Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5527148, 15 pages
https://doi.org/10.1155/2021/5527148

Research Article

Mutual Positioning Method in Unknown Indoor Environment Based on Visual Image Semantics

Lin Ma ,[1,2] He Dong,[1,2] and Bin Wang[1,2]

[1] *School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China*
[2] *Science and Technology on Communication Networks Laboratory, Shijiazhuang, China*

Correspondence should be addressed to Lin Ma; malin@hit.edu.cn

Received 31 January 2021; Revised 16 February 2021; Accepted 25 February 2021; Published 8 March 2021

Academic Editor: Xin Liu

In our society, realizing intelligent positioning in indoor environments is important to build a smart city. Currently, mutual positioning requirements in the unknown indoor environment are growing fast. However, in such environment, we can obtain neither outdoor radio signal nor the indoor images in advance for online positioning. Therefore, how to achieve mutual positioning becomes an interesting problem. In this paper, we propose a vision-based mutual positioning method in an unknown indoor environment. First, two users take images of the unknown indoor environment, use semantic segmentation network to identify the semantic targets contained in the images, and upload the generated semantic sequence to the user shared database in real time. Then, every time two users reupload a semantic sequence due to a change of location, it is necessary to retrieve whether another user has uploaded the same semantic sequence in the shared database. If the retrieval is successful, it means that two users have seen the same scene. Finally, two users select a target from the two user images taken based on the same scene to establish a three-dimensional coordinate system, respectively, calculate their own position coordinates in this coordinate system, and realize mutual positioning through position coordinate sharing. Experiment results show that our proposed method can successfully realize mutual positioning between two users in an unknown indoor environment, while ensuring high positioning accuracy.

1. Introduction

1.1. Motivation. In our daily life, people often enter and exit some completely unknown indoor places such as office buildings and shopping malls. When multiple users are in different positions in an unknown indoor environment, they are eager to know where others are relative to them. Therefore, mutual positioning among users in an unknown environment has an important practical significance and a very broad development prospect. However, due to the lack of prior information of environmental layout in the unknown indoor environment, it is difficult for people to determine their current positions. Therefore, in the process of mutual positioning, users need to conduct information interaction and share the information they can currently obtain so as to determine the position of each other.

With the rapid development of smart phones, they are generally equipped with high-pixel image acquisition sensors, which can be used to collect images of indoor scene more conveniently and quickly. In the process of mutual positioning, users need to take pictures of the scene they can currently see and share them to others in real time. After users take images of an unknown indoor environment, this paper uses semantic segmentation network to identify the semantic targets contained in the images and uploads the generated semantic sequence to the shared database in real time. Based on the semantic sequences that users upload, we can judge whether they have seen the same scene. When it is determined that users can see the same scene, then they choose a target as the benchmark to establish the location connection among them. Users solve their position coordinates relative to the target, respectively, and finally realize mutual positioning through the sharing of coordinate information.

1.2. Related Works. With the popularization of smart phones and the rapid improvement of terminal processing speed, visual positioning technology has become a research hotspot in recent years and has been widely concerned by researchers. At present, indoor positioning methods in indoor environment are mainly divided into four categories: indoor positioning method based on wireless signals [1–3], based on inertial navigation [4], based on geomagnetic information [5], and based on vision [6–8]. In these approaches, indoor positioning method based on vision has obvious advantages because it not only has the advantages of low deployment cost, strong autonomy, and high positioning accuracy but also the principle of image collection in the method is very similar to that of human eyes observing the surrounding environment. Therefore, this paper uses visual positioning method to achieve mutual positioning among users. The current mutual positioning technology is mainly researched for outdoor scenes to realize mutual positioning and navigation of intelligent vehicles and drones. In [9, 10], the mutual positioning technology was applied to the development of intelligent transportation systems, using the Android operating system that is provided by Google attached to the smartphone that can provide access to the original GNSS measurement. Silantyev et al. [11] studied the method of mutual positioning in unmanned aerial vehicle (UAV) group control systems. Each UAV has a global navigation satellite system (GNSS) receiver and its own radar station with active response.

However, in the indoor environment, mutual positioning based on unknown environment is still a relatively new field, and there are not many reference materials. Therefore, this paper starts from the research status of visual positioning technology and image semantic segmentation technology [12–14] and proposes a method for mutual positioning between users in an unknown environment. In this paper, a visual positioning method based on identification is adopted to solve the user's position coordinates in the current coordinate system by taking the same object that can be seen between users as the center of the coordinate system. Zhang et al. [15] proposed a joint BA framework to consider other constraints from detected road traffic signs and solved the problem of error drift by correlating several image frames together to optimize the camera attitude and simultaneously extract 3D map points. In order to focus on the camera localization problem using visual semantic information, a coarse to a fine mechanism in [16] was aimed at localizing the camera position. Through simulation experiment, this proposed framework was not only useful for visual localization but also useful for other advanced tasks of robot. In [17], aiming at the problem of excessive or insufficient exposure of image areas due to the rapidly changing lighting conditions during feature detection, synchronous video stream fusion in HDR stereo cameras can significantly improve the problem of low matching accuracy of identification areas due to changes in light environment and effectively improve the positioning accuracy. The authors in [18–20] proposed a localization method based on monocular vision in dynamic environment, aiming at the disturbance of dynamic objects to the estimation results of visual odometer of visual SLAM system in dynamic environment.

In order to recognize the semantic identifiers in the image correctly, it is necessary to use the image semantic segmentation technology to process the user image. Sun et al. [21] put forward a novel RGB and thermal data fusion network FuseSeg to achieve the excellent performance of semantic segmentation in urban scenes, which can be better applied to urban scenes. The novel end-to-end network for multimode salient target detection in [22] transforms the challenge of RGB-T salient detection into the problem of CNN feature fusion. This method has good segmentation effect in the environment of cluttered background and insufficient light. Wei et al. [23] adopted the image-level annotation method for the training data, used the classification network to obtain the significant area target in the image according to the primary and secondary level, improved the pixel accuracy of the significant area, solved the situation of semantic segmentation edge blur, and got a good classification effect. Liu [24] proposed a computationally efficient lightweight image semantic segmentation network based on multilevel feature parallel network (LSSN), which comprehensively improves the real-time performance of semantic segmentation algorithms and accuracy.

At present, according to the different camera working methods in the visual positioning method, the positioning system can be divided into monocular visual positioning [25], binocular visual positioning [26–28], and depth visual positioning [29]. Since both monocular cameras and depth cameras have certain limitations in current applications, this paper uses binocular vision positioning technology to determine the user's position coordinates. The binocular vision positioning technology conforms to the process of human beings perceiving the position of surrounding objects through both eyes, and the left and right cameras represent the left and right eyes of humans. By observing the imaging difference of the same object between the left and right cameras, we can obtain the depth information of the target. Therefore, this paper uses binocular vision positioning method to solve the distance between the selected target and users, then calculate the position coordinates of users relative to the target, and finally realize the mutual positioning between users through the coordinate information sharing between users.

1.3. Contributions and Paper Organization. In this paper, we investigate the problem of mutual positioning among users in an unknown indoor environment. The major contributions of this paper are summarized as follows:

(1) We propose a method to represent images by using semantic objects contained in corresponding images in order to reduce the time of image matching and the storage capacity in the database. We use R-FCN to detect semantic targets in images. Therefore, network training should be carried out on R-FCN before positioning. Some common objects in indoor

environment, such as doors, windows, and hydrants, should be photographed from multiple angles so that the final trained R-FCN network has a high target recognition accuracy

(2) We propose an image retrieval algorithm based on image semantic sequence to solve the problem of fast matching among user images. When we need to determine whether users have seen the same scene, we need to retrieve whether the pictures taken by users are similar. Traditional image retrieval methods need to extract all feature points in the image and observe the matching degree of image feature points, which not only takes up a large storage capacity but also takes a long time. In this paper, R-FCN is used to identify the semantic target of user images and finally generate a corresponding semantic sequence of images. Users upload the corresponding semantic sequence to the shared database after each image shooting and judge whether they have seen the same scene by retrieving the same or similar semantic sequence from the database, which greatly reduces the retrieval time

(3) We propose a binocular vision localization algorithm based on image semantic target to solve the problem of mutual location among users. When we are sure that users can see the same scene, we select a semantic target from this scene as the center to establish the coordinate system. At this time, each user has their own corresponding coordinate in this coordinate system, and they establish position correlation with the selected target. We can identify the pixel coordinates of the selected target corners in the image by using the semantic segmentation network. By using the difference between the left and right images of the same corner in the binocular camera, we can solve its depth information and then solve the 3D coordinates of the user relative to the target. Finally, when each user has figured out their coordinates relative to the same position, they can realize mutual positioning by sharing coordinate information

The rest of the paper is organized as follows. Section 2 formulates the problems of mutual positioning and illustrates the system model for the proposed scheme; Section 3 gives a detailed introduction to the algorithm proposed in this paper based on the user mutual positioning problem; Section 4 evaluates the system performance of the proposed method by experimental simulation; in Section 5, we present a brief conclusion in which we summarize key contributions of our work.

2. Problem Formulation

2.1. Problem Statement. This paper mainly studies the mutual positioning method based on unknown environment. Figure 1 shows the mutual positioning of two users as an example. Initially, user 1 and user 2 are in two different positions in an unknown indoor environment. Besides, they do not understand the indoor environment and cannot see each other due to obstructions. Therefore, at this time, they cannot determine each other's position.

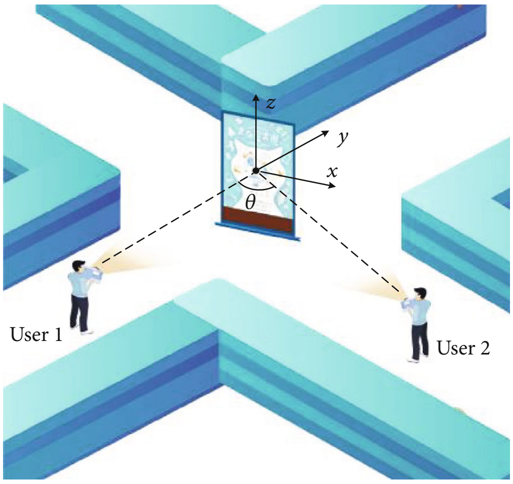


Figure 1: Mutual positioning in an unknown environment.

In order to determine the position of each other, each user should share information about the scene he can see whenever he walks a certain distance. According to the scene information that users can obtain, we can judge whether they can see the same scene and they will conduct mutual positioning in this scene when it is determined that they can see the same scene. When performing mutual positioning, we choose a target in this scene as the center for the establishment of a three-dimensional coordinate system; then, two users have their own corresponding position coordinates in this coordinate system, and they can finally realize mutual positioning through position information sharing.

2.2. System Model. Since both users are unknown to the current indoor environment, they need to judge their current position based on the semantic target category contained in the scene image they can currently see. If the scenes that two users have, respectively, seen contain the same kind and number of semantic target, it means that they both have seen this scene, so they can perform mutual positioning in this scene. Therefore, every time a user shoots the current scene, this user needs to use the R-FCN semantic segmentation network to identify the semantic targets contained in the current scene and generate a corresponding semantic sequence. Two users upload the generated semantic sequence to the shared database through the shared network each time and search in the database whether another user has also uploaded the same semantic sequence. If the retrieval is successful, it means that another user has also seen this scene, so they can conduct mutual positioning in this scene. If the retrieval fails, it means that they are currently far apart and they need to continue walking along the current track until the retrieval is successful.

When performing mutual positioning, it is necessary to ensure that two users are in the same coordinate system. Therefore, after the semantic targets in the user images are identified, we select a target in both user images and take this target as the center for the establishment of three-dimensional coordinate system. At this time, both users are in this coordinate system and realize mutual positioning by solving and sharing position information. Therefore, the mutual positioning method proposed in this paper mainly includes two modules: image searching and matching module and mutual positioning module. The flowchart is shown in Figure 2.

As can be seen from Figure 2, during mutual positioning, user 1 and user 2 first take pictures of the indoor environment in front of them, then they put the images into the R-FCN semantic segmentation network which is used to identify the semantic targets contained in the user images, and finally generate the semantic sequence corresponding to the images.

In the image searching and matching module, user 1 and user 2 upload the semantic sequence generated by the currently photographed picture to the shared database through a shared network such as ad hoc network or Wi-Fi network. Every time a user uploads a semantic sequence to a shared database, we record the time that user uploads it and retrieve the semantic sequence with those semantic sequences already stored in the shared database to find whether there is a similar semantic sequence that satisfies our requirements. If the retrieval is successful, it means that another user has seen the same scene, and then, we can find the corresponding images taken by users using binocular cameras according to the uploading time of the selected semantic sequence. Therefore, we determine the scene to be positioned according to the image, and then, two users can determine the position of each other based on this scene.

In the mutual positioning module, in order to establish the position connection between the two users, it is necessary to select a target seen by both users and with abundant feature points as the center for the establishment of coordinate system in this scene. Both users need to restore their position coordinates relative to the target according to the images they have taken. Therefore, users first use the semantic segmentation network to calculate the pixel coordinates of the corner points of the selected target in the image. Then, they use the difference between the pixel coordinates of the corner points in the left and right images to solve the depth information and solve the position coordinates of the user relative to the target and the turning angle. Finally, they realize mutual positioning by sharing position information.

3. Proposed Method

3.1. Principle of Image Semantic Segmentation. The semantic segmentation network used in this paper is R-FCN [30, 31], which is a two-stage target detection model. R-FCN developed from Faster R-CNN, followed the idea of fully convolutional network (FCN), and solved the contradiction between location insensitivity of classification network and location sensitivity of detection network. This network is composed of the fully convolutional network (FCN), the region proposal network (RPN), and the region of interest (ROI) subnetwork. FCN is used for feature extraction of input original image to generate feature map, RPN generates regions of interest (ROI) [32, 33] according to the extracted features, and ROI subnet locates and classifies target areas according to features extracted by FCN and ROI output by RPN. R-FCN first uses FCN to convert the original image into a corresponding feature map and then uses RPN to filter the foreground information on the feature map and frame the area that belongs to the object. At present, it is only a dichotomy operation, which can only determine whether the region belongs to foreground or background, but cannot know its specific classification information. Finally, the specific classification and location of the target are realized through the position-sensitive score chart. The structure of R-FCN model is shown in Figure 3.

It can be seen from Figure 3 that when a user image is input into the R-FCN semantic segmentation network, this network will detect the semantic information and record the types of semantic objects contained in the user image and finally generate a semantic sequence corresponding to the image.

The workflow of R-FCN is shown in Figure 4; the deep residual network performs a full convolution operation on the original user image and obtains a corresponding $W \times H \times 1024$-dimensional feature map, where W and H, respectively, represent the width and height of the feature map, and they are the result of the actual input image being reduced according to a certain ratio.

ROI subnet and RPN are the results obtained by convolving the feature map output by ResNet-50 again; ROI subnet uses $k \times k \times (c+1)$ convolution kernels, where k represents the number of equal divisions of the rectangular frame of the candidate area in the length and width directions. Generally, $k = 3$, that is, each ROI is divided into 9 equal parts. c represents the final number of categories; there are a total of $(c+1)$ categories because of some background information. The ROI subnet convolves the $W \times H \times 1024$-dimensional feature map output by ResNet-50 to generate a new $W \times H \times 9(c+1)$-dimensional feature map, which is called a position-sensitive score map. The position-sensitive score map has $(c+1)$ layers; each layer corresponds to a category, and for each layer, the ROI obtained by RPN is divided into 9 subregions [34]. The meaning of the division is that the ROI should contain all parts of category $c_i (i = 1, 2, \cdots, c+1)$ in each region, and when all subregions have high response values to the corresponding parts of a target, then the classifier will judge the ROI as this category. Each part of the target and each subregion of ROI have a one-to-one mapping relationship.

The ROI area extracted by RPN contains the four attributes of horizontal and vertical coordinates, length, and width, which means that different ROI areas can correspond to different positions on the score map. In addition, each ROI is divided into 9 subareas, each of which contains multiple position-sensitive score values. Because too much data will interfere with subsequent classification operations, data needs to be compressed by pooling operations. For each

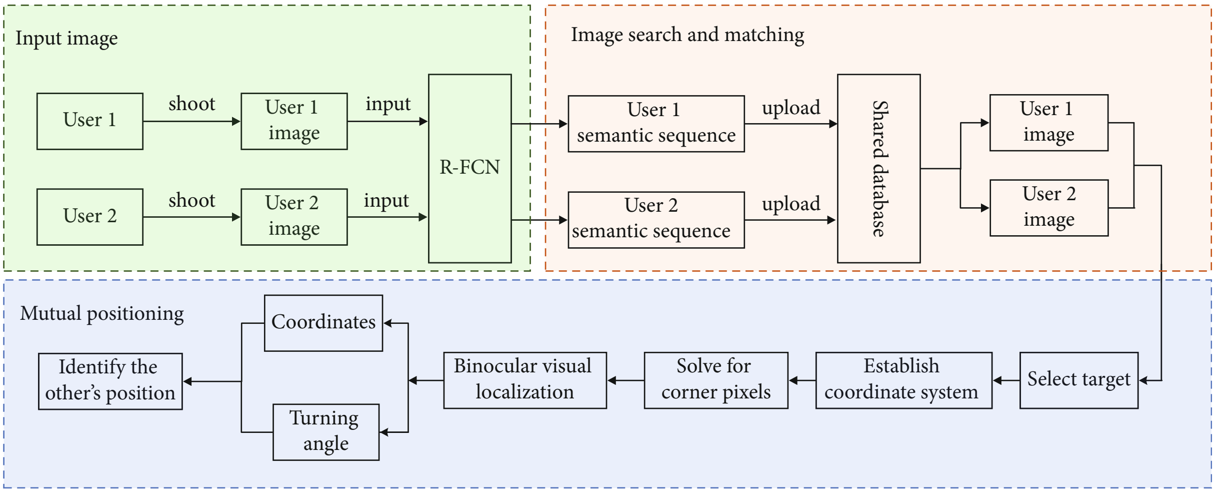


Figure 2: Process of indoor positioning method based on unknown environment.

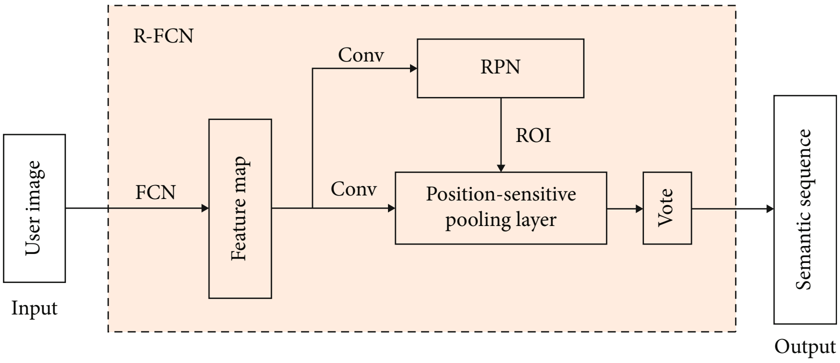


Figure 3: R-FCN model structure.

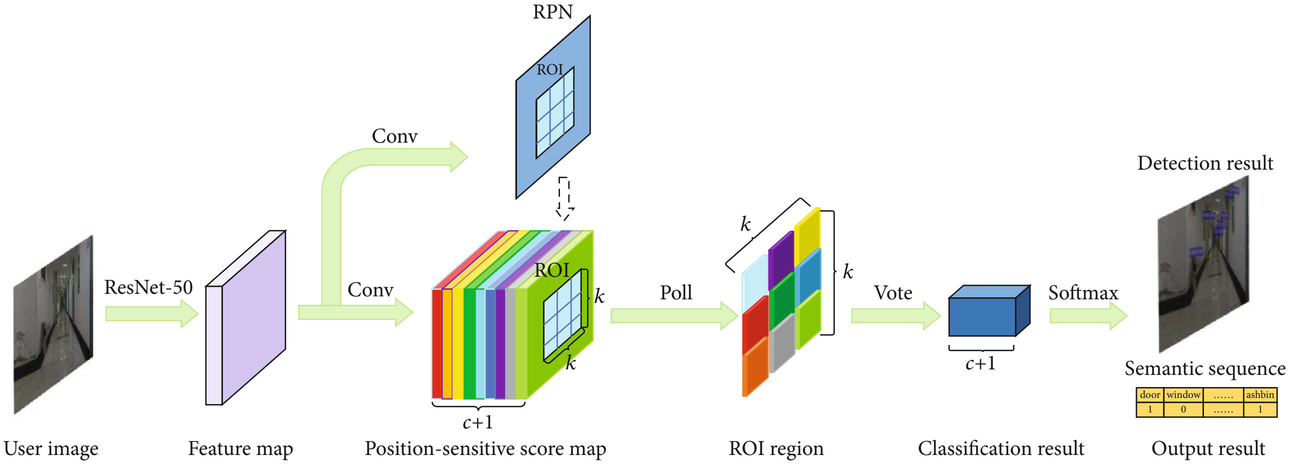


Figure 4: R-FCN workflow.

subregion $\mathrm{bin}(i, j), 0 \le i, j \le k-1$, the following pooling operation is performed:

$$r_c(i, j \mid \Theta) = \sum_{(x,y)\in \mathrm{bin}(i,j)} \frac{z_{i,j,c}(x + x_0, y + y_0 \mid \Theta)}{n}, \qquad (1)$$

where $r_c(i, j \mid \Theta)$ is the pooling response of subregion $\mathrm{bin}(i, j)$ to c categories, $z_{i,j,c}$ is the position-sensitive fraction graph corresponding to subregion $\mathrm{bin}(i, j)$, (x_0, y_0) represents the pixel coordinates in the upper left corner of the target candidate box, x and y are the offset of the current pixel coordinate from the upper left pixel coordinate, Θ represents all the learning parameters of the network, and n is the number of pixels in subregion.

After the pooling operation, the 9 subregions have become 9 position-sensitive scores, which, respectively, represent the scores of the location corresponding to the 9 spatial orientations of the category. Then, the scores of these 9 subregions can be summed to get the ROI belonging to the category score. Finally, for the $(c+1)$ categories, the output of the pooling layer is summed according to the dimensions to obtain a $(c+1)$-dimensional vector:

$$r_c(\Theta) = \sum_{i,j} r_c(i, j \mid \Theta). \tag{2}$$

This vector is then substituted into the Softmax equation to obtain the probability that the target belongs to each category using the Softmax regression class method:

$$s_c(\Theta) = \frac{e^{r_c(\Theta)}}{\sum_{c'=0}^{c} e^{r_{c'}(\Theta)}}. \tag{3}$$

After calculating all the probabilities, R-FCN classifies each ROI according to the principle of maximum probability, and finally, we can know the category information of each ROI framed target. In order to determine the accuracy and optimal number of iterations during network training, the relevant loss function needs to be set. When the final training output value of the loss function is less than the threshold specified in advance, it indicates that the network training result is better. The loss function of the R-FCN network uses a multiobjective loss function, while considering the loss of classification and the loss of location. Therefore, the equation can be derived:

$$L\left(s, t_{x,y,w,h}\right) = L_{\text{cis}}(S_{c*0}) + \lambda[c*>0]L_{\text{reg}}(t, t*), \tag{4}$$

where $c*$ stands for ground truth, L_{cis} represents the loss of classification cross entropy, L_{reg} represents the loss of position, and $t*$ represents the location of ground truth. $[c*>0]$ means that if the classification is correct, its value is 1; if the classification is wrong, its value is 0, that is, no position loss is performed for the wrong classification. λ represents the super parameter, and if its value is 1, it means that the classification loss is as important as the position loss. During the training, if the final loss function is less than the specified threshold or the number of iterations reaches the upper limit, the training will be stopped. At this time, all parameters in the R-FCN model have been adjusted to appropriate values, which can be used for target detection and classification operations.

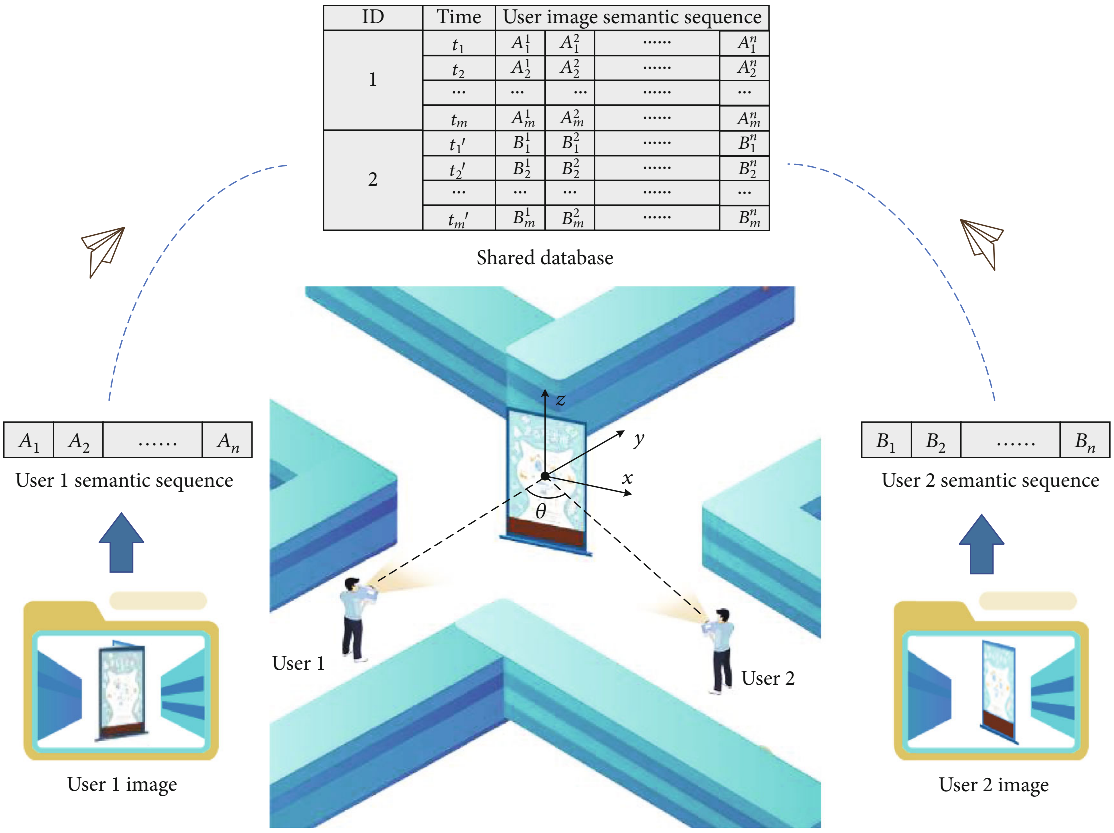
3.2. Fast Matching Retrieval between User Images. The precondition for two users to conduct mutual positioning is to ensure they can see the same scene, so each user should share the scene he can see with the other at regular intervals. However, image contains a large amount of data due to its rich information content. If user image retrieval is based on the image database, it will not only take up a large storage capacity of the database but also take a long time to match the feature points of the image. Therefore, this paper proposes an image semantic sequence retrieval method based on shared database. When two users conduct mutual positioning, they first shoot the scene in front of them and then use semantic segmentation network to identify the semantic target of the image they shoot and generate the corresponding semantic sequence of the current user image. The generated sequence is uploaded to the shared database through shared network, such as ad hoc network and Wi-Fi network. Finally, we retrieve the semantic sequence uploaded by users and the sequences stored in the shared database. If we find that another user has uploaded the same semantic sequence, it means that they have seen the same scene. Then, we conduct mutual positioning between them based on this scene. If the retrieval fails, it means they have not seen the same scene, so they should continue to walk along the current track until the retrieval is successful. The process of image information interaction between users is shown in Figure 5.

Assuming that the trained R-FCN semantic segmentation network can recognize n types of semantic targets, for any user image I, the following semantic sequence format can be generated.

In Figure 6, C_i represents the name of the identified target, and A_i represents the number of times that the corresponding category appears in the user image. Each category occupies a bit in the semantic sequence, and they are arranged in order from high to low according to the richness of feature points contained in the target. For example, in an indoor environment, the poster on the wall contains the most feature points, and then, the category corresponding to the highest bit in the semantic sequence is the poster; the door contains the least feature points, so the category corresponding to the lowest bit in semantic sequence is the door. Therefore, although each user image contains different semantic information, the resulting semantic sequence format is the same.

When users share the image information they can see, both users need to put the scene images they have currently taken into the R-FCN semantic segmentation network for semantic target recognition at regular intervals and then upload the semantic sequence to the shared database through the sharing network. The shared database stores the semantic sequences transmitted by users and records the time when database receives the sequences. The data format stored in the shared database is shown in Figure 7.

It can be seen from Figure 7 that the shared database will identify the user's ID according to the IP address of the sender and store the received data in the data storage space of the corresponding user. t_i in the figure represents the time when the shared database receive the semantic sequence uploaded by the user. It should be noted that the time when user 1 and user 2 upload the semantic sequence does not need to be strictly consistent, and the time interval between the upload of the semantic sequence can also be not the same. A_i^j represents the j bit of the semantic sequence uploaded by user 1 for the i time, which stores the number of times that the semantic category corresponding to the bit appears in



ID	Time	User image semantic sequence			
1	t_1	A_1^1	A_1^2	......	A_1^n
	t_2	A_2^1	A_2^2	......	A_2^n
	...	...	...	......	...
	t_m	A_m^1	A_m^2	......	A_m^n
2	t_1'	B_1^1	B_1^2	......	B_1^n
	t_2'	B_2^1	B_2^2	......	B_2^n
	...	...	...	......	...
	t_m'	B_m^1	B_m^2	......	B_m^n

Figure 5: The process of image information interaction between users.

C_1	C_2	......	C_n
A_1	A_2	......	A_n

Figure 6: Semantic sequence format.

User ID	Time	User image semantic sequence			
1	t_1	A_1^1	A_1^2	......	A_1^n
	t_2	A_2^1	A_2^2	......	A_2^n
	...	...	...	......	...
	t_m	A_m^1	A_m^2	......	A_m^n
2	t_1'	B_1^1	B_1^2	......	B_1^n
	t_2'	B_2^1	B_2^2	......	B_2^n
	...	...	...	......	...
	t_m'	B_m^1	B_m^2	......	B_m^n

Figure 7: The format of the data stored in the shared database.

the image captured by user 1 at the current time. B_i^j represents the information uploaded by user 2, which has the same meaning as A_i^j.

Whenever a user uploads a new semantic sequence, we will retrieve it with all the semantic sequences uploaded by another user. The criteria for determining that the user images corresponding to two semantic sequences are taken in the same scene during retrieval are as follows.

Assume that the corresponding semantic sequence of user 1 image is $\mathrm{Sem}_1 = [a_1, a_2, \cdots, a_n]$ and the corresponding semantic sequence of user 2 image is $\mathrm{Sem}_2 = [b_1, b_2, \cdots, b_n]$. When the two semantic sequences meet,

$$\sum_{i=1}^{n} |a_i - b_i| \leq D, \tag{5}$$

where D is the distance threshold, reflecting the similarity degree of the two semantic sequences. It is believed that the user images corresponding to the current two semantic sequences are shot in the same scene, and two users can conduct mutual positioning in this scene. Through the experiment, it can be found that the smaller the threshold D, the higher the matching accuracy of the user's image. In order to consider the difference between two users when observing the same scene from different perspectives, there may be a problem that one user can see the target and the other user cannot see it due to the obstruction of obstacles. Therefore, the distance threshold D is set as 2 in this paper.

Therefore, if user 1 uploads a semantic sequence of an image to the shared database, we find that user 2 has previously transmitted a semantic sequence that meets the distance threshold D requirement. At this time, we return the time that user 2 passed in this semantic sequence to him, and then, user 2 can find the scene image uploaded at that moment. Then, two users can conduct mutual positioning based on the photos taken in the scene. If the retrieval fails, it means that they have not seen the same scene at present, and they need to continue walking along the current path until the retrieval is successful to start mutual positioning.

3.3. Binocular Visual Localization Based on Semantic Target.

When it is determined that users can see the same scene, they can locate each other based on this scene. First of all, each user will obtain the upload time of the semantic sequence that meets the threshold requirements if the image retrieval is successful, and they can find the corresponding left and right photos taken by binocular camera according to the upload time of the semantic sequence. Then, users choose a target in the corresponding scene as the positioning baseline according to the photos and establish their location connection with the target as the center. Finally, using the semantic segmentation network, each user can know the pixel coordinates of the corner points of the selected semantic target in the left and right images, and they can use the binocular ranging algorithm to solve the coordinates of the target relative to their own and the turning angle.

For the binocular vision system, it is assumed that the coordinates of a point $P(X, Y, Z)$ in the world coordinate system in the left camera coordinate system are expressed as $(U_l, V_l, W_l)^{\mathrm{T}}$, and the coordinates in the right camera coordinate system are expressed as $(U_r, V_r, W_r)^{\mathrm{T}}$; then, according to the conversion relationship between the world coordinate system and the camera coordinate system, the coordinate mapping relationship between the left and right camera coordinate systems can be established, and the expression is shown as

$$\begin{bmatrix} U_r \\ V_r \\ W_r \\ 1 \end{bmatrix} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}_3^{\mathrm{T}} & 1 \end{bmatrix} \begin{bmatrix} U_l \\ V_l \\ W_l \\ 1 \end{bmatrix}, \tag{6}$$

where $\mathbf{R}$ is the rotation matrix between the left and right cameras, which is a 3×3dimensional matrix, and $\mathbf{t}$ is the translation vector between the left and right cameras, which is a 3-dimensional vector. The two jointly constitute the external parameters of the binocular camera, and the respective expressions of the two parameters are shown as

$$\mathbf{R} = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \\ r_7 & r_8 & r_9 \end{bmatrix}, \qquad \mathbf{t} = \begin{bmatrix} t_x & t_y & t_z \end{bmatrix}^{\mathrm{T}}. \tag{7}$$

At this time, we substitute the coordinate $(U_r, V_r, W_r)^{\mathrm{T}}$ of the point P in the right camera coordinate system obtained by Equation (6) into the projection relationship between the right camera coordinate system and the image coordinate system, and the following relationship can be obtained by using the mapping relation between the coordinates in the camera coordinate system and the pixel coordinate system:

$$\rho_r \begin{bmatrix} x_r \\ y_r \\ 1 \end{bmatrix} = \begin{bmatrix} f_r r_1 & f_r r_2 & f_r r_3 & f_r t_x \\ f_r r_4 & f_r r_5 & f_r r_6 & f_r t_y \\ r_7 & r_8 & r_9 & t_z \end{bmatrix} \begin{bmatrix} W_l \dfrac{u_l}{f_l} \\ W_l \dfrac{v_l}{f_l} \\ W_l \\ 1 \end{bmatrix}, \tag{8}$$

where ρ_r is the proportionality coefficient, (x_r, y_r) is the coordinate of point P mapped to the image coordinate system of the right camera, and (u_l, v_l) is the coordinate of point P mapped to the pixel coordinate system of the left camera. We can eliminate ρ_r by combining the equations in the first and third lines of Equation (8) so that the coordinate expression of the point in the left camera coordinate system can be solved:

$$\begin{cases} U_l = W_l \dfrac{u_l}{f_l}, \\ V_l = W_l \dfrac{v_l}{f_l}, \\ W_l = \dfrac{f_l(f_r t_x - x_r t_z)}{x_r(r_7 u_l + r_8 v_l + r_9 f_l) - f_r(r_1 u_l + r_2 v_l + r_3 f_l)}. \end{cases} \tag{9}$$

It can be seen from Equation (9) that the semantic segmentation network can identify the pixel coordinate (u_1^l, v_1^l) of the upper left corner point of the selected target in the left image and the pixel coordinate (u_1^r, v_1^r) of the upper left corner point of the target in the right image. Then, according to the zoom and translation relationship between the pixel coordinate system and the image coordinate system, we can calculate the coordinate (x_1^l, y_1^l) of the upper left corner of the target in the image coordinate system in the left image and the coordinate (x_1^r, y_1^r) of the upper left corner of the target in the image coordinate system in the right image; we put this parameter into Equation (9) to solve the coordinate (U_l^1, V_l^1, W_l^1) of the upper left corner of the selected target in the left camera coordinate system, which completes the binocular ranging work.

Therefore, next we need to solve the turning angle between the left camera coordinate system and the reference coordinate system established with the selected target. Both the reference coordinate system $O_r - XYZ$ and the left camera coordinate system $O_{cl}\text{-}UVW$ meet the criteria for establishing the left-handed coordinate system, and we stipulate that the clockwise direction is the positive direction of rotation. At the same time, we ensure that the $X - Y$ plane of the reference coordinate system and the $U - W$ plane of the left camera coordinate system are parallel to the ground.

Therefore, according to the theory of coordinate system conversion relations, the conversion of any point $(U, V, W)^{\mathrm{T}}$ in the left camera coordinate system to a point in the reference coordinate system can be expressed as the following form:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \mathbf{R}_r \begin{bmatrix} U - U_0 \\ W - W_0 \\ V - V_0 \end{bmatrix}. \tag{10}$$

It can be seen from Equation (10) that the abovementioned coordinate conversion can be divided into two steps. The first step is the translation of the coordinate system. We shift the origin of the left camera coordinate system O_{cl} to the origin of the reference coordinate system O_r and use the translation vector $\mathbf{t}_r = (U_0, W_0, V_0)^{\mathrm{T}}$ to represent the translation relationship between the two. This vector is also the position coordinate of the upper left corner of the target in the left camera coordinate system. The second step is the rotation of the coordinate system. We rotate plane $U - W$ in the left camera coordinate system by θ degrees clockwise about axis V to plane $X - Y$ in the reference coordinate system. The rotation relationship between the two coordinate systems can be represented by a rotation matrix $\mathbf{R}_r$, whose expression is shown as

$$\mathbf{R}_r = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{11}$$

where θ is the user's turning angle to be solved. Now, we select the upper edge of the target to select any point Q, whose coordinate in the reference coordinate system is $(0, Y_Q, 0)$ and whose coordinate in the left camera coordinate system is (U_Q, W_Q, V_Q). Then, according to Equation (10), the following corresponding relation can be obtained:

$$\begin{bmatrix} 0 \\ Y_Q \\ 0 \end{bmatrix} = \mathbf{R}_r \begin{bmatrix} U_Q - U_0 \\ W_Q - W_0 \\ V_Q - V_0 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} U_Q - U_0 \\ W_Q - W_0 \\ V_Q - V_0 \end{bmatrix}. \tag{12}$$

Equation (12) contains three equations, and then, we can obtain the following relationship by combining these three equations simultaneously:

$$\begin{bmatrix} U_Q \\ W_Q \\ V_Q \end{bmatrix} = \begin{bmatrix} U_0 + Y_Q \sin\theta \\ W_0 + Y_Q \cos\theta \\ V_0 \end{bmatrix}, \tag{13}$$

where $(U_0, W_0, V_0)^{\mathrm{T}}$ is the coordinate of the upper left corner of the selected target in the left camera coordinate system.

The next step is to convert the Q point to the image coordinate system, which is a process from the three-dimensional coordinate system to the two-dimensional coordinate system. This process is consistent with the pinhole imaging model, and the depth information is lost in the conversion process. Now, assuming that the coordinate converted to the image coordinate system is (x_Q, y_Q), the following relation can be solved according to the transformation relationship between the camera coordinate system and the image coordinate system:

$$\begin{cases} x_Q = \dfrac{U_Q f}{W_Q}, \\ y_Q = \dfrac{V_Q f}{W_Q}, \end{cases} \tag{14}$$

where f is the focal length of the camera, which can be solved by the calibration of the camera's internal parameters. By substituting Equation (13) into Equation (14), we can obtain the equation of coordinate (x_Q, y_Q) in the image coordinate system:

$$x_Q - \frac{U_0 \cos\theta - W_0 \sin\theta}{V_0 \cos\theta} y_Q - f \tan\theta = 0. \tag{15}$$

It can be seen from Equation (15) that the coordinate Q $(0, Y_Q, 0)$ of any point on the upper edge of the selected target in the reference coordinate system has been converted to the coordinate (x_Q, y_Q) in the image coordinate system.

Through the semantic segmentation network, we can solve and get the pixel coordinates (u_1^l, v_1^l) and (u_2^l, v_2^l) of the upper left and right corner points of the selected target in the left image as well as the pixel coordinates (u_1^r, v_1^r) and (u_2^r, v_2^r) of the upper left and right corner points of the selected target in the right image. Thus, the linear equation of the upper edge of the selected target in the pixel coordinate system can be calculated. The coordinate transformation relationship between pixel coordinate system and image coordinate system is shown as

$$\begin{cases} u = \dfrac{x}{dx} + u_0, \\ v = \dfrac{y}{dy} + v_0. \end{cases} \tag{16}$$

Therefore, we convert the linear equation of the upper edge of the selected target in pixel coordinate system to the image coordinate system, which can be expressed in the following form:

$$x_Q + b y_Q + c = 0. \tag{17}$$

It can be seen from Equation (17) that the expression form of the line along the upper edge of the selected target in the image coordinate system is the same as that of Equation (15). Thus, the following relation can be obtained:

$$c = -f \tan\theta, \tag{18}$$

where the focal length f is known. Therefore, the value of user turning angle θ can be solved as

$$\theta = \arctan\left(-\frac{c}{f}\right). \tag{19}$$

Since the user's current position is the origin $O_{cl}(0, 0, 0)^{\mathrm{T}}$ in the left camera coordinate system, put it into Equation (10) together with the turning angle θ obtained by solving Equation (19), and the user's current position coordinates in the reference coordinate system can be obtained:

$$\begin{bmatrix} X_p \\ Y_p \\ Z_p \end{bmatrix} = -\mathbf{R}_r \begin{bmatrix} U_0 \\ W_0 \\ V_0 \end{bmatrix} = \begin{bmatrix} -U_0 \cos\theta + W_0 \sin\theta \\ -U_0 \sin\theta - W_0 \cos\theta \\ -V_0 \end{bmatrix}. \tag{20}$$

The above is the derivation of all the equations of binocular vision localization algorithm based on semantic objective proposed in this paper. In this algorithm, we firstly use semantic segmentation network to solve the pixel coordinates of each corner point of the selected semantic target in the left and right images. Then, based on these pixel coordinates, we use the coordinate system transformation relation derived above to solve the current user's position coordinates and turning angle relative to the selected target in the indoor scene. Finally, we complete the visual positioning service for each user.

After users have solved their own position coordinates in the coordinate system and the information of turning angle relative to the coordinate center, the position coordinates of other users relative to themselves can be solved through the sharing of the position information, thus completing the mutual positioning service among users. Taking two users as examples, it is assumed that the current position coordinate of user 1 relative to the selected target is $P_1(X_1, Y_1, Z_1)$ and the turning angle is θ_1. User 2 calculates that the current position coordinate relative to the selected target is $P_2(X_2, Y_2, Z_2)$ and the turning angle is θ_2. The schematic diagram is shown in Figure 8.

As can be seen from Figure 8, after two users have, respectively, determined the position coordinates in the current reference coordinate system and the turning angle relative to the selected target, the distance D between the two users can be expressed as

$$D = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2}. \tag{21}$$

Since when solving the turning angle information, we believe that when the user is on the right side of the selected target, its value is positive; otherwise, it is negative, so the angle θ between the two users relative to the selected target can be expressed as

$$\theta = |\theta_1 - \theta_2|. \tag{22}$$

Therefore, according to the above equation derivation, two users can figure out the distance and angle between them and determine the position of each other by sharing their own position information, thus realizing the mutual positioning service between users in an unknown indoor environment.

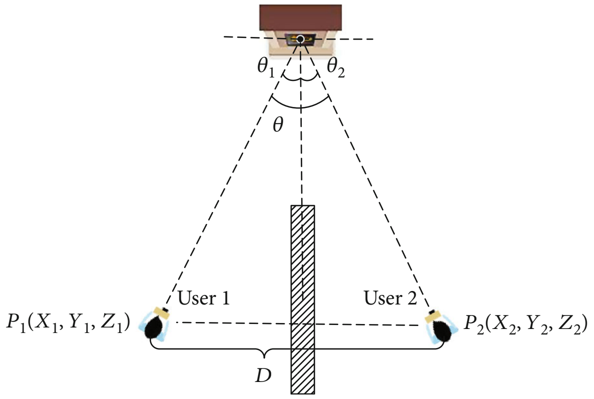


Figure 8: Schematic diagram of mutual positioning between users.

4. Implementation and Performance Analysis

4.1. Experiment Setup. In order to verify the feasibility of the method proposed in this paper, we need to select an experimental scenario for testing. The experimental environment in this paper is the corridor on the 12th floor of Building 2A of Harbin Institute of Technology Science Park. The experiment environment is shown in Figure 9.

It can be seen from the schematic diagram that the experimental scene contains multiple corners. When two users stand on both sides of the corner, they cannot see each other due to obstructions, but they can observe the same scene at the same time. Therefore, this scenario conforms to the required background conditions and is suitable for verifying the feasibility of the method proposed in this paper.

Therefore, after the indoor positioning scene model is established, we collect images in the scene by binocular camera and use the collected images to verify the effectiveness of the method proposed in this paper.

4.2. Experiment Results. Before positioning, it is necessary to accurately recognize the semantic information contained in the user's image, so as to judge whether two users can observe the same scene through the semantic sequence corresponding to the image. This paper uses R-FCN for image semantic segmentation, so R-FCN needs to be trained. This paper takes pictures every 0.5 meters when collecting images in the experimental scene. In addition, for the semantics of doors and windows that have multiple states (open, closed), it is necessary to take images of different states so that the semantics can be accurately recognized. At the same time, we also need to add some training sets with different indoor light brightness to ensure that we will not be affected by the light when we perform target recognition.

After the images are taken and formed into a data set, each picture in the training data set needs to be semantically annotated. This paper divides the semantics of the corridor

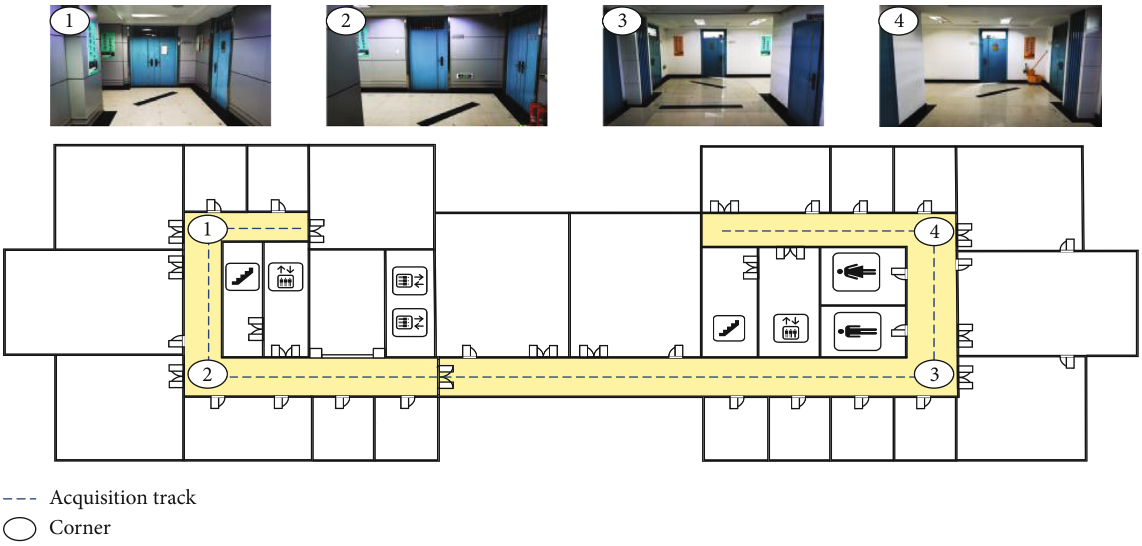
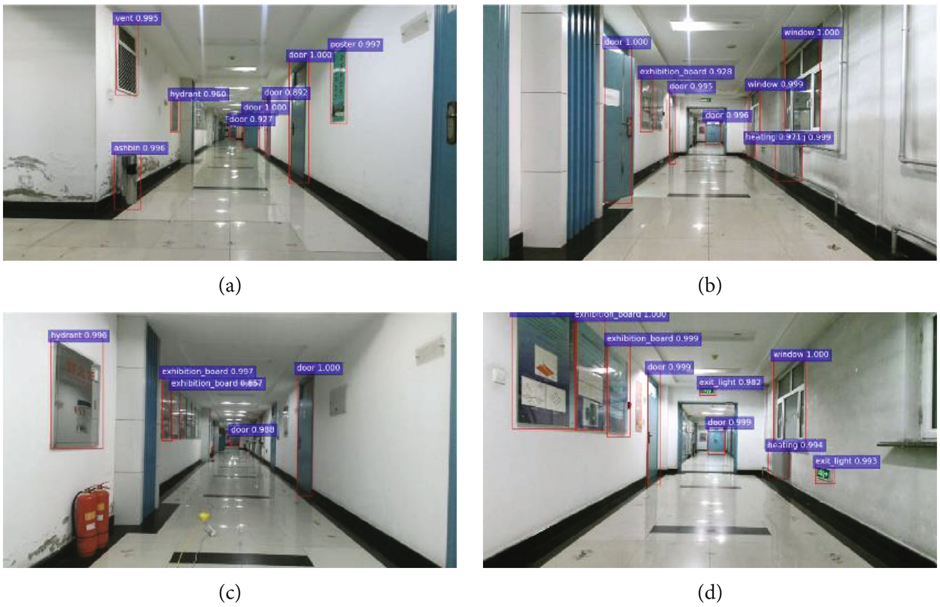


Figure 9: Schematic diagram of experimental scene.



(a)

(b)

(c)

(d)

Figure 10: Semantic target detection results in user images. (a) Corner of corridor. (b) Front half of corridor. (c) Back half of corridor. (d) Middle part of corridor.

into 10 categories, namely, door, window, heating, hydrant, ashbin, vent, poster, exhibition board, exit light, and background categories. When all the images are marked, they are put into the network model for training. When the network training is completed, in order to verify the accuracy of R-FCN in image semantic segmentation, it is necessary to shoot several test images to verify the accuracy of R-FCN target recognition. The output result of user images after R-FCN is shown in Figure 10.

It can be seen from Figure 10 that the red box is the semantic target identified by R-FCN, and the corresponding blue box above indicates the category to which the semantic target belongs and the probability of belonging to this semantic category. We can see that most of the semantic targets can

Table 1: The recognition accuracy of R-FCN for different targets.

Semantic target	Correct number	Wrong number	Unrecognized number	Accuracy
Door	724	7	6	98.23%
Exhibition board	678	5	7	98.26%
Poster	375	3	4	98.17%
Window	297	4	3	97.70%
Heating	266	0	2	99.25%
Hydrant	102	2	2	96.23%
Exit light	86	0	3	96.63%
Vent	47	2	0	95.92%
Ashbin	39	0	1	97.50%

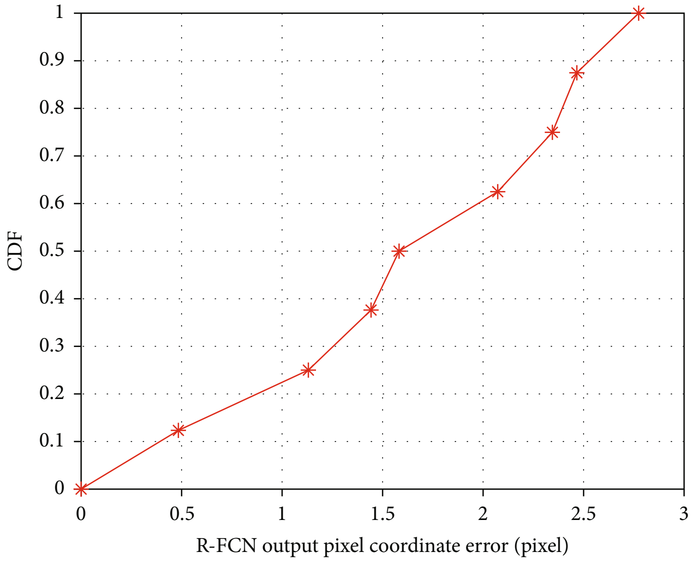


Figure 11: The CDF curve of pixel coordinate error output by R-FCN.

be correctly identified, indicating that this network has good target recognition capabilities, so we can use this network model to identify semantic targets in user images. In order to verify the accuracy of R-FCN recognition for each category, this paper performs semantic recognition on a large number of test images. The final results are shown in Table 1.

From the statistical data in Table 1, we can see that the recognition accuracy of the 9 types of semantic targets by R-FCN is above 95%, indicating that the trained network model has high recognition accuracy. Therefore, this paper can use this network model to identify semantic targets in user images. The data sets adopted in model training in this paper were all collected in experimental scenes. In order to improve the application scope of the method proposed in this paper, we will expand the types and numbers of training sets in the future so that the trained network model can identify common semantic targets such as doors and windows in any experimental environment. When the semantic segmentation network recognizes the semantic target in the user image, it will also give the pixel coordinates of the upper left corner and the lower right corner of each semantic target. In order to verify the accuracy of the pixel value corresponding to the semantic target output by R-FCN, we select 50 images for verification and compare the pixel value of the corresponding corner point of each semantic target output with the real pixel value in the image. The final result is shown in Figure 11.

In the process of determining their own position, users need to convert the indoor scene they see at the current moment into image information to share with other users. In this paper, an image data retrieval method based on image semantic sequence is proposed. Every time a user takes a scene image, he will use the trained semantic segmentation network to identify the semantic target contained in the image and generate the corresponding semantic sequence, which will be uploaded to the shared database. Through the retrieval and matching of semantic sequences in the shared database, we can judge whether the same scene can be seen by other users. Compared with traditional image retrieval methods, the method proposed in this paper not only reduces the time of image information uploading and retrieval but also reduces the storage capacity of the database. The performance comparison of the two methods is shown in Table 2.

Table 2: Performance comparison of the two upload methods.

Upload method	Upload time (ms)	Storage capacity (B)	Retrieval time (ms)	Precision
User image	1580	152K	6584	98%
Semantic sequence	0.06	10	0.24	96%

In the experimental test, in order to ensure that the performance comparison of the two methods is more convincing, this paper uses the same method to upload the two data formats; we both use the UDP transmission protocol to transmit semantic sequences and user images through the ad hoc network. As can be seen from Table 2, compared with uploading user images, the method of uploading semantic sequences consumes less time in data uploading and data retrieval and occupies less storage capacity in the database, but is slightly worse in retrieval accuracy. However, the method proposed in this paper still meets the requirements for accuracy of image information retrieval. Therefore, after weighing the performance of image information uploading and image information retrieval, it is concluded that the image retrieval method based on semantic sequence proposed in this paper is better than the image retrieval method based on user image.

In order to verify whether the mutual positioning method proposed in this paper has good positioning accuracy, we conduct a positioning test in the experimental scene. We collect a total of 50 pairs of user images of the same scene taken by two users from different angles and analyze the final results of mutual positioning based on user images. We verify

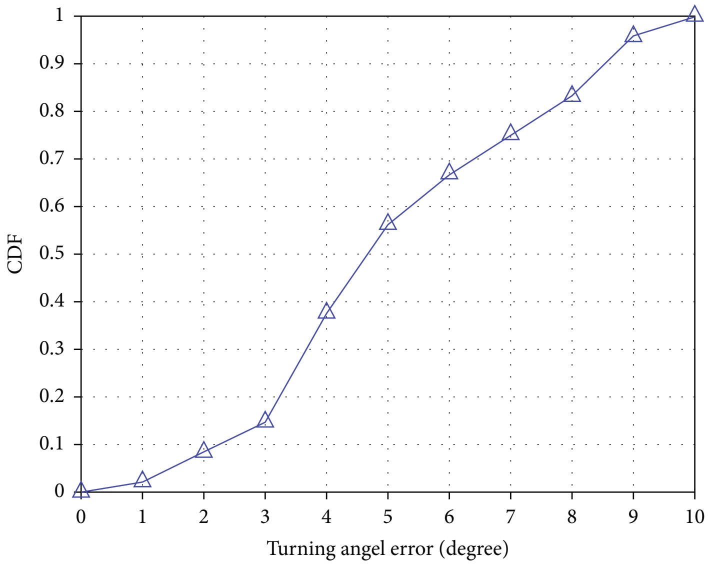


FIGURE 12: The CDF curve of user turning angle error.

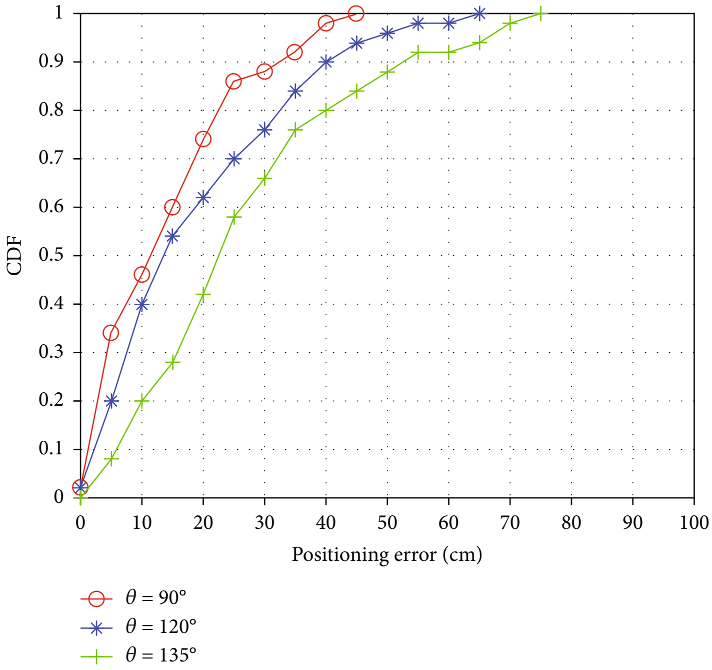


FIGURE 13: The CDF curve of mutual positioning error.

the solving accuracy of the user's turning angle relative to the selected target, and the simulation result is shown in Figure 12.

It can be seen from Figure 12 that the error generated in this paper when calculating the user's turning angle has a 90% probability within ±9°, and the maximum error is about 10°. Considering that the experimental equipment also produces some small errors during measurement, the above-mentioned error is very small and has little effect on the positioning results. Therefore, the calculation results of the turning angle meet the accuracy requirements. Users do not need to fix the shooting angle. This is because the proposed algorithm in this paper can restore the current user's rotation angle relative to the selected target through the user image. The following verifies the accuracy of the algorithm proposed in this paper to calculate the distance between users using the semantic targets in the scene. The simulation result is shown in Figure 13.

In the positioning test, it is necessary to ensure that the angle between the users and the target is an obtuse angle, so as to reflect the application value of the mutual positioning method proposed in this paper in real scenarios. In addition, due to the limitation of the experimental scene, this paper sets the angle as 90°, 120°, and 135°, respectively. It can be seen from Figure 13 that when the angle θ between the two users and the target is smaller, the mutual positioning accuracy is higher. This is because the more similar the viewing angles of the two users are, the better the matching effect of the user images will be, and the higher the final mutual positioning accuracy will be. It can be seen from the figure that the positioning error of the mutual positioning method proposed in this paper can be controlled within 50 cm with 90% probability, which has a good positioning effect.

5. Conclusions

This paper proposes a mutual positioning method between users in an unknown indoor environment. In terms of image information sharing between users, the image semantic sequence retrieval algorithm based on shared database proposed in this paper can greatly reduce the time of image information uploading and matching. In addition, this algorithm can also reduce the storage capacity of the database while ensuring retrieval accuracy. In terms of positioning between users, the binocular vision positioning algorithm based on semantic target proposed in this paper can calculate the user's position coordinates and turning angle information relative to the selected target, and mutual positioning between users can be achieved through position sharing. Through the analysis of simulation results on the 12th floor of Building 2A of Harbin Institute of Technology Science Park, we conclude that the proposed method can realize the mutual positioning between users in unknown indoor environment, and the positioning error can be within 1 m, which can provide users with an accurate positioning service.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Disclosure

The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Conflicts of Interest

The authors declare no conflict of interest.

Authors' Contributions

Lin Ma provided the conception; He Dong made the analysis and experiment; Bin Wang reviewed and edited this paper.

Acknowledgments

This research was funded by the National Natural Science Foundation of China (61971162 and 41861134010) and Foundation of Science and Technology on Communication Networks Laboratory (HHX20641X002).

References

[1] P. Davidson and R. Piché, "A survey of selected indoor positioning methods for smartphones," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 1347–1370, 2017.

[2] Y. Cheng and T. Zhou, "UWB indoor positioning algorithm based on TDOA technology," in *10th International Conference on Information Technology in Medicine and Education (ITME)*, pp. 777–782, Qingdao, China, 2019.

[3] J. Wang, X. Yang, and Z. Luo, "An indoor Wi-Fi positioning approach optimized by virtual node," in *5th International Conference on Computer and Communication Systems (ICCCS)*, pp. 609–612, Shanghai, China, 2020.

[4] Y. Zhao, Z. Yang, C. Song, and D. Xiong, "Vehicle dynamic model-based integrated navigation system for land vehicles," in *25th Saint Petersburg International Conference on Integrated Navigation Systems (ICINS)*, pp. 1–4, St. Petersburg, Russia, July 2018.

[5] M. Liu, P. Wang, J. Guo, Y. Niu, T. Shi, and C. Wang, "Research on geomagnetic navigation and positioning algorithm based on full-connected constraints for AUV," in *OCEANS-Marseille*, pp. 1–5, Marseille, France, June 2019.

[6] A. Xiao, R. Chen, D. Wu, and Y. Chen, "Positioning in large indoor spaces using smartphone camera based on static objects," in *Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS)*, pp. 1–7, Wuhan, 2018.

[7] Y. Xia, C. Xiu, and D. Yang, "Visual indoor positioning method using image database," in *Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS)*, pp. 1–8, Wuhan, 2018.

[8] F. Xiao, "Indoor robot visual positioning system based on floor features," in *3rd International Conference on Robotics and Automation Engineering (ICRAE)*, pp. 97–101, Guangzhou, 2018.

[9] R. S. Kulikov, A. A. Chugunov, D. V. Tsaregorodcev, N. I. Petukhov, and A. P. Malyshev, "Investigating of the accuracy of vehicles mutual positioning using smartphones," in *International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE)*, pp. 1–4, Moscow, Russia, 2020.

[10] R. Kulikov, A. Chugunov, A. Sizyakova, and E. Zakharova, "Relative mutual positioning using smartphones," in *International Conference on Engineering and Telecommunication (EnT)*, pp. 1–4, Dolgoprudny, Russia, 2019.

[11] A. B. Silantyev, D. S. Tereshchenko, L. N. Kazakov, and E. A. Selyanskaya, "Adaptive system of mutual positioning for controlling the groups of unmanned aerial vehicles," in *Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO)*, pp. 1–6, Minsk, 2018.

[12] Y. Duan, X. Tao, C. Han, and J. Lu, "Semantic conditional random field for object based SAR image segmentation," in *25th IEEE International Conference on Image Processing (ICIP)*, pp. 2625–2629, Athens, 2018.

[13] M. Akbari, J. Liang, and J. Han, "DSSLIC: deep semantic segmentation-based layered image compression," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2042–2046, Brighton, United Kingdom, 2019.

[14] Y. Xing, J. Wang, X. Chen, and G. Zeng, "Coupling two-stream RGB-D semantic segmentation network by idempotent mappings," in *IEEE international conference on image processing (ICIP)*, pp. 1850–1854, Taipei, Taiwan, 2019.

[15] Y. Zhang, J. Yang, H. Zhang, and J. N. Hwang, "Bundle adjustment for monocular visual odometry based on detected traffic sign features," in *IEEE International Conference on Image Processing (ICIP)*, pp. 4350–4354, Taipei, Taiwan, 2019.

[16] W. Zhang, G. Liu, and G. Tian, "A coarse to fine indoor visual localization method using environmental semantic information," *IEEE Access*, vol. 7, pp. 21963–21970, 2019.

[17] A. Albrecht and N. F. Heide, "Improving feature-based visual SLAM in person indoor navigation with HDR imaging," in *IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 369–373, Weihai, China, 2019.

[18] B. Jang and H. Kim, "Indoor positioning technologies without offline fingerprinting map: a survey," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 508–525, 2019.

[19] X. Liu and X. Zhang, "NOMA-based resource allocation for cluster-based cognitive industrial Internet of things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 8, pp. 5379–5388, 2020.

[20] X. Liu, X. Zhai, W. Lu, and C. Wu, "QoS-guarantee resource allocation for multibeam satellite industrial Internet of things with NOMA," *IEEE Transactions on Industrial Informatics.*, vol. 17, no. 3, pp. 2052–2061, 2021.

[21] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: semantic segmentation of urban scenes based on RGB and thermal data fusion," *IEEE Transactions on Automation Science and Engineering*, vol. 99, pp. 1–12, 2020.

[22] Q. Zhang, N. Huang, L. Yao, C. Shan, J. Han, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Transactions on Image Processing*, vol. 29, pp. 3321–3335, 2020.

[23] Y. Wei, J. Feng, X. Liang, M. M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: a simple classification to semantic segmentation approach," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6488–6496, Honolulu, HI, 2017.

[24] Y. Liu, "Design of visual gaze target locating device based on depth camera," in *IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, pp. 1–5, Tianjin, China, 2019.

[25] S. Yang, R. Jiang, H. Wang, and S. S. Ge, "Road constrained monocular visual localization using Gaussian-Gaussian cloud model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 12, pp. 3449–3456, 2017.

[26] F. Wang, X. Chen, C. Tan, J. Li, and Y. Zhang, "Hexagon-shaped screw recognition and positioning system based on binocular vision," in *37th Chinese Control Conference (CCC)*, pp. 5481–5486, Wuhan, 2018.

[27] X. Liu and X. Zhang, "Rate and energy efficiency improvements for 5G-based IoT with simultaneous transfer," *IEEE Internet of Things Journal*, vol. 6, no. 4, pp. 5971–5980, 2019.

[28] X. Liu, X. Zhang, M. Jia, L. Fan, W. Lu, and X. Zhai, "5G-based green broadband communication system design with simultaneous wireless information and power transfer," *Physical Communication*, vol. 28, pp. 130–137, 2018.

[29] Y. Sun, X. Liang, H. Fan, M. Imran, and H. Heidari, "Visual hand tracking on depth image using 2-D matched filter," in *UK/China Emerging Technologies (UCET)*, pp. 1–4, Glasgow, UK, 2019.

[30] Z. Zhigang, L. Huan, D. Pengcheng, Z. Guangbing, W. Nan, and Z. Wei-Kun, "Vehicle target detection based on R-FCN," in *Chinese Control And Decision Conference (CCDC)*, pp. 5739–5743, Shenyang, 2018.

[31] Q. Chen, F. Shen, Y. Ding, P. Gong, Y. Tao, and J. Wang, "Face detection using R-FCN based deformable convolutional networks," in *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 4165–4170, Miyazaki, Japan, 2018.

[32] J. Tang, Y. Mao, J. Wang, and L. Wang, "Multi-task enhanced dam crack image detection based on faster R-CNN," in *IEEE 4th International Conference on Image, Vision and Computing (ICIVC)*, pp. 336–340, Xiamen, China, 2019.

[33] X. Mou, X. Chen, J. Guan, B. Chen, and Y. Dong, "Marine target detection based on improved faster R-CNN for navigation radar PPI images," in *International Conference on Control, Automation and Information Sciences (ICCAIS)*, pp. 1–5, Chengdu, China, 2019.

[34] J. Xu, L. Song, and R. Xie, "Two-stream deep encoder-decoder architecture for fully automatic video object segmentation," in *IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, St. Petersburg, FL, 2017.