IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 4, 2020, accepted March 25, 2020, date of publication April 6, 2020, date of current version April 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2986044

Continuous Indoor Visual Localization Using a Spatial Model and Constraint

XING ZHANG[1,2,3,4], JING LIN[1,2,3,4], QINGQUAN LI[1,2,3,4], TAO LIU[5], AND ZHIXIANG FANG[6]

[1]Guangdong Key Laboratory of Urban Informatics, School of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China
[2]MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area, Shenzhen University, Shenzhen 518060, China
[3]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China
[4]Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China
[5]College of Resources and Environment, Henan University of Economics and Law, Zhengzhou 450002, China
[6]State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430072, China

Corresponding author: Xing Zhang (xzhang@szu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 41301511, Grant 41801376, and Grant 41771473, in part by the National Key Research Development Program of China under Grant 2016YFB0502203, in part by the Natural Science Foundation of Guangdong Province under Grant 2018A030313289, in part by the Shenzhen Scientific Research and Development Funding Program under Grant JCYJ20170818144544900 and Grant JCYJ20180305125033478, in part by the Open Research Fund of State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, under Grant 18S03, and in part by the Key Research Projects of Henan Higher Education Institutions under Grant 19A420004.

ABSTRACT Visual localization is an accurate and low-cost indoor localization solution. A bottleneck for visual localization is the computation efficiency of continuous image searching and matching. In this paper, an indoor visual localization method is proposed to realize continuous and accurate indoor localization based on image matching. This method uses smartphones to collect multi-sensor data, including video frames and inertial readings. To improve the computation efficiency of the proposed visual localization method, a spatial model is developed to optimize the spatial organization of geo-tagged images in a dataset. Several spatial constraint-based image searching strategies are also designed to further reduce the computation time. Based on the spatial model and spatial constraint-based strategies, a visual localization algorithm is proposed. The experimental results show that the localization errors of the image querying, continuous offline localization and online localization of this method are approximately 0.4 m, 0.7 m and 0.9 m, respectively. This method can achieve an accuracy of 1.3 m, even under a random camera opening condition. The average computation time (i.e.. the average time needed to provide a location estimation result) is approximately 0.59 s. The results indicate that the proposed method can realize efficient and continuous indoor localization with high localization accuracy.

INDEX TERMS Indoor positioning, visual localization, image matching, spatial model.

I. INTRODUCTION

The localization of people in large indoor environments, such as shopping malls, office buildings or large parking garages, has become a common issue for many industry and commercial applications. Due to the shielding effect caused by obstacles (e.g., buildings), it is difficult to obtain accurate localization results from GPS in indoor spaces. During the past decade, various indoor localization technologies, such as Wi-Fi [1], Bluetooth [2], ultrasonic [3], radio frequency (RFID) [4], ultra-wideband (UWB) [5], and magnetic fields [6], have been developed. Currently, some technologies, for example, UWB or Bluetooth, can achieve good localization performance in indoor spaces. However, the requirement of extra devices and infrastructures limits the large-scale application of these systems. Among various indoor localization technologies, visual localization is an accurate and low-cost indoor localization solution. It uses a camera (e.g., from a smartphone) to collect video frames from the environment. The collected video frames can either be used to achieve a relative localization based on an SFM (structure from motion) scheme [7] or be compared with geo-tagged data in a database to find the best matching result [8]. Thus, visual localization systems can be easily deployed in various indoor or outdoor environments and do not rely on extra infrastructures or devices.

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

Image matching-based location is mostly used in visual place recognition. The aim of visual place recognition is to decide whether or not an image of a place has already been seen by a human or robot [9]. Usually, an image will be visually compared with database images (with known place names or locations) to find the best matching result. Image matching methods use image feature descriptors, such as SIFT, SURF or Gist [10]–[12], to visually describe the scene of the image. Different methods, such as machine learning or visual words [13], measure the likelihood or confidence that the current visual input matches database images and then find the best matching result. However, these studies mainly focus on the outdoor environment and large-scale place recognition. It has not been considered whether the required computation time for visual place recognition can support continuous localization or navigation. Some studies include three-dimensional information from LIDAR or R-GBD [14], [15] cameras to improve the accuracy of place recognition. However, this will also increase the device requirements and costs of a localization system.

For indoor environments, some studies [16]–[18] used smartphone cameras to detect visual features from the environment and match the detected data with previously geo-tagged features to realize indoor localization. For example, in [19], an indoor positioning system was designed that uses common static objects, e.g., doors or windows, to locate users. A smartphone-based visual positioning method was also proposed in [20]. This method uses a mobile mapping system to generate a high-precision 3D photorealistic map of an indoor space and then matches the smartphone image with the generated map to obtain the position of the smartphone. These methods perform well in indoor environments with high location accuracy. However, the localization results of some studies are separate query locations and not continuous trajectories. In [21], an image-based indoor trajectory estimation method was proposed, which recovered the pose of the camera from the 2D-3D correspondences between the 2D image positions and the 3D points of the scene model. The 3D scene model was previously constructed by using a SFM pipeline. The computation time is another bottleneck of visual feature matching-based methods. Generally, the localization accuracy of visual feature (e.g., geo-tagged image) matching can be improved if the spatial density of the visual features increases. The reason for this improvement is that the location error can be corrected more frequently if there are more geo-tagged visual features. However, a higher spatial density of visual features also means that the required computation time for feature matching will greatly increase. It is difficult to balance the relation between the spatial density of visual features and localization accuracy.

As a well-known imaging technology, the SFM method can be used to recover the relative camera pose and 3D structure from a set of camera images. Previous studies have utilized the SFM method to recover the geometry of trajectories in indoor spaces [22], [23]. An advantage of the SFM method is that the heading estimation error of the SFM is significantly smaller than that of the PDR-based estimation (from the gyroscope). However, the location error of the SFM-based method will accumulate continuously as the walking time increases. In addition, the initial location of a trajectory is still required for an SFM-based method, which limits its application in wayfinding and navigation.

The present study proposes a visual-based method that can realize continuous indoor localization. We first collected a geo-tagged image dataset of an indoor environment. The spatial density of the images in this dataset is relatively high (approximately 2 m). To reduce the required computation time of image matching, a spatial model of geo-tagged images is proposed. This model optimizes the spatial organization of geo-tagged image data considering the relationship between image similarity and spatial distance. A visual localization algorithm is also developed in this study. Several spatial constraint-based visual search strategies are defined to further increase the efficiency of the localization algorithm. To realize continuous indoor localization, an SFM-based method is employed in this algorithm. The image matching and SFM location results are fused to increase the accuracy and frequency of the indoor localization algorithm. The relationship between the spatial density of geo-tagged images and localization accuracy is also studied in this work.

The paper is structured as follows Section II reviews the related work. Section III presents the spatial model of geo-tagged images and the visual localization algorithm. The experimental results and discussion are given inSections IV and, respectively. Section VI concludes this wor.

II. RELATED WORK

Visual information is valuable for various positioning systems because it contains rich environmental information around a camera and does not rely on any additional infrastructure. In outdoor environments, previous studies have employed an image matching method to locate moving objects or query images [24]–[29]. For example, in [24], an image matching-based localization system was designed that provides satisfactory localization performance with low computational complexity. This system uses a key frame selection method and a simple tree scheme to achieve fast image search. In [25], a geo-registration approach that can estimate and geo-tag the location of query images (e.g., images of famous landmarks) by matching the images with geo-located aerial images was proposed. The localization method presented in [26] can topologically localize a vehicle on a previously travelled road by using image feature matching. There are also studies that have localized a vehicle or device by estimating camera pose from images [27]–[29]. For example, the method presented in [28] can localize and track mobile devices in urban outdoor environments. This method estimates the absolute camera orientation from straight line segments and the camera translation by matching a semantic segmentation of an image with a city map model. In [29], a deep learning architecture was developed to achieve image-based localization of a camera or an autonomous

system. This approach combines a CNN with LSTM units for camera pose regression, which leads to better localization performance.

In indoor environments, visual feature matching has been used in many indoor localization methods [30]–[37]. These studies mainly used cameras to collect query images in indoor environments and match them with a previously built image database. The location of the ''closest'' image is treated as the localization result of a query image. Some studies used a trained CNN to match the query images taken by the cameras to estimate the location of mobile devices [30]–[33]. In [34], an alternative approach was proposed that leverages environmental reference objects, for example, store logos, for indoor localization. An image matching algorithm was designed to automatically identify the chosen reference objects in photographs. In [35], a sparse 2.5D georeferenced image database was generated using an ambulatory backpack-mounted system. The query images can be matched against the image database to retrieve the best-matching database image for indoor localization. However, most of these methods are not designed for continuous indoor localization. In addition, whether or not the required computation time can support continuous indoor localization and navigation has not been considered yet. Efforts have also been devoted to improving the efficiency of visual localization. In [8], a sorting hat approach was proposed to filter out uncorrelated feature pairs of images to improve the efficiency of image matching. The improved RANSAC method proposed in [36] can reduce the iterations and running time for indoor visual-based localization. However, it remains unclear how to improve the efficiency of high-accuracy continuous indoor visual localization.

Considering the required computation time, it is difficult to realize continuous indoor localization using the image matching method alone. A way to increase the frequency of visual localization is to include a pedestrian dead reckoning (PDR) method in the system. PDR techniques take advantage of the measurements from a micro-electro-mechanical system (MEMS) to calculate one's current location by adding the estimated displacement to the previously determined location. However, due to the drift noise of the MEMS, the location error accumulates as time goes on. Numerous studies have been presented to reduce the accumulative error of PDR. For example, in [37], several algorithms, including robust step detection algorithms, adaptive stride length estimation algorithms and heading estimation algorithms, were presented to reduce the sensor drift error of PDR. The system proposed in [38] utilized modified Kalman filtering to fuse acceleration, angular rate and magnetic field sensor data to provide a long-term stable orientation solution. This system also uses zero velocity updating and body movement monitoring to reduce localization error. There also have been studies that use inertial information to detect pedestrian activity [39], [40]. The detected activities are used as landmarks to reduce the accumulative error of PDR. Other studies have employed different information sources, e.g., Bluetooth, to correct the error of PDR [41], [42]. However, the requirement of extra devices also increases the difficulty of system deployment.

The SFM method is another way to realize continuous visual localization. This method is often used to recover the relative camera pose and 3D structure from a set of camera images. In [43], a structure-from-motion framework was designed to handle ''generalized'' cameras and works at an unprecedented scale by exploiting a good relative pose along vehicle paths. The SFM method has been used in photogrammetric measurements to solve for the camera pose and scene geometry simultaneously [44], [45]. In [46], iMoon built a 3D model of an indoor environment by using SFM technology that supports image-based localization. Reference [23] employed an SFM method to estimate the trajectory of a moving camera in an indoor environment. However, a problem is that the initial location of a camera should be given as an input for SFM-based trajectory recovery. In addition, the error of heading estimation accumulates as time passes, which significantly affects the accuracy of indoor localization.

In summary, visual information has the potential to improve the performance of indoor localization. However, there are still some problems with image-based indoor localization, such as the accumulative error, computation time, and continuity of estimated trajectory. In this study, a visual localization method is proposed to realize continuous indoor localization. A spatial constraint-based localization algorithm is designed based on the integration of image matching and the SFM method. A spatial model of geo-tagged images is also developed in this study and can significantly reduce the computation time of the proposed algorithm.

III. METHODOLOGY

Figure 1 shows a block diagram of the proposed method. This method uses the built-in sensors of a smartphone to collect sensor data, including video frames and inertial readings. During the offline phase, we use a multi-constrained image matching method to extract and describe the visual features from collected geo-tagged images. A spatial model is developed to optimize the spatial organization of the geo-tagged images. During the online phase, several spatial constraint-based image search strategies are designed to increase the efficiency of image matching. A visual localization algorithm is proposed to realize continuous indoor localization by integrating both image matching and the SFM method.

A. MULTI-CONSTRAINED IMAGE MATCHING

The basic idea of the visual localization method is to match a sequence of video frames (collected by a smartphone camera) to the geo-tagged images in an indoor environment. The location of a matched geo-tagged image will be used to provide a localization result for a query image. In this study, a multi-constrained image matching method is used to find the correspondence between a query image and a geo-tagged image on the pixel scale. The best fitting image from the

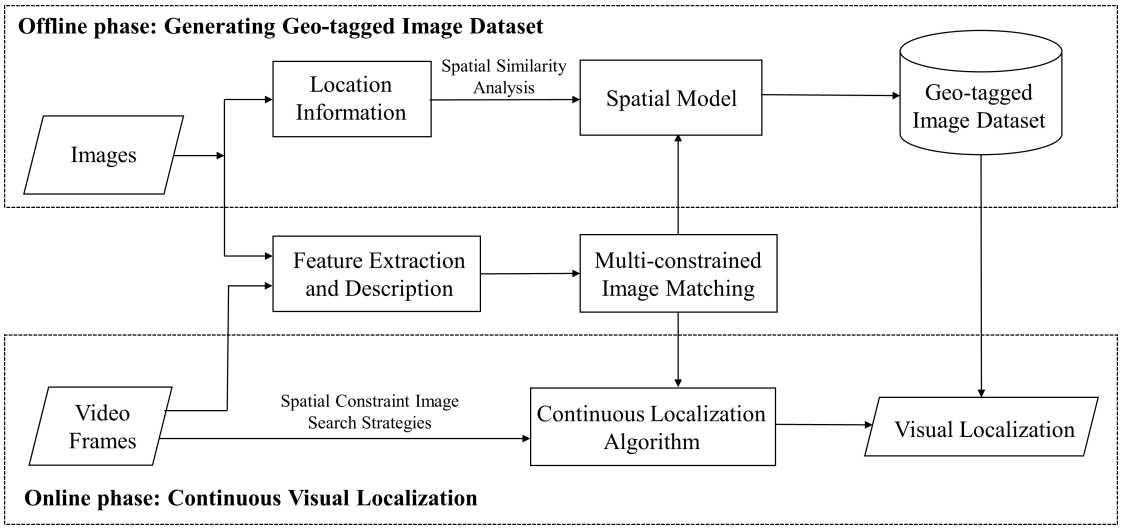


FIGURE 1. Block diagram of the proposed method.

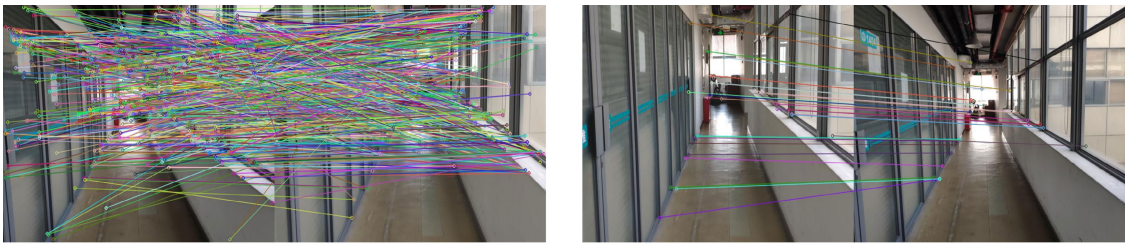


FIGURE 2. Matching result of two images. (a) matching result of the SIFT method; (b) matching result of the improved method.

geo-tagged database will be selected as the matching result of a query image. This approach uses a scale-invariant feature transform (SIFT) [47] algorithm to detect and describe local features from images, which are important for establishing the correspondence among pixels. The SIFT algorithm applies the Gaussian differential function shown below to select key locations at maxima and minima in scale space:

$$g(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} \tag{1}$$

where σ is the standard deviation of the normal distribution, which is equal to $\sqrt{2}$, and x is the fuzzy radius, which refers to the distance from the template element to the centre of the template. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame. The features achieve partial invariance to local variations, such as affine or 3D projections, by blurring image gradient locations.

As shown in Figure 2 (a), during image matching, many incorrectly matched keypoint pairs may exist if using only the SIFT algorithm. To improve the accuracy of image matching, two different constraints are employed to reduce the incorrectly matched keypoint pairs:

- Symmetry constrain. When a keypoint from image a has been matched with multiple keypoints from image b, the redundant matches should be removed from the results. In this case, the two images are matched to each other two times: (1) from a to b and (2) from b to a. The common parts of the two matchesare used as the final matching result.
- Ratio constraint. For a keypoint from image a, it best matching keypoint from image b can be calculated as the Euclidean distance d between feature vectors of the two keypoints. A matched keypoint pair is not treated as a successful matchwhen the ratio of the smallest distance d_1 to the second smallest distance d_2 is higher than a threshol r.

Based on these two constraints, a RANSAC algorithm [48] is employed to further improve the matching result. It can calculate mathematical model parameters from data and obtain effective sample data according to a set of sample data containing abnormal data. In this study, this algorithm randomly extracts four sample data points from the matching result of keypoint pairs and calculates a homography matrix. The outliers can be removed by iterating the RANSAC algorithm until the maximum number of inliers of the homography matrix has been obtained. As shown in Figure 2 (b), most of

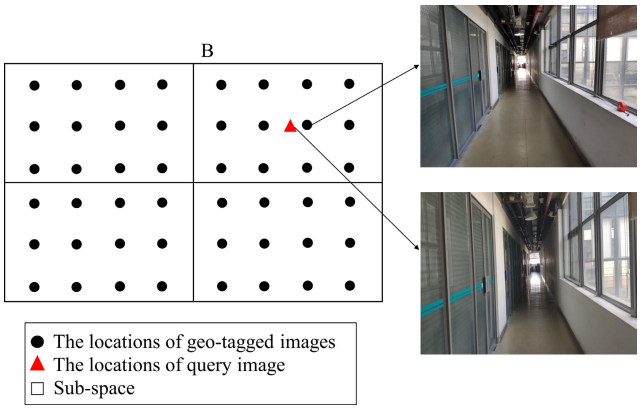


FIGURE 3. Subdivision of an indoor space for image matching.

the incorrectly matched keypoint pairs have been removed by using the RANSAC algorithm and the two constraints.

B. SPATIAL MODEL OF INDOOR GEO-TAGGED IMAGES

For a large indoor environment, image matching-based visual localization is quite time-consuming because a query image should be matched with each image in the environment to find the best fitting result. To improve the efficiency of visual localization, a spatial model of geo-tagged images is proposed in this study. In this model, a whole indoor space is divided into many sub-spaces. For example, a room or a corridor can be treated as a sub-space. A sub-space may contain a series of geo-tagged images that fall into its spatial extent. When searching for the spatial location of a query image, this model will first estimate the possible sub-space that may contain the query image and then match the query image with the geo-tagged images belonging to these sub-spaces (but not all the geo-tagged images in the whole space). For example, as shown in Figure 3, the whole space consists of four sub-spaces. Each sub-space includes a series of geo-tagged images (black dots). If a query image (red triangle) is located in sub-space B, the geo-tagged images from sub-space B will be treated as the matching candidates of the query image.

By subdividing a whole indoor space, spatial indexing can be generated for geo-tagged images. As shown in Figure 4, an indoor space consists of a series of sub-spaces with corresponding sub-space IDs. Each sub-space is associated with a list of geo-tagged image IDs that fall into its spatial extent. Each geo-tagged image can only belong to one sub-space. The attributes of a geo-tagged image include its coordinates, direction, visual features and sub-space ID.

In the spatial model, a graph G(S, E) is defined to represent the spatial adjacency relationship of sub-spaces in an indoor environment. G(S, E) includes a node set S along with an edge set E. Each node in S refers to a sub-space, and an edge between two nodes represents that the two nodes are spatially adjacent. As shown in Figure 5 (a), the whole space consists of five different sub-spaces. The spatial adjacency relationship of these sub-spaces can be represented as the graph in Figure 5 (b).

The attribute of a node in G(S, E) includes the sub-space ID, its neighbouring nodes, and the transfer image IDs. Here, transfer images represent the images located at the border of a sub-space. From one node to its neighbouring node, at least one transfer image will be passed through. The aim of defining the transfer image is to accurately determine the current sub-space of a moving smartphone. The current sub-space will not be changed to its neighbours until the current query image (from the smartphone) has been matched with a transfer image of the sub-space. For example, as shown in Figure 6, sub-space A and B are two spatially adjacent nodes. Each sub-space contains three transfer images located on its border. When the current image (from the smartphone camera) has been successfully matched with a transfer image from sub-space A, the following camera image will be matched with all geo-tagged images from sub-space A and B. The transformation (from A to B) will not be conducted until the current image has been matched with a non-transfer image from sub-space B. In this way, the determination of sub-space transformation can be more accuracy for indoor localization.

Spatial subdivision is a key issue in spatial modelling of geo-tagged images. Typically, a room or a corridor can be treated as a sub-space. However, for large buildings (e.g., shopping malls, museums or supermarkets), a continuous open space should also be divided into sub-spaces to reduce the time required for image searching and matching. Generally, if the sub-space ID of a query image is known, the efficiency of image matching will be increased when the spatial size of the sub-space is smaller. However, considering the localization error, it will be difficult to accurately determine the sub-space of a query image if the sub-spaces is too small. Therefore, it is important to determine a suitable spatial size for subdivision. A basic principle of spatial subdivision is to increase the visual similarity of geo-tagged images within the same sub-spaces. In this study, we tested the relationship between image similarity and spatial distance. The similarity of two images is defined as the number of matched feature point pairs:

$$VS(i,j) = \begin{cases} N(i,j) & if\ N(i,j) > N_0 \\ 0 & if\ N(i,j) < N_0 \end{cases} \quad (2)$$

where $VS(i,j)$ is the image similarity between images i and j, $N(i,j)$ is the number of matched feature point pairs between images i and j, and N_0 is a threshold. The two images are more similar to each other when they have a higher VS value. If $N(i,j)$ is smaller than N_0, the two images are not similar. In this study, N_0 is set to 20 according to the result of image matching testing.

To test the relationship between image similarity and spatial distance, fifteen sets of image sequences were collected in different types of indoor spaces, including a room, a corridor and an indoor open space. Each type of space has five sets of image sequences. The image sequences were collected in different areas in a building by a participant walking and using a smartphone. For each image sequence, the first image is matched with the following images on the sequence one by

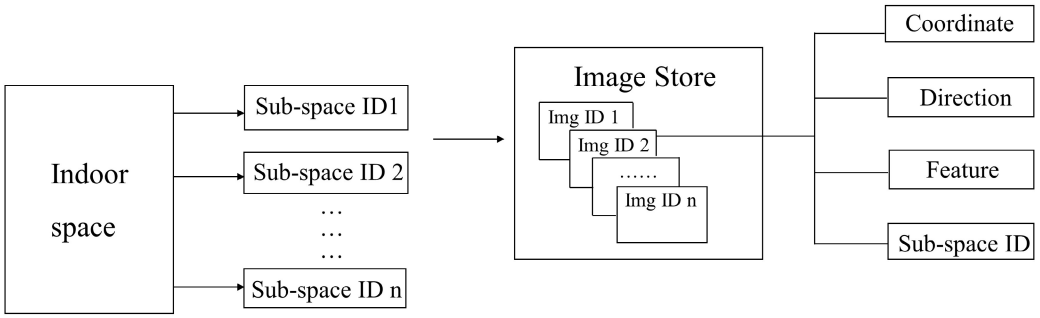


FIGURE 4. An example of spatial indexing for geo-tagged images.

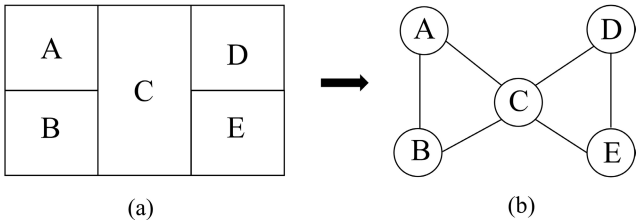


FIGURE 5. The spatial adjacency relationship of sub-spaces: (a) five sub-spaces in an indoor space; (b) spatial adjacency relationship of the sub-spaces.

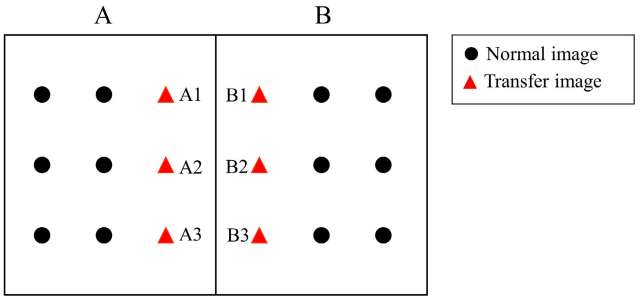


FIGURE 6. Transfer images of two adjacent sub-spaces.

one. Figure 7 shows an example of the relationship between image similarity and spatial distance under each condition. Figure 7 reveals a similar decreasing trend among the three types of spaces: a short spatial distance leads to an extremely large *VS* value; when distance increases, the *VS* value drops quickly and ultimately approaches zero. The rate of decrease varies in different environments. In addition, the *VS* value of the corridor is obviously smaller than those of the open space and room, possibly due to the relatively poor texture of the corridor area. According to the observations, in this study, some strategies have been designed for spatial subdivision and sub-space construction:

- The average length of the sub-space is set to 6 m for the corridor and 8 m for the other spaces in this study.
- The length of the sub-space can be appropriately increased (or reduced) for spaces with rich (or poor) texture.
- Separate space (e.g., a small room) can be given higher priority for sub-space construction.
- The common boundary of two adjacent sub-spaces should be as short as possible for better sub-space determination during indoor localization.

C. CONTINUOUS INDOOR VISUAL LOCALIZATION

1) SPATIAL CONSTRAINT-BASED IMAGE SEARCH STRATEGIES

This study intends to realize continuous visual localization when a person is walking indoors. The frequency of image matching-based visual localization depends on the spatial distribution density of geo-tagged images in an environment. This approach can provide high-frequency localization results when the spatial density of geo-tagged images is high (e.g., 0.5 or 1 m). However, for environments with a low spatial density of geo-tagged images (e.g., 10 m), the frequency of visual localization is insufficient to achieve continuous indoor localization. To increase the commonality, both visual and inertial data collected by smartphones are integrated in this algorithm to achieve continuous indoor localization. The location estimation results from visual and inertial sensors are termed the visual estimation and inertial estimation, respectively. An inertial estimation can be calculated continuously by using a PDR method. A visual estimation can be obtained whenever the current image is successfully matched with a geo-tagged image. In most cases, considering the accumulative error of PDR, a visual estimation is more reliable than an inertial estimation. Therefore, visual estimations can be used as landmarks to correct the accumulative error of PDR. However, the error of an inertial estimation will rapidly accumulate when image matching fails or the spatial density of geo-tagged images is low. The accumulative error will obviously reduce the accuracy of continuous indoor localization. To further improve the performance of this method, an SFM-based method is employed to estimate the heading angle of the PDR when a visual estimation cannot be obtained.

It is necessary to consider both the accuracy and efficiency of the image matching-based visual localization method. This method can achieve a high accuracy of localization results when image matching is successful. However, it may require a long time to match each video frame image collected by a

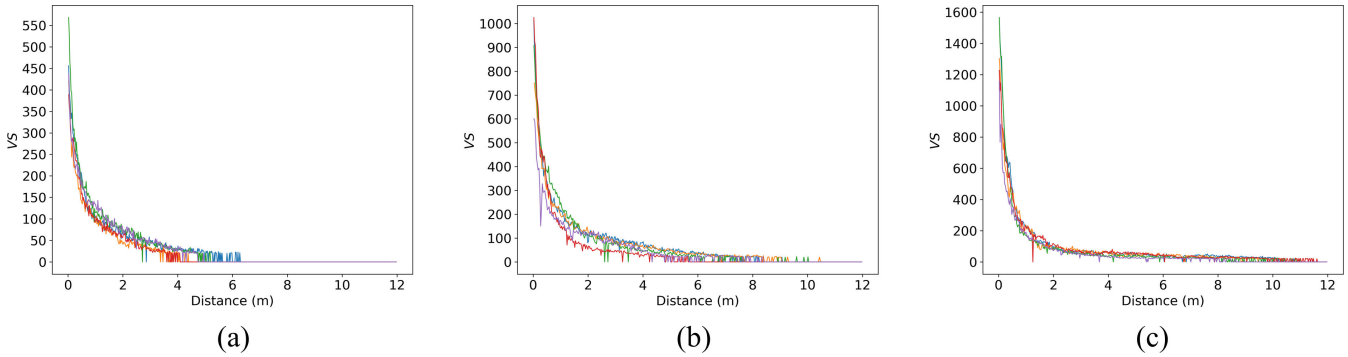


FIGURE 7. The relationship between the distance and spatial similarity in the three scenes: (a) corridor; (b) open space; (c) room.

smartphone camera (when one is walking) with geo-tagged images in the environment. In other words, the computing speed of image matching may be much higher than the walking speed of a person. To solve this problem, this method utilizes spatial constraint information to reduce the search space of image matching. The main idea is that the geo-tagged images spatially closer to the current image will be given higher priority during image matching. Two different image searching strategies are defined, including global search and local search.

Local search is used when the distance between the current location and the location of the last visual observation is smaller than a threshold D0, which is set to 5 m considering the size of the sub-space. It is not necessary to match the current image to all geo-tagged images in the environment. The geo-tagged images from the current sub-space that the person is in will be first used as candidates to implement a local search. These candidates are sorted by a variable C(i), which can be calculated as follows:

$$C\,(i) = \frac{|A_c - A_i|}{180} + \frac{D_i}{D_0} \tag{3}$$

where $C(i)$ represents the spatial difference between the current image and geo-tagged image i, A_c is the azimuth of the current image, A_i is the azimuth of the geo-tagged image i, and D_i is the distance between the current location and the location of the geo-tagged image i. Images with high $C(i)$ will be given low priority during the local search.

Global search is used when the current location is unknown or the distance between the current location and the location of the last visual observation is higher than the threshold D0. For example, when the localization algorithm begins (the initial location is assumed to be unknown) or when local search fails, it is necessary to match the current image (collected by camera) with images from all the sub-spaces in the environment. As shown in Figure 8 (a), the current image is first matched with one geo-tagged image from each sub-space to find the most possible sub-space of the current image. Then, the current image is matched with all geo-tagged images from this sub-space to find the best-fitting image. Specifically, when the local search fails, the global search begins from the

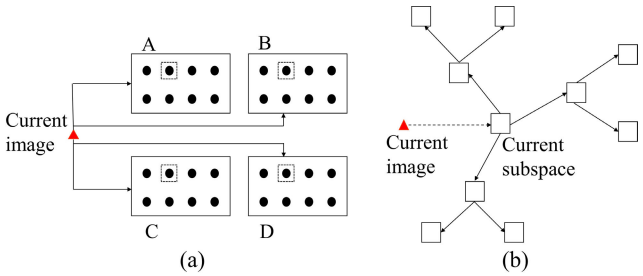


FIGURE 8. Spatial constraint-based visual search strategies: (a) global search; (b) local search.

neighbouring sub-spaces of the current sub-space according to graph G and then turn to their neighbouring subs-paces until a successful match is found (as shown in Figure 8 (b)).

2) CONTINUOUS VISUAL LOCALIZATION ALGORITHM

This algorithm integrates both visual and inertial estimations to achieve continuous indoor localization. Visual estimations can be obtained by using image matching based on the proposed image search strategies. A PDR method is utilized to calculate the inertial estimations that can increase the frequency of the localization algorithm. To improve the heading estimation performance of this algorithm, an SFM-based method is employed to estimate the PDR heading angle using video frames. A schematic diagram of the SFM-based method is shown in Figure 9. The grey and white blocks represent two matched geo-tagged images and a series of video frames, respectively. The smartphone camera is calibrated using the Matlab Camera Calibrator to estimate the parameters of the intrinsic matrix. The fundamental matrix F of adjacent frames can be calculated by the keypoint pairs computed before:

$$\left[u_i', v_i', 1\right] \begin{bmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{bmatrix} \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = 0 \tag{4}$$

where $m_i(u_i, v_i, 1)^T$ and $m_i'(u_i', v_i', 1)$ are the homogeneous keypoints of the matched keypoint set $\{m_i, m_i' | i = 1, 2, \ldots n\}$. Given eight or more pairs of matched keypoints, it is possible to linearly solve matrix F. After obtaining the fundamental

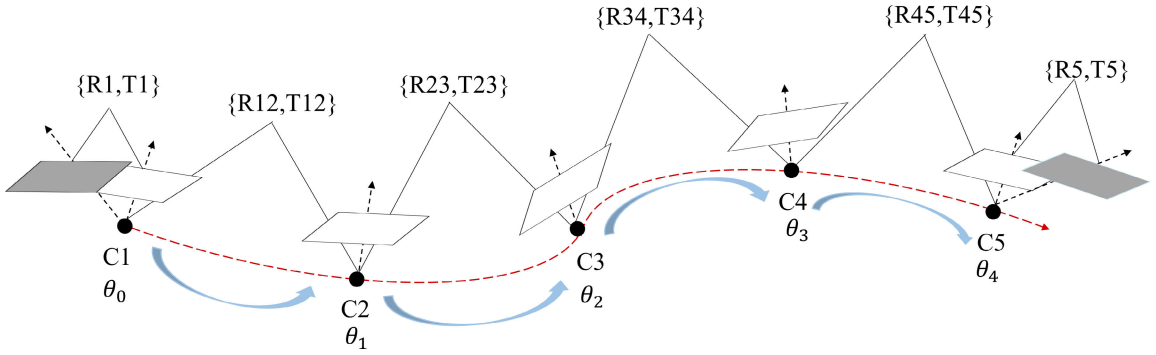


FIGURE 9. Details of the SFM-based heading angle estimation.

matrix, the essential matrix E can be calculated and decomposed to estimate the pose of the camera. The relationship between the fundamental matrix and the essential matrix can be described as follows:

$$E = K^T FK \tag{5}$$

where $\boldsymbol{K}$ is the intrinsic matrix of the camera of a smartphone. By utilizing singular value decomposition (SVD) of E, the rotation matrix R and translation vector T can be calculated. According to the rotation matrix R of the two adjacent images, the heading angle change can be expressed by:

$$R = \begin{bmatrix} cos\Delta\theta & 0 & sin\Delta\theta \\ sin\Delta\vartheta sin\Delta\theta & cos\Delta\vartheta & -sin\Delta\vartheta cos\Delta\theta \\ -cos\Delta\vartheta sin\Delta\theta & sin\Delta\vartheta & cos\Delta\vartheta cos\Delta\theta \end{bmatrix} \tag{6}$$

where $\Delta\theta$ is the heading angle change of the smartphone at instant t and $\Delta\vartheta$ is the pitch angle change of the smartphone at instant t. If the initial heading angle is θ_0, the heading angle of the smartphone at instant t can be calculated as:

$$\theta_t = \theta_0 + \sum_{i=1}^{t} \Delta\theta_i \tag{7}$$

where θ_t is the heading angle of the smartphone at instant t.

Although the SFM-based method can improve the heading estimation performance of PDR, the heading error still accumulates as the walking time increases. To solve this problem, this algorithm also uses geo-tagged images to eliminate the accumulation error of the heading angle. Once a video frame is successfully matched with a geo-tagged image, the heading angle of the geo-tagged image can be used to correct the heading angle of the smartphone:

$$\theta_t = \begin{cases} \theta_{t-1} + \Delta\theta\,(t-1,t) & if\ no\ match \\ \theta_g(t) - \Delta\theta_g\,(t) & if\ match\ success \end{cases} \tag{8}$$

where θ_t is the heading angle of smartphone at instant t, θ_{t-1} is the heading angle at instant $t-1$, $\Delta\theta\,(t-1,t)$ is the heading angle change from instant $t-1$ to t, and $\theta_g(t)$ is the heading angle of a successfully matched geo-tagged image at instant t. $\Delta\theta_g\,(t)$ is the heading angle change between the image at instant t and the geo-tagged image, which can be estimated by calculating the rotation matrix R of the two images.

Based on the heading estimation results, a PDR method can be used to continuously estimate the location of a smartphone. To improve the practicability of this approach, the initial location is assumed to be unknown for the PDR method. Thus, the estimation result of the PDR is the relative coordinates of a smartphone until a video frame has been successfully matched with a geo-tagged image. The location of the geo-tagged image can provide the absolute coordinates for a smartphone:

$$\begin{cases} x_t = x_g\,(t) + \sum_{i=k}^{t} d_i \cdot \cos\theta_i \\ y_t = y_g\,(t) + \sum_{i=k}^{t} d_i \cdot \sin\theta_i \end{cases} \tag{9}$$

where $(x_t,\ y_t)$ are the coordinates of the smartphone at instant t, $(x_g\,(t), y_g\,(t))$ represents the coordinates of the latest matched geo-tagged image until instant t, θ_i is the heading angle of the smartphone at instant i, and d_i is the distance between instant t-1 and t. Therefore, the coordinates of the smartphone will be corrected whenever a video frame image is successfully matched with a geo-tagged image. The accumulative error of PDR can be reduced continuously by using image matching. A peak detection algorithm is used for step detection. It compares the detected peak of the acceleration value with the preset threshold value. If the acceleration value is greater than the threshold value, the user is judged to have taken a step. A linear model is utilized to calculate step length L_k according to the variation in step frequency while walking, which is calculated as follows:

$$L_k = A + B * SF_k \tag{10}$$

where L_k is the length of the kth step of a trajectory, SF_k is the step frequency, and A and B are the parameters.

An algorithm has been developed to achieve continuous indoor localization, and the details of the algorithm are described as follows:

The input of the algorithm includes geo-tagged image dataset D, Graph G, the collected video frames and inertial data (from smartphone). Graph G represents the spatial adjacency relationship of sub-spaces. The collected video frames are matched with the geo-tagged images in dataset D based on the image searching strategies (described in Section III C).

Algorithm 1

input: Indoor geo-tagged image dataset: D (L, A, V, P), the four variables represent location, direction, image id and sub-space id attributes of geo-tagged image
input: The collected video frame up to current time t: $f_{1:t}$
input: The collected inertial sensor data up to current time t: $s_{1:t}$
input: Graph G(S, E) which represents the spatial adjacency relationship of sub-spaces
output: The location of each step
definition: E_{mn} = the Euclidean distance between the location m and n
definition: Z = the set of image id of geo-tagged images in global search
for f_a in $f_{1:t}$ do
  R, T= findEssentialMat(f_a, f_{a-1})
  $s_{t(f_a)}$ = Update Heading Angle Through R
    for img_i in D (L, A, V, P) do
      if V(img_i) $\notin$ Z || (f_{a-1} matched with normal image & $P_{(f_{a-1})} \neq P_{img_i}$)||$s_{t(f_a)}$−A($img_i$) > Threshold || ($f_{a-1}$ matched with transfer image & $P_{(f_{a-1})}$ not in S)
        then continue
      end if
      if f_a is matched with img_i
        L(f_a) = L(img_i)
        break
      end if
    end for
end for
$L_{original}$ = L(f_a)-$L_{t(f_a)\_pdr}$(start from the origin point)
$L_{0:t(f_a)\_pdr}$ = Update Position Through $L_{original}$
for f_b in $f_{b:t}$ do
  R, T= findEssentialMat(f_b, f_{b-1})
  $s_{t(f_b)}$ = Update Heading Angle Through R
  for img_j in D (L, A, V, P) do
    if (f_{b-1} matched with normal image & $P_{(f_{b-1})} \neq P_{img_j}$)||$s_{t(b)}$−A($img_j$) > Threshold || ($f_{b-1}$ matched with transfer image & $P_{(f_{b-1})}$ not in S)||$E_{L(img_j)L_{t(f_k)\_pdr}}$ > d(Δt)
      then continue
    end if
    if f_b is matched with img_j
      L(f_b) ← L(img_j)
      $L_{a:t(f_b)\_pdr}$ = Update Position Through L(f_b)
    end if
  end for
end for

The heading angle is calculated by using inertial readings and is further corrected by an SFM method. A PDR method is used to estimate the location of the smartphone. The location of the smartphone will be frequently corrected whenever a successful image matching result has been found. By using geo-tagged images along the path, the accumulative error can be reduced continuously, and the location accuracy can be improved.

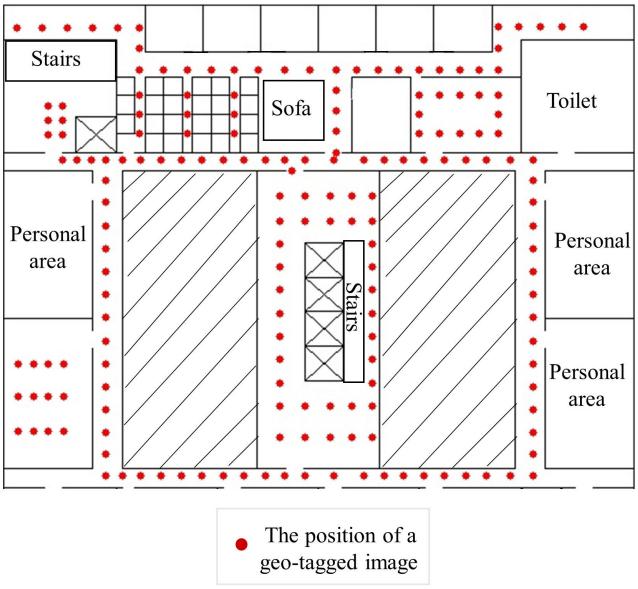


FIGURE 10. The floor plan of the experimental environment.

IV. EVALUATION

In this section, we evaluate the performance of the proposed visual localization method employing sensor data and video frames collected by smartphones.

A. EXPERIMENT SETUP

Two experiments were conducted in an office building with a floorplan of 52.5 m × 40 m, as shown in Figure 10. We collected a geo-tagged image dataset in the experimental area. The average spacing of image sampling points is approximately 2 m. At each sampling point, several images in different directions (interval of 90°) were collected and geo-tagged. There are 484 geo-tagged images in this dataset. As shown in Figure 10, the red dots represent the positions of geo-tagged images in the experiment area. The resolution of the all the geo-tagged images was resized to 640 × 480. The visual features of the geo-tagged images were extracted by using the multi-constrained image matching method. The geo-tagged images were organized and indexed according to the defined spatial model. The subdividing result of the experiment area is shown in Figure 11(a). There are 35 sub-spaces in the environment. Figure 11(b) shows the spatial adjacency relationship of the sub-spaces.

During the experiments, a smartphone (Xiaomi 8) was used to collect sensor data as a participant walked along three representative routes. The distance of the three testing trajectories are 55 m, 79 m and 80 m, respectively. The smartphone was held in front of the participant (keeping the camera forward facing), who walked at a normal pace. The collected video frames were also resized to 640 × 480 to increase the computation efficiency. The initial location of each route was assumed to be unknown for the localization algorithm. We set some markers at known coordinates along

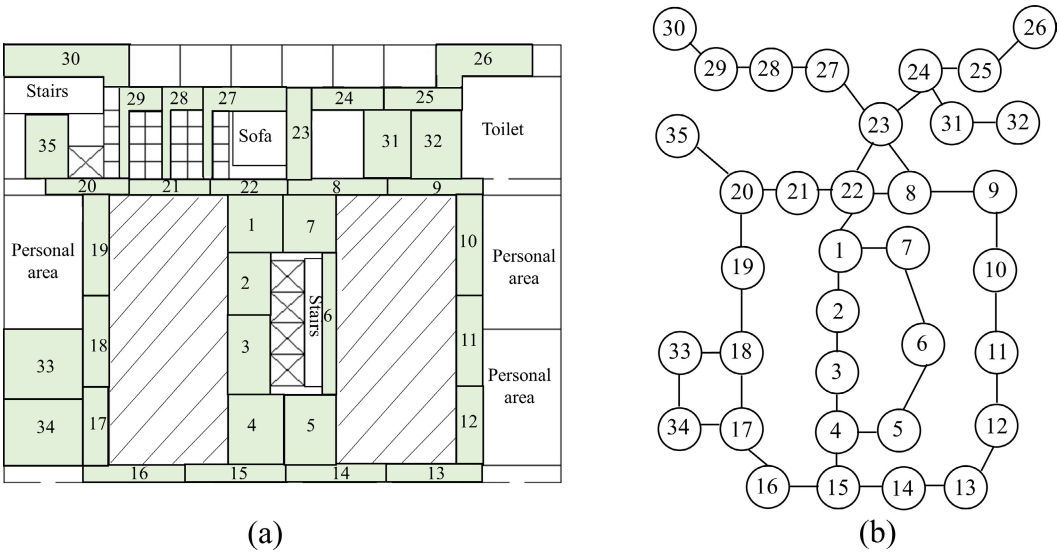


FIGURE 11. The spatial adjacency relationship of sub-spaces: (a) the sub-spaces in the experimental environment; (b) the spatial adjacency relationship of the sub-spaces.

the routes to collect the ground truth data. The collected data include video frames, acceleration data and gyroscope readings. The sampling frequencies of inertial and visual data are 100 Hz and 30 FPS, respectively. The localization experiments were performed on a laptop with an i7 9750H CPU (2.6 GHz).

The first experiment evaluated the effect of the proposed spatial model on reducing computation time. In the experiment, the collected video frames were matched with the images in the generated dataset to find the best-fitting results. The locations of successfully matched geo-tagged images were used as the localization outputs. The PDR and SFM methods were not employed in this experiment. We evaluated the localization accuracy and computation time of the image matching method using the proposed spatial model. The second experiment tested the performance of the continuous localization algorithm under two conditions: offline localization and online localization. During offline localization, the three trajectories were spatially recovered by the proposed algorithm. A SFM-based PDR method is employed to estimate the continuous location of the smartphone. The collected video frames were matched with the images in the dataset. All the successfully matched geo-tagged images were used as landmarks to correct the PDR location error. The online localization experiment is similar to the offline localization experiment. A main difference is that the successfully matched geo-tagged images can be used only to correct the current location of a smartphone and cannot be used to correct the previous localization result of a trip. The localization result was provided in an online manner. In this experiment, we also tested the influence of the spatial density of geo-tagged images on the localization performance.

B. PERFORMANCE OF THE SPATIAL MODEL OF GEO-TAGGED IMAGES

Computation time is a key consideration of visual localization. In this experiment, we tested the efficiency of the spatial model proposed in section III. Considering the high frequency of the smartphone camera (30 FPS), key frames (extracted from every 30 normal frames) were used as query images and were matched with the geo-tagged image dataset. The image matching efficiency of this model is compared to that of a normal image matching strategy and a Perceptual Hash algorithm (Phash) [49] based image matching strategy. The normal image matching strategy matches a query image to all geo-tagged images to find the best-fitting result. The Phash method generate a fingerprint string for each image by calculating the mean hash based on low frequency using Discrete Cosine Transform (DCT). The similarity between images can be determined by comparing the strings. The images judged to be dissimilar will not be matched with a query image.

The matching efficiency is evaluated by using the average computation time of each successfully matched query image from a trajectory. As shown in Figure 12, each dot represents the image matching time of the corresponding query image (key frame). The average computation times (of each query image) of the three trajectories are 32.6 s, 36.65 s and 42.54 s using a normal matching strategy. When using the Phash-based matching strategy, the average computation times of the three trajectories is 8.91 s, 8.84 s, 7.90 s, respectively. When using the proposed spatial model, the average computation times of the three trajectories decrease to 1.86 s, 1.69 s and 1.59 s, respectively. The matching efficiency can be significantly improved by optimizing the spatial organization of geo-tagged images.

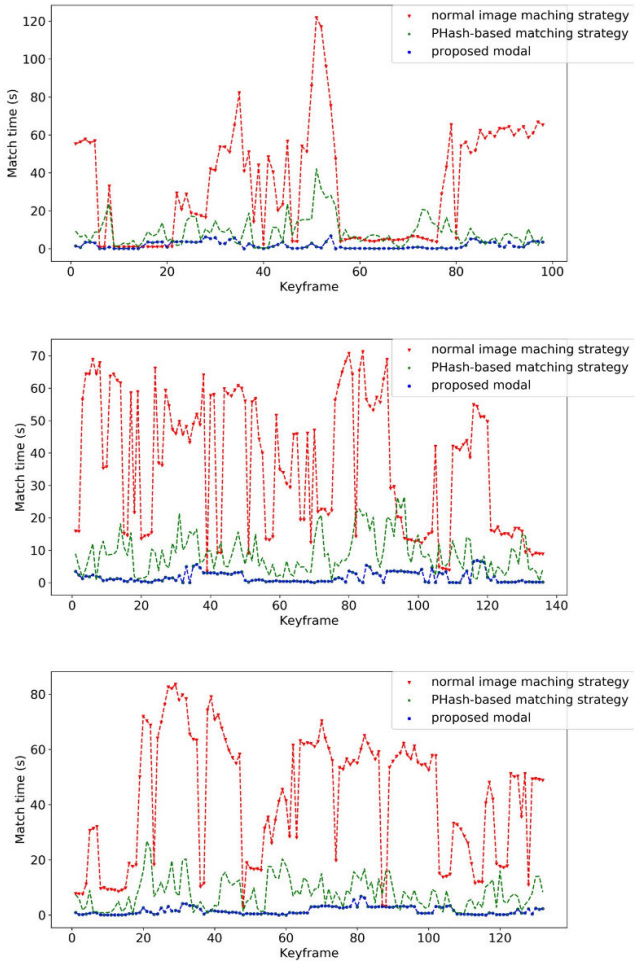


FIGURE 12. The computation time of each query image from three trajectories. Each dot represents the image matching time of a query image. Blue dot is the calculated matching time by using the spatial model. Red dot is the calculated matching time by using a normal matching strategy. Green dot is the calculated matching time by using a Phash-based matching strategy.

In addition to computation efficiency, we also evaluated the matching accuracy of the proposed model by using two indexes: matching accuracy and location error. Matching accuracy refers to the probability of correct matching, which can be calculated as follows:

$$Matching\ accuracy = \frac{M}{N} \cdot 100\% \tag{11}$$

where M is the number of correctly matched query images and N is the total number of query images from a trajectory. Here, the location error refers to the mean localization error of all query images from a trajectory. As shown in Table 1 (condition 1), the matching accuracies of the three trajectories are 73.33%, 82.92% and 64.44%. The location errors of the three trajectories are 0.40 m, 0.30 m and 0.35 m, respectively. The results indicate that the incorrect cases (query images) have been matched to the geo-tagged images that are spatially adjacent to the ground truth image. Although the matching accuracy may be affected by the visual similarity of adjacent geo-tagged images, this method can still achieve a high localization performance accuracy.

TABLE 1. Evaluation result of the image matching method based on the spatial model.

Trajectory	Condition 1		Condition 2	
	Matching accuracy	Location error (m)	Matching accuracy	Location error (m)
1	73.33%	0.40	76.67%	0.47
2	82.92%	0.30	73.17%	0.35
3	64.44%	0.35	62.22%	0.43

To further test whether the diversity of mobile devices affects the localization accuracy, another type of smartphone (ZUK Z1) was used to collect the video frames along the same three routes. The results are shown in Table 1 (condition 2). The matching accuracies of the three trajectories are 76.67%, 73.17% and 62.22%, and the location errors are 0.47 m, 0.35 m and 0.43 m, respectively. The results showed that the matching and localization performance was not obviously affected when using another type of smartphone. The influence of device diversity on localization accuracy is relatively small for visual positioning.

C. PERFORMANCE OF CONTINUOUS VISUAL LOCALIZATION

In Section IV B, we tested the accuracy of the image matching-based method. However, the localization result of image matching is not continuous. The frequency of the localization result is affected by the spacing or density of the geo-tagged images. In this section, we tested the performance of the continuous visual localization method proposed in Section III C. The same three trajectories were used as experimental data. The performance of this method was evaluated in two manners: offline localization and online localization. For offline localization, the localization result is provided after a whole trip is finished. All successfully matched geo-tagged images can be used to correct the localization error of a trajectory. For online localization, a successfully matched geo-tagged image can only be used to correct the current location of a smartphone but not the location error of the previous trip.

1) OFFLINE LOCALIZATION PERFORMANCE

The offline localization performance is shown in Figure 13. The average offline localization error of all the trajectories is approximately 0.73 m. As shown in Figure 13, the localization errors of trajectories 1, 2 and 3 are 0.84 m, 0.54 m and 0.83 m, respectively. These errors are slightly higher than the results shown in Table 1 (image matching-based method). The reason for this difference is that an SFM-based PDR method (defined in Section III C) is employed to estimate the continuous location of the smartphone. Although the accu-

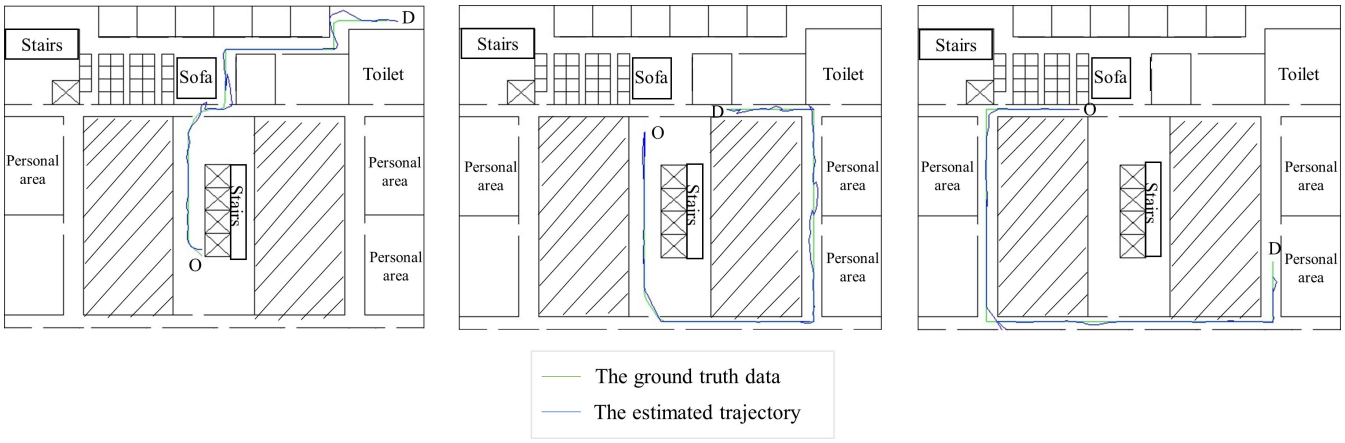


FIGURE 13. The estimation results of three trajectories.

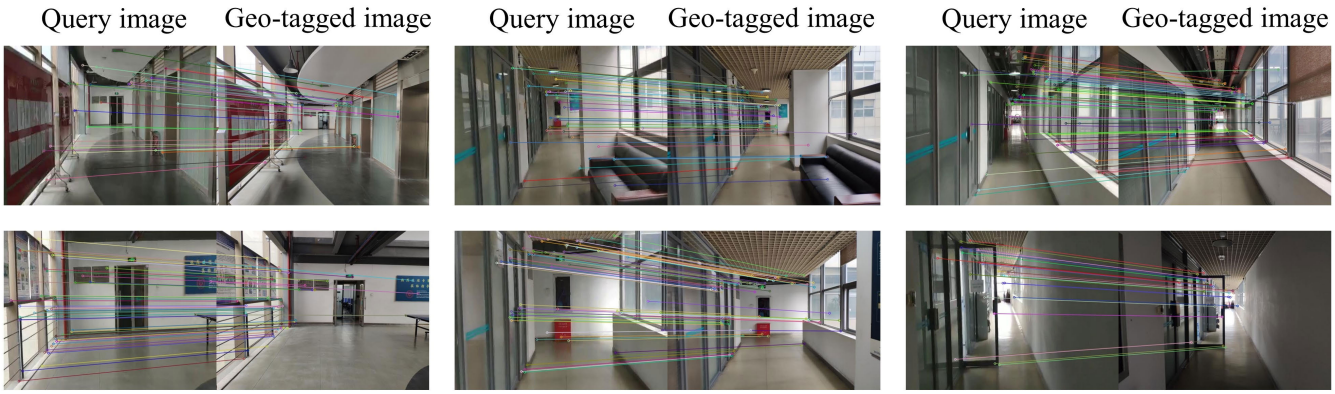


FIGURE 14. Examples of image matching along the three trajectories.

racy is slightly reduced, this method can achieve a continuous localization result but not separate query locations. The average spacing of sampling points from the trajectories decreases from 2 m to 0.5 m by using this method. The estimated trajectories can be seen from Figure 13. The locations of the three estimated trajectories are very close to the ground truth data.

Figure 14 shows some examples of image matching along the trajectories. Most of the query images from the trajectories have been successfully matched to the correct geo-tagged images. The six pairs of query and geo-tagged images are visually similar. The location of the matched geo-tagged images was used to correct the location of the query images from the smartphone. A small location error arose because the location of a query might not be completely the same as the matched geo-tagged image. The error was mainly determined by the spatial density of the geo-tagged images. Here, the error was considered and included in the mean location accuracy.

2) ONLINE LOCALIZATION PERFORMANCE

The performance of online localization was evaluated under four different conditions: (1) the PDR method using map constraint information; (2) visual localization using PDR; (3) visual localization using SFM-based PDR; (4) visual localization using SFM-based PDR under a random camera opening condition. In the first condition, a normal PDR method was used to estimate the location of the smartphone using inertial reading from the gyroscope and accelerometer. The initial location of the smartphone was assumed to be already known for this condition. In addition, map information was employed to correct the location error of PDR. In the section condition, the proposed visual localization method (without the SFM-based PDR method) was used. A normal PDR method was also conducted to realize continuous localization when visual estimations were not obtained. In the third condition, the proposed visual localization method (with the SFM-based PDR method) was used. In the fourth condition, the proposed visual localization method was used under a random camera opening condition. It was assumed that the smartphone camera might not remain open all the time; users may open smartphone cameras to start a localization process and close the camera when they have obtained enough location information. This condition was used to simulate a more practical situation in which a user might randomly open or close the smartphone camera. When the camera was opened,

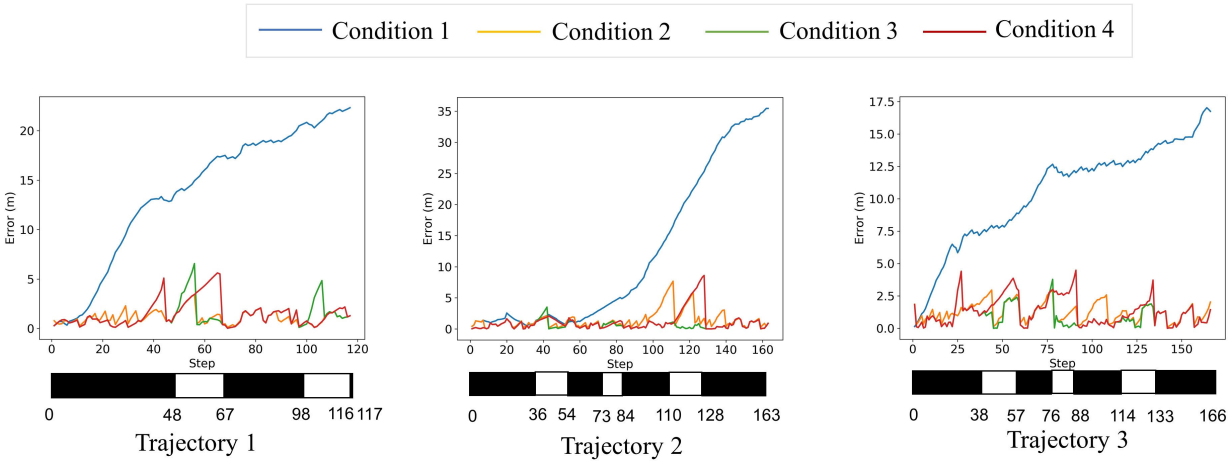


FIGURE 15. The localization errors of three trajectories under different conditions. In the bar below, black means the camera is open, and white means the camera is closed.

the SFM-based visual localization method was used to realize continuous localization. When the camera was closed, a PDR method was employed to estimate the location of the smartphone (from the last visual estimation) until the camera was opened again.

Figure 15 shows the performance of online localization under different conditions. Generally, the average location error (of three trajectories) under each condition is 12.05 m, 1.35 m, 0.95 m and 1.31 m, respectively. The errors under conditions 2-4 are clearly smaller than the error under condition 1. This indicates that the accumulative error of image matching-based visual localization is considerably smaller than that of the PDR method. The error of condition 3 is smaller than that of condition 2, showing that the employed SFM-based PDR method can further improve the heading estimation error of PDR. In addition, condition 4 achieves a relatively high location accuracy, even under a random camera opening condition.

In this experiment, the location results are provided in an online manner. The image matching and SFM-based PDR methods were performed continuously. The total computation times of the three trajectories (i.e., condition 3) are 95.89 s, 82.11 s and 72.24 s, respectively, which are shorter than the corresponding walking times of the trajectories (98 s, 136 s and 132 s). To further evaluate the efficiency of this algorithm, we calculated the average computation time needed to provide each location estimation (e.g., a step) during a trip. The average computation time of three trajectories is 0.59 s. The average computation time of successful image matching-based estimation was also considered and referred to the average time needed to provide a location result from successful image matches. A shorter image matching time means that the location of the geo-tagged images can be used to correct the cumulative error of PDR more frequently. In this experiment, the average times for (successful) image matching for the three trajectories are 3.46 s, 2.52 s and 2.35 s, respectively. This indicates that the smartphone location error can be frequently corrected by geo-tagged images.

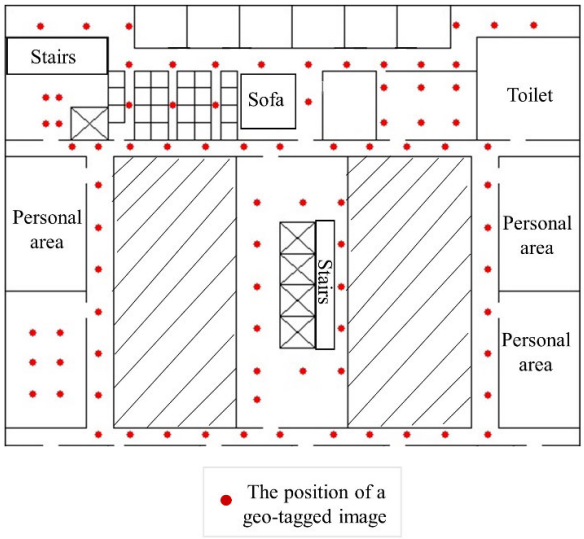


FIGURE 16. A geo-tagged image dataset with a low spatial density.

The spatial density of geo-tagged images is another factor that may affect the performance of visual localization. A higher density of geo-tagged images may improve the accuracy of visual localization. However, it will also increase the workload for image collection. In this experiment, we also tested the influence of the spatial density of geo-tagged images on localization performance. As shown in Figure 16, some geo-tagged images have been removed from the original image dataset. There are 218 images in the new low-density dataset. The spatial density of the geo-tagged images is calculated as the mean number of images per square metre. The spatial density results of the images in the original and low-density datasets are 0.23 and 0.10, respectively. We tested both the online and offline localization performance by using the low-density dataset. As shown in Table 2, the online localization errors of the three trajectories are 1.84 m, 1.24 m and 1.20 m, slightly higher than the errors of the original dataset (1.36 m, 0.61 m,

TABLE 2. The online and offline localization performance of two datasets.

Mean error (m)	Online		Offline	
	Low-density dataset	Original dataset	Low-density dataset	Original dataset
Trajectory 1	1.84	1.36	1.28	0.84
Trajectory 2	1.24	0.61	0.90	0.54
Trajectory 3	1.20	0.89	0.85	0.83

0.89 m). Similarly, the offline localization error of the low-density dataset is also slightly higher than that of the original dataset. The results revealed that a higher spatial density of geo-tagged images can improve the localization accuracy of the proposed visual localization method. However, the localization accuracy of the low-density datasets is still acceptable for many indoor localization applications. For large indoor spaces, it is practical to reduce the spatial density of geo-tagged images for fast system deployment.

V. DISCUSSION

This study proposes a continuous visual localization approach that can realize accurate indoor localization. The localization errors of image querying, continuous offline localization and online localization are approximately 0.4 m, 0.7 m and 0.9 m, respectively. This approach can achieve an accuracy of 1.3 m, even under a random camera opening condition. An advantage of this method is that successfully matching geo-tagged images can frequently reduce the accumulative error of PDR. In other words, each a matched geo-tagged image can serve as a landmark in indoor environments. Compared to common indoor landmarks (e.g., elevators, stairs or intersections), geo-tagged images can be taken from most positions in an indoor environment, which greatly increases the number of available landmarks for accurate indoor localization. In addition, this method can provide a continuous localization result with a relatively short computation time (the average time for a location result is approximately 0.59 s), which makes it more suitable for wayfinding or navigation applications.

The computation efficiency is an essential bottleneck for image matching-based visual localization. Higher frequency image matching requires more computation time, which reduces the practicality of visual localization. This study proposes a spatial model to optimize the spatial organization of geo-tagged images. Several spatial constraint-based image searching strategies have also been designed to further reduce the computation time for continuous visual localization. According to the experimental results (experiment B), the average computation time of each successfully matched frame is approximately 1.8 s, which is obviously smaller than that of a normal matching strategy (approximately 40 s). For online localization (experiment C), the average computation time of each trajectory is shorter than the corresponding walking time. This demonstrates that the spatial organization of geo-tagged images and image searching strategies are important for improving the efficiency of image matching-based visual localization.

The average computation time of online localization has also been considered in this study. The average computation time (i.e., the average time needed to provide a location estimation result) is approximately 0.59 s according to the experimental results (experiment C). This is clearly a shorter time than the average time needed for successful image matching-based location estimation (approximately 2.78 s). Consequently, this approach employs an SFM-based PDR method to reduce the average computation time of image matching-based localization. A considerable part of the calculation process of the SFM-based method (e.g., keypoint extraction) is included in the process of the image matching-based algorithm. Thus, including the SFM method will not obviously increase the computation time but can improve both the computation efficiency and accuracy of the localization algorithm.

A limitation of visual localization is that it requires the camera to remain open. Compared to other indoor localization solutions (e.g., Wi-Fi localization), it increases the usage requirement of an extra device (e.g., a user's smartphone). However, considering the relatively high accuracy of visual localization, it is more suitable for localization applications that require stable localization accuracy (navigation services for visually impaired people) than for those that do not have such a requirement. The visual localization results can also be integrated with an augmented reality (AR) application to provide visually augmented localization and navigation guides, especially helpful for people with a poor sense of direction. In addition, both the PDR and SFM methods are integrated into the algorithm. The experimental results showed that the algorithm can achieve a relatively high accuracy (approximately 1.3 m), even with a camera that randomly closed. When the camera was closed, the PDR method continuously estimated the location of the smartphone. After the camera was opened again, the successfully matched geo-tagged images served as landmarks to correct the accumulative error of PDR. In addition, visual localization does not require extra environmental infrastructure, which reduces the difficulty of system deployment. Another limitation of the proposed visual localization method is that it requires a collected geo-tagged image dataset. In further work, we intend to develop an efficient image collecting and geo-tagging method for visual localization based on our previous work [22], which proposes a crowdsourcing-based approach to geo-tag collected data (Wi-Fi, image, noise, PM 2.5, etc.) in indoor spaces. In this way, the visual localization method can be efficiently deployed in various indoor environments.

VI. CONCLUSION

In this paper, we proposed an efficient visual localization method that can achieve continuous indoor localization without auxiliary infrastructure or previous knowledge of the initial location. The experimental results showed that the

matching of geo-tagged images can continuously correct the accumulative error of PDR. To reduce the computation time required for frequent image matching, a spatial model was designed to optimize the spatial organization of geo-tagged images. We also developed a spatial constraint-based visual localization algorithm. The experimental results showed that the computation time of image matching and continuous visual localization can be considerably reduced by adopting a spatial model and spatial constraint-based image searching strategies. The offline and online localization experiments revealed a relatively high location accuracy (approximately 0.7 m and 0.9 m, respectively). In future work, we intend to develop an efficient indoor image collecting and geo-tagging method to reduce the system deployment workload of visual localization.

REFERENCES

[1] C.-H. Lim, Y. Wan, B.-P. Ng, and C.-M. See, "A real-time indoor WiFi localization system utilizing smart antennas," *IEEE Trans. Consum. Electron.*, vol. 53, no. 2, pp. 618–622, Jul. 2007.

[2] Y. Zhuang, J. Yang, Y. Li, L. Qi, and N. El-Sheimy, "Smartphone-based indoor localization with Bluetooth low energy beacons," *Sensors*, vol. 16, no. 5, p. 596, 2016.

[3] M. Hazas and A. Hopper, "Broadband ultrasonic location systems for improved indoor positioning," *IEEE Trans. Mobile Comput.*, vol. 5, no. 5, pp. 536–547, May 2006.

[4] A. Athalye, V. Savic, M. Bolic, and P. M. Djuric, "Novel semi-passive RFID system for indoor localization," *IEEE Sensors J.*, vol. 13, no. 2, pp. 528–537, Feb. 2013.

[5] A. De Angelis, S. Dwivedi, and P. Handel, "Characterization of a flexible UWB sensor for indoor localization," *IEEE Trans. Instrum. Meas.*, vol. 62, no. 5, pp. 905–913, May 2013.

[6] K. P. Subbu, B. Gozick, and R. Dantu, "LocateMe: Magnetic-fields-based indoor localization using smartphones," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 4, p. 73, Sep. 2013.

[7] N. Snavely, S. M. Seitz, and R. Szeliski, "Modeling the world from Internet photo collections," *Int. J. Comput. Vis.*, vol. 80, no. 2, pp. 189–210, Nov. 2008.

[8] J.-Y. Huang, S.-H. Lee, and C.-H. Tsai, "A fast image matching technique for the panoramic-based localization," in *Proc. IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. (ICIS)*, Okyaman, Japan, Jun. 2016, pp. 1–6.

[9] S. Lowry, N. Sunderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, and M. J. Milford, "Visual place recognition: A survey," *IEEE Trans. Robot.*, vol. 32, no. 1, pp. 1–19, Feb. 2016.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understand.*, vol. 110, no. 3, pp. 346–359, Jun. 2008.

[12] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[13] T. Sattler, B. Leibe, and L. Kobbelt, "Fast image-based localization using direct 2D-to-3D matching," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 667–674.

[14] A. Canedo-Rodríguez, V. Alvarez-Santos, C. V. Regueiro, R. Iglesias, S. Barro, and J. Presedo, "Particle filter robot localisation through robust fusion of laser, WiFi, compass, and a network of external cameras," *Inf. Fusion*, vol. 27, pp. 170–180, Jan. 2016.

[15] W. Yuan, Z. Li, and C.-Y. Su, "RGB-D sensor-based visual SLAM for localization and navigation of indoor mobile robot," in *Proc. Int. Conf. Adv. Robot. Mechatronics (ICARM)*, Macau, China, Aug. 2016, pp. 82–87.

[16] M. Werner, M. Kessel, and C. Marouane, "Indoor positioning using smartphone camera," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat.*, Guimaraes, Portugal, Sep. 2011, pp. 1–6.

[17] D. Wu, R. Chen, and L. Chen, "Visual positioning indoors: Human eyes vs. smartphone cameras," *Sensors*, vol. 17, no. 11, p. 2645, 2017.

[18] H. Sadeghi, S. Valaee, and S. Shirani, "A weighted KNN epipolar geometry-based approach for vision-based indoor localization using smartphone cameras," in *Proc. IEEE 8th Sensor Array Multichannel Signal Process. Workshop (SAM)*, A Coruña, Spain, Jun. 2014, pp. 37–40.

[19] A. Xiao, R. Chen, D. Li, Y. Chen, and D. Wu, "An indoor positioning system based on static objects in large indoor scenes by using smartphone cameras," *Sensors*, vol. 18, no. 7, p. 2229, 2018.

[20] T. Wu, J. Liu, Z. Li, K. Liu, and B. Xu, "Accurate smartphone indoor visual positioning based on a high-precision 3D photorealistic map," *Sensors*, vol. 18, no. 6, p. 1974, 2018.

[21] Y. Zhou, X. Zheng, R. Chen, H. Xiong, and S. Guo, "Image-based localization aided indoor pedestrian trajectory estimation using smartphones," *Sensors*, vol. 18, no. 1, p. 258, 2018.

[22] T. Liu, X. Zhang, Q. Li, and Z. Fang, "A visual-based approach for indoor radio map construction using smartphones," *Sensors*, vol. 17, no. 8, p. 1790, 2017.

[23] T. Liu, X. Zhang, Q. Li, Z. Fang, and N. Tahir, "An accurate visual-inertial integrated geo-tagging method for crowdsourcing-based indoor localization," *Remote Sens.*, vol. 11, no. 16, p. 1912, 2019.

[24] J. Son, S. Kim, and K. Sohn, "A multi-vision sensor-based fast localization system with image matching for challenging outdoor environments," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8830–8839, Dec. 2015.

[25] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz, "Accurate geo-registration by ground-to-aerial image matching," in *Proc. 2nd Int. Conf. 3D Vis.*, Tokyo, Japan, Dec. 2014, pp. 525–532.

[26] D. M. Bradley, R. Patel, N. Vandapel, and S. M. Thayer, "Real-time image-based topological localization in large outdoor environments," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Edmonton, AB, Canada, Aug. 2005, pp. 3670–3677.

[27] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele, "Real-time image-based 6-DOF localization in large-scale environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1043–1050.

[28] C. Arth, C. Pirchheim, J. Ventura, D. Schmalstieg, and V. Lepetit, "Instant outdoor localization and SLAM initialization from 2.5D maps," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 11, pp. 1309–1318, Nov. 2015.

[29] F. Walch, C. Hazirbas, L. Leal-Taixe, T. Sattler, S. Hilsenbeck, and D. Cremers, "Image-based localization using LSTMs for structured feature correlation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 627–637.

[30] G. Yanga and Y. Liang, "An indoor localization method of image matching based on deep learning," in *Proc. 2nd Int. Conf. Mech., Electron., Control Automat. Eng. (MECAE)*, Mar. 2018, pp. 103–108.

[31] W. Sui and K. Wang, "An accurate indoor localization approach using cellphone camera," in *Proc. 11th Int. Conf. Natural Comput. (ICNC)*, Zhangjiajie, China, Aug. 2015, pp. 949–953.

[32] S. Xu, W. Chou, and H. Dong, "A robust indoor localization system integrating visual localization aided by CNN-based image retrieval with Monte Carlo localization," *Sensors*, vol. 19, no. 2, p. 249, 2019.

[33] I. Ha, H. Kim, S. Park, and H. Kim, "Image retrieval using BIM and features from pretrained VGG network for indoor localization," *Building Environ.*, vol. 140, pp. 23–31, Aug. 2018.

[34] R. Gao, Y. Tian, F. Ye, G. Luo, K. Bian, Y. Wang, T. Wang, and X. Li, "Sextant: Towards ubiquitous indoor localization service by photo-taking of the environment," *IEEE Trans. Mobile Comput.*, vol. 15, no. 2, pp. 460–474, Feb. 2016.

[35] J. Choi and G. Friedland, *Multimodal Location Estimation of Videos and Images*. Cham, Switzerland: Springer, 2015.

[36] K. Wan, L. Ma, and X. Tan, "An improvement algorithm on RANSAC for image-based indoor localization," in *Proc. Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Paphos, Cyprus, Sep. 2016, pp. 842–845.

[37] J. Qian, J. Ma, R. Ying, P. Liu, and L. Pei, "An improved indoor localization method using smartphone inertial sensors," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat.*, Montbeliard-Belfort, France, Oct. 2013, pp. 1–7.

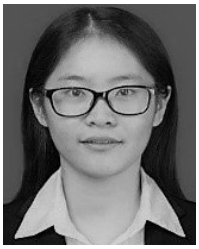
[38] R. Zhang, F. Hoflinger, and L. Reindl, "Inertial sensor based indoor localization and monitoring system for emergency responders," *IEEE Sensors J.*, vol. 13, no. 2, pp. 838–848, Feb. 2013.

[39] B. Zhou, Q. Li, Q. Mao, W. Tu, and X. Zhang, "Activity sequence-based indoor pedestrian localization using smartphones," *IEEE Trans. Human-Machine Syst.*, vol. 45, no. 5, pp. 562–574, Oct. 2015.

[40] B. Zhou, Q. Li, Q. Mao, W. Tu, X. Zhang, and L. Chen, "ALIMC: Activity landmark-based indoor mapping via crowdsourcing," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2774–2785, Oct. 2015.

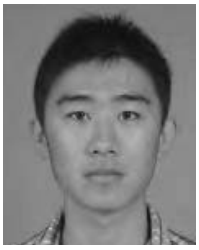
[41] Z. Chen, Q. Zhu, and Y. C. Soh, "Smartphone inertial sensor-based indoor localization and tracking with iBeacon corrections," *IEEE Trans. Ind. Informat.*, vol. 12, no. 4, pp. 1540–1549, Aug. 2016.
[42] X. Wu, R. Shen, L. Fu, X. Tian, P. Liu, and X. Wang, "IBILL: Using iBeacon and inertial sensors for accurate indoor localization in large open areas," *IEEE Access*, vol. 5, pp. 14589–14599, 2017.
[43] B. Klingner, D. Martin, and J. Roseborough, "Street view motion-from-structure-from-motion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 953–960.
[44] M. A. Fonstad, J. T. Dietrich, B. C. Courville, J. L. Jensen, and P. E. Carbonneau, "Topographic structure from motion: A new development in photogrammetric measurement," *Earth Surf. Processes Landforms*, vol. 38, no. 4, pp. 421–430, Mar. 2013.
[45] M. J. Westobya and J. Brasington, "'Structure-from-motion' photogrammetry: A low-cost, effective tool for geoscience applications," *Geomorphology*, vol. 179, pp. 300–314, Dec. 2012.
[46] J. Dong, Y. Xiao, M. Noreikis, Z. Ou, and A. Ylä-Jääski, "iMoon: Using smartphones for image-based indoor navigation," in *Proc. ACM Conf. Embedded Netw. Sensor Syst.*, Seoul, South Korea, Nov. 2015, pp. 85–97.
[47] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, vol. 2, 1999, p. 1150.
[48] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.
[49] J. Oostveen, A. A. C. Kalker, and J. A. Haitsma, "Visual hashing of digital video: Applications and techniques," *Proc. SPIE*, vol. 4472, pp. 121–131, Dec. 2001.

XING ZHANG received the B.E. and Ph.D. degrees in geographic information science from Wuhan University, Wuhan, China. He is currently with the Guangdong Key Laboratory of Urban Informatics, Shenzhen University. His research interests include pedestrian navigation, indoor localization, ubiquitous computing, and intelligent transportation.



JING LIN received the B.E. degree in geodesy and survey engineering from the Central South University of Forestry and Technology, Changsha, China. She is currently pursuing the master's degree with the Guangdong Key Laboratory of Urban Informatics, Shenzhen University. Her research interests include pedestrian navigation, indoor localization, and intelligent transportation.

QINGQUAN LI received the Ph.D. degree in geographic information system (GIS) and photogrammetry from the Wuhan Technical University of Surveying and Mapping, Wuhan, China, in 1998. He is currently a Professor with Shenzhen University, Guangdong, China, and Wuhan University, Wuhan. His research interests include 3-D and dynamic data modeling in GIS, location-based service, surveying engineering, integration of GIS, global positioning systems and remote sensing, intelligent transportation systems, and road surface checking.



TAO LIU received the Ph.D. degree in geodesy and survey engineering from Wuhan University, Wuhan, China, in 2017. He is currently a Lecturer with the College of Resources and Environment, Henan University of Economics and Law, China. His research interests include indoor localization and mapping, mobile computing, and rural and urban planning.

ZHIXIANG FANG received the M.Sc. and Ph.D. degrees from Wuhan University, in 2002 and 2005, respectively. He is currently a Professor with the State Key Laboratory for Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University. His research interests include spatial-temporal modeling of human behavior, space-time GIS for transport, and intelligent navigation.