Visual Inertial Map Matching for Indoor Positioning using Architectural Constraints

Author 1, Author 2, Author 3, Author 4, Author 5, Author 6, Author 7, Author 8

***Abstract*—Deployment-independent indoor localization methods, such as inertial tracking and vision-based tracking have been popular for years for they require no extra infrastructure cost. However, most systems require accurate initial positions from either the users or other external signals, making it difficult to meet on-demand localization requirements. To solve this issue, in this paper we propose a novel and special map matching system which matches the topology of the floor plan to the spatial structure extracted from the image of the environment taken by the camera of the mobile phone. The proposed system utilizes Convolution Neural Networks (CNN) to extract the spatial structure from images and Siamese Network for spatial structure matching. Extensive experiments have been conducted in three different buildings to demonstrate that our system can provide accurate positions without extra infrastructures and manual initial positions.**

***Index Terms*—Indoor Localization, map-matching, CNN, data fusion**

I. Introduction

Map-based indoor localization that reconciles the observations with the constraints provided by the maps in order to estimate the most trajectory, has drawn much attention in both civilian and military fields during the last decade [1]–[4]. Unlike outdoor localization, it cannot adopt the Global Navigation Satellite System (GNSS) owing to poor penetration. From the perspective of the pre-deployment hardware, indoor localization methods can be divided into two types: deployment-dependent and deployment-independent positioning [5], [6]. A disadvantage to all deployment-dependent positioning methods is the tediousness and delicacy of the deployment process. In a typical deployment processes, a trainer should perform a careful survey of the environment [7], [8]. This includes going to a certain locations, collecting the value of some type of signals and repeating this for probably a large number of points whose locations are known by accurate measurements. This implies several hours or even days of data collection for radio-map of buildings. Most deployment-dependent algorithms need to update databases or maintain equipments [9]–[11]. This requires a lot of manpower and material investment. additionally, deployment-dependent positioning will fail in specific scenarios where the facility cannot be deployed in advance [12]. Therefore, this work focuses on map-based deployment-independent indoor positioning.

As the key to providing accuracy indoor location, the map can be viewed as a geometric information providing spatial constraints. A floor plan of a building calibrates the movement of a user. For example, people can only enter the room through the door, not through the wall. However, most existing map matching methods rely on the external information of prior site survey or initial positions provided by users [13], [14], which makes it challenging to meet on-demand localization requirements. Researchers have investigated the possibility of reducing reliance for initial positions. [15] modelled the architectural plan as a topological graph and then realized the indoor topology matching, proving the effectiveness of topology matching. [16] adopted visual information of map landmark to achieve map matching by leveraging machine learning methods. However, we cannot ignore the fact that maps with landmarks will bring extra limitation to positioning [17]–[19]. [20] and [21] using building structure make it possible to achieve indoor localization through easily accessible information.

Inspired by the state-of-the-art researches, we explore a novel strategy for performing deployment-independent indoor positioning in this work. Here, we adopt easily accessible information such as floor plan and simple real-time images captured by smartphones to realize the user positioning by utilizing object recognition, video feature matching [22], and topological graphic matching all based on deep learning. In particular, topological graphic matching is based on the three-dimensional structure recovered from the floor plan and the spatial information directly extracted from the environmental picture.The goal of our proposed system is to complete indoor pedestrian positioning with similar accuracy as existing techniques. Unlike exiting techniques, our system does not require initial positions information, which is more in line with the actual use conditions of pedestrians.

We propose a system that uses target recognition and the Siamese Network for map matching.Our system requires floor plan,Inertial Navigation System data and video data as input.Among them, INS data is processed by topology matching to obtain the candidate trajectories.Candidate trajectories will be processed by cursory selecting and intensive selecting using video data, and finally the output trajectory will be obtained. In summary, the main contributions of this paper are as follows:

- On-demand positioning system without initial constraint: We design an indoor positioning system only utilizing map, Inertial Navigation System (INS) and videos to provide ON-DEMAND positioning service. Neither prior site survey nor initial positions from external information is needed.
- Novel matching method guiding pedestrian positioning: A novel method based on architectural perspective theory is designed to recover three-dimensional structure from floor plan. Besides, a Siamese Network has been em-

ployed to match the recovered three-dimensional structure with spatial information extracted from smartphone cameras.
- Extensive experiments validation: The proposed system demonstrates its performance in terms of tracking accuracy and robustness with a huge number of real-world experiment data in three different experiment scenarios.

The remainder of this paper is organized as follows:Section II reviews some related works, and Section III describes the proposed architecture in detail. The experimental results and analysis are performed in Section IV. Finally, we conclude the paper in Section V.

II. Related works

This paper focuses on deployment-independent indoor positioning technology. Researchers have proposed many effective approaches to solve this problem. In this section, we discuss various algorithms and their differences from others. Here, we mainly focus on the characteristics of building spatial structure and indoor localization.

A. The Characteristics of Building Spatial Structure

Most indoor scenes satisfy the Manhattan world Assumption where most planar directions belong to one of the three orthogonal directions. Under this assumption, there are many related studies on perspective line-drawing in architectural space. Guzman [23] focusing on segment graphs divided the sets of line drawings into parts of different objects. Huffman [24] and Clowes [25] proposed the standardized methods for defining concave, convex, and occlusive surfaces in perspective, allowing us to restore the three-dimensional description of objects from perspective line drawings. Many researches on perspective line maps based on Manhattanworld Assumption have been done by Kosecka et al. [26]. Kosecka and Wei [27] developed a perspective restoration method for the rectangular structure based on the vanishing point concept. On the other side, there are many state-of-the-art methods focusing on the relationship between an image and its corresponding three-dimensional structure. David et al. [20] focuses on vein, edge segmentation line, and position of the image. They used the three-dimensional structure of indoor image to restore the vertical and horizontal directions of each areas of the image to obtain a simple 3D model. Y. Li and S. T. Birchfield [21] completed an indoor floors segmentation scheme based on a single image without calibrating the camera. These methods imply that it is feasible to use architectural constrains to assist map matching.

B. Indoor Localization

Indoor positioning methods have been studied for decades, and many excellent approaches have been proposed. Gu Fuqiang [1] proposed a new indoor positioning scheme to improve the positioning accuracy. The emerging deep learning, a data-driven approach, has widely used in improving, reconstructing or innovating indoor positioning technology. Simon T [28] introduced an approach of indoor localization based on confidence interval fuzzy model by using fingerprint maps based on Bluetooth signals to improve the performance of the traditional K-nearest neighbor indoor localization method. Yuan Y [29] that adopted AdaBoost integrated learning method, perform multi-sensor data fusion to achieve an average accuracy of 1.39m with lower computational overhead.There are also many researches on data fusion positioning based on video information. Specifically, these methods rely on landmark information to achieve indoor positioning. Extra prior information based on designed landmarks can be used as a guidance of the location on real maps to assist the accurate positioning. [30] compares landmarks with image tags restored in the database by using the Scale-Invariant Feature Transform (SIFT) feature. [31] introduces a geometric localization method by leveraging indoor objects such as doors, elevators, cabinets, and characters of landmarks. [16] proposes a learning-based landmark recognition system by detecting continuous landmark sequences and using Markov Model to infer the trajectory of users. However, the performance of these landmark-based methods is severely limited by the environment.

In the field of deployment-independent indoor localization, the following three methods have their own characteristics and provide some ideas for our system. We will introduce their advantages and disadvantages in detail. The main content includes the specific conditions, costs and positioning performance of each method. 1) [MapCraft] [14]MapCraft is a robust and responsive technology. The response is less than 10ms on the Android platform. Even in the case of a large amount of noisy sensor data, it can have a good tracking and positioning effect. The advantage of the MapCraft system is that it has good positioning accuracy, supports almost all indoor scenes, and does not need to be trained again in advance. It has the characteristics of simple use and wide application scenarios. The system uses conditional random field technology (good application in natural language processing). In the system verification experiment, the RMS error is 1.14m-1.83m, and the 97th quantile is 2.37-4.53m. In terms of use conditions, the system only needs to pre-process the floor plan, provide the initial point of pedestrian walking, and locate the scene to support entering and leaving the room. However, we believe that it is difficult for pedestrians to provide an initial point in positioning. Imagine a pedestrian enters a building that has never been visited before. Since multiple points in the map are similar, pedestrians want to confirm their current location through the surrounding environment as the initial point of walking which is very difficult. In the light of this actual problem, it reflects the advantage of this system-indoor positioning does not need to provide an initial point of walking. 2) [VLSIL] [16] VLSIL provides an indoor topological positioning method based on mobile phone video. The system uses highly abstract landmark information to indicate location. The VLSIL system does not need to be retrained in a similar environment. The system analyzes the sequence of images in the video, uses CNN classification technology (specifically using AlexNet) to match the landmarks on each route on the

map, and selects the route with the most certain number of matches to select the final trajectory. The landmarks used in the experiment include fire extinguishers, stairs, door frames, elevators, toilet signs, etc., mainly single-object landmarks. In the experiment, 7 paths are set as candidate paths, and each path includes different road signs. The experimental results and analysis show that in a trajectory route, an average of 9 road signs can be observed to determine the trajectory. The VLSIL system also supports map matching without the need for initial point conditions, but it is costly in terms of positioning conditions, such as the need to mark landmarks on the map and set possible paths, which is difficult to complete in actual positioning as it requires advance site research and mark the detailed location of landmarks on the map, such as fire extinguishers, signboards, stairs, etc. For this type of landmarks, they cannot be automatically marked on the map by automated procedures, and a large number of manual marking operations are required, indicating that the system is costly and has low versatility. Through the comparison of conditions, it shows that our system has the advantage of not needing to investigate in advance, not using object landmarks but using scene landmarks (including door frames, building convex corners, building concave corners, etc.). The difference is that the former requires on-site investigation. The latter can be detected using automated procedures. Furthermore, because we use inertial navigation information, we can locate the specific location of the pedestrian by using the specific features of the inertial navigation trajectory after cursory selection and intensive selection, rather than just screening the pedestrian's walking route. 3) [PF-net] [32] The article applies Particle Filter Network (PF-net) to the visual positioning of robots, and proposes a network architecture that matches 3D real scene images with 2D maps. Experiments on the simulation data set of the House3D data set show that it is effective in matching environments with furniture and other obstructions. It evaluates the tracking effect on 820 trajectories of 47 untrained buildings, and the RMSE is 40.5 cm under the condition of using RGB cameras. PF-net can match the photos taken and the plane map in the room with obstacles, which is a huge advantage of this network, but it is difficult to complete in the real environment-it uses 45,000 trajectories from 200 buildings (virtual environment) for training. The images in the virtual environment are all clear and the background environment is not very different. We believe that such experiments are difficult to carry out in a real environment for two reasons: one is that the cost is very high, it is difficult to collect a large number from 200 buildings in a real environment, and it requires a one-to-one trajectory corresponding to the truth value, and the other is that the real situation is not considered the impact of image noise on training, the image background in real pictures may have inconsistencies in lighting, blurry photos, and inconsistent photo quality, making the network effect worse or even difficult to converge. In these aspects, the advantages of our system are demonstrated through comparison-using real image training instead of virtual building environment, and image processing methods such as histogram equalization and Gaussian blurring are used to reduce the noise impact of real images at different times and architectural backgrounds. On the other hand, due to the limited number of images collected in the real environment, we cannot directly match the screenshot fragments of the flat map with the real image, so this system innovatively converts the two-dimensional map into a simulated three-dimensional image. Using the deep learning matching method to match it with the real image, we believe that this can reduce the number of images required for training, and pre-processing in advance will help the network converge quickly.

III. Model

The system framework designed in this paper is shown in Fig. 1. The system is divided into three parts: data preprocessing, cursory selection, and intensive selection. The data preprocessing includes extraction and modeling of map topology models, inertial navigation trajectory generation. The cursory selection includes the feature vector conversion of pedestrian trajectory, matching with the map topology model. The intensive selection part includes door frame recognition and matching, building space structure matching. We use an example to illustrate the detailed steps: pedestrians use mobile phones to shoot video data and wear inertial measurement unit(IMU). When pedestrians enter the building, system loads the preprocessed map model of the building. Pedestrians walk a certain distance to generate an inertial navigation trajectory. The system uses the reconstructed trajectory to match the trajectory in the map topology model obtaining the candidate trails. Perform sampling processing on the video data taken by pedestrians, recognize the door frame of the image in the video to form a feature vector. At the same time, conduct door frame analysis on the candidate trajectory in the floor plan to match the feature vector with the previously obtained feature vector to further remove the mismatched trajectory. Image processing is conducted on the video image, and the part of floor plan corresponding to the discrete points of the candidate trajectory is converted into a simulated three-dimensional structure image, the obtained image is matched with the processed real video image. The output trajectory and corresponding location are finally obtained according to the degree of matching.

A. Preprocessing and cursory selection

Floor plan preprocessing

The floor plan used by this system is a two-dimensional flat map, and the map information does not contain semantic information (no need to manually mark door frames, corners, landmarks, etc.). The advantage of doing so is that there is no need to investigate the building in advance, thereby saving costs. In other words, it is feasible and convenient to use this system in new buildings. The preprocessing process of the building map includes extracting the pedestrian passage part of the map, marking the topological points according to the

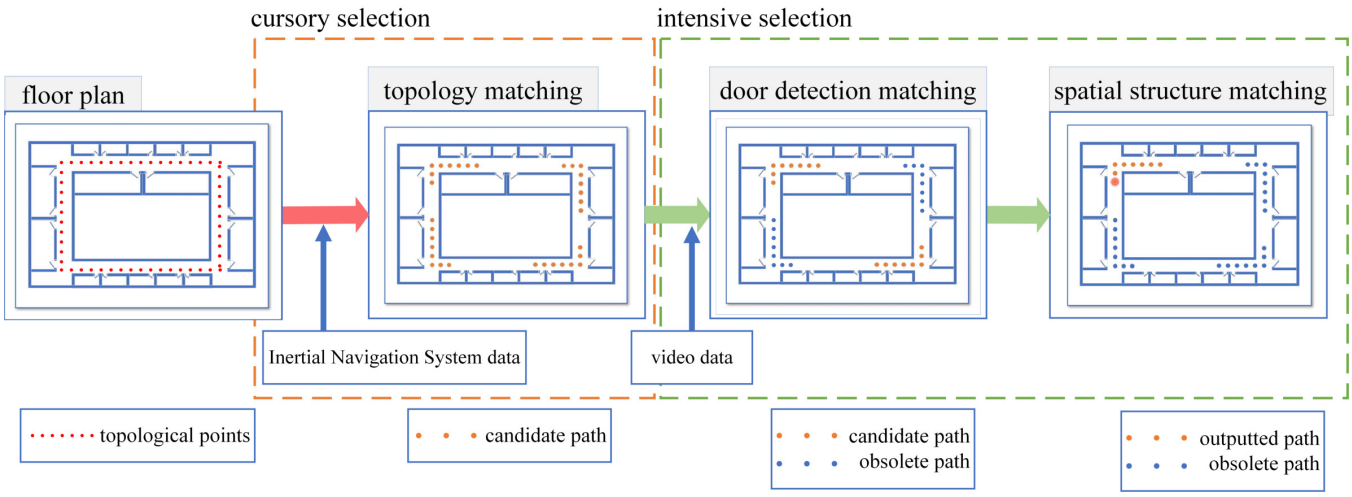


Fig. 1. System architecture

TABLE I
SPECIFICATIONS OF NGIMU.

	Accelerometer	Gyroscope	Magnetometer
Dynamic Range	$\pm 16g$	$\pm 4000\pi rad/h$	$\pm 1300\mu T$
Accuracy	$490\mu g$	$1.2\pi rad/h$	$0.3\mu T$
Dynamic Range	400Hz	400Hz	20Hz

distance of each half step (0.3m) of the pedestrians and converting them into a mathematical topological model, as shown in Fig. 12(a), 12(b), and 12(c).The storage relationship of topological points is $[(x0, y0) -> [(x1, y1) \dots (xn, yn)]], n = 1, 2, 3 \cdots]$ where $x0, y0$ represent the coordinates of the topological point in the x direction and y direction on the map, the unit is pixel, xn and yn are the topological points closest to $x0$ and $y0$ in each direction.

Trajectory generation

The IMU used in this system is the Next Generation Intertial measurement unit (NGIMU) produced by X-IO Technologies, which includes a three-axis gyroscope, a three-axis accelerometer and a magnetometer. In the experiment, the NGIMU is mounted on the foot. The module specifications are shown in the TABLEI.The sampling frequency of the inertial sensor is set 400 Hz. Because the general motion of the pedestrian will not be too intense, and does not involve high-speed rotation and vibration, so the NGIMU meets the calculation needs of this article.

When pedestrians walking, the IMU on the foot changes its posture continuously with the movement of the foot. The posture matrix of the strapdown inertial navigation system is

$$C_b^n(t) = C_b^n(t - \Delta t) \cdot \frac{2I + [\Omega_t] \times \Delta t}{2I - [\Omega_t] \times \Delta t} \tag{1}$$

where $[\Omega_t] = \begin{bmatrix} 0 & -\omega_z(t) & \omega_y(t) \\ \omega_z(t) & 0 & -\omega_x(t) \\ -\omega_y(t) & \omega_x(t) & 0 \end{bmatrix}$ represents the skew-symmetric matrix of the gyroscope at time t, w_x, w_y, w_z three-axis gyroscope readings, $\Delta t = 1/400$ is the sampling time interval, and I is the unit matrix.

After getting the updated attitude matrix, we can get the system navigation acceleration:

$$a_n(t) = C_b^n(t) \cdot a_b(t) - [0, 0, g]^T \tag{2}$$

where gravity is $g = [0, 0, -9.81]$. This system uses the inertial measurement unit, and the sampling frequency is 400 Hz. Therefore, it can be assumed that the acceleration and velocity from $t - \Delta t$ to t are constant. Take the average value from $t - \Delta t$ to t to get the velocity vector and position vector at this time:

$$v(t) = v(t - \Delta t) + a_n(t) \cdot \Delta t \tag{3}$$

$$p(t) = p(t - \Delta t) + [v(t - \Delta t) + v(t)] \cdot \Delta t/2. \tag{4}$$

With the zero-veocity detection algorithm based on a fixed threshold, ZUPT utilizing error-state Kalman filter (ESKF) can be used to correct and calibrate this. This is not the focus of this article, and will not be expanded. This system processes the inertial navigation data of pedestrians to obtain the inertial navigation trajectory. We need to further process the trajectory and match it with the known map topology information to obtain a cursory selected trajectory.

Trajectory feature vector conversion and matching

After the pedestrian inertial navigation trajectory is obtained by the above method, the system converts the trajectory into a feature vector

$$V = \{L^{'}, ..., d_n, L_n, ..., d_N, L_N\} \tag{5}$$

where N represents the total number of pedestrians turns in the real trajectory; d_n is the nth turn, this article takes 10 to the right and -10 to the left; L_n represents the cumulative

distance of the nth turn from the starting point of the trajectory; $L^{'}$ represents the total length of the trajectory.

Performing point traversal on the spatial topological model generated by the map, several paths with length $L^{'}$ are generated, which are converted to the above feature vectors. The Euclidean distance between feature vector of reconstructed trail V and selected trails $V^{'}$ is

$$d = \|V - V^{'}\|. \tag{6}$$

Trajectory whose vector Euclidean distance is less than the threshold (the threshold is 10 in this paper) is used as the candidate trajectory, which is the output result of cursory selection. The results obtained in this section are used as input for the subsequent intensive selection. Part B will explain the specific process and methods of intensive selection in detail.

B. Intensive selection

Video processing and analysis

The system needs to match the real information contained in the video with the map information contained in the candidate path obtained in the previous step, and output the correct track to get the pedestrian location. First, the video needs to be sampled appropriately according to cursory selection of the topological trajectory, so that the real image and the topological point of the trajectory can be one-to-one correspondence. A low sampling rate will result in the loss of rich door frame information and building spatial structure information, and a too high sampling rate will cause duplication of information, such as repeated door frame recognition, etc. The distance of the topological points in the map is approximately equal to half the pedestrian step length, which is the appropriate interval in sampling. It is obtained that the number of topological points contained in each candidate track is N, and the total length of video recording is T, then the video sampling rate is $f = N/T$. The sampled image sequence is used as input for subsequent analysis.

Door frame detection and matching

There are two reasons why this system relies on door frame detection. One is that the door frame can be detected by the pixel detection algorithm in the two-dimensional map, which helps to reduce the cost. There is no need to investigate the site in advance when we use the system in a new building. The second is that the current target detection algorithms used are mature, especially the YOLO algorithm based on deep learning, which helps us get more accurate information from real images.

The YOLOv3 [33] algorithm used in this paper improves the effect of feature extraction by adding residual network structures to achieve a deeper network main frame. On the COCO data set, using mAP-50 and single image detection time as evaluation indicators, YOLOv3 has state-of-the-art performance. In the case of similar detection speed, the accuracy of YOLOv3 is improved by at least 10%, and in the case of accuracy, the detection speed of YOLOv3 is 3-4 times that of models such as RetinaNet-101. We use YOLOv3 to detect the sequence images obtained by video sampling, and obtain the door frame information vectors on the left and right sides of the pedestrian walking.

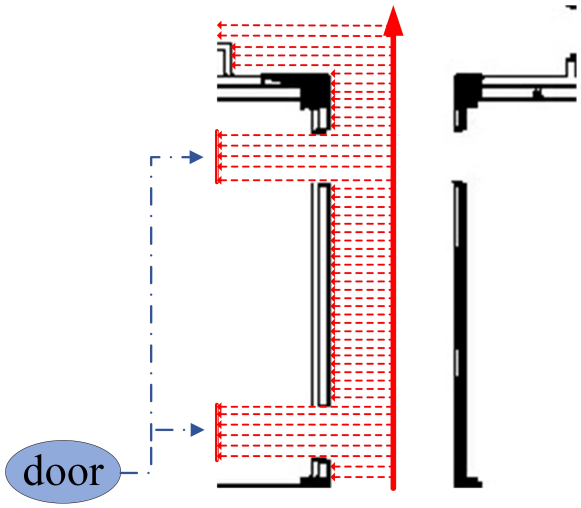


Fig. 2. Door frame detection from the floor plan

$$V_{r/l} = (L_0, L_1, ..., L_n) \tag{7}$$

where r/l represents the left and right sides, and L_i represents the cumulative distance of the ith door frame from the initial point of the track.

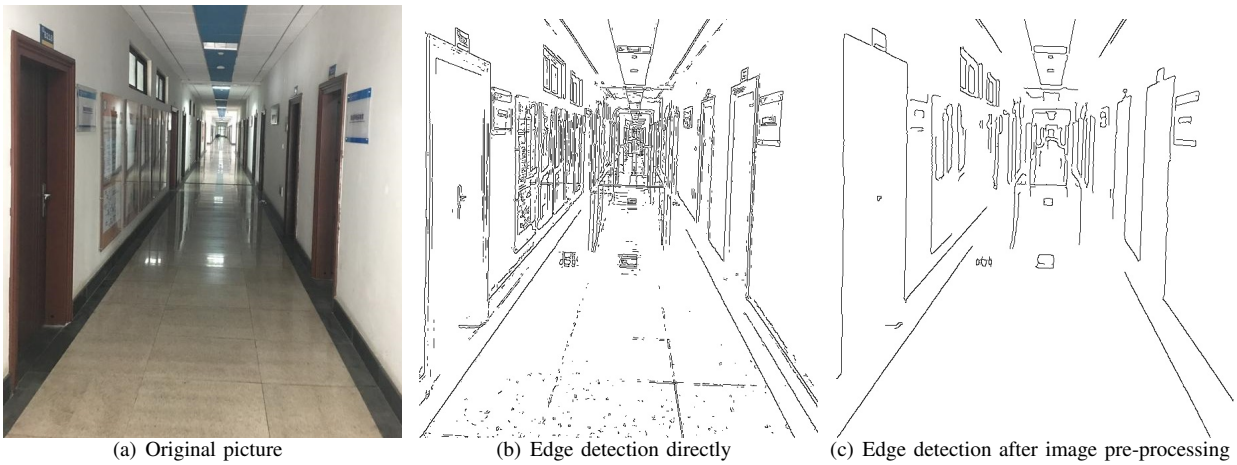
For the floor plan analysis part, as shown in Fig. 2, we analyze the door frame information of the two-dimensional map by analyzing the pixels, and also get the door frame information vector. Calculate the Euclidean distance between the vectors and remove the candidate trajectories that exceed the threshold. At this time, the output result of the first step of the system's intensive selection is obtained.

Spatial structure matching

*1) **Image processing**:* Due to the influence of the light intensity of the scene, the video taken by pedestrians has strong noise directly on the edge of the image, which is not conducive to subsequent matching. We do image processing with histogram equalization and Gaussian blur before edge extraction. The difference between whether or not to preprocess is shown in Fig. 3(a), 3(b), and 3(c).The processed image has less noise, and the image display is closer to the real spatial structure information of the scene, which is also conducive to matching in the subsequent steps.

In the edge extraction part, the system uses the Canny edge detection algorithm, which mainly includes two steps, suppressing noise and determining the edge position, as follows:

(1) First, use a Gaussian filter to convolution filter the input image to reduce the influence of noise on the gradient calculation. Because the gradient amplitude near the noise pixel is large, the edge detection operator is easy to misdetect the noise pixel as an edge pixel. The Gaussian filter is as follows:



(a) Original picture (b) Edge detection directly (c) Edge detection after image pre-processing

Fig. 3. Image processing with histogram equalization and Gaussian blur before edge extraction.

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right) \tag{8}$$

(2) Using the first-order difference operator to calculate the gradient amplitude components in the horizontal and vertical directions, so as to obtain the gradient amplitude M and the gradient direction θ of the image:

$$M = \sqrt{G_x^2 + G_y^2} \tag{9}$$

$$\theta = \arctan\left(G_y / G_x\right) \tag{10}$$

(3) Non-maximum suppression: traverse each pixel on the gradient amplitude image M[i, j], and interpolate the gradient amplitude of two adjacent pixels in the gradient direction of the current pixel. If the gradient amplitude of the current pixel is greater than or equal to these two values, the current pixel is a possible edge point, otherwise the pixel point is a non-edge pixel, and the edge of the image is refined into a pixel width, and the gradient magnitude image M [i, j] is processed to obtain the image NMS [i, j] by non-maximum suppression.

(4) Double threshold detection and edge connection: The Canny edge detection method uses high threshold T_h and low threshold T_l to extract edges, traverses the image NMS [i, j], and uses high threshold and low threshold for thresholding to obtain edge images E1 and E2 and E1 are strong edge points, there may be discontinuities, and E2 is weak edge point. Track the edge in E1. When the edge reaches the end point, search for edge points in the neighborhood of the corresponding position of the image E2 to connect the discontinuity in the strong edge E1, and constantly search and track the edge to connect the discontinuity in the edge of the high threshold image E1.

*2) **Generate simulated three-dimensional images from the floor plan**:* Further, we need to convert the floor plan information around the topological point into a simulated three-dimensional image. The purpose is to match the structural information contained in the map with the real information obtained by edge detection. As shown in Fig.4, when processing,

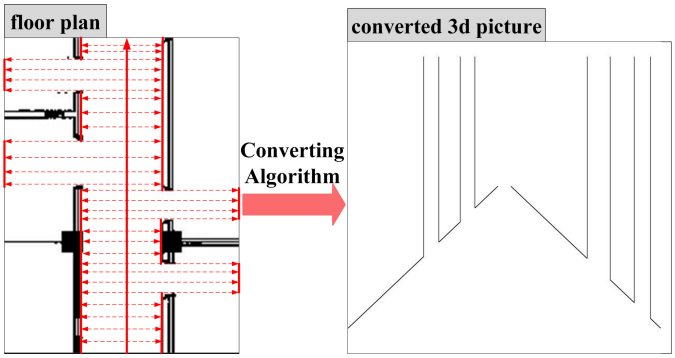


Fig. 4. Getting converted 3d picture from floor plan.

Algorithm 1 Feature extraction of spatial structure.

1: Input partially map of the front passage of point P;
2: **for** all pixels in that map **do**
3: Take the next pixel in the direction of motion as the initial point, and calculate whether there are pixels on both sides of the initial point and record the state: S_w;
4: According to the current S_w and the previous state, the structure of the left and right sides is judged as wall, concave angle, convex angle and empty, and the result is stored in L_l or L_r;
5: **end for**
6: According to the L_l and L_r, Draw analog images based on perspective theory;
7: end

we use the current topological point as the observation point, the movement direction as the observation direction, and use the architectural perspective principle to generate a simulated three-dimensional image. The specific algorithm is shown in Algorithm 1. There are two reasons why we do not use the original two-dimensional map image for matching. One is that the original two-dimensional image contains information outside the pedestrian's field of view, such as information within the room, which is one of the interference factors

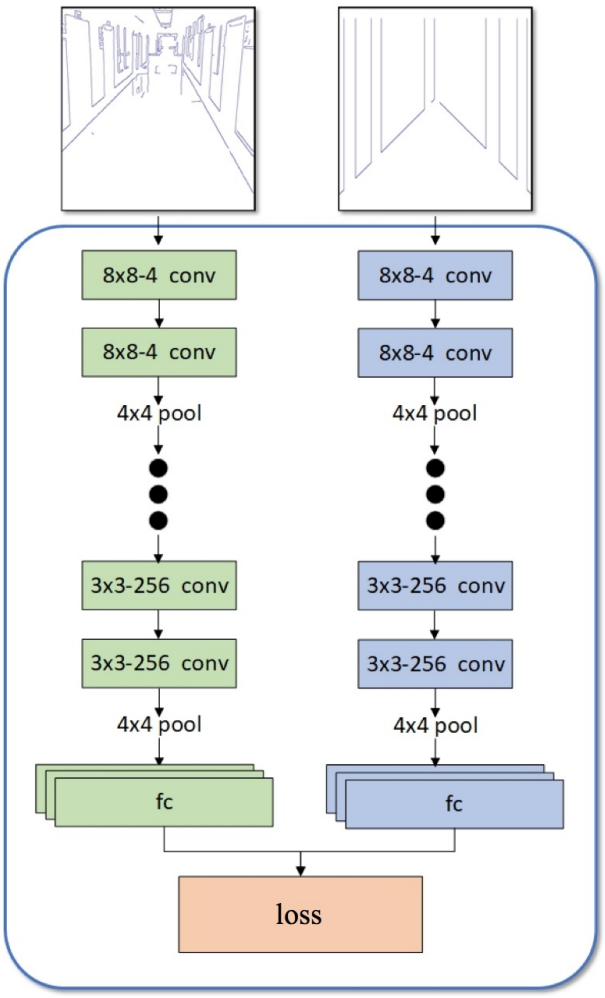


Fig. 5. The model of Siamese Network

for our matching. Second, the background difference between the two-dimensional map and the real edge extraction map is too large. If deep learning matching is used, an extremely large amount of data is required for training. The paper [32] uses a two-dimensional map to match the real environment, but the experiment is in a virtual software environment. Pedestrian images are simulated images generated by software and require a huge amount of data.

3) Siamese network: The spatial structure of buildings around specific topological points on the map can be regarded as unique landmarks. The image obtained through the previous steps contains spatial structural features (concave corners, convex corners of buildings, etc.). This system uses the Siamese network to match the real image obtained after the edge detection and the generated simulated three-dimensional image.

As shown in Fig. 5, the Siamese Network consists of 16 convolutional layers with a parametric rectified linear unit (PReLU) activation following each layer, 8 pooling layers, and 6 fully connected layers. Input images are in the size of $630 \times 600 \times 1$, and dropout is adopted to avoid overfitting considering the small-scale of our dataset. Note that two branches of the neural network are trained separately which means parameters of two subnetworks are not shared. The output of each branch represents the feature vector comprising high-dimension features. The similarity between two feature vectors is measured according to the loss function [34]

$$L = \frac{1}{2N} \sum_{i=0}^{N} \left(y_i d_i^2 + (1 - y_i) \max \left(M - d_i, 0 \right)^2 \right) \qquad (11)$$

where N indicates the batchsize, y_i indicates the label of the sample, and d_i is the Euclidean distance between two feature vectors. In this paper, we consider the loss L below the threshold $M = 10$ as the acceptable pair of trajectories.

As the result, the features of images which fit the matching rules get closer and closer in the feature space as training proceeds. On the contrary, the features of images which do not fit the matching rules get farther and farther. We use Euclidean distance to reflect the similarity of feature vectors. With the training positive data, the Euclidean distance between two groups of vectors would reduce; however, the Euclidean distance would increase with the negative data. In this study, the parameters of two subnetworks are not shared, and they are trained and used separately. The reason is that our input data is the matching group of real images and simulated feature images. The two have different backgrounds and scene information. Therefore, independent parameters are suitable for this condition.

4) Trajectory matching: In order to determine the final path from the candidates, we use the well trained Siamese Network to output the matching degree of the architectural structure image obtained from the video and the simulated three-dimensional structure picture obtained from the map. The matching probability P_k of each candidate path and the truth value of the trajectory is obtained by

$$P_k = \left(\sum_{i=1}^{N} p_i \right) / N \qquad (12)$$

where p_i is the class matching result at ith step, N means total steps of the real trajectory. The candidate path with maximum probability is taken as the final matching path, and its endpoint is the current location of the user.

IV. Evaluation

Sites and participants

In order to verify the positioning performance of this system in new buildings without training, we selected three buildings in the experiment. The experimental buildings are all reinforced concrete structures, namely the laboratory building, office building, and administrative building.The specific conditions of the experimental site including the area,etc., are shown in the table II.We build a topological model of the building as shown in Fig.12(a), 12(b), and 12(c).

TABLE II
EXPERIMENTS SITES.

	Laboratory building	Office	Administrative building
Area(m^2)	6756	3601	2829
The planned route length(m)	261	142	154
Number of door frames	80	37	32
The horizontal spacing of passages(m)	3	2.7	-

There are three participants employed to collect data with foot-mounted IMU and hand-held mobile phones to shoot video during walking. The mobile phones including iPhone X, iPhone Xr and iPhone 6s Plus. Pedestrians try to ensure that the camera shooting range is directly in front, to facilitate better collection of building door frame information and spatial structure information.

In this experiment, information of 468 trajectories with the average length of 50m were collected, including the raw IMU data, and corresponding video information. During the experiment, pedestrians try their best to walk along the topological route, and the experiment does not record the true position they passed during the walking process. In order to mark the true start point and endpoint of each trajectory, we recorded the true position by analyzing videos and maps.

Network training

The neural networks used in our system include the YOLOv3 model used to identify the door frame and the Siamese Network model used to match the processed real image with the simulated generated image. This paper uses the existed YOLOv3 model and adds 140 pedestrian-view door frame pictures. On the validation set, the recognition accuracy is 95.5%, which meets the requirements of the system for door frame recognition.

The Siamese Network model is the core framework in our positioning system. It is implemented in TensorFlow, and an NVIDIA TITAN XP GPU is used for training. Image data of the train set is collected from 4 buildings which are different sites from validation set and following test experiments. The image data includes the processed real image data of the pedestrian perspective and the simulated three-dimensional image conversion map in the field of view of the two-dimensional map where the topological point is located. Our training data set has a total of 16.1k pairs of pictures, and the ratio of positive samples to negative samples is 1:3. In the validation set, we constructed 1.0k pairs of pictures. The test set consists of 672 pictures, and the real pictures tested are partly taken from pictures taken in different environments and time periods.

During the training of the Siamese Network, the batchsize is set to 32, the learning rate is set to 0.001, and the training loss function curve changes as shown in Fig. 6(a). When the training round reaches 25000, the loss is 0.603. The F1 value of the Siamese Network on the test set is 76.63%, and the ROC curve and PR diagram are shown in Fig. 6(b) and 6(c). The neural network has not achieved high-performance in the test set. The reasons are: 1) The matching images have different backgrounds, and the shooting of real images will be affected by lighting, angles, random obstacles, etc. 2) The interference lines extracted from billboards and windows in the shot images can not be automatically analyzed from floor plan without manual labeling. 3) There are not enough buildings used in training, and the train set is relatively small.

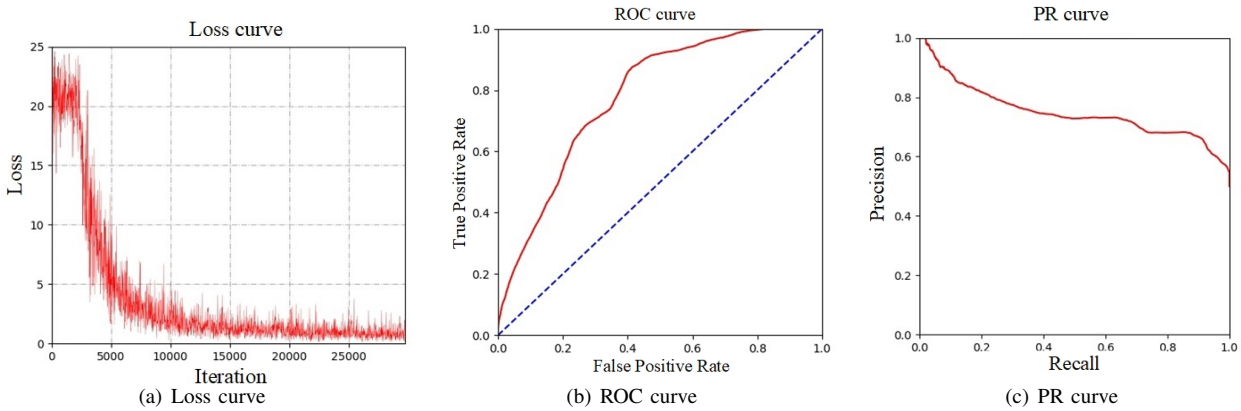
In terms of time consumption, the time consumption of each group of images is 0.255s when the Siamese Network in this article matches the Intel Core i7-8700 CPU and 3.20GHz environment. As shown in Fig.7, we compare the parameter amount and time complexity of the Siamese Network with well-known networks, including ResNet (RES-50), GoogleNet, AlexNet, and SqueezeNet. The results show that the complexity of the Siamese Network in this paper is equivalent to that of GoogleNet.

Distance required for positioning

The main purpose of our system design is to quickly obtain positioning without an initial point. We select three points A, B, and C in the building shown in Fig. 8, and walk 60m in the direction shown in the Fig. , and we count each point how long is needed to get the correct positioning. We start from each starting point 20 times, and the result is shown in Fig. 9. The abscissa is the distance in which the model calculates the correct positioning, and the ordinate is the number of times obtained by statistics. Take the point A to the direction in Fig. 9 as an example. In this experiment, the correct positioning was obtained at the 42 meters for 7 times. The results show that the distance required for positioning at points A, B, and C are 42m, 19m, and 25m, respectively. The median required distance are 41.5m, 18.5m, and 24m, respectively. It indicates that the distances required to obtain positioning from different starting points are different. Our analysis believes that the distance required from each point is related to the particularity of the path in the map. If the door frame information and building space structure information of a path have multiple similar results in the map, as the system cannot accurately determine the final matching location, then definite location can not be obtained at this time. As shown in Fig. 10, the selected two area A and B are very similar in the detected door frame and the structure information analysed through the video. This is the difficulty encountered in obtaining positioning using this system. Therefore, we believe that when the trajectory is special in the map, such as the unique topological matching result and the unique spatial structure information contained in the corner, the positioning can be obtained more easily.

Positioning performance

The goal of this experiment is to verify that the system has a good positioning performance in buildings with rich door frame information and building spatial structure information. Without prior knowledge of initial location, our algorithm has the capacity for accurate positioning relying on floor plan, inertial data, and shot images which are not required to be taken in advance for Siamese Network training, and only the preprocessing of topology modeling of the floor plan is



(a) Loss curve (b) ROC curve (c) PR curve

Fig. 6. Loss curve,ROC curve and PR curve.

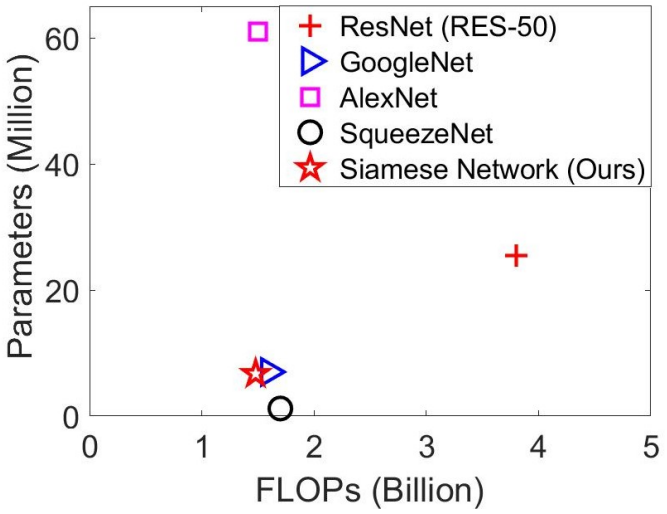


Fig. 7. FLOPs and Parameters

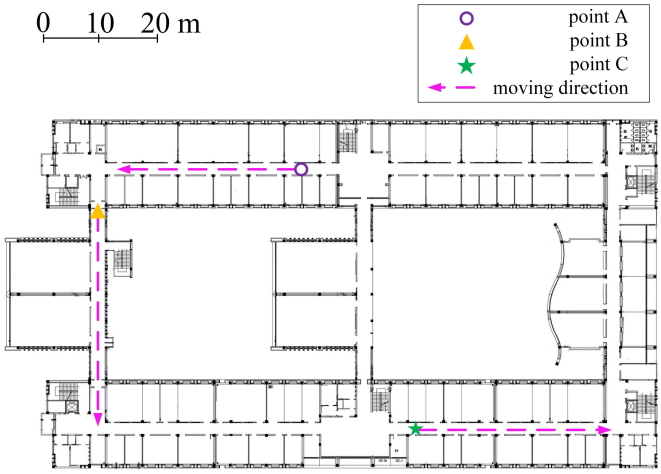


Fig. 8. Experiment–distance required for positioning from A,B and C.

TABLE III
EXPERIMENT RESULTS IN THREE SITES.

	Laboratory building	Office	Administrative building
RMS error(m)	1.81	1.43	1.20
85 percentile(m)	2.73	2.30	2.07
Choosing the right path(%)	91.4%	95.0%	96.3%
Number of door frames	80	37	32
The length of corridor	261	142	154

required. The system inputs the collected inertial navigation trajectory and pedestrian image sequence as data input. The inertial navigation trajectory is obtained by processing the original inertial navigation data by a zero-speed update algorithm. The image sequence is obtained from processing the original video data by sampling, histogram equalization, Gaussian blur and edge detection mentioned above.

In this experiment, the positioning success rate is defined as the correct candidate trajectory output by the system, that is, a trajectory whose starting point and ending point are closest to the pedestrian's actual walking is selected among multiple candidate trajectories. As shown in TABLE III, the results of this experiment show that the positioning success rate is above 91.4%, indicating that the system can be applied to buildings with rich door frame information, rich building spatial structure information, and no site investigation in advance. The positioning RMS error, the 85th quantile of error, the positioning success rate, the number of door frames, and the length of corridor in the experimental building are shown in the table. The cumulative distribution function of the experimental positioning errors in the three buildings is shown in Fig. 11. It can be seen from the Fig. that the best positioning performance among the three different types of buildings is the administrative building. From the analysis on the map, this building floor plan is not a centrally symmetric structure. The similarity of each candidate trajectory is low, and the system matching difficulty is low. In the real map matching experiment, because the door frame information and spatial structure information around some trajectories to be selected are very similar, the trajectory selection is wrong and cannot be matched correctly.

As shown in Fig.12(d)12(e)12(f),we show the positioning

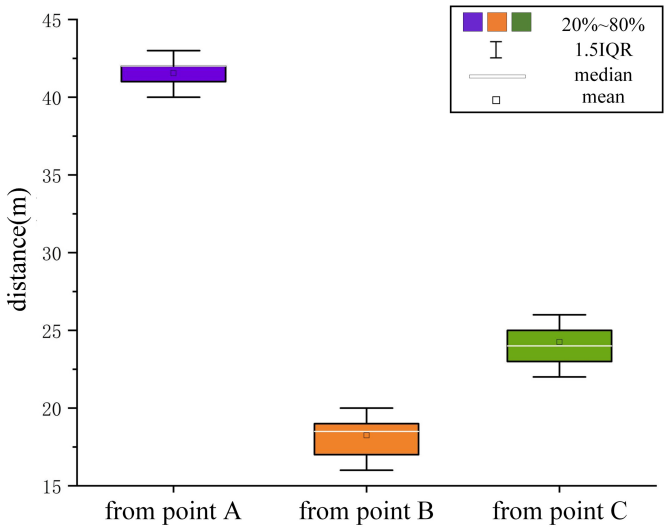


Fig. 9. The distance required to obtain the location from point A,B and C.

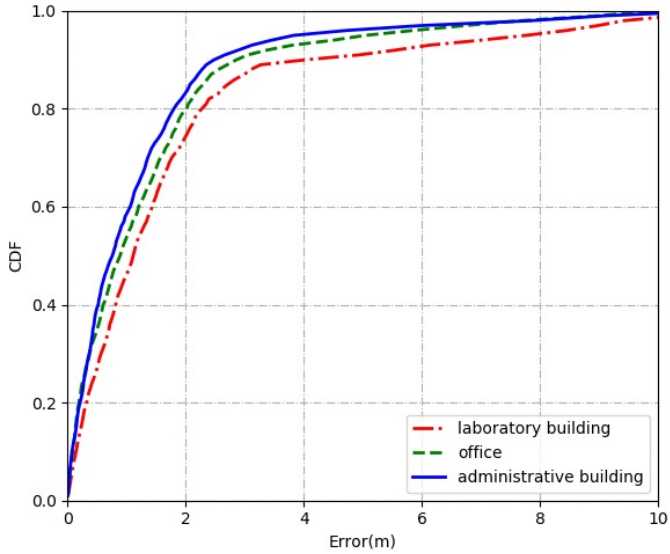


Fig. 11. Error CDFs in three sites.

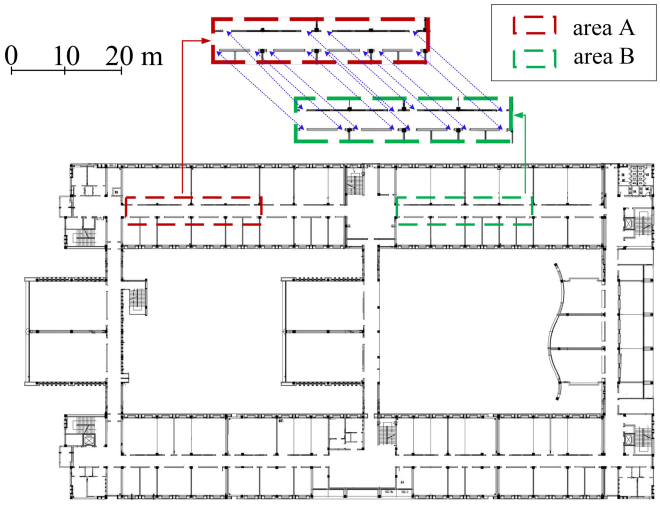
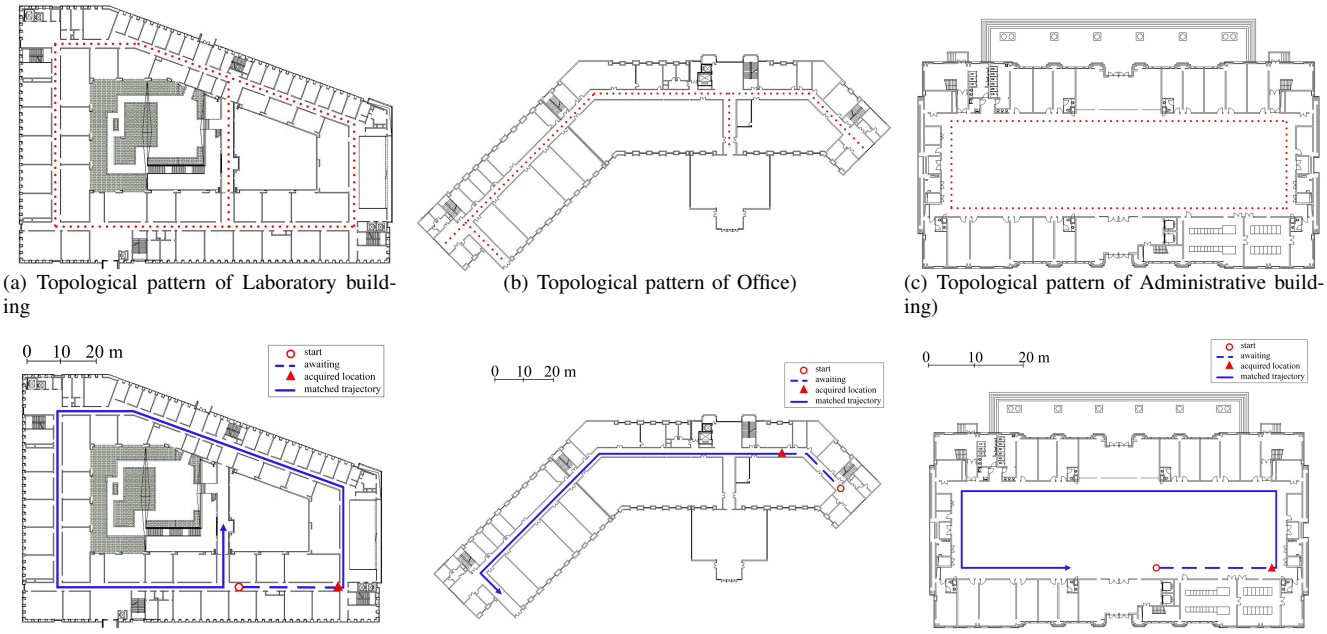


Fig. 10. The characteristics of area A and area B are similar.

performance of starting from a certain position in three buildings. Since this system solves indoor positioning under the condition of unknown initial point, we do not input the starting position into the system as known data during the experiment. Participants in the experiment walk along the path, and the experiment intercepts the trajectory and video data for map matching. When the input data can output accurate positioning, the end point of the trajectory data is used as the acquired location. Theoretically speaking, the walking distance required to complete the positioning is related to the actual starting point of the pedestrian, whether the building space structure information of each location on the map is similar, and whether the door frame interval in the trajectory is similar. The experimental results show that the more special the building map, the richer the information contained in the video, and the less information about the similarity of the door frame and the spatial structure, the easier it is to complete the positioning during short-distance walking.

Discussion

Through the above experiments result and comparison with other indoor positioning systems, we have analyzed the advantages of our system: 1) No need to provide initial points for pedestrians; 2) Low cost, no need to manually mark landmarks on the map and advance on-site investigation of buildings; 3) The Siamese Network in the system uses real video data as input and has strong versatility; 4) Converting a flat two-dimensional map into a simulated three-dimensional image (Algorithm 1)is a new method of acquiring map information, using less data and faster convergence. On the other hand, in the actual use of the system, we also analyzed the limitations of the system. 1) It is currently more suitable for buildings with compact passages as shown in Fig.12(a)12(b), and we are also experimenting in the buildings shown in Fig. 12(c). This kind of building is characterized by large open space and unpredictable pedestrian walking space. The walking path is specified in our experiment. However, in actual positioning, pedestrians may walk in an open area, and it is difficult to collect information such as road signs and building space structure through video. 2) Currently, pedestrians entering the room and walking outside of the set topology area are not supported. The system is designed to be able to walk in the corridor to obtain rapid positioning. 3) Since there is no limitation of the initial point, when there are more trajectories to be selected in the path matching, it will increase the amount of calculation. This situation has an impact especially on maps with many building passages and large building areas. Through the above analysis, we believe that this system is suitable for obtaining the position of pedestrians through short-term positioning. This position can be used as the initial point of a general positioning method, such as the initial point of methods such as [14] [15].Through such a combination, the general indoor positioning method using map and inertial



(a) Topological pattern of Laboratory building

(b) Topological pattern of Office)

(c) Topological pattern of Administrative building)

(d) Getting outputted location in 28.8 meters at Laboratory building.

(e) Getting outputted location in 24.2 meters at Office.

(f) Getting outputted location in 24.5 meters at Administrative building.

Fig. 12. Experiments in three different sites without training.

navigation information can solve the problem that the initial point is required, and retain the advantages of these methods of high accuracy, rapid response, and strong versatility.

V. Conclusion

In conclusion, the focus of this paper is to achieve the purpose of indoor localization by using easily available information under the condition that the initial position of users cannot be provided. We have designed an indoor map matching system based on the INS trajectory and video information. This system does not require any other data training. Many experiments show that our system has favorable generalization ability. The positioning system designed in this paper does not require the initial position and has the characteristics of accurate positioning and high efficiency. In the future, we will focus on using more visual information to assist positioning and improve positioning accuracy and efficiency.

References

[1] Gu Fuqiang, Hu Xuke and Ramezani Milad. Indoor Localization Improved by Spatial Context - A Survey. *ACM Comput Surv*, 2019.

[2] F. Zafari, A. Gkelias and K. K. Leung, A Survey of Indoor Localization Systems and Technologies.*IEEE COMMUN SURV TUT*, vol. 21, no. 3, pp. 2568-2599, thirdquarter 2019.

[3] Mendoza-Silva, G.M., Torres-Sospedra, J., Huerta, J.,A Meta-Review of Indoor Positioning Systems. *Sensors* 2019, 19, 4507.

[4] Morar, A., Moldoveanu, A., Mocanu, I., Moldoveanu, F., Radoi, I.E., Asavei, V.,A Comprehensive Survey of Indoor Localization Methods Based on Computer Vision. *Sensors* 2020, 20, 2641.

[5] F. Zafari, I. Papapanagiotou, and K. Christidis, Microlocation for Internet-of-Things-EquippedSmartBuildings,*IEEE INTERNET THINGS*, vol. 3, no. 1, pp. 96–112, 2016.

[6] K. Al Nuaimi and H. Kamel, A survey of indoor positioning systems and algorithms, in 2011 *IIT*, Abu Dhabi, United Arab Emirates, 2011, pp. 185–190.

[7] Zhi-An Deng, Guofeng Wang, Danyang Qin, Zhenyu Na, Yang Cui, and Juan Chen. Continuous Indoor Positioning Fusing WiFi, Smartphone Sensors and Landmarks. *Sensors*16.9 (2016): 1427. Web.

[8] Chen, Zhenghua, Han Zou, Hao Jiang, Qingchang Zhu, Yeng Chai Soh, and Lihua Xie. Fusion of WiFi, Smartphone Sensors and Landmarks Using the Kalman Filter for Indoor Localization.*Sensors* 15.1 (2015): 715-732. Web.

[9] Idrees, Affan, Zahid Iqbal, and Maria Ishfaq.An efficient indoor navigation technique to find optimal route for blinds using QR codes. *2015 ICIEA*. IEEE, 2015.

[10] Fusco, Giovanni, and James M Coughlan.Indoor Localization Using Computer Vision and Visual-Inertial Odometry.*ICCHP* (2018): 86-93. Web.

[11] Chaccour, Kabalan, and Georges Badr. Computer Vision Guidance System for Indoor Navigation of Visually Impaired People.*IS* (2016): 449-54. Web.

[12] F. Vedadi and S. Valaee, Automatic Visual Fingerprinting for Indoor Image-Based Localization Applications. *IEEE TSMCS*, pp. 1–13, 2017.

[13] K.-C. Lan and W.-Y. Shih, On Calibrating the Sensor Errors of a PDR-Based Indoor Localization System. *Sensors*, vol. 13, no. 4, pp. 4781–4810, Apr. 2013.

[14] Z. Xiao, H. Wen and A. Markham, Lightweight map matching for indoor localisation using conditional random fields.*IPSN-14*, Berlin, 2014, pp. 131–142.

[15] S. Shahidi and S. Valaee, Graph matching for crowdsourced data in mobile sensor networks.*2014 IEEE 15th International Workshop on SPAWC*, Toronto, ON, Canada, 2014, pp. 414–418.

[16] Q. Li, J. Zhu and T. Liu, Visual Landmark Sequence-based Indoor Localization. *in GeoAI*, New York, NY, USA, 2017, pp. 14–23.

[17] Kosecka, J., Liang Zhou, Barber, and Duric. Qualitative Image Based Localization in Indoors Environments. *IEEE CVPR*, 2003. Proceedings 2 (2003): II. Web.

[18] Lu, Guoyu, Yan Yan, Nicu Sebe, and Chandra Kambhamettu. Indoor Localization via Multi-view Images and Videos.*COMPUT VIS IMAGE UND* 161 (2017): 145-60. Web.

[19] Piciarelli, Claudio. Visual Indoor Localization in Known Environments.*IEEE SIGNAL PROC LET* 23.10 (2016): 1330-334. Web.

[20] D. C. Lee, M. Hebert and T. Kanade, Geometric reasoning for single image structure recovery. *2009 IEEE CVPR*, Miami, FL, 2009, pp. 2136–2143.

[21] Y. Li and S. T. Birchfield, Image-based segmentation of indoor corridor floors for a mobile robot. *2010 IEEE/RSJ*, 2010, pp. 837–843.

[22] L. Chen, F. Rottensteiner and C. Heipke, Invariant descriptor learning using a Siamese convolutional neural network. *XXIII ISPRS Congress, Commission III 3 (2016)*, Nr. 3, 2016, vol. 3, pp. 11–18.

[23] A. Guzman. Decomposition of a visual scene into three-dimensional bodies. *Proceedings of Fall Joint Computer Conference*, 1968

[24] D. A. Huffman. Impossible objects as nonsense sentences. *Machine Intelligence*, 1971.

[25] M. B. Clowes. On seeing things. *Artificial Intelligence*, 1971.

[26] J. Košecká and W. Zhang, Video compass. *European conference on computer vision*, 2002, pp. 476–490.

[27] J. Košecká and W. Zhang.Extraction, matching, and pose recovery based on dominant rectangular structures. *Computer Vision and Image Understanding*, vol. 100, no. 3, pp. 274–293, 2005.

[28] S. Tomažič, D. Dovžan, and I. Škrjanc.Confidence-Interval-Fuzzy-Model-Based Indoor Localization. *IEEE Transactions on Industrial Electab1tab1tronics*, vol. 66, no. 3, pp. 2015–2024, 2018.

[29] Y. Yuan, C. Melching, Y. Yuan, and D. Hogrefe.Multi-device fusion for enhanced contextual awareness of localization in indoor environments. *IEEE Access*, vol. 6, pp. 7422–7431, 2018.

[30] Y. Bai, W. Jia, H. Zhang, Z.-H. Mao, and M. Sun.Landmark-based indoor positioning for visually impaired individuals. *2014 12th ICSP*, 2014, pp. 668–671.

[31] M. Serrão, J. M. Rodrigues, J. I. Rodrigues, and J. H. du Buf.Indoor localization and navigation for blind persons using visual landmarks and a GIS.*Procedia Computer Science*, vol. 14, pp. 65–73, 2012.

[32] Karkus, Peter, David Hsu, and Wee Sun Lee. Particle filter networks with application to visual localization. *arXiv preprint* arXiv:1805.08975 (2018).

[33] J. Redmon and A. Farhadi.YOLOv3: An Incremental Improvement.*arXiv:1804.02767 [cs]*, Apr. 2018.

[34] S. Chopra, R. Hadsell and Y. Lecun.Learning a Similarity Metric Discriminatively, with Application to Face Verification. *IEEE CVPR*, 2005.