the distance-base metric only when improved localization performance is the primary objective of the fingerprint-based localization system. It is also recommended for use in small to medium-sized fingerprint databases for clustering and localization.

INDEX TERMS Clustering, distance-based metrics, fingerprint similarity metric, inverse weighted, pattern-based metrics.

I. INTRODUCTION

As the demand for accurate indoor positioning solutions grows, fingerprint-based indoor localization has emerged as a promising technique. Fingerprint-based localization is a type of localization technique that uses position-dependent signal parameters (PDSPs), such as received signal strength (RSS) or channel state information (CSI), obtained from spatially deployed wireless access points (APs), to determine the location of a target located within an indoor environment [1], [2]. The localization process of the fingerprint-based system is in two phases, namely the offline and online phases [2]. The offline phase involves the generation of a fingerprint database, also known as a radio map. This involves first collecting RSS or CSI measurements for the spatially deployed wireless APs at several locations known as reference locations (RL) within the indoor environment. The vector representation of all RSS measurements collected

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

from the wireless APs at a single RL is known as a fingerprint vector, and each fingerprint vector is mapped to the RL and stored in a database [2], [3]. The second phase of the fingerprint-based localization process, known as the online phase, involves the determination of the position of a target using the instantaneously acquired fingerprint vector. This involves searching through the fingerprint database using an algorithm called the localization matching algorithm to find a fingerprint vector with the highest degree of correlation to the instantaneously acquired fingerprint vector [2], [3]. The RL of this fingerprint identified in the database is returned as the estimation position of the target.

The density of the fingerprint database is one of the factors that affects the localization accuracy of the fingerprint-based localization system [4], [5]. The higher the density, the higher the localization accuracy; however, the longer the localization time. Fingerprint database clustering techniques such as k-means have been proposed as a way to reduce localization time while retaining the highly dense fingerprint database [6], [7]. Clustering is the process of dividing fingerprint vectors into clusters based on a shared feature known as the fingerprint similarity metric. The fingerprint similarity metric is an important factor in determining the performance of the fingerprint-based localization system [8], [9]. This is because both the clustering algorithm used to cluster the fingerprint database and the localization matching algorithm used to scan through the clustered database relied on fingerprint similarity metrics for their core operations. The distance-based fingerprint similarity metric is widely used in both clustering and localization matching algorithms [9], [10], [11], [12]. The distance-based similarity metric quantifies the similarity of two fingerprint vectors by calculating their distance. The shorter the distance, the more similar the fingerprints are considered to be. However, distance-based metrics do not take into account the behavior of each RSS measurement in the fingerprints being compared, nor do they capture the fingerprints' non-linear relationships. When looking for fingerprint similarities, these factors must be considered. As an alternative to the traditional distance-based fingerprint similarity metric, this paper proposes a pattern-based fingerprint similarity metric for use with both clustering and localization matching algorithms. Unlike distance-based metrics, pattern-based similarity metrics focus on the qualitative resemblance of fingerprint vector patterns, providing a more complete representation of their structural features [8].

This paper makes the following contributions: (a) develop and improve a pattern-based fingerprint similarity metric using the exponential inverse weighted (IW) normalization method; (b) investigate whether the improved pattern-based fingerprint similarity metric can outperform traditional distance-based metrics in fingerprint-based localization systems. The reminder of the paper is organized as follows: Section II presents a review of related works and an overview of the pattern-based fingerprint similarity measure, while Section III gives a mathematical description of the proposed pattern-based similarity metric. The simulation result and discussion are presented in Section IV, followed by the conclusion and recommendation for future work in Section V.

II. REVIEWS OF RELATED WORK

This section of the paper first presents a review of the literature on fingerprint similarity metrics used in fingerprint database clustering and localization matching algorithms. This is followed by an overview of pattern-based fingerprint similarity metric measures.

A. REVIEW OF RELATED WORKS ON FINGERPRINT SIMILARITY METRICS

As mentioned earlier, the fingerprint similarity metric plays an important role in the performance of the fingerprint-based localization system. Several research works have used different fingerprint similarity metrics with either clustering or localization matching algorithms [9], [10], [11], [13], [14], [15]. The authors in [9] and [13] use cosine and Canberra distances, respectively, as a fingerprint similarity metric with the k-means clustering algorithm. In [14], the clustering performance of the k-means clustering algorithm was determined using thirteen different fingerprint similarity metrics, which include squared Euclidean (sqeuclidean), Manhattan, Minkowski, Chebyshev, Sorensen, Soergel, Kulezynski d, Canberra, Lorentzia, wave hedges, divergence, and Clark squared. The authors concluded that, based on the databases they considered, the Manhattan and Minkowski distance metrics generated the best clusters. The Euclidean, Canberra, and Chebyshev distances were used as fingerprint similarity metrics with the k-means clustering algorithm in [10], and based on the database used, the Chebyshev distance had the best clustering performance. The authors of [11] and [15] used Euclidean, Manhattan, and Chebyshev distances as similarity metrics to evaluate the clustering performance of the k-means and k-medoids algorithms, respectively. While the authors in [11] concluded that the best fingerprint similarity metric is determined by the nature of the fingerprint distribution in the database, the authors in [15] concluded that Manhattan and Euclidean distances produced the best clusters.

Rather than calculating the distances between fingerprint vectors to determine their similarity, the authors in [5], [16], and [17] used the wireless APs closest to each fingerprint vector as fingerprint similarity metrics. In [16] and [17], the wireless APs that are closest to each fingerprint are used to cluster the fingerprint vectors. That is, two fingerprints belong to the same cluster if the wireless AP closest to the RL from which they were obtained is identical. The closest wireless AP is the one with the highest RSS value in the fingerprint vector. The authors of [5] expand on the work of [16] by utilizing the two closest wireless APs. Table 1 gives a summary of the fingerprint similarity metrics used in clustering fingerprint databases in related works.

Looking at the fingerprint similarity metrics used with the localization matching algorithm, most research works used distance-based metrics as the fingerprint similarity metrics.

TABLE 1. Summary of fingerprint similarity metrics used in clustering by related works.

Reference	Algorithm	Similarity Metric	Type of Similarity Metric
[13]	k-means	Cosine distance	Distance-based metric
[9]	k-means	Canberra distance	Distance-based metric
[14]	k-means	sqeuclidean, Manhattan, Minkowski, Chebyshev, Sorensen, Soergel, Kulezynski d, Canberra, Lorentzia, wave hedges, divergence, and Clark squared distances	Distance-based metric
[10]	k-means	Euclidean, Canberra and Chebyshev distances	Distance-based metric
[15]	k-medoids	Euclidean, Manhattan and Chebyshev distances	Distance-based metric
[11]	k-means	Euclidean, Manhattan and Chebyshev distances	Distance-based metric
[5]	Closest AP	Two closest APs	Distance-based metric
[16]	Closest AP	Used the closest AP	Distance-based metric
[17]	Strongest AP	Used the closest AP	Distance-based metric

TABLE 2. Summary of fingerprint similarity metrics used with localization matching algorithms in related works.

Reference	Algorithm	Similarity Metric	Type of Similarity Metric
[18]	Wk-NN	Euclidean and Shepard distances	Distance-based metric
[19]	SLAM	Euclidean, Manhattan, Chebyshev, Cosine, Spearman, variable, and correlation	Distance-based metric
[20]	RBF	Spearman distance, Spearman's footrule distance, Jaccard coefficient, hamming distance and Canberra distance	Distance-based metric
[21]	RBF	Lorentzian, Hamming, Jaccard, Wave Hedges, and Canberra distances	Distance-based metric

For instance, in [18], the Euclidean distance is used as the fingerprint similarity metric with the weighted k-nearest neighborhood (Wk-NN) algorithm. The localization performance of the simultaneous localization and mapping (SLAM) algorithm was determined using eight different distance-based similarity metrics, which include Euclidean, Manhattan, Chebyshev, cosine, Spearman, variable, and correlation [19]. Also, in [20], the localization performance of the rank-based fingerprinting (RBF) localization algorithm was determined using Spearman distance, Spearman's footrule distance, Jaccard coefficient, hamming distance, and Canberra distance as fingerprint similarity metrics. Based on the database they considered, the authors concluded that Spearman's footrule distance resulted in the best localization accuracy. Similarly, the authors in [21] evaluated the localization performance of the RBF localization algorithm using Lorentzian, Hamming, Jaccard, Wave Hedges, and Canberra distances as fingerprint similarity metrics. They concluded that Lorentzian distance as a fingerprint similarity metric resulted in the best localization performance. A summary of fingerprint similarity metrics used with localization matching algorithms in related work is presented in Table 2.

From Tables 1 and 2, most clustering or localization algorithms used by other researchers are directly or indirectly based on the distance metric. Furthermore, the performance of each distance metric with either the clustering or localization algorithm varies with fingerprint database characteristics; as such, there is no one-size-fits-all fingerprint similarity metric. A type of fingerprint similarity metric that has not been fully explored for use with fingerprint-based localization is the pattern-based similarity metric, such as the context similarity coefficient (CSC). An attempt has been made in [8] to use CSC as a fingerprint similarity metric with the affinity propagation clustering (APC) algorithm. The authors evaluated the performance of the APC algorithm with the CSC similarity metric using four different databases and compared it to the Euclidean distance metric. The results show that the use of the CSC with the APC algorithm generated more well-separated clusters than with the Euclidean distance. Thus, in this paper, an improved version of the CSC-based similarity metric is proposed and used with both clustering and localization matching algorithms. The k-means algorithm is the most commonly used clustering algorithm; however, in this paper, an improved version of the k-means algorithm, which is the k-means algorithm with the k-means++ cluster initialization process, is used. A detailed description of the improved k-means (ik-means) clustering algorithm can be found in [6] and [7]. As for the localization matching algorithm, the k-NN algorithm, which is the most commonly used, is considered, and a detailed description of the localization matching using the k-NN algorithm can be found in [18]. In the next subsection, an overview of the pattern-based fingerprint similarity metric is presented.

B. OVERVIEW OF PATTERN-BASED SIMILARITY METRICS

As previously stated, most clustering and localization algorithms traditionally employed distance-based fingerprint similarity metrics. The distance-based metrics do not consider the behavior of each RSS measurement in the fingerprint vector during the similarity measure. Furthermore, they do not consider the linear and non-linear relationships between the two fingerprint vectors [8], [22]. Because these two factors are critical when determining the similarity between two or more fingerprint vectors, the pattern-based similarity metric is used, as it takes both into account.

Given four fingerprint vectors, each containing three RSS measurements, their RSS pattern distribution is shown in Figure 1. From Figure 1, it can be seen that fingerprint vectors #1 and #2 follow the same RSS pattern distribution, while fingerprint vectors #3 and #4 follow the same RSS pattern distribution. Using the pattern-based similarity measure, it means that fingerprint vectors #1 and #2 have the highest degree of similarity, while fingerprint vectors #3 and #4 have the highest degree of similarity. Within the context of clustering, fingerprint vectors #1 and #2 will be in the same cluster, while fingerprint vectors #3 and #4 will also be in the same cluster. Like-wise, within the context of localization matching with the k-NN algorithm, for instance, if fingerprint vector #1 is given as the input, its closest neighborhood will be fingerprint vector #2.

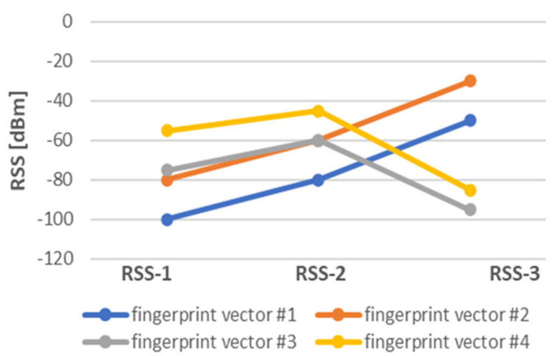

FIGURE 1. RSS pattern distributions for each fingerprint vector.

In the next section, the proposed pattern-based fingerprint similarity metric based on the improvement of the context similarity coefficient (CSC) is presented.

III. PROPOSED IMPROVED CSC-BASED FINGERPRINT SIMILARITY METRIC

This subsection provides a mathematical description of the proposed improved CSC-based fingerprint similarity metric.

Given two fingerprint vectors, $\mathbf{f}_i$ and $\mathbf{f}_j$, containing N number RSS measurements as shown in (1) and (2), respectively:

$$\mathbf{f}_i = [rss_i(1), rss_i(2), \ldots, rss_i(N-1), rss_i(N)] \quad (1)$$

$$\mathbf{f}_j = \left[rss_j(1), rss_j(2), \ldots, rss_j(N-1), rss_j(N)\right] \quad (2)$$

The similarity value between fingerprint vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ based on the CSC is calculated as follows [8], [22]:

Step 1: Obtain $\mathbf{t_{ij}}$ as shown in (3), which is based on the vector addition of $\mathbf{f}_i$ and $\mathbf{f}_j$.

$$\mathbf{t_{ij}} = [t(1), t(2), \ldots, t(N)] \quad \text{for } 1 \leq n \leq N \quad (3)$$

where: $t(n) = rss_i(n) + rss_j(n)$

Step 2: Calculate the probability of the outcome, p_{f_i}, for fingerprint vector $\mathbf{f}_i$ using (4).

$$p_{f_i} = \frac{\sum_{n=1}^{N} \mathbf{f_i}(n)}{\sum_{n=1}^{N} \mathbf{t_{ij}}(\mathrm{n})} \quad (4)$$

Step 3: Determine the expectation value, $<rss_i(n)>$, for each RSS measurement in fingerprint vector $\mathbf{f_i}$ using (5).

$$< rss_i(n) >= p_{f_i} \times \mathbf{t_{ij}}(n) \quad \text{for } 1 \leq n \leq N \quad (5)$$

Step 4: Determine the error for each RSS measurement in the fingerprint vector $\mathbf{f}_i$ using (6).

$$error_{f_i}(n) = \frac{< rss_i(n) > - \mathbf{f}_i(n)}{\sqrt{\mathbf{t_{ij}}(n) \times p_{f_i} \times (1 - p_{f_i})}} \quad (6)$$

Step 5: The CSC-based similarity value between fingerprint vectors $\mathbf{f}_i$ and $\mathbf{f}_j$ is calculated using (7).

$$d_{csc}\left(\mathbf{f_i}, \mathbf{f_j}\right) = \frac{\sum_{n=1}^{N} \left(\left(error_{f_i}(n)\right)^2 \times \sqrt{\mathbf{t_{ij}}(n)}\right)}{\sum_{n}^{N} \sqrt{\mathbf{t_{ij}}(n)}} \quad (7)$$

The CSC-based similarity value for fingerprint vectors $\mathbf{f}_i$ and $\mathbf{f}_j$, calculated using (7), represents the degree of similarity based on the correlation between each RSS measurement in the two fingerprint vectors. A low CSC value indicates a high degree of similarity, whereas a higher value indicates a high degree of dissimilarity. Given the nature of the individual RSS measurements in each fingerprint vector, the CSC value calculated in (7) will be numerically large. Several large CSC-based similarity values are generated during cluster assignment in the clustering process and fingerprint matching in the localization matching process. These high numerical similarity values can be difficult to understand and compare directly. Furthermore, large fingerprint similarity values can obscure smaller differences, affecting the performance of the clustering or localization matching algorithm. Normalizing similarity values is a good practice that can help with the interpretability, computational efficiency, and comparability of similarity-based analyses. This paper proposes using the inverse weighted (IW) method to normalize the CSC-based similarity value.

The IW normalization is a technique used to adjust data points, which in this case are the CSC similarity values, by assigning different weights to the values inversely proportional to their importance. Smaller CSC similarity values indicate a high degree of similarity, which is considered to be of high importance during similarity determination. The IW normalization technique is particularly useful in scenarios such as this, where less importance is given to larger CSC similarity values and more importance is given to smaller

CSC similarity values. Given M numbers of CSC-based similarity values obtained using Eq. (7), the normalized IW-CSC similarity measure value is calculated as follows:

Step 1: Compute the weights of each CSC-based similarity value using the exponential function as follows:

$$w_m = e^{-\lambda \times d_{csc}} \left(\mathbf{f_i}, \mathbf{f}_j\right)_m$$
$$m \in [1, M],\ i \in [1, N],\ j \in [1, N],\ i \neq j \quad (8)$$

where the $d_{csc}\left(\mathbf{f_i}, \mathbf{f_j}\right)$ in (8) is the CSC-based similarity metric obtained in (7), and the "λ" is a positive constant that regulates the rate at which the weights decrease in relation to the CSC-based similarity metric values. A higher "λ" value causes faster decay, implying that smaller CSC values are given greater importance. A lower value of "λ" results in a slower decay, emphasizing the importance of large CSC values.

Step 2: Normalize the weights from Step 1 to sum to 1 by dividing each weight by the sum of all the weights as shown in (9).

$$\hat{w}_m = \frac{w_m}{\sum_{m=1}^{M} e^{-\lambda \times d_{csc}\left(\mathbf{f_i}, \mathbf{f_j}\right)_m}} \quad (9)$$

Step 3: Multiply each weight in (9) with its corresponding CSC-based similarity value to obtain the exponential IW normalized variant as shown in (10).

$$d_{iw-CSC}\left(d_{csc}\left(\mathbf{f_i}, \mathbf{f_j}\right)_m\right)_m = d_{csc}\left(\mathbf{f_i}, \mathbf{f_j}\right)_m \times \hat{w}_m \quad (10)$$

The exponential IW normalized variant of (7) shown in (10) is the proposed IW-CSC-based similarity metric, which is to be used in place of the CSC-based similarity metric in (7) for similarity measure determination by the clustering and fingerprint localization matching algorithms. In clustering or fingerprint localization matching, the smaller the CSC value, the more important it is in the cluster assignment or nearest neighbor determination phase; thus, a large value of $\lambda = 2$ is considered in (10).

An overview of the fingerprint-based localization system based on the proposed IW-CSC-based fingerprint similarity metric is shown in Figure 2.

In the next section of the paper, the clustering and localization performances of the proposed system shown in Figure 2 are determined.

IV. SIMULATION RESULT AND DISCUSSION

The clustering performance of the ik-means algorithm and the localization performance of the k-NN algorithm, both using the proposed IW-CSC-based fingerprint similarity metric, are determined and presented in this section of the paper. The simulation setup and parameters are presented first, followed by the time comparison of fingerprint similarity measure determination. Lastly, the clustering and localization performance analysis of the ik-means and k-NN algorithms with the proposed IW-CSC-based metric are presented.

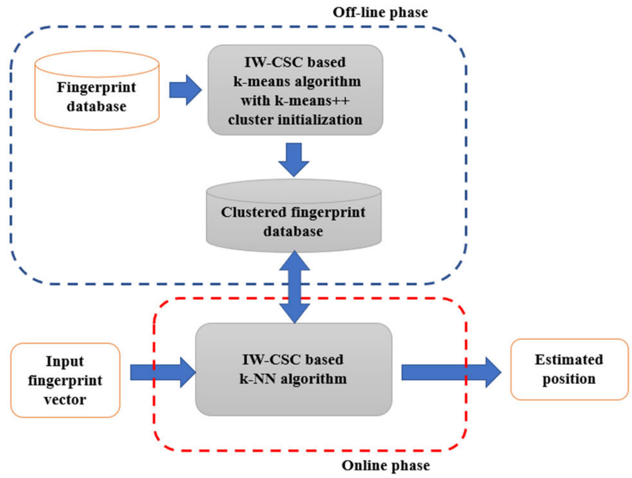

FIGURE 2. Overview of the proposed IW-CSC based fingerprint-based localization system.

A. SIMULATION SETUP AND PARAMETERS

The performance of the proposed fingerprint-based system is determined using four experimentally generated and publicly available fingerprint databases found in [23], [24], [25], and [26] with characteristics listed in Table 3.

TABLE 3. Fingerprint database characteristics.

Databases	Database characteristics	
	RSS per fingerprint	Number of fingerprints
MSI_IndoorLoc [23]	11	4973
PIEP_UM_IndoorLoc [24]	8	1000
IIRC_IndoorLoc [25]	4	194
SEUG_IndoorLoc [26]	3	49

There are several distance-based fingerprint similarity metrics used with either the ik-means clustering or k-NN algorithm; however, the sqeuclidean, Manhattan, and cosine distances are the most commonly used and will be considered in this paper for comparison with the proposed IW-CSC-based similarity metric measure. The distance similarity values for the sqeuclidean, Manhattan, and cosine are calculated mathematically using (11), (12) and (13), respectively [13], [14].

$$d_{sqeucl}\left(\mathbf{f_i}, \mathbf{f_j}\right) = \sum_{n=1}^{N}\left(\mathbf{f}_i(n) - \mathbf{f}_j(n)\right)^2 \quad (11)$$

$$d_{manh}\left(\mathbf{f_i}, \mathbf{f_j}\right) = \sum_{n=1}^{N}\left|\mathbf{f}_i(n) - \mathbf{f}_j(n)\right| \quad (12)$$

$$d_{cosine}\left(\mathbf{f_i}, \mathbf{f_j}\right) = 1 - \frac{\sum_{i}^{N}\left(\mathbf{f}_i(n) \times \mathbf{f}_j(n)\right)}{\sqrt{\sum_{i}^{N}\left(\mathbf{f}_i(n)\right)^2} \times \sqrt{\sum_{i}^{N}\left(\mathbf{f}_j(n)\right)^2}} \quad (13)$$

The distance-based fingerprint similarity metrics presented in (9) to (11), as well as the proposed pattern-based fingerprint similarity metric, IW-CSC, will be used with both the ik-means and the k-NN algorithms. For each fingerprint similarity metric, the clustering performance of the ik-means algorithm and the localization performance of the K-NN algorithm are determined and compared. The silhouette score will be used as the clustering performance metric, while the position root mean square error (RMSE) will be used as the localization performance metric. As for the fingerprint similarity measure determination time, the big O notation will be used. The entire simulation results are generated using an ASUS computer with the following specifications: an Intel (R) Core (TM) i3-2400 CPU running at 1.8 GHz, 8 GB of RAM, the Windows 11 operating system, and MATLAB R2023a.

B. FINGERPRINT SIMILARITY MEASURE DETERMINATION TIME COMPARISON

In this subsection, the time computation complexity (CC) to determine the similarity measure value between two fingerprint vectors is determined and compared. Using the big O notation, the time CCs of the sqeuclidean distance, Manhattan distance, cosine distance, and the proposed IW-CSC metric are determined and presented in Table 4. Note that the 'N' denotes the number of RSS measurements in each fingerprint vector.

TABLE 4. Time CC comparison of fingerprint similarity metrics.

Fingerprint similarity metric	Time CC	Type of similarity metric
Sqeuclidean	$O(N)$	
Manhattan	$O(N)$	Distance based
Cosine	$O(N)$	
IW-CSC	$O(5N)$	Pattern-based

Table 4 shows the time CCs used to determine the similarity value of two fingerprint vectors. The distance-based metrics, that is, the sqeuclidean, Manhattan, and cosine distances, all have a time CC of $O(N)$, whereas the proposed fingerprint similarity metric has a time CC of $O(5N)$. This means that, given two fingerprint vectors, the proposed metric will take five times longer to calculate the similarity value than any of the distance-based fingerprint similarity metrics considered. The time CC of each fingerprint similarity metric is directly proportional to the localization time of the fingerprint-based localization system. The higher the time CC, the longer the localization process. This is because using a fingerprint similarity metric with a high time CC causes the localization matching algorithm to take longer to scan the clustered fingerprint database. Despite the fact that the fingerprint database is clustered, the proposed IW-CSC-based similarity metric will take longer to scan through the clustered fingerprint database, resulting in a longer localization time. The longer localization time will undoubtedly be shorter than the localization time of a non-clustered fingerprint database. The overall localization time will be reduced, though not as much as the localization time achieved by the k-NN algorithm with the sqeuclidean, Manhattan, and cosine distances as fingerprint similarity metrics.

C. CLUSTERING PERFORMANCE COMPARISON AND ANALYSIS

In the earlier subsection, it was determined that the use of the proposed IW-CSC-based fingerprint similarity metric will result in a longer localization time in comparison to the distance-based metrics considered. In this section using the silhouette score as the clustering performance metric, the clustering performance of the ik-means algorithm with the sqeuclidean distance, Manhattan distance, cosine distance, and the proposed similarity metric, IW-CSC, as fingerprint similarity metrics are determined and compared in this section of the paper. The silhouette score is used to evaluate the quality of clusters generated by any clustering algorithm and is a measure of how well clusters are separated from one another and how well fingerprints are assigned to their respective clusters.

Silhouette scores range from −1 to 1, with a higher score value of 1 indicating well-separated clusters and fingerprints are well-assigned to their respective clusters. A clustering algorithm is considered to have very good clustering performance when it has a silhouette score of 0.7 and above. A silhouette score between 0.7 and 0.25 indicates that clusters are fairly well separated with few overlapping clusters, while a silhouette score of 0.25 and below indicates that clusters are weakly separated. Table 5 shows the silhouette score comparison of the ik-means algorithm for the four fingerprint databases using the sqeuclidean distance, Manhattan distance, cosine distance, and IW-CSC as fingerprint similarity metrics. For each fingerprint database, k = 3 and k = 5, which indicate the number of clusters generated by the ik-means algorithm, are considered. The entries in Table 4 with green highlights indicate the highest silhouette score for each number of clusters generated in each fingerprint database considered.

The highest silhouette score from Table 5 is 0.39, while the lowest is 0.11. This indicates that all the clusters generated by the ik-means algorithm using the different fingerprint similarity metrics for all four fingerprint databases are not well-separated. There is a high probability that some fingerprint measurements are misassigned. Comparing the silhouette scores obtained by the ik-means algorithm with k = 3 and k = 5, the silhouette scores obtained using k = 3 are higher than those obtained using k = 5. This is irrespective of the fingerprint database and the fingerprint similarity metric. What this shows is that even though the clusters generated using k = 3 and k = 5 are generally not well-separated, clusters generated using k = 3 are fairly well-separated compared to those generated using k = 5. As a result, for the localization performance comparison with the k-NN algorithm, the number of clusters generated by the ik-means clustering algorithm, regardless of fingerprint similarity metric, will be k = 3. This is to ensure that all similarity metrics are compared at their peak performance.

Extending the result discussion to determine which fingerprint similarity metric has the best clustering performance for each database with k = 3, the sqeuclidean distance has the highest silhouette score of about 0.34 in the MSI_IndoorLoc database. The IW-CSC, which is the similarity metric proposed in this paper, has the lowest silhouette score of about 0.26. This means that the sqeuclidean distance metric generated fairly well-separated clusters compared to the proposed metric in the MSI_IndoorLoc database. In the PIEP_UM_IndoorLoc database, all four-fingerprint similarity metrics have the same silhouette score of about 0.3, indicating equal clustering performance. As for the SEUG_IndoorLoc database, the cosine distance metric has the highest silhouette score of about 0.38. This is followed by the sqeuclidean distance and IW-CSC metrics both with a silhouette score of 0.36. In the IIRC_IndoorLoc database, the sqeuclidean distance metric has the highest silhouette score of 0.39, which is followed by the IW-CSC with a silhouette score of 0.36.

Overall, considering all four fingerprint databases, with k = 3, the well-separated clusters are generated by the sqeuclidean distance metric. The proposed fingerprint similarity metric, IW-CSC, came in second, generating fairly well-separated clusters compared to the cosine and Manhattan distance metrics. The Manhattan distance metric generated the least well-separated clusters amongst all the fingerprint similarity metrics in all four fingerprint databases. Having a well-clustered fingerprint database does not automatically translate to better localization performance. The type of localization matching algorithm as well as the choice of fingerprint similarity metric used by the localization matching algorithm also play an important role in the localization accuracy of the fingerprint-based localization system. As such, in the next subsection, the localization performance of the fingerprint-based localization system using the fingerprint databases clustered by the ik-means algorithm with each of the four fingerprint similarity metrics is determined and compared.

D. LOCALIZATION PERFORMANCE COMPARISON AND ANALYSIS

In this subsection, the localization performances of each of the four fingerprint databases clustered using the ik-means algorithm with the sqeuclidean distance, Manhattan distance, cosine distance, and IW-CSC as fingerprint similarity metrics are determined. As earlier mentioned, the k-NN localization matching algorithm is considered, and its fingerprint similarity metric is chosen to be the same as the fingerprint similarity metric used in clustering the database. That is, if the ik-mean algorithm with the sqeuclidean distance metric is used to cluster a database, then the k-NN algorithm with the sqeuclidean distance metric is also used to determine the localization performance. Since k = 3 for the ik-means algorithm, k = 3 is also considered for the k-NN localization algorithm. Using position RSME as localization performance metrics, the localization performance on each of the four fingerprint databases is determined and presented in Table 6. A graphical illustration of the localization performance comparison can be seen in Figure 3. The entries in Table 6 with the green highlights indicate the fingerprint similarity metric with the least position RMSE.

For all four fingerprint databases considered, the k-NN algorithm with the proposed fingerprint similarity metric, which is IW-CSC, has the lowest position RMSE. In the MSI_IndoorLoc database, the IW-CSC fingerprint similarity metric has the least position RMSE of 0.3 m^2 with the k-NN algorithm. The sqeuclidean, Manhattan, and cosine distance-based fingerprint similarity metrics have about the same position RMSE of about 6.3 m^2. The percent improvement achieved by the k-NN algorithm with the IW-CSC fingerprint similarity metric over the other three distance-based metrics is, on average, about 95%.

In the PIEP_UM_IndoorLoc database, the k-NN algorithm with the IW-CSC as fingerprint similarity metric also has the least position RMSE of about 0.3 m^2, with the cosine distance metric having the highest with about 13.7 m^2. The overall localization performance improvement achieved by the k-NN algorithm with the IW-CSC over the other three distance-based fingerprint similarity metrics is about 97% on average. Extending the analysis to the SEUG_IndoorLoc and IIRC_IndoorLoc databases, the k-NN algorithm with the IW-CSC as fingerprint similarity metric has the lowest position RMSE in both databases. In the SEUG_IndoorLoc database, the k-NN algorithm with the IW-CSC has a position RMSE of about 0.6 m^2, while in the IIRC_IndoorLoc database, it has a position RMSE of about 0.4 m^2. The k-NN algorithm with the cosine distance metric has the highest position RMSE of about 1.4 m^2 in the SEUG_IndoorLoc database, while in the IIRC_IndoorLoc database, the k-NN algorithm with the Manhattan distance metric has the highest position RMSE of about 5.1 m^2. The overall localization performance improvements achieved by the k-NN algorithm with the IW-CSC as fingerprint similarity metric over

TABLE 5. Silhouette score comparison of ik-means algorithm with different fingerprint similarity metrics for k = 3 and k = 5 on four fingerprint databases.

Similarity metric	MSI_IndoorLoc		PIEP_UM_IndoorLoc		SEUG_IndoorLoc		IIRC_IndoorLoc	
	k=3	k=5	k=3	k=5	k=3	k=5	k=3	k=5
Sqeuclidean	0.34	0.29	0.30	0.30	0.36	0.35	0.39	0.36
Manhattan	0.33	0.26	0.30	0.30	0.30	0.28	0.33	0.29
Cosine	0.33	0.27	0.30	0.29	0.38	0.39	0.34	0.30
IW-CSC	0.26	0.17	0.30	0.23	0.36	0.11	0.36	0.32

TABLE 6. Position RMSE error comparison.

Similarity metric	Position RMSE (m^2)			
	MSI_IndoorLoc	PIEP_UM_IndoorLoc	SEUG_IndoorLoc	IIRC_IndoorLoc
Sqeuclidean	6.3	12.4	1.2	3.3
Manhattan	6.3	13.4	1.2	5.1
Cosine	6.2	13.7	1.4	4.2
IW-CSC	0.3	0.3	0.6	0.4

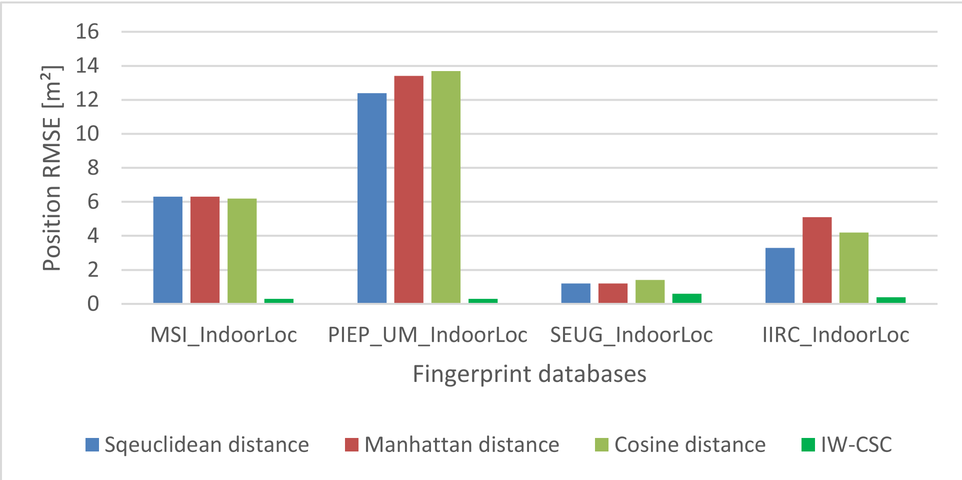

FIGURE 3. Localization performance comparison of k-NN algorithm with different fingerprint similarity metrics.

the other three distance-based fingerprint similarity metrics are 52% and 90% in the SEUG_IndoorLoc and the IIRC_IndoorLoc databases, respectively.

The proposed pattern-based fingerprint similarity metric, when used with the ik-means algorithm, has a clustering performance that is slightly below the clustering performance of the best distance-based metric considered, as shown in the results analysis in Section IV-C. However, despite having a lower clustering performance, it resulted in the best localization performance, which is at least 52% higher than the localization performances of the three distance-base fingerprint similarity metrics considered. The distance-based metrics determine the similarity between two fingerprint vectors by only looking at how close they are to each other, ignoring the spatial relationship between the fingerprint vectors. Furthermore, they fail to consider how each RSS measurement behaves within the fingerprint vector and how it varies between wireless APs. The pattern-based fingerprint similarity metrics take all these into account. They not only consider the distance between fingerprint vectors but also how each RSS measurement behaves relative to others within the fingerprint vector during fingerprint similarity determination. Furthermore, in fingerprint localization, the RSS measurement patterns are crucial for accurately matching a fingerprint vector to a specific location within an environment. For these reasons, the proposed IW-CSC-base metric has better localization performance when compared to the distance-based metrics considered.

In summary, as an alternative to the distance-base metric, the proposed IW-CSC-based fingerprint similarity metric is ideal for use with the k-NN algorithm on fingerprint databases clustered using the ik-means algorithm, which also employs the IW-CSC as a fingerprint similarity metric. It has superior localization performance with a clustering performance that is moderate in comparison to the distance-base metrics, only that it takes a longer time to determine the location of an indoor user. The proposed IW-CSC-based metric is not ideal for applications where near-real-time localization is the primary objective. However, the longer localization times can be easily solved by implementing the fingerprint-based localization system's online phase on high-power computational hardware or through cloud computing. In scenarios where better localization accuracy is the primary object, the proposed metric is an ideal alternative to the distance-based metrics. Overall, it is recommended that the proposed IW-CSC as a fingerprint similarity metric be used to cluster and perform localization on small to medium-sized fingerprint databases. This is to have a moderate localization time that could be acceptable for a near-real-time localization system.

V. CONCLUSION AND RECOMMENDATION FOR FUTURE WORKS

This paper proposes a pattern-based fingerprint similarity metric as an alternative to the widely used distance-based fingerprint similarity metric for fingerprint database clustering and localization processes of the fingerprint-based localization system. The proposed fingerprint similarity metric is based on the CSC-based fingerprint similarity metric, which has been IW normalized. The proposed IW-CSC-based metric is used with the ik-means and k-NN algorithms for fingerprint database clustering and localization, respectively. The ik-means and k-NN algorithms' clustering and localization performances are determined and compared using the proposed IW-CSC as fingerprint similarity metrics, as well as three commonly used distance-based fingerprint similarity metrics, namely sqeuclidean, Manhattan, and cosine distances. The simulation results show that the ik-means algorithm with the IW-CSC-based fingerprint similarity metric has the least clustering performance, as evidenced by the lowest silhouette score. Furthermore, the proposed IW-CSC-based fingerprint similarity metric measure has the slowest clustering time, which is five times slower than the distance-based metrics considered. This translates to a slow localization time. It has superior localization performance, but its clustering performance is slightly lower than the best distance-based metric considered. In summary, the proposed IW-CSC-based metric is not ideal for use in a near-real-time localization system but is an ideal alternative to the distance-based metric where improved localization performance is required. It is recommended that, for a moderate localization time, the proposed IW-CSS-based fingerprint similarity metric be used for clustering and localization in small to medium-sized fingerprint databases. Future work will focus on improving the localization time of the IW-CSC-based fingerprint similarity metric for use in a near-real-time application.

ACKNOWLEDGMENT

The authors acknowledge funding from the Excellence project run at the Faculty of Informatics and Management, University of Hradec Kralove, Czech Republic. Special thanks to Ing. Daniel Schmidt for his help with data collection and preparation.

REFERENCES

[1] A. S. Yaro, F. Maly, P. Prazak, and K. Malý, "Outlier detection performance of a modified Z-score method in time-series RSS observation with hybrid scale estimators," *IEEE Access*, vol. 12, pp. 12785–12796, 2024, doi: 10.1109/access.2024.3356731.

[2] A. S. Yaro, F. Maly, and P. Prazak, "A survey of the performance-limiting factors of a 2-Dimensional RSS fingerprinting-based indoor wireless localization system," *Sensors*, vol. 23, no. 5, p. 2545, Feb. 2023, doi: 10.3390/s23052545.

[3] S. Shang and L. Wang, "Overview of WiFi fingerprinting-based indoor positioning," *IET Commun.*, vol. 16, no. 7, pp. 725–733, Apr. 2022, doi: 10.1049/cmu2.12386.

[4] A. S. Yaro, F. Maly, P. Prazak, and K. Malý, "Improved fingerprint-based localization based on sequential hybridization of clustering algorithms," *Emerg. Sci. J.*, vol. 8, no. 2, pp. 394–406, Apr. 2024, doi: 10.28991/esj-2024-08-02-02.

[5] A. S. Yaro, F. Malý, and K. Malý, "A two-nearest wireless access point-based fingerprint clustering algorithm for improved indoor wireless localization," *Emerg. Sci. J.*, vol. 7, no. 5, pp. 1762–1770, Oct. 2023, doi: 10.28991/esj-2023-07-05-019.

[6] D. Arthur and S. Vassilvitskii, "K-Means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2007, pp. 1027–1035.

[7] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[8] A. S. Yaro, F. Malý, and K. Malý, "Improved indoor localization performance using a modified affinity propagation clustering algorithm with context similarity coefficient," *IEEE Access*, vol. 11, pp. 57341–57348, 2023, doi: 10.1109/ACCESS.2023.3283592.

[9] M. Raeisi and A. B. Sesay, "A distance metric for uneven clusters of unsupervised K-means clustering algorithm," *IEEE Access*, vol. 10, pp. 86286–86297, 2022, doi: 10.1109/ACCESS.2022.3198992.

[10] S. Gultom, S. Sriadhi, M. Martiano, and J. Simarmata, "Comparison analysis of K-means and K-medoid with ecluidience distance algorithm, chanberra distance, and Chebyshev distance for big data clustering," *IOP Conf. Ser., Mater. Sci. Eng.*, vol. 420, Oct. 2018, Art. no. 012092, doi: 10.1088/1757-899x/420/1/012092.

[11] M. E. M. Gonzales, L. C. Uy, J. A. L. Sy, and M. O. Cordel, "Distance metric recommendation for k-means clustering: A meta-learning approach," in *Proc. IEEE Region 10 Conf. (TENCON)*, Nov. 2022, pp. 1–6, doi: 10.1109/TENCON55691.2022.9978037.

[12] S. Kapil and M. Chawla, "Performance evaluation of K-means clustering algorithm with various distance metrics," in *Proc. IEEE 1st Int. Conf. Power Electron., Intell. Control Energy Syst. (ICPEICES)*, Jul. 2016, pp. 1–4, doi: 10.1109/ICPEICES.2016.7853264.

[13] Md. K. Khan, S. Sarker, S. M. Ahmed, and M. H. A. Khan, "K-cosine-means clustering algorithm," in *Proc. Int. Conf. Electron., Commun. Inf. Technol. (ICECIT)*, Sep. 2021, pp. 1–4, doi: 10.1109/ICECIT54077.2021.9641480.

[14] M. K. Gupta and P. Chandra, "Effects of similarity/distance metrics on K-means algorithm with respect to its applications in IoT and multimedia: A review," *Multimedia Tools Appl.*, vol. 81, no. 26, pp. 37007–37032, Nov. 2022, doi: 10.1007/s11042-021-11255-7.

[15] A. S. Sunge, Y. Heryadi, and Y. Religia, "Comparison of distance function to performance of K-medoids algorithm for clustering," in *Proc. Int. Conf. Smart Technol. Appl. (ICoSTA)*, Feb. 2020, pp. 1–6, doi: 10.1109/ICoSTA48221.2020.1570615793.

[16] D. Quezada-Gaibor, J. Torres-Sospedra, J. Nurmi, Y. Koucheryavy, and J. Huerta, "Lightweight Wi-Fi fingerprinting with a novel RSS clustering algorithm," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Nov. 2021, pp. 1–8, doi: 10.1109/IPIN51156.2021.9662612.

[17] B. Ezhumalai, M. Song, and K. Park, "An efficient indoor positioning method based on Wi-Fi RSS fingerprint and classification algorithm," *Sensors*, vol. 21, no. 10, p. 3418, May 2021, doi: 10.3390/s21103418.

[18] F. Li, M. Liu, Y. Zhang, and W. Shen, "A two-level WiFi fingerprint-based indoor localization method for dangerous area monitoring," *Sensors*, vol. 19, no. 19, p. 4243, Sep. 2019, doi: 10.3390/s19194243.

[19] H. D. Satea, A. A. Ibrahem, M. Faiq, Z. A. Abbood, and S. R. Ahmed, "Similarity measurement's comparison with mapping and localization in large-scale," in *Proc. AIP Conf.*, 2023, Art. no. 030029, doi: 10.1063/5.0119964.

[20] J. Machaj, P. Brida, and R. Piché, "Rank based fingerprinting algorithm for indoor positioning," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat.*, Sep. 2011, pp. 1–6, doi: 10.1109/IPIN.2011.6071929.

[21] G. Minaev, A. Visa, and R. Piché, "Comprehensive survey of similarity measures for ranked based location fingerprinting algorithm," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Sep. 2017, pp. 1–4, doi: 10.1109/IPIN.2017.8115922.

[22] A. Kulkarni, V. Tokekar, and P. Kulkarni, "Discovering context of labeled text documents using context similarity coefficient," *Proc. Comput. Sci.*, vol. 49, pp. 118–127, Jan. 2015, doi: 10.1016/j.procs.2015.04.235.

[23] A. Moreira, I. Silva, F. Meneses, M. J. Nicolau, C. Pendao, and J. Torres-Sospedra, "Multiple simultaneous Wi-Fi measurements in fingerprinting indoor positioning," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Sep. 2017, pp. 1–8, doi: 10.1109/IPIN.2017.8115914.

[24] J. Torres-Sospedra, A. Moreira, G. M. Mendoza-Silva, M. J. Nicolau, M. Matey-Sanz, I. Silva, J. Huerta, and C. Pendão, "Exploiting different combinations of complementary sensor's data for fingerprint-based indoor positioning in industrial environments," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Sep. 2019, pp. 1–8, doi: 10.1109/IPIN.2019.8911758.

[25] T. Alhmiedat, "Fingerprint-based localization approach for WSN using machine learning models," *Appl. Sci.*, vol. 13, no. 5, p. 3037, Feb. 2023, doi: 10.3390/app13053037.

[26] S. Sadowski, P. Spachos, and K. N. Plataniotis, "Memoryless techniques and wireless technologies for indoor localization with the Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 11, pp. 10996–11005, Nov. 2020, doi: 10.1109/JIOT.2020.2992651.

ABDULMALIK SHEHU YARO was born in Kaduna, Nigeria, in 1990. He received the B.Eng. degree in electrical engineering from Ahmadu Bello University (ABU), Zaria, and the M.Eng. degree in telecommunications engineering and the Ph.D. degree in electrical engineering from the University of Technology Malaysia (UTM), in 2014 and 2018, respectively. He is currently a Senior Researcher with the Department of Informatics and Quantitative Methods, University of Hradec Králové, Czech Republic, and an Assistant Professor with the Department of Electronics and Telecommunications Engineering, ABU. His articles have appeared in more than 20 international peer-reviewed journals and conference proceedings. His research interests include wireless localization and sensor networks.

FILIP MALY was born in Czech Republic, in 1981. He received the master's degree in information management from the University of Hradec Králové, Czech Republic, in 2005, and the Ph.D. degree in information and knowledge management, in 2009. Since 2014, he has been an Associate Professor in systems engineering and informatics with the University of Hradec Králové, where he is currently a Senior Researcher and a Teacher with the Department of Informatics and Quantitative Methods. His research interests include indoor localization, wireless sensor networks, programming languages, and operating systems.

PAVEL PRAZAK was born in Czech Republic, in 1968. He received the master's degree in the teaching of mathematics and physics and the Ph.D. degree in general problems of mathematics and informatics from the Faculty of Mathematics and Physics, Charles University, Prague, in 1996 and 2006, respectively. Since 2014, he has been an Associate Professor in systems engineering and informatics with the University of Hradec Králové, Czech Republic, where he is currently the Head of the Department of Informatics and Quantitative Methods, Faculty of Informatics and Management. His research interests include indoor localization, dynamical systems and optimization in economics, and applications of statistics.

KAREL MALÝ was born in Czech Republic, in 1973. He received the master's degree in cybernetics and the Ph.D. degree in artificial intelligence and biocybernetics from the Faculty of Electrical Engineering, Czech Technical University in Prague (CTU), Czech Republic, in 1998 and 2005, respectively. Since 2000, he has been an Assistant Professor and a Researcher with the Department of Informatics and Quantitative Methods, University of Hradec Králové, Czech Republic. His research interests include artificial intelligence, robotics, indoor localization, programming languages, and Unix operating systems.