Sensors and Materials, Vol. 34, No. 1 (2022) 337–348
MYU Tokyo

S & M 2814

Indoor Visual Positioning Method Based on Image Features

Xun Liu,[1] He Huang,[1*] and Bo Hu[2]

[1]School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, No. 15, Yongyuan Road, Huangcun Town, Daxing District, Beijing 102616, China
[2]National Geomatics Center of China, No. 28, Lianhuachi West Road, Haidian District, Beijing 100830, China

(Received July 21, 2021; accepted October 25, 2021)

Keywords: indoor visual positioning, ORB feature, bag-of-visual-words model, term frequency–inverse document frequency, efficient perspective-n-point

In this study, we propose an indoor visual positioning method based on image features. RGB-D camera data are used to establish an image database used for positioning. The 3D coordinates of pixels are obtained from an RGB image and depth information, and then the oriented fast and rotated brief (ORB) features of the image are extracted. The bag-of-visual-words model is used in combination with the *K*-means algorithm and a *k*-dimensional tree structure to classify storage and expressions in the dictionary. In the positioning process, the positioning image is obtained using a camera with known parameters, and the term frequency–inverse document frequency model is used to achieve image feature indexing to match the most similar image. Finally, using the matching feature points in the image, an efficient perspective-n-point method and a bundle adjustment method are used to calculate the camera pose information on the positioning image to complete indoor positioning. Experiments on real scenes verify the feasibility of the proposed method and its positioning accuracy. The results presented in this study provide a useful reference in the research and application of vision-based indoor positioning.

1. Introduction

With the popularization of Internet of Things technology and the development of information technology, people's demand for location services in indoor environments has increased significantly. Accurate and fast indoor positioning methods are fundamental to meet user demands and needs. Owing to the maturity of computer vision theories, the measurement, recognition, and tracking of objects can be easily achieved using cameras and computer software. Therefore, research on image processing has become a hotspot in many fields. One of the features of cameras is non-contact measurement, which can obtain 3D information of objects. The output image of a camera reflects the 3D world in a 2D form, and rich image features can be obtained in a scene with a rich image texture. Therefore, it is possible to find the same feature points of two images describing the same scene by mining the image feature information, calculating the camera pose, and realizing the relative positioning between the two position points.

*Corresponding author: e-mail: huanghe@bucea.edu.cn
https://doi.org/10.18494/SAM3562

ISSN 0914-4935

There are a variety of methods for indoor visual positioning, including the method of adding or marking markers in an indoor scene in advance and then realizing the positioning through marker recognition.[1,2] The positioning accuracy of this method is high, but the markers are not removable, require easy storage, and are greatly affected by changes in the indoor environment. As another type of method, by establishing various types of map databases, the data collected using various types of sensors are compared with the information in databases to obtain positioning information. This type of method involves establishing an image location fingerprint database, dividing an indoor space into grids in advance, recording the data information collected in each grid, and obtaining positioning information based on the consistency between the collected data information of the locator and the grid information.[3–5] This method has a high positioning speed but limited accuracy, and any improvement in accuracy significantly increases the workload of database construction. It also requires the establishment of a 3D image database, the pre-calculation of the spatial position of every feature point, the comparison of feature points in the positioning image to form the corresponding feature point pair, and then the solution of the position information of the visual sensor to complete the positioning process.[6–9] The accuracy of this type of method is related to the richness of environmental features. Other methods complete the positioning and mapping at the same time, such as simultaneous localization and mapping (SLAM). These methods mainly rely on adjacent frame images collected using a vision sensor to calculate the camera pose using an epipolar constraint by visual odometry and to obtain its positioning information, and then the map is constructed using the collected images.[10–12] These methods do not require a database to be established in advance and the positioning is relatively simple, but visual SLAM using a single sensor is not robust enough and the cumulative error is large at long distances.

In recent research on indoor visual positioning, the main challenge has been how to achieve accurate and fast indoor positioning. In this study, we propose an indoor positioning method based on image features. The feasibility and position accuracy of the proposed method are verified by experiments on real scenes. The method presented in this study can provide a useful reference for vision-based indoor positioning research and applications.

2. Indoor Positioning Method Based on Image Features

The proposed indoor visual positioning method uses a Kinect V2 camera as a carrier, uses feature information on the camera image, combines the feature points of the image and its depth information, and realizes indoor visual positioning based on the image features. The proposed method consists of two parts, database construction and positioning realization. The framework of the proposed method is shown in Fig. 1.

In the first part, Kinect V2 camera data are used to construct the image database. The 3D coordinates of pixels are obtained from an RGB image and depth information, whereupon the features of the image are extracted and an index is created according to the feature type. In the second part, the positioning image is obtained using a camera with known parameters, and its features are extracted to match it with the most similar image. Finally, using the matching feature points in the image, the camera pose information on the positioning image is calculated to complete indoor positioning.

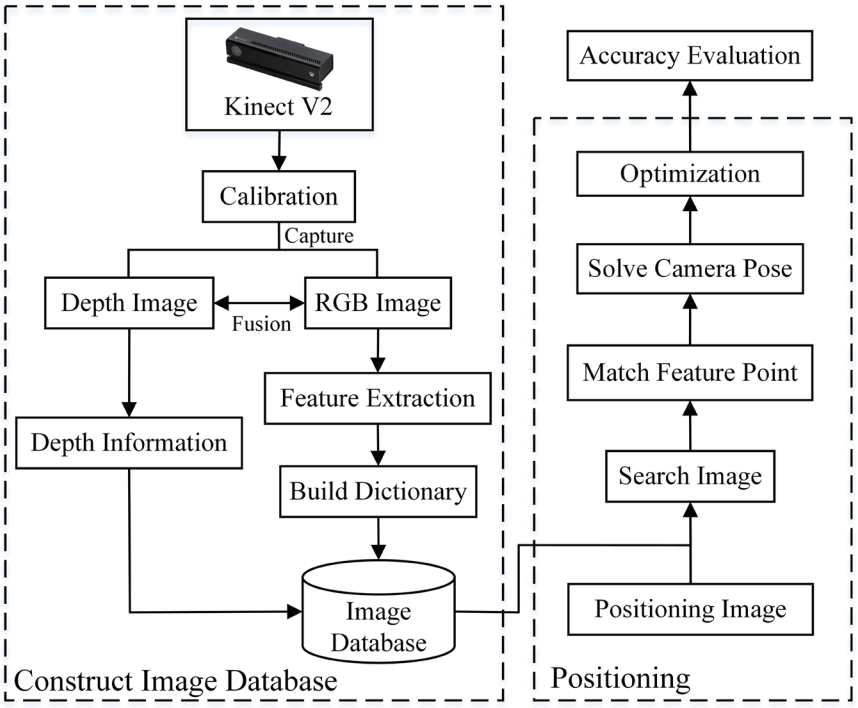


Fig. 1. Framework of the proposed indoor visual positioning method based on image features.

2.1 Image database construction

(a) Image information fusion

An RGB-D camera obtains RGB image information and depth image information separately. Although their resolutions are different, they need to be registered together. Therefore, the RGB and depth cameras are calibrated at the same time using the classic calibration method.[13] First, the internal parameters of the two cameras are obtained, and then the conversion parameters between the coordinate systems of the two cameras are calculated using the matching points in the captured images. Finally, the image registration and coordinate transformation of the RGB and depth images are completed, as shown in Fig. 2.

(b) Image feature extraction

In this study, we extract the oriented fast and rotated brief (ORB) feature of an image. This feature is selected because low image quality, high noise, and image rotation, translation, and zooming have very little impact on this feature, and its extraction is accurate and fast.[14] In addition, it has the characteristics of repeatability, distinguishability, and high efficiency, which can be beneficial in indoor visual positioning.

(c) Image classification and expression

After extracting image features, they should be accurately classified and stored in a database, and an orderly index structure is constructed to reduce the computational complexity and time cost of feature matching and retrieval. The problem of image classification can be solved using

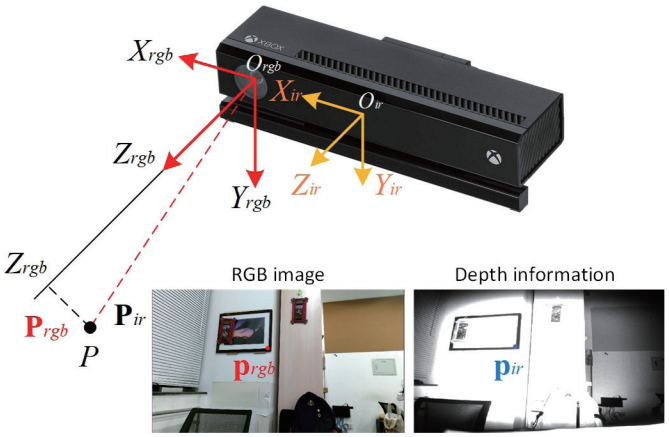


Fig. 2. (Color online) Kinect V2 camera registration.

models and algorithms related to text classification. The bag-of-words (BoW) model was first proposed in 1998 by Joachims. The BoW model puts all words contained in several texts into a bag, regardless of the word order in the texts, and all words are independent of each other. In this way, the text can be expressed using the index number of the dictionary to form a vector map. The bag-of-visual-words (BoVW) model treats a visual image as a text, where image features correspond to words in the text. The general process of converting an image into words is to convert all image feature points into different types of visual words through clustering, and finally, all visual words are combined into a dictionary, and every image is stored and expressed in the form of an image histogram according to the corresponding dictionary content. As shown in Fig. 3, a user can query images by training a classifier to match and retrieve similar images. In this study, we apply the classification principle to image matching and retrieval in indoor visual positioning, which can significantly reduce the time cost and computational complexity of traditional image retrieval based on feature point extraction and matching.

The image database construction includes the following five steps:

(1) Extract ORB features from all images.
(2) Use the *K*-means algorithm to cluster ORB feature descriptors, generate words from the clustering results, and store them in the *K-d* tree dictionary.
(3) Compare the similarity between the words clustered by the ORB features in an image and the words in the dictionary; if the predefined threshold is exceeded, the word exists in the image.
(4) By analogy, construct a visual image histogram that shows which words the image consists of and store the corresponding index structure.
(5) Repeat Steps (1)–(4) until histograms for all images in the database are generated and labeled.

2.2 Indoor positioning

The positioning process can begin after the construction of the indoor image database. The most similar image is first retrieved from the database on the basis of the features of the positioned image, after which the feature points in the two images are matched. Finally, the pose information is calculated and optimized using the matching results.

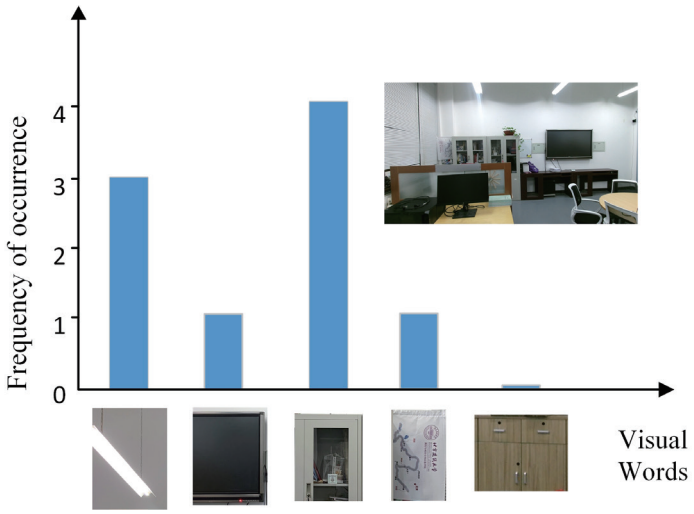


Fig. 3. (Color online) Image histogram.

(a) Image retrieval and similarity calculation

In this study, in the construction of the BoVW model, by indexing the extracted features, images with similar words to the query image can be retrieved and matched. One of the commonly used methods for text retrieval is the term frequency–inverse document frequency (TF-IDF) method. The TF-IDF model represents a widely used weighting algorithm, which is often used for information retrieval and data mining. We use the concept of this model and apply it to the classification and retrieval of image features. The idea of the term frequency (TF) is that if a word often appears in an image, its discrimination is high, and the idea of the inverse document frequency (IDF) is that the lower the frequency of a word in the dictionary, the higher the discrimination will be when classifying images.

In the BoVW model, the ratio of the number of features in a leaf node w_i to the total number of features is denoted as the IDF part. Assuming that the total number of features is n and the number of word w_i is n_i, the IDF of a word is calculated as

$$IDF_i = \log \frac{n}{n_i}. \tag{1}$$

Furthermore, the TF part refers to the frequency of a certain feature in an image. Assuming that word w_i appears n_i times in image A and the total number of words appearing in image A is n, then TF_i is calculated as

$$TF_i = \frac{n_i}{n}. \tag{2}$$

Therefore, the weight of word w_i is equal to the product of TF and IDF, which can be expressed as

$$\eta_i = TF_i \times IDF_i. \tag{3}$$

After considering the weights of image A, its feature points can correspond to many words in its BoVW model, which can be expressed as

$$A = \left\{ \left(w_1, \eta_1\right), \left(w_2, \eta_2\right), \ldots, \left(w_N, \eta_N\right) \right\} \triangleq \boldsymbol{v}_{\boldsymbol{A}}. \tag{4}$$

Since similar features may be part of the same class, there can be a large number of zeros in vector $\boldsymbol{v}_{\boldsymbol{A}}$, but $\boldsymbol{v}_{\boldsymbol{A}}$ can be used to describe image A by considering the BoW model. This vector is a sparse vector, its non-zero parts indicate which words are contained in the image, and the value of these parts represents the TF-IDF value; namely, the similarity between the image to be retrieved and the image in the image database can be found using the TF-IDF value.

(b) Feature point matching

Feature matching is performed on feature descriptors, and every feature point corresponds to an ORB feature descriptor. The distances between an image and all the descriptors in the image are measured, and the matching points are determined using the measured values. Every two points with the smallest distance between them are considered as feature matching points, so all distance values are sorted and selected in advance. The degree of similarity between two feature points is reflected in the distance of the descriptors, so the distance measurement adopts different norms according to different descriptor types. For instance, the ordinary Euclidean distance metric is suitable for storing single-precision or double-precision floating-point number-type descriptors, while for a binary descriptor of the rBRIEF, the metric norm is the Hamming distance, which means that two strings with identical length correspond to different numbers of bits. Therefore, the Hamming distance is used to measure the ORB features to achieve feature point matching. To avoid mismatches in the matching process, the PROSAC algorithm[11,15] is used. This algorithm first sorts data by similarity and selects the relationship with the highest similarity as a subset and then extracts sample data for model parameter estimation, eliminates mismatches, and ensures the correctness of matching points with the same characteristics.

(c) Camera pose calculation

After image retrieval, multiple 3D–2D matching point pairs of a user's position image are obtained. The 2D point denotes the position of the user's position image feature point in the image pixel coordinate system, and the 3D point denotes the 3D position coordinate in the world coordinate system corresponding to the 2D feature point. The pose information obtained using a local camera can be obtained by 3D–2D point mapping. The perspective-n-point (PnP) method can be used for solving the 3D–2D point pair motion. Compared with the 2D–2D epipolar constraint method, the PnP method can more effectively solve the problems of initialization, pure curl, and scale. This method does not require the use of epipolar constraints and can provide an accurate motion estimation using only a few matching points, which makes it an important attitude estimation method. Next, the EPnP algorithm[16] is applied to the 3D–2D matching point

pairs obtained in the image matching retrieval stage to estimate the user location. The EPnP algorithm expresses the coordinate information on points in the world coordinate system as a weighted sum of four non-coplanar control points. By determining the coordinates of the four control points in the camera coordinate system, the camera pose is obtained. The workflow of the EPnP algorithm is shown in Fig. 4.

(d) Pose optimization

This study uses the bundle adjustment (BA) method to optimize the obtained transformation parameters. The BA method is a pose solution method that comprehensively considers the constraints between multiple points in an image. By adjusting the pose parameters and feature point positions, the reprojection error of all matching points is minimized. The BA method simultaneously optimizes the positions of feature points and pose parameters, and in this study, it is applied to the pose results of the PnP method, thereby improving the accuracy of pose estimation. After obtaining the optimized pose parameters by the BA method, by combining the translation part in the parameters with the camera position coordinates of the optimal matching image retrieved from the database, the spatial coordinates of the positioning points are calculated.

3. Experiment and Results

3.1 Image database construction

We used the second-floor corridor and Room 223 of Building F, Beijing University of Civil Engineering and Architecture, as experimental sites. The point-by-point interval sampling method was used to collect indoor images to build a database. First, a point in the corridor was

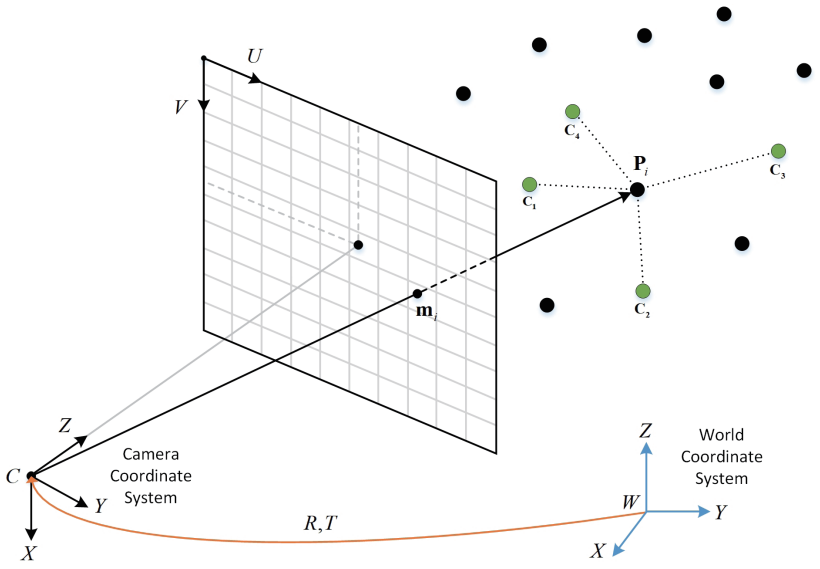


Fig. 4. (Color online) Workflow of the EPnP algorithm.

selected to establish the world coordinate system, as shown in Fig. 5 (the coordinate axis perpendicular to the ground is not shown in the plan). Then, an image acquisition station was set near the center of Room 223, and one station was placed at every 4 m in the corridor; it was ensured that all indoor scenes were within the effective detection range of the Kinect V2 camera used.

A total of 11 measuring stations were arranged in Room 223 and the second-floor corridor, and a total of 132 sets of color and depth images were collected. The image data of Room 223 are shown in Fig. 6. Using the method proposed in this paper, an indoor image database storing image feature information, depth information, and geographic location information was constructed.

3.2 Indoor positioning results and accuracy evaluation

First, using the results of the database image retrieval, the matching image that was most similar to the positioning image was found. The matching image results of eight location image collection points in Room 223 and ten location image collection points in the second-floor corridor were retrieved. As shown in Fig. 7, matching points were selected for each pair of retrieved images according to the matching degree. The 3D position information on the matching image in the 3D space was obtained from the image database, and 3D–2D matching points of the positioning image were obtained. The EPnP method was used to obtain the rotation and translation matrix of the user's camera, and finally, the positioning was realized.

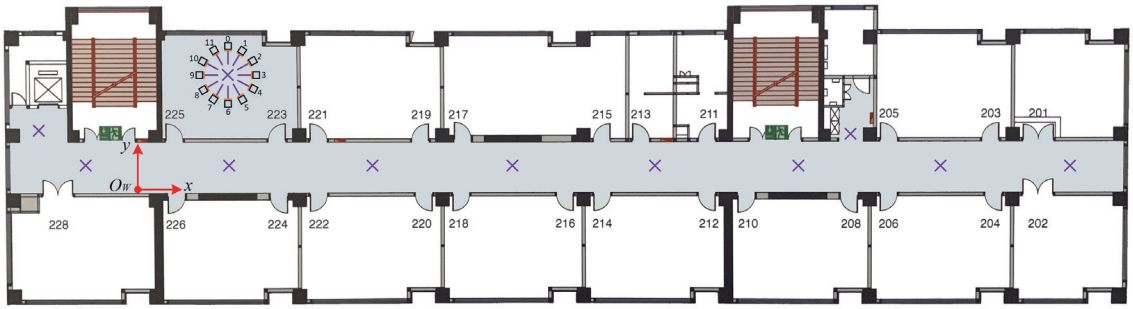
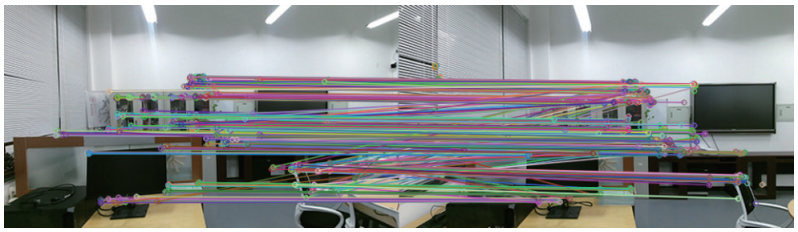


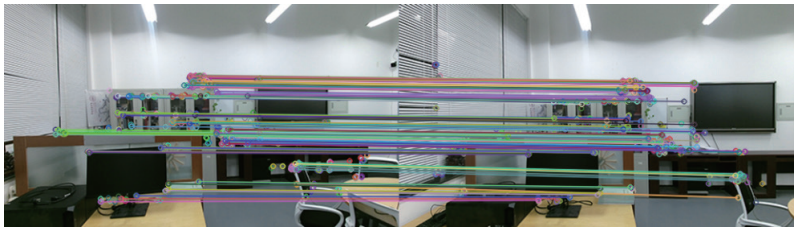
Fig. 5. (Color online) Schematic diagram of the experimental site.

(a) (b)

Fig. 6. (Color online) Image data of the Room 223 environment. Contrast and brightness have been adjusted in the original depth images to assist readers' interpretation. (a) RGB images. (b) Depth images.



(a)



(b)

Fig. 7. (Color online) ORB feature extraction and matching experiment. (a) Feature matching of two images. (b) Optimization of feature matching results.

To demonstrate the reliability of the pose calculation results and evaluate the indoor positioning accuracy, the total station measurement method was used to transform the pose calculation results into the database-building coordinate system. The accuracy and robustness of indoor positioning results obtained by different algorithms in different scenarios were compared. In the experiment, the positioning error of images at different locations in the same scene was used to calculate the root mean square error (RMSE) to evaluate the positioning accuracy. The RMSE was calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(M_i - G_i)^2}{n}}, \tag{5}$$

where M_i is a measured value and G_i is the corresponding ground truth value.

(a) Point position error calculation

In the two indoor scenes, the polar constraint method and EPnP algorithm were used for visual positioning, and the plane coordinate error between the measurement result and the ground truth value was calculated. Statistical analysis and calculations were performed for all location points, and the results are shown in Table 1. The RMSE of the point error was used to evaluate the accuracy of the positioning method. The RMSEs of the two algorithms were calculated to obtain the positioning accuracy separately.

We compare and analyze the errors of the two sets of positioning results. In the indoor environment of Room 223, the RMSE values obtained by the EPnP algorithm and epipolar constraint method were 0.139 and 0.164 m, respectively. In the second-floor corridor environment, the RMSE values of the EPnP algorithm and epipolar constraint method were 0.129 and 0.161 m, respectively.

Table 1
Point positioning errors.

	Point	Error (m)	
		EPnP algorithm	Epipolar constraint method
Room 223	1	0.059	0.069
	2	0.107	0.108
	3	0.056	0.082
	4	0.151	0.236
	5	0.156	0.185
	6	0.189	0.204
	7	0.175	0.176
	8	0.154	0.175
	RMSE (m)	0.139	0.164
Second-floor corridor	9	0.151	0.231
	10	0.148	0.149
	11	0.167	0.165
	12	0.147	0.238
	13	0.064	0.066
	14	0.136	0.131
	15	0.145	0.176
	16	0.098	0.108
	17	0.123	0.134
	18	0.072	0.129
	RMSE (m)	0.129	0.161

(b) Accuracy analysis and evaluation

According to the results, the EPnP algorithm improved the accuracy of indoor visual positioning compared with the extreme constraint method by 15.2 and 19.9% in Room 223 and the second-story corridor, respectively. The positioning error of the pose estimation algorithm was within 20 cm and the positioning effect was good. By comparing the performance of the two algorithms, it can be concluded that the overall error of the EPnP algorithm in indoor positioning was lower than that of the extreme constraint algorithm. Thus, the 3D–2D pose calculation that used the depth information was more accurate than the 2D–2D pose calculation based on the epipolar constraints of two images, and the indoor positioning accuracy was higher.

4. Conclusion

In this study, an indoor visual positioning method was proposed. The BoVW model was used to facilitate the retrieval of an optimal matching image, and the positioning process was completed using the spatial coordinates of the optimal matching image and pixels. This method can reduce the time required to build a 3D database model. The experimental results indicated that the EPnP algorithm, which combines the 3D information of the spatial point and the 2D image feature point information to calculate the pose, can solve the initialization, pure rotation, and scale problems that exist in the traditional 2D–2D epipolar geometric constraint method, thus improving the positioning accuracy. The overall positioning accuracy of the proposed positioning method was within 20 cm.

However, the proposed method has certain limitations. Internal parameters must be obtained in advance for the camera used to collect positioning images, and the positioning accuracy is affected by the number of image feature points. Therefore, in scenes with no features or a relatively small number of repetitive scenes, these limitations will cause matching errors and positioning failures. In future research, some real-time calibration methods need to be used to solve the problem of pre-calibration, and graphical features instead of point features should be used to increase the stability of matching.

References

1 H. Sadeghi, S. Valaee, and S. Shirani: 2014 IEEE 25th Annu. Int. Symp. Personal, Indoor, and Mobile Radio Communication (IEEE, 2014) 2024–2028. https://doi.org/10.1109/PIMRC.2014.7136504
2 J. Lin, C. Su, J. Zheng, and X. Yu: Mech. Sci. Technol. Aerosp. Eng. **34** (2015) 1675. https://doi.org/10.13433/j.cnki.1003-8728.2015.1107
3 F. Vedadi and S. Valaee: IEEE Trans. Syst. Man Cybern. Syst. **50** (2020) 305. https://doi.org/10.1109/TSMC.2017.2695080
4 J. Z. Liang, N. Corso, E. Turner, and A. Zakhor: Int. Conf. Indoor Positioning & Indoor Navigation (2014) 1–9. https://doi.org/10.1109/IPIN.2013.6817866
5 H. Xue, L. Ma, and X. Tan: Int. Wireless Commun. Mobile Comput. Conf. (2016) 650–654. https://doi.org/10.1109/IWCMC.2016.7577133
6 G. Feng, L. Ma, and X. Tan: J. Sens. (2017) 1–18. https://doi.org/10.1155/2017/8037607
7 H. Sadeghi, S. Valaee, and S. Shirani: 2014 IEEE 8th Sensor Array and Multichannel Signal Process. Workshop. (IEEE,2014) 37–40. https://doi.org/10.1109/SAM.2014.6882332
8 M. Li, R. Chen, X. Liao, B. Guo, W. Zhang, and G. Guo: Remote Sens. **12** (2020) 869. https://doi.org/10.3390/rs12050869
9 Y. Chen, R. Chen, M. Liu, A. Xiao, D. Wu, and S. Zhao: Sensors **18** (2018) 2692. https://doi.org/10.3390/s18082692
10 J. Xin, J. Gou, X. Ma, K. Huang, and Y. Zhang: Robot **36** (2014) 560. https://doi.org/10.13973/j.cnki.robot.2014.0560
11 M. Cao, L. Hu, P. Xiong, J. Peng, and L. Zeng: Chin. J. Sens. Actuator **32** (2019) 1706. https://doi.org/10.3969/j.issn.1004-1699.2019.11.018
12 J. Zhang and S. Singh: J. Field Rob. **35** (2018) 1242. https://doi.org/10.1002/rob.21809
13 Z. Zhang: IEEE Trans. Pattern Anal. Mach. Intell. **22** (2000) 1330. https://doi.org/10.1109/34.888718
14 M. Bansal, M. Kumar, and M. Kumar: Multimedia Tools Appl. **80** (2021) 18839. https://doi.org/10.1007/s11042-021-10646-0
15 O. Chum and J. Matas: 2005 IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognition (2005) 220–226. https://doi.org/10.1109/CVPR.2005.221
16 V. Lepetit, F. Moreno-Noguer, and P. Fua: Int. J. Comput. Vision **81** (2009) 155. https://doi.org/10.1007/s11263-008-0152-6

About the Authors

Xun Liu received his B.S. degree from Jilin University, China, in 2017. He is studying for his M.S. degree at Beijing University of Civil Engineering and Architecture, China. His research interest is in visual navigation and positioning.

He Huang received his B.S. degree from Wuhan University, China, in 2000 and his M.S. and Ph.D. degrees from Sungkyunkwan University, South Korea, in 2004 and 2010, respectively. Since 2010, he has been a lecturer and associate professor at Beijing University of Civil Engineering and Architecture, China. His research interests are in autonomous driving, high-precision navigation maps, and visual navigation and positioning.

Bo Hu obtained his master's degree from Beijing University of Civil Engineering and Architecture in 2021, where he majored in surveying and mapping engineering. At present, he is an assistant engineer mainly engaged in research on the operation, maintenance, and expansion of a surveying and mapping archive management system, and the standardized construction of surveying and mapping digital archives.