Hierarchical Image Retrieval Method Based on Bag-of-Visual-Word and Eight-point Algorithm with Feature Clouds for Visual Indoor Positioning

Tainhui Zhang, Shizeng Guo, Lin Ma, Weixiao Meng
School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China
malin@hit.edu.cn

***Abstract*—The advance of 5G has brought great convenience and opportunities for the application of visual indoor positioning in the Internet of Things, augmented reality, emergency rescue and other important fields, but its positioning performance still needs to be improved. In order to increase the accuracy and efficiency of the vision-based indoor localization system, we propose hierarchical image retrieval algorithm based on Bag-of-Visual-Word and eight-point algorithm with feature clouds for position estimation which can effectively refine the efficiency and accuracy of image retrieval and location estimation. The experimental results show that the proposed method can dramatically improve the real-time performance and accuracy of the visual positioning system.**

Keywords—visual indoor positioning, image retrieval, Bag-of-Visual Word, eight-point algorithm

I. INTRODUCTION

With the development of modern technology and industrialization, people's demand for the perception of the surrounding environment is becoming stronger and stronger. Smartphone-based indoor positioning services have changed people's daily life greatly [1, 2]. Most of the traditional indoor positioning systems employ ultra-wideband, WIFI, RFID, bluetooth and so on. However, these methods cannot be promoted on a large scale for their extra deployment overhead, limited positioning range, poor anti-interference ability and many other real factors. In contrast, the vision-based positioning system has low positioning cost, high positioning accuracy and strong interactivity, which make it an indoor positioning technology with tremendous application potential in the Internet of Things, augmented reality, emergency rescue and other important fields [3, 4].

Generally, the visual indoor positioning system utilizes the real-time user image to estimate the user's location based on fingerprint recognition, and it consists of offline database creation and online localization corresponding to two stages[5]. In the offline stage, the main task is to build a visual map, which is the premise of image retrieval and location estimation. In the online stage, users submit a picture of the position by smart devices, then the positioning system will extract image features, retrieve to obtain similar database images, and finally estimate the user's location with the algorithm. The performance of the visual positioning system largely depends on the image retrieval performance and the accuracy of the location estimation algorithm, which are hot topics in current research [6]. However, the flow of people, transfer of markers, changes in light and any other environment changes always lead to certain differences between images and features collected at different moments. These cases will worsen the image retrieval and location estimation performance. Therefore, it is necessary to research how to improve the performance of image retrieval and the accuracy of position estimation to achieve better positioning performance.

Most of the research on refining image retrieval performance focuses on two aspects: improvement of the image features and optimization of the matching algorithm. The study of [7] shows that the speeded-up robust features can be well employed in the complicated indoor environment. Global features could be used for urban localization and they could optimize storage and computational efficiency [8]. The bag-of-word method widely used in text recognition has also been tried for image retrieval [9]. Multiclass support vector machines could be used for classifying the SIFT features to improve retrieval efficiency [10]. The epipolar geometry method is one of the classical methods for position estimation, and [11] uses the eight-point algorithm to resolve the fundamental matrix which is a critical part of epipolar geometry, but the method is not robust enough. In [12], a method of choosing the eight points has been proposed, which makes full use of the uniform distribution of feature points to avoid some match errors, however, it is too simple to satisfy most changeable cases.

Inspired by the above image retrieval and location estimation algorithms, and to enhance the performance of the visual indoor positioning system, we propose a hierarchical image retrieval algorithm based on Bag-of-Visual-Word with high retrieval speed and accuracy and an eight-point algorithm with feature clouds which takes a full account of the relationship and wholeness of feature points to improve robustness. Replacing global feature rough retrieval with Bag-of-Visual Word and considering the overall matching property of feature points are core ideas of our method. We have implemented the proposed method in our lab and a mall to evaluate the positioning performance. The experiment results show that they could achieve good performance in image retrieval and position estimation. The remainder of this paper is organized as follows. In section II, we would introduce the visual indoor positioning system and give the framework of the hierarchical image retrieval method and eight-point algorithm with feature clouds that we proposed. And the theoretical derivation and realization of the method will be demonstrated in section III. Section IV will provide the implementation and performance analysis to evaluate our method. And conclusion will be drawn finally.

II. System Overview

A. Epipolar Geometry Localization Algorithm

The epipolar geometric constraint relationship is a description of the position relationship between different cameras. Assuming that the same point can be recorded as (X, Y, Z) in the first coordinate system and (X', Y', Z') in the second coordinate. Then the relationship can be expressed by a rotation matrix $\mathbf{R}$ and transfer vector $\mathbf{t}$ as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{R} \begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} + \mathbf{t} \tag{1}$$

In epipolar geometry, there exists a point-to-line mapping between the first view and the second view, and it can be described by the fundamental matrix. Points can be noted as $\tilde{\mathbf{m}}$ and $\tilde{\mathbf{m}}'$ in the two pixel coordinates of the cameras, and they satisfy the following equation:

$$\tilde{\mathbf{m}}^{\mathrm{T}} \mathrm{F} \tilde{\mathbf{m}}' = 0 \tag{2}$$

where F represents the fundamental matrix.

Assuming that the two images have n pairs of matching points and can be expressed as a set:

$$\left\{ (\mathbf{m}_i, \mathbf{m}_i') \mid i = 1, 2, 3, ..., \mathrm{n} \right\} \tag{3}$$

where $\mathbf{m}_i = (u_i, v_i, 1)^{\mathrm{T}}$ and $\mathbf{m}_i' = (u_i', v_i', 1)^{\mathrm{T}}$.

According to (2), these pairs of points can form a system of linear equations as follows:

$$\mathbf{U}_n \mathbf{f} = 0 \tag{4}$$

where $\mathbf{f} = (F_{11}, F_{21}, F_{31}, F_{12}, F_{22}, F_{32}, F_{13}, F_{23}, F_{33})^{\mathrm{T}}$ and F_{ij} is the element of the fundamental matrix. Therefore $\mathbf{U}_n$ can be expressed as:

$$\mathbf{U}_n = \begin{bmatrix} u_1u_1' & u_1v_1' & u_1 & v_1u_1' & v_1v_1' & v_1 & u_1' & v_1' & 1 \\ u_2u_2' & u_2v_2' & u_2 & v_2u_2' & v_2v_2' & v_2 & u_2' & v_2' & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ u_1u' & u_nv_n' & u_n & v_nu_n' & v_nv' & v_n & u_n' & v_n' & 1 \end{bmatrix} \tag{5}$$

As (2) is a homogeneous equation set with $\det(\mathbf{F}) = 0$ and $\mathbf{F}$ is a singular matrix, we can easily know $\mathrm{rank}(\mathbf{F}) = 2$ which makes it convenient to solve the fundamental matrix.

There are many different algorithms for solving the fundamental matrix such as linear algorithms, iterative methods and robust methods. But it is usually solved by the eight-point algorithm, one of the linear algorithms[13]. If there are exactly 8 pairs of matching points, the fundamental matrix can be solved directly. For cases over 8 pairs of points, as most matching points meet (2), we only need to calculate a F by minimizing the cost as follows:

$$\min_{\mathrm{F}} \sum_i \left(\tilde{\mathbf{m}}_i^{\mathrm{T}} \mathrm{F} \tilde{\mathbf{m}}_i' \right)^2 \tag{6}$$

And it can also be rewritten in another way as:

$$\min_{\mathbf{f}} \left\| \mathbf{U}_n \mathbf{f} \right\|^2 \tag{7}$$

We could set a $\|\mathbf{f}\| = 1$ constraint to avoid getting zero, and then find the solution that satisfies the above condition. After eigen-decomposition of $\mathbf{U}_n^{\mathrm{T}} \mathbf{U}_n$, we get the eigenvectors corresponding to the smallest eigenvalue as $\mathbf{f}$.

B. System Framework

In our proposed work, the whole system can be divided into two parts which are the offline stage and the online stage. The structure of the system is shown in Fig. 1.

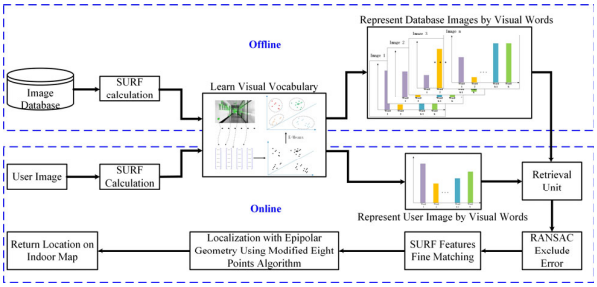

Fig. 1. System framework

The task of the offline phase is to build a database that consists of image information and geographic coordinate. The key part of the online phase is to perform a series of image processing and matching to find the most similar database images, and then use the epipolar geometry localization algorithm to calculate the geographic location of the images to be located.

The efficiency and accuracy of image retrieval determine the real-time performance and accuracy of the visual positioning system. To reduce the positioning time and improve the accuracy of image retrieval, we propose a hierarchical image retrieval algorithm based on Bag-of-Visual-Word.

In the offline stage, we calculate SURF features of database images and learn visual vocabulary, then we use the frequency of each word to represent the database images. By this means we can obtain a visual map with Bag-of-Visual Word and SURF features.

While, in the online stage, the system would extract SURF features of the user image and act a rough matching process between the pictures based on the word frequency using the retrieval unit, which calculates the matching level based on the similarity of the visual word frequency. By rough matching, we select certain database images as input to fine matching. After RANSAC fine matching, we get at least $k\,(k \geq 2)$ database images similar to the user image.

To reduce the error caused by environment change, we propose an eight-point algorithm with feature clouds to choose the eight pairs of matching points and solve the fundamental matrix. It acts as the following steps: feature clouds extraction, feature clouds matching, getting optimal feature cloud according to the cost, matching points search within the feature clouds, obtaining the rest optimal points and finally getting 8 pairs of matching feature points output as the input data to the epipolar geometry localization process. Then we can present a localization result to our users in a quick time.

III. The Proposed Method

A. The Hierarchical Image Retrieval Algorithm

To facilitate the subsequent calculation, we first process the extraction of SURF features. Gaussian filtering is performed before calculating the Hessian matrix for scale consistency and noise cancelation. Then we get the Hessian matrix as:

$$\mathrm{H}(\mathbf{x},\sigma)=\begin{bmatrix}\mathrm{L}_{\mathrm{xx}}(\mathbf{x},\sigma) & \mathrm{L}_{\mathrm{xy}}(\mathbf{x},\sigma)\\ \mathrm{L}_{\mathrm{xy}}(\mathbf{x},\sigma) & \mathrm{L}_{\mathrm{yy}}(\mathbf{x},\sigma)\end{bmatrix} \tag{8}$$

where $\mathrm{L}_{\mathrm{xx}}(\mathbf{x},\sigma)$ is the Gaussian second-order derivative to the image point in the image I, $\mathrm{L}_{\mathrm{xy}}(\mathbf{x},\sigma)$ and $\mathrm{L}_{\mathrm{yy}}(\mathbf{x},\sigma)$ have similar meanings.

In feature point detection, as the Hessian matrix describes the curvature of the function, its determinant represents the magnitude of the grayscale variation around the pixel point. Therefore, we can obtain the stable extreme value points of the image with the determinant.

Assuming that the SURF features we extract can be expressed as follows:

$$\boldsymbol{X}=[x_1,x_2,...,x_M],\boldsymbol{X}\in\mathbb{R}^{D\times M} \tag{9}$$

Bag-of-Visual-Word uses feature vector quantization to form a dictionary as:

$$\begin{aligned} q:\mathbb{R}^{D\times M}&\rightarrow[1,k]\\ x&\rightarrow q(x)\end{aligned} \tag{10}$$

where function q gives the mapping of the feature vector to an index. K-means clustering is generally used in the visual bag-of-word model and could obtain some cluster centers as the indexes of the mapping. And the dictionary can be described as:

$$\boldsymbol{B}=[\boldsymbol{b}_1,\boldsymbol{b}_2,...,\boldsymbol{b}_N],\boldsymbol{B}\in\mathbb{R}^{D\times N} \tag{11}$$

where the index $\boldsymbol{b}_i$ is commonly named word. So, two similar feature vectors have a similar probability distribution over visual words, and we set a subset of feature vectors belonging to the word $\boldsymbol{b}$ as follows:

$$\boldsymbol{X}_b=\{x\in\boldsymbol{X}:q(x)=\boldsymbol{b}\} \tag{12}$$

Therefore the similarity measurement function between two images can be given as:

$$\begin{aligned}&K(\boldsymbol{X},\boldsymbol{Y})=\\ &\gamma(\boldsymbol{x})\gamma(\boldsymbol{y})\sum_{b\in\boldsymbol{B}}(TF\sim IDF(\boldsymbol{b}))M(\boldsymbol{X}_b,\boldsymbol{Y}_b)\end{aligned} \tag{13}$$

where $TF\sim IDF$ are term frequency and inverse document frequency of the word $\boldsymbol{c}$. The normalization factor $\gamma(\boldsymbol{x})$ can be expressed as:

$$\gamma(\boldsymbol{x})=\left(\sum_{b\in\boldsymbol{B}}(TF\sim IDF(\boldsymbol{b}))\right)M(\boldsymbol{X}_b,\boldsymbol{X}_b)^{-\frac{1}{2}} \tag{14}$$

where $M(\cdot)$ is an optional kernel function, $K(\boldsymbol{X},\boldsymbol{Y})$ measures the similarity between the two images and ranks it among all retrieval images.

So far, we get rough matching output images I_{BOVW} as input to fine matching, then we calculate the confidence levels for each input database image and user image as:

$$N_{\mathrm{norm}}=\frac{N_{\text{RANSAC-in}}}{\min(N_{\mathrm{A}},N_{\mathrm{B}})} \tag{15}$$

where N_{A} and N_{B} are the number of the feature points of the two images. $N_{\text{RANSAC-in}}$ denotes the number of interior points estimated by the random sampling consistency algorithm. The optimal images can be obtained by N_{norm} which indicates the confidence level.

B. The Eight-point Algorithm with Feature Clouds

Compared to the normalized eight-point algorithm proposed in [13] to solve the fundamental matrix, the key of our proposed method is to modify the algorithm to choose the eight pairs of points. Supposing that the user image is I_{query} and the optimal database image is $I_{database}$, and we have already kept the two images' SURF features.

We partition each image into 9 blocks averagely denoted as R_{qi}, R_{di}, $1\le i\le 9$, for alternative feature cloud regions, and save the number of feature points in each region as N_{qi}, N_{di}, $1\le i\le 9$. Afterwards, we can express the percentage of feature points with P_{qi} and P_{di} as:

$$P_{qi}=\frac{N_{qi}}{\sum_{1}^{9}N_{qi}},1\le i\le 9 \tag{16}$$

Similarly, we can define P_{di} as:

$$P_{di}=\frac{N_{di}}{\sum_{1}^{9}N_{di}},1\le i\le 9 \tag{17}$$

Set a threshold value p, if $P_{qi}>p$ or $P_{di}>p$, the corresponding region would be a feature cloud C_{qi} or C_{dj}, and keep the number of feature points in the cloud as $N_{C_{qi}}$ and $N_{C_{dj}}$. Then, matching the feature clouds C_{qi} with C_{dj} and give the ratio W_{ij} to express the similarity between the two feature clouds as:

$$W_{ij}=\frac{N_{con}}{\max(N_{C_{qi}},N_{C_{dj}})} \tag{18}$$

where N_{con} represents the number of match points of the feature cloud.

According to (18), we get the match feature clouds and keep 4 pairs of most similar matching feature points within the clouds as: $\{(\mathbf{m}_{ci},\mathbf{m}'_{ci})\mid i=1,\cdots,4\}$, while based on the similarity of SURF features, we obtain the rest 4 pairs of matching feature points outside of the match feature cloud as: $\{(\mathbf{m}_{co},\mathbf{m}'_{co})\mid i=1,\cdots,4\}$.

So far, we obtain the 8 pairs of feature points for fundamental matrix solving. With the points, we can easily get the transfer vector of database images to the user image as $\mathbf{t}_{\mathrm{total}}$ which could be described as a straight line in the world coordinate. Theoretically, the intersection of two straight lines produces a point, which is the geographic

location of the user. To reduce the error, we use n lines to calculate the optimal location by minimizing the sum of the distance from the point to each line as:

$$\min_{x,y} \sum_i d_i(x,y) \tag{19}$$

where $d_i(x,y)$ denotes the distance from the point (x,y) to the ith line. And it could be shown as:

$$d_i(x,y) = \frac{|a_i x + b_i y + c_i|}{\sqrt{a_i^2 + b_i^2}} \tag{20}$$

where $a_i x + b_i y + c_i = 0$ represents the ith transfer line.

Considering the accuracy of the transfer lines, we use matching points of SURF features to weight the distances and the result could be rewritten as:

$$\min_{x,y} \sum_i d_i(x,y) N_i \tag{21}$$

where N_i expresses the number of matching points of the database image corresponding to the ith line. Then we obtain the optimal location of the user with our proposed algorithm.

IV. Implementation and Performance Analysis

A. Experiment Environment

In order to test the proposed hierarchical image retrieval method, we experiment on 476 indoor images of the UKB dataset and test the location performance in our lab which is located in the Technology Innovation Building, Science Park of Harbin Institute of Technology, China. It is a typical office environment and the total length of the target corridor is 57 meters. The floor plan is illustrated in Fig. 2, and we took database images at the blue points and collect user images at the red ones.

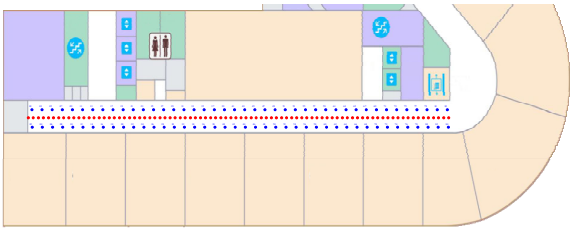

Fig. 2. Experiment floor plan of the office

To verify the proposed eight-point algorithm with feature clouds, we test it in a mall with a complex and changeable environment. The floor plan and acquisition route are shown in Fig. 3.

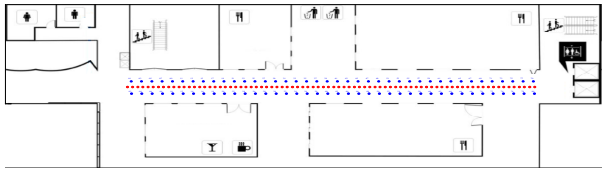

Fig. 3. Experiment floor plan of the mall

B. Proposed Retrieval Method Simulation Results

To analyze the rough retrieval method, we upload each image in the UKB dataset and process the rough retrieval to get k matching images among the rest 475 images. There are only 3 correct images in the dataset, if the match result contains more than 2 correct images, it can be seen as a successful matching. And we change k to test the retrieval performance of GIST rough matching and Bag-of-Visual-Word rough matching. And the accuracy result is provided in Fig. 4.

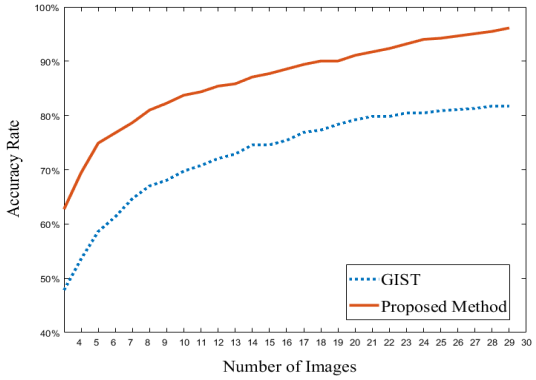

Fig. 4. Accuracy rate of rough retrieval

As shown in Fig. 4, when the acquired number of images $k = 9$, the proposed method has achieved an accuracy of over 80%. Compared to the traditional GIST method, its accuracy rate improves by nearly 15%. So the proposed method could effectively improve retrieval accuracy and efficiency. Besides, we test the location performance in the office environment and the results are shown in Table I.

TABLE I. Localization Performance

Method	Localization Performance		
	Localization error (CDF=80%)(m)	*Average error(m)*	*Average time(s)*
Proposed method	0.555	0.567	0.26
GIST+RANAC	0.723	0.845	0.38
GIST+SUEF	0.830	0.998	0.33

From Table I, we know that the positioning accuracy and efficiency of bag-of-visual-word are superior to the GIST matching. And our proposed hierarchical image retrieval algorithm cuts down the location error by 43% and shortens positioning time by over 20% compared to the traditional GIST and SURF image retrieval algorithm.

C. Proposed Eight-point Algorithm with Feature Clouds

In order to discuss the accuracy of solving fundamental matrix, we use Err_i to express the calculation error which could be defined as:

$$Err_i = \mathbf{x}_i'^{\mathrm{T}} \mathrm{F} \mathbf{x}_i \tag{22}$$

where $\mathbf{x}_i$ and $\mathbf{x}_i'$ denote the ith pair of SURF matching points of database image and user image in the corresponding pixel coordinate.

And we take the image extracted with SURF features shown in Fig. 5 as an example and calculate its Err_i as shown in Fig. 6.

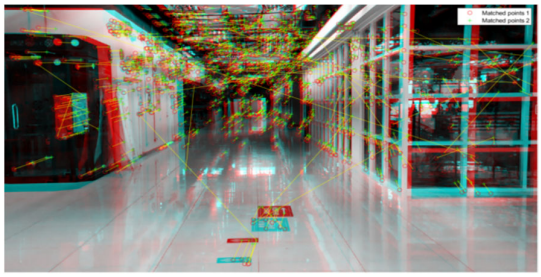

Fig. 5. Scene and SURF features of the mall

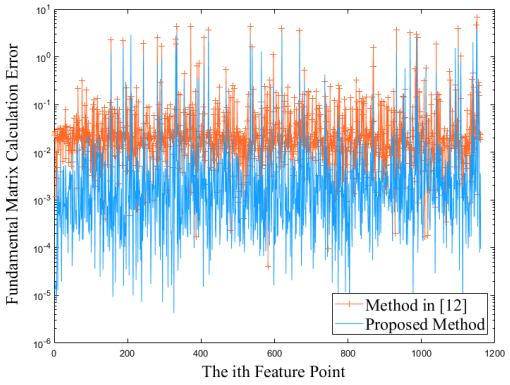

Fig. 6. Positioning error

As seen in Fig. 6, for most matching points in Fig. 5, the error of the fundamental matrix solved by our proposed method is less than the method in [12]. What's more, we test location accuracy in the mall of Fig. 3 and the result is given in Fig. 7.

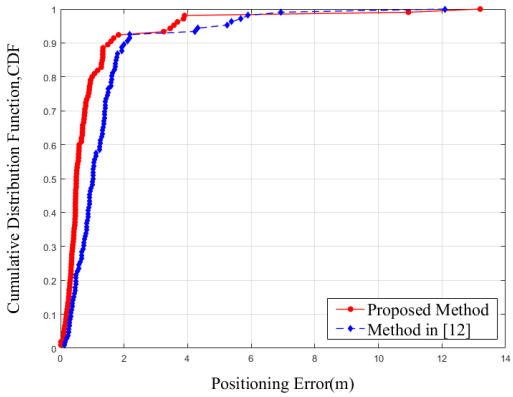

Fig. 7. Positioning error

From Fig. 7, we know that the proposed method can significantly improve location accuracy. It is worth noting that the CDF of 1 meter would increase to 80% if our method is used. And the CDF between 1 meter and two meters has been improved significantly. Therefore, our proposed method has a critical effect to improve the positioning accuracy while the error of solving the fundamental matrix is reduced.

V. Conclusion

In order to enhance the real-time performance and accuracy of image retrieval of visual positioning systems and improve the robustness of location estimation, we propose a hierarchical image retrieval method based on Bag-of-Visual Word and an eight-point algorithm with feature clouds for position estimation. The key parts of the proposed method are to replace global feature rough retrieval with Bag-of-Visual Word and consider the overall matching property of feature points. We test the proposed method in office and mall environments, and the experimental results show good performance in terms of real-time performance and localization accuracy.

Acknowledgment

This paper is supported by National Natural Science Foundation of China (61971162, 41861134010), and National Aeronautical Foundation of China (2020Z066015002).

References

[1] Y. You and C. Wu, "Hybrid Indoor Positioning System for Pedestrians With Swinging Arms Based on Smartphone IMU and RSSI of BLE," IEEE Transactions on Instrumentation and Measurement, vol. 70, 2021, pp. 1-15.

[2] S. Mekruksavanich, P. Jantawong and A. Jitpattanakul, "Deep Learning-based Action Recognition for Pedestrian Indoor Localization using Smartphone Inertial Sensors," 2022 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering, 2022, pp. 346-349.

[3] T. L. N. Nguyen, T. D. Vy, K. -S. Kim, C. Lin and Y. Shin, "Smartphone-Based Indoor Tracking in Multiple-Floor Scenarios," IEEE Access, vol. 9 2021, pp. 141048-141063.

[4] J. Dong, M. Noreikis, Y. Xiao and A. Ylä-Jääski, "ViNav: A Vision-Based Indoor Navigation System for Smartphones," IEEE Transactions on Mobile Computing, vol. 18, no. 6, 2019, pp. 1461-1475.

[5] Y. Xia, C. Xiu and D. Yang, "Visual Indoor Positioning Method Using Image Database," 2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS), 2018, pp. 1-8.

[6] J. Z. Liang, N. Corso, E. Turner and A. Zakhor, "Image Based Localization in Indoor Environments," 2013 Fourth International Conference on Computing for Geospatial Research and Application, 2013, pp. 70-75.

[7] Y. Zhuang, N. Jiang, H. Hu and F. Yan, "3-D-Laser-Based Scene Measurement and Place Recognition for Mobile Robots in Dynamic Indoor Environments," IEEE Transactions on Instrumentation and Measurement, vol. 62, no. 2, 2013, pp. 438-450.

[8] A. C. Murillo, G. Singh, J. Kosecká and J. J. Guerrero, "Localization in Urban Environments Using a Panoramic Gist Descriptor," IEEE Transactions on Robotics, vol. 29, no. 1, 2013, pp. 146-160.

[9] A. A. Olaode and G. Naghdy, "Elimination of Spatial Incoherency in Bag-of-Visual Words Image Representation Using Visual Sentence Modelling," 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ), 2018, pp. 1-6.

[10] S. K. Sundararajan, B. S. Gomathi and D. S. Priya, "Continuous set of image processing methodology for efficient image retrieval using BOW SHIFT and SURF features for emerging image processing applications," 2017 International Conference on Technological Advancements in Power and Energy (TAP Energy), 2017, pp. 1-7.

[11] H. Sadeghi, S. Valaee and S. Shirani, "A weighted KNN epipolar geometry-based approach for vision-based indoor localization using smartphone cameras," 2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop (SAM), 2014, pp. 37-40.

[12] L. Ma, H. Xue, T. Jia and X. Tan, "A Fast C-GIST Based Image Retrieval Method for Vision-Based Indoor Localization," 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), 2017, pp. 1-5.

[13] W. Chojnacki and M. J. Brooks, "Revisiting Hartley's normalized eight-point algorithm," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 9, 2003, pp. 1172-1177