Hindawi
Journal of Sensors
Volume 2017, Article ID 8037607, 18 pages
https://doi.org/10.1155/2017/8037607

Research Article

Visual Map Construction Using RGB-D Sensors for Image-Based Localization in Indoor Environments

Guanyuan Feng,[1,2] Lin Ma,[1] and Xuezhi Tan[1]

[1] *Communication Research Center, Harbin Institute of Technology, Harbin 150080, China*
[2] *Chair of Media Technology, Technical University of Munich, 80333 Munich, Germany*

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

Received 8 June 2017; Revised 27 August 2017; Accepted 13 September 2017; Published 18 October 2017

Academic Editor: Calogero M. Oddo

RGB-D sensors capture RGB images and depth images simultaneously, which makes it possible to acquire the depth information at pixel level. This paper focuses on the use of RGB-D sensors to construct a visual map which is an extended dense 3D map containing essential elements for image-based localization, such as poses of the database camera, visual features, and 3D structures of the building. Taking advantage of matched visual features and corresponding depth values, a novel local optimization algorithm is proposed to achieve point cloud registration and database camera pose estimation. Next, graph-based optimization is used to obtain the global consistency of the map. On the basis of the visual map, the image-based localization method is investigated, making use of the epipolar constraint. The performance of the visual map construction and the image-based localization are evaluated on typical indoor scenes. The simulation results show that the average position errors of the database camera and the query camera can be limited to within 0.2 meters and 0.9 meters, respectively.

1. Introduction

The emergence of wireless communication and the Global Positioning System (GPS) has ignited the idea of Personal Navigation Systems (PNSs). A PNS has the positioning and navigation functions to provide location information to individuals via portable devices. In outdoor environments, GPS sensors are the most common and efficient navigation instruments. For indoor spaces, however, because satellite signals cannot penetrate buildings, a smartphone with a GPS sensor is incapable of providing reliable position services to a pedestrian. Therefore, indoor localization methods that do not depend on satellite signals have become a research hotspot in recent years.

In the recent decade, many indoor localization algorithms have been proposed and implemented. Of these algorithms, most have concentrated on radio signal-based methods, making use of radio access networks. These algorithms are mainly based on measuring the distances to access points by means of the angle of arrival (AOA), the time of arrival (TOA), the carrier phase of arrival (POA), or the received signal strength indicator (RSSI), and so forth [1, 2]. With the increase of Wi-Fi access points deployed in public areas, researchers have put focus on Wi-Fi-based indoor localization [3–5]. However, the localization system based on Wi-Fi signals usually cannot provide users with stable results because moving objects or persons affect the signal strength. In addition, Bluetooth beacons are also used for indoor localization [6–8]. The effective range of Bluetooth signals is approximately 5–10 meters, which leads to frequent beacon switching and results in high power consumption.

For Wi-Fi or Bluetooth-based localization systems, expensive infrastructure needs to be installed and maintained. An evident disadvantage of these localization systems is that Wi-Fi hotspots or Bluetooth beacons as anchor points cannot be moved after the initial calibration. Moreover, if high localization accuracy is required by indoor navigation systems, for example, submeter accuracy, Wi-Fi hotspots or Bluetooth beacons must be densely distributed in public areas [9]. However, in some public buildings, such as railway stations and airports, high-density infrastructures are costly and inconvenient to achieve. Due to these problems, an economical and stable indoor localization method is needed to provide users with quality location-based services.

The emergence of vision-based methods has made it possible to achieve good performance in indoor localization in an economical and practical way. Therefore, more and more studies have focused on the field of localization and navigation based on visual methods using smartphones [9–17]. Vision-based localization takes advantage of image retrieval to search for the similar geo-tagged database images for a query image, and then the location of the query image can be estimated. In the phase of visual localization, the homography matrix [10] and the epipolar constraint [11] are typically used to calculate the location of the query images based on the positions of the matched database images. Therefore, the performance of the visual localization system depends on the position accuracy of the database images. However, some existing visual localization methods [11–14] put their focus on the pose estimation of the query camera and do not mention the method of pose acquisition for the database cameras. In these methods, the database images and their positions are captured and recorded manually. However, in this manner, the acquisition of the database images and their positions is a heavy burden in the phase of database generation, especially for large-scale indoor spaces. TUMindoor [15] is a representative indoor database automatically created by a mapping trolley for visual location recognition [16, 17]. For the TUMindoor database, data acquisition equipment including cameras and laser scanners is triggered roughly every 1.2 meters, but this manner is inapplicable for handheld RGB-D sensors because the walking distance of the operator is difficult to measure.

A practical way to collect the pose of the database camera is by Simultaneous Localization and Mapping (SLAM) technology, which is widely used in the field of robot path planning and indoor 3D reconstruction. In vision-based SLAM, stereo camera pairs or a monocular camera is used to estimate the 3D structure and simultaneously deduce the pose of the cameras [18–21]. In stereo SLAM system [22–25], the 3D points are triangulated for every stereo pair, and the relative motion of the camera pair is estimated by a 3D-to-3D feature registration. In contrast to stereo SLAM, in monocular SLAM systems [26–29], the 3D structure of the scene and the relative motion of the camera are computed by 2D bearing data.

In recent years, with the emergence of consumer-level RGB-D sensors, RGB and depth cameras are utilized cooperatively to build dense 3D maps of indoor environments [30–33]. Compared with previous SLAM methods, RGB-D SLAM is able to preserve the full structural features of indoor scenes and present a superior dense 3D map. In RGB-D SLAM systems, the Iterative Closest Point (ICP) [34] algorithm is a typical method to align the sequential point clouds captured by depth sensors. Then, the aligned point clouds are used to estimate the consecutive frame status (the rotation matrix and the translation vector) and generate the 3D map [32]. In addition, joint optimization combining visual feature matching and shape-based registration is employed to improve the accuracy of the 3D map [33]. The advantage of RGB-D SLAM is significant, as it recovers indoor structures adequately and places the visual features at their position in the map. The existing algorithms of RGB-D indoor mapping pay more attention to indoor scene recovery, but fewer of the algorithms focus mainly on the optimization of the RGB-D camera poses. Moreover, hardly any algorithms optimize the indoor 3D map for visual localization.

Therefore, in this paper, a novel method of visual map construction is proposed, and on the basis of the visual map, image-based localization is investigated to estimate the location of the query camera. Kinect [35] is a popular sensor that captures visual images along with per-pixel depth information, which has the potential to be used in visual map construction. The visual map is similar to a dense 3D map for indoor environments, but it is improved according to the actual demands on visual localization. In contrast to [15], there is no need to measure the walking distance during the construction of the visual map. The proposed visual map as a database contains three main elements: first, the floor plans recovered from the reconstructed 3D model, second, the database images captured by the visual sensor (database camera) of Kinect, and third, the poses of the camera views in the database. In the process of localization, the database images and the poses of the database camera are used to achieve the estimation of the query camera location.

In this paper, a novel construction method of the visual map is proposed based on local and global optimizations. The local optimization takes full advantage of the visual features and depth values to refine the transformation matrix of the RGB-D camera poses. In the global optimization phase, an RGB-D detection method is presented to accurately detect the loop closure, which is beneficial for improving the performance of the graph-based global consistency optimization. Then, on the basis of the visual map, an image-based localization method under the epipolar constraint is introduced. This is a new idea of the utilization of the indoor dense 3D map in that the user's location is estimated by the positional relationship between the query camera and the database camera. Compared with the existing database-assisted localization systems, such as the single view-based [9, 10] or the CBIR-based [15, 16] localization methods relying on the plane-to-plane homography method which requires that the visual features for localization should distribute in a plane, our localization method utilizes the epipolar constraint to associate database camera poses with query camera poses regardless of scene structures. The performance of the proposed localization method depends on the accuracy of the visual map. Therefore, the optimization of the visual map based on the requirements of the localization is necessary and indispensable.

The main contribution of this paper can be stated as four aspects: (1) a local optimization method of visual map construction is proposed based on multiple constraints; (2) in the global optimization, an RGB-D-based detection method is presented to accurately determine loop closures; (3) an image-based localization method is introduced, taking advantage of the proposed visual map; and (4) combining the visual map construction and the image-based localization, a novel integrated localization system is obtained, which contains the offline stage of database generation and the online

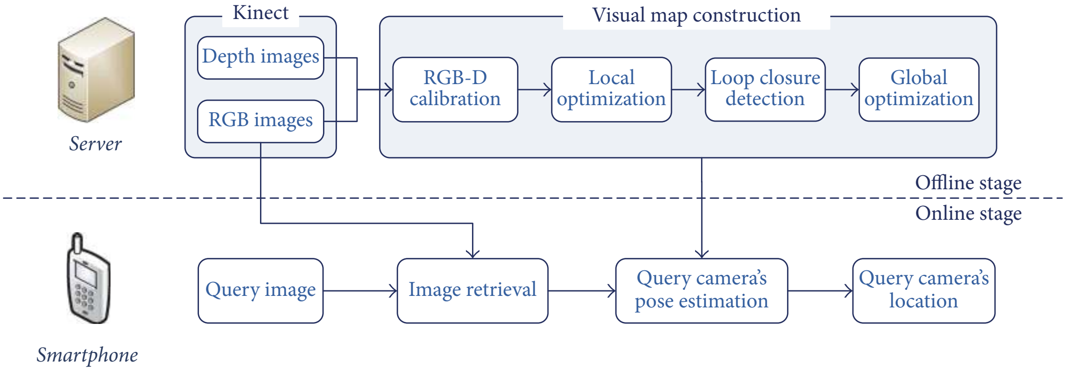


Figure 1: The framework of visual map construction and image-based localization system.

stage of location estimation. The framework of the proposed visual map construction and image-based localization system is shown in Figure 1.

The remainder of this paper is organized as follows. Section 2 provides a detailed discussion of the local optimization of the visual map construction. Section 3 describes the loop closure detection algorithm and the global optimization algorithm based on RGB-D images. Section 4 studies the image-based localization algorithm, including the image retrieval and location estimation of the query camera. Section 5 investigates the performance of the proposed system. Finally, our conclusions are presented in Section 6.

2. Local Optimization Based on MC-ICP for the Visual Map

Because the proposed visual localization method is carried out on the visual map, the precision of the visual map determines the localization accuracy to a large extent. Therefore, a multiple constraints ICP (MC-ICP) optimization algorithm is proposed in this paper based on the RGB-D images captured by Kinect. As a local optimization, MC-ICP aligns the RGB image with the corresponding depth image simultaneously captured by Kinect and then minimizes the errors of the point cloud registration in pixel space and 3D space.

The ICP algorithm is a conventional and effective method for point cloud registration, but due to the lack of joint optimization by shape and visual information, the transformation between the sequential point clouds hardly achieves the optimum value. Indeed, the ICP algorithm does not utilize data associations provided by the matched visual features in RGB images, and thus the transformation obtained by the classical ICP algorithm inevitably contains errors, which will severely reduce the precision of the visual map and consequently lead to the deterioration of the localization performance. Inspired by this problem, on the basis of the classical ICP alignment, an optimized transformation of the camera poses can be obtained by taking full advantage of the matched visual features and the corresponding depth values.

As shown in Figure 2, Kinect has an infrared emitter and two cameras: an RGB camera and an infrared camera. The two cameras are used to simultaneously capture an RGB image and a depth image in the same scene. The working principle of Kinect is that there is a fixed pattern of speckles generated by the IR emitter projecting onto some object, and after that the speckles on the object are compared with the reference patterns of speckles which have already been stored in Kinect. By this means, the 3D distance between Kinect and the object can be estimated, and then the 3D distance is presented as the disparity image by the infrared camera.

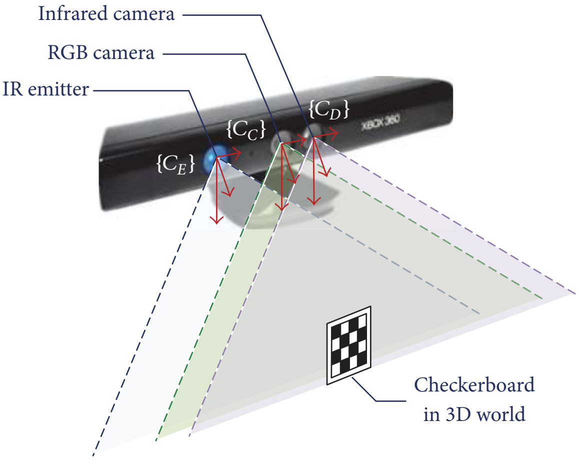


Figure 2: The schematic diagram of Kinect.

Although Kinect is calibrated during manufacturing with a proprietary algorithm, the precision of the manufacturer's calibration cannot satisfy the requirement of the visual map construction, because the depth distortion is not considered in the manufacturer's calibration [36]. Therefore, Kinect should be calibrated before the visual map construction to achieve a better accuracy. In this paper, the method proposed in [37] is utilized to calibrate Kinect before the visual map construction.

The classical ICP algorithm relies only on the correspondence of the 3D point cloud shapes, which frequently leads to a local optimum and then cannot obtain the rigid body transformation. Therefore, a multiconstraint 3D point registration algorithm is proposed on the basis of the classical ICP algorithm.

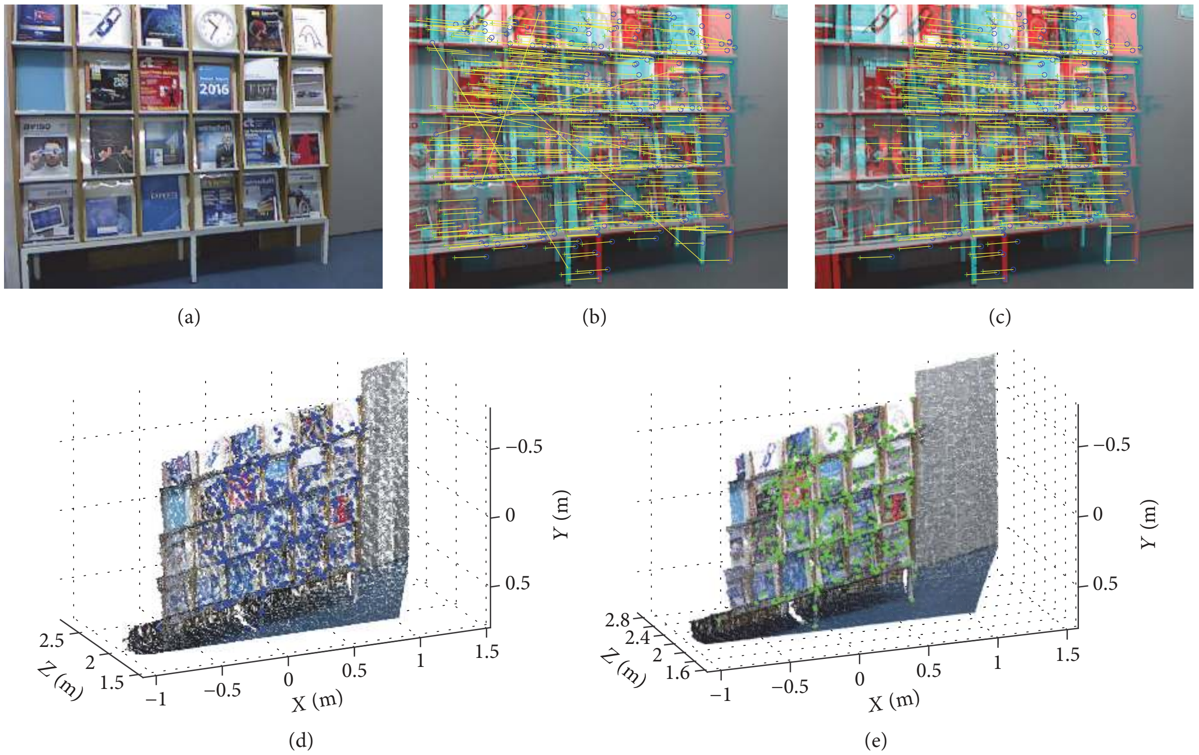


Figure 3: The matched visual features in RGB images and 3D space.

The first step of the proposed algorithm is to match the feature descriptors on the reference image I_r and the current image I_c. Here, the reference and current images represent the two successive RGB images in the time domain, and there are also two successive depth images simultaneously captured by Kinect. The most common feature descriptors for image matching are the Scale-Invariant Feature Transform (SIFT) [38] and the Speeded Up Robust Features (SURF) [39]. The ORB descriptor as an efficient alternative to SIFT or SURF is tested to be rotation-invariant and resistant to noise. According to existing research, the ORB descriptor is effective and stable for visual localization [40]. Therefore, in this paper, the ORB descriptors $\mathbf{F}_r$ and $\mathbf{F}_c$ are extracted from the reference and current images, and because the descriptors are binary, the hamming distance is used to match the descriptors in $\mathbf{F}_r$ and $\mathbf{F}_c$. Figure 3(b) shows a result of descriptor matching between the reference image and the current image in an indoor scene as depicted in Figure 3(a).

However, Figure 3(b) shows clearly that there are some mismatched pairs between the two feature descriptor sets. To get rid of the outliers, a best fitting transformation matrix is estimated by the Maximum Likelihood Estimation SAmple Consensus (MLESAC) algorithm [41]. The final result of the transformation matrix maps the inliers $\mathbf{M}_r$ ($\mathbf{M}_r \in \mathbf{F}_r$) on the reference image to the inliers $\mathbf{M}_c$ ($\mathbf{M}_c \in \mathbf{F}_c$) on the current image, which is shown in Figure 3(c). With this method, one-to-one matching can be obtained between the feature descriptors on the reference image and those on the current image. On the basis of the descriptor matching and the result of the RGB-D camera calibration, there will be two 3D point sets, $\mathbf{N}_r$ and $\mathbf{N}_c$, which contain the matched points in the reference point cloud $\mathbf{P}_r$ and the matched points in the current point cloud $\mathbf{P}_c$. As an example, the matched 3D points in sets $\mathbf{N}_r$ and $\mathbf{N}_c$ are separately shown in Figures 3(d) and 3(e) with a 3D view.

In contrast to the alignment of images without the requirement of initialization, the alignment of point clouds using the ICP algorithm has shown that if the two point clouds are already nearly aligned, the ICP algorithm can effectively avoid convergence at an incorrect local optimum [34]. Therefore, an initialization method is used before implementing the ICP algorithm by the similar method as Perform_RANSAC_Alignment proposed in [33]. But, in our initialization method, the RANSAC algorithm is replaced by the MLESAC algorithm, and visual features are described by ORB descriptors instead of SIFT descriptors.

The second step of the MC-ICP algorithm is to align the point clouds by the classical ICP algorithm with the initialization result. The ICP algorithm attempts to align two point clouds by means of the nearest local minimum of a mean-square distance metric [34]. Given two 3D point sets, which are the reference point cloud $\mathbf{P}_r$ and the current point cloud $\mathbf{P}_c$, the algorithm iteratively strives to obtain $\mathbf{T}_{\text{ICP}}$ using transformation matrix $\mathbf{T}_{c,r}$ by the objective function:

$$\mathbf{T}_{\text{ICP}} = \underset{\mathbf{T}_{c,r}}{\arg\min} \left(\sum_{p_i^r \in \mathbf{P}_r, p_i^c \in \mathbf{P}_c} \left\| p_i^r - p_i^c \cdot \mathbf{T}_{c,r} \right\|^2 \right), \qquad (1)$$

where p_i^r and p_i^c are the homogeneous coordinates of the associated points used in the interaction of the ICP algorithm.

$\mathbf{T}_{c,r}$ is the transformation matrix that converts the current point coordinates into the reference point coordinates. The transformation matrix $\mathbf{T}_{c,r}$ can be represented as

$$\mathbf{T}_{c,r} = \begin{bmatrix} \mathbf{R}_{c,r} & \mathbf{0} \\ \mathbf{t}_{c,r} & 1 \end{bmatrix} \in \mathbb{SE}_3, \tag{2}$$

where the Euclidean group $\mathbb{SE}_3$ contains the rotation matrix $\mathbf{R}_{c,r}$ and the translation vector $\mathbf{t}_{c,r}$ in three-dimensional Euclidean space:

$$\mathbb{SE}_3 := \left\{ \mathbf{R}_{c,r}, \mathbf{t}_{c,r} \mid \mathbf{R}_{c,r} \in \mathbb{SO}_3, \mathbf{t}_{c,r} \in \mathbb{R}^3 \right\}, \tag{3}$$

where $\mathbf{R}_{c,r}$ and $\mathbf{t}_{c,r}$ reflect the 3D motion between the current point coordinates and the reference point coordinates.

The third step of the MC-ICP algorithm is the optimization procedure of the transformation $\mathbf{T}_{\mathrm{ICP}}$ obtained by the classical ICP algorithm. Because the color and depth features are not taken into account simultaneously, there are not enough constraints for the point cloud registration. The result of the classical ICP-based registration inevitably contains some errors. Therefore, a multiconstraint optimization method is proposed, which optimizes the point cloud registration in both the 3D domain and the image domain. The objective function f_1 and solution $\mathbf{T}_1$ can be denoted as

$$\begin{aligned} f_1 = \frac{1}{n_{\mathrm{mat}}} \sum_{i=1}^{n_{\mathrm{mat}}} & \left(\alpha_1 \left(x_i^c r_{11} + y_i^c r_{21} + z_i^c r_{31} + t_1 - x_i^r \right) \right. \\ & + \alpha_2 \left(x_i^c r_{12} + y_i^c r_{22} + z_i^c r_{32} + t_2 - y_i^r \right) \\ & \left. + \alpha_3 \left(x_i^c r_{13} + y_i^c r_{23} + z_i^c r_{33} + t_3 - z_i^r \right) \right), \end{aligned} \tag{4}$$

$$\begin{gathered} \mathbf{T}_1 = \underset{(r_{ij}, t_k)}{\arg\min} \left(f_1 \right), \\ \left(i = 1, 2, 3;\ \ j = 1, 2, 3;\ \ k = 1, 2, 3 \right), \end{gathered} \tag{5}$$

where n_{mat} is the number of matched feature descriptors; the point $p_i^c = (x_i^c, y_i^c, z_i^c)$ denotes the 3D position of the point in set $\mathbf{N}_c$; the point $p_i^r = (x_i^r, y_i^r, z_i^r)$ denotes the 3D position of the point in set $\mathbf{N}_r$; and $(\alpha_1, \alpha_2, \alpha_3)$, which satisfy $\alpha_1 + \alpha_2 + \alpha_3 = 1$, are the weights to control the contributions in the X, Y, and Z directions. The rotation matrix $\mathbf{R}_{c,r}$ and the translation vector $\mathbf{t}_{c,r}$ as parts of the transformation matrix $\mathbf{T}_{c,r}$ can be described as

$$\begin{gathered} \mathbf{R}_{c,r} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix}, \\ \mathbf{t}_{c,r} = \begin{bmatrix} t_1 & t_2 & t_3 \end{bmatrix}. \end{gathered} \tag{6}$$

The optimization function shown in (5) can be regarded as a nonlinear least squares problem, where the number of variables is less than the number of matched descriptors. In practical terms, the matched descriptors are much greater in number than the variables in function (4). Therefore, this function can be optimized by the trust-region-reflective algorithm [42], taking $\mathbf{T}_{\mathrm{ICP}}$ as an initial value to achieve solution $\mathbf{T}_1$.

On the basis of the above 3D domain optimization, a pixel-level optimization in the image domain is proposed in this paper. The 3D coordinates of the matched descriptors are transformed into 2D coordinates in the image plane, and then the distance of the matched descriptors can be measured at pixel level. The objective function to measure the distance between the matched descriptors at pixel level can be defined as

$$\begin{aligned} f_2 = \sum_{i=1}^{n_{\mathrm{mat}}} & \left(\beta_1 \left(\frac{f_{\mathrm{RGB}}}{z_i^{Tc} p_x} x_i^{Tc} + u_o - u_i^r \right) \right. \\ & \left. + \beta_2 \left(\frac{f_{\mathrm{RGB}}}{z_i^{Tc} p_y} y_i^{Tc} + v_o - v_i^r \right) \right), \end{aligned} \tag{7}$$

where f_{RGB} is the focal length of the RGB camera, which is obtained by the RGB camera calibration. The pixel sizes p_x and p_y are used to convert the distance in the 3D coordinate system to the image coordinate system. The principle point coordinates (u_0, v_0) present the intersection position of the principle axis and the image plane. The directional weights β_1 and β_2, which satisfy $\beta_1 + \beta_2 = 1$, control the contributions of the X and Y directions in the image domain to the overall error. In (7), the 2D coordinates (u_i^r, v_i^r) are the position of the matched descriptor on the reference image, and the 3D position $(x_i^{Tc}, y_i^{Tc}, z_i^{Tc})$ is transformed from $p_i^c = (x_i^c, y_i^c, z_i^c)$ by the transformation matrix $\mathbf{T}_{c,r}$. Then, the transformed 3D position $(x_i^{Tc}, y_i^{Tc}, z_i^{Tc})$ can be calculated by

$$\begin{aligned} x_i^{Tc} &= x_i^c r_{11} + y_i^c r_{21} + z_i^c r_{31} + t_1, \\ y_i^{Tc} &= x_i^c r_{12} + y_i^c r_{22} + z_i^c r_{32} + t_2, \\ z_i^{Tc} &= x_i^c r_{13} + y_i^c r_{23} + z_i^c r_{33} + t_3. \end{aligned} \tag{8}$$

To obtain the body transformation, which is constrained in both the 3D domain and the image domain, the optimization should be treated as a multiobjective problem in the two domains. Based on the concept of the Tolerant Lexicographic Method [43] for multiobjective optimization, the solution can be determined as

$$\begin{gathered} \mathbf{T}_2 = \underset{(r_{ij}, t_k)}{\arg\min} \left(\frac{f_2 \left(r_{ij}, t_k \right)}{n_{\mathrm{mat}}} \right), \\ \left(i = 1, 2, 3;\ \ j = 1, 2, 3;\ \ k = 1, 2, 3 \right), \\ \left(r_{ij}, t_k \right) \in \left\{ \left(r_{ij}, t_k \right) \mid f_1 \le f_1^* + \varepsilon \right\}, \end{gathered} \tag{9}$$

where $f_1^* = \min(f_1)$ denotes the minimum values corresponding to the 3D objective function. For (9), it is a solution for a multiobjective optimization problem that constrains the transformation matrix in both the 3D space domain and the image space domain. This optimization can be treated as a nonlinear least squares problem and solved by the trust-region-reflective algorithm. However, sometimes there is no feasible solution for $\mathbf{T}_2$ when $f_1 \le f_1^*$, and therefore it is

```
$\mathbf{F}_r$ ←Extract the ORB feature descriptors from the reference RGB image $I_r$;
$\mathbf{F}_c$ ←Extract the ORB feature descriptors from the current RGB image $I_c$;
$\mathbf{F}_{\text{inlier}}$ ←Match the feature descriptors in sets $\mathbf{F}_r$ and $\mathbf{F}_c$, and exclude the
    outliers via the MLESAC algorithm;
$\mathbf{T}_{\text{ICP}}$ ←Perform the ICP algorithm with the initialization result and the inputs:
    reference point cloud $\mathbf{P}_r$ and current point cloud $\mathbf{P}_c$;
$\mathbf{T}_1$ ←Perform 3D domain optimization via the Trust-region-reflective algorithm
    with the input: $\mathbf{T}_{\text{ICP}}$;
$f_{\text{bound}} = f_1^*$, $i_{\text{iter}} = 0$;
while ($i_{\text{iter}} \le i_{\max}$);
    ($f_2, \mathbf{T}_2$) ←Perform pixel-level optimization via the Trust-region-reflective
        algorithm with the constraint $f_{\text{bound}}$;
    if $\mathbf{T}_2 \neq \Phi$
        The optimization result $\mathbf{T}_2^* = \mathbf{T}_2$;
        break;
    else
        $f_{\text{bound}} = f_{\text{bound}} + \varepsilon$;
    end
    $i_{\text{iter}} = i_{\text{iter}} + 1$;
end
```

Algorithm 1: The multiconstraint ICP algorithm.

necessary to expand the constraint by providing a tolerant increment ε in the process of iterative solving. The detailed process of the proposed MC-ICP algorithm is described in Algorithm 1.

3. Loop Closure Detection and Global Optimization

Although the MC-ICP algorithm is proposed as local optimization in the previous section to minimize the errors of the RGB-D alignment, drift caused by the alignment of the associated frames is inevitable, which will ultimately lead to inaccuracies for the visual map, especially for long paths of the camera. Therefore, global optimization is essential to reduce the cumulative errors after the RGB-D alignment. The main process of global optimization can be divided into three parts: keyframe selection, loop closure detection, and graph-based optimization.

For loop closure detection, the core task is to find the location that the camera has once visited, and then the path from the previously visited location to the revisited location forms a loop closure. In common research on loop closure, the visual method plays the main role in matching two image frames to determine the loop closure [44, 45]. However, in some cases, the image matching determined by feature descriptors does not definitely mean that the two images were taken in the same location; that is, high similarity of the vision cannot guarantee that the camera is at the same position. Figure 4 shows an example of two images that are matched but that were taken in different locations 1.54 meters apart. In this example, it can be clearly seen that similar images may be taken in different positions while the feature descriptors in the images are matched, so it is unreliable to definitely determine that the camera came back to the previous position. For small-scale indoor environments, the false detection will directly lead to poor performance of the global optimization. Therefore, to precisely detect the loop closure, a novel detection method is proposed based on the RGB-D constraint.

3.1. Keyframe Selection. Keyframe selection is an essential step of loop closure detection. If the current frame matches all the previous frames the camera has captured, the complexity of the image matching will make the loop closure detection time-consuming. To address the challenge of complexity, keyframes as a subset of the image consequence are introduced, and they contain enough visual features to detect the loop closure. A common way is selecting the keyframe every n frames [46, 47] or within a fixed distance interval [48], but some essential visual information in nonkeyframes may be lost in this way. Another feasible way for keyframe detection is based on visual overlap determined by feature matching [49, 50], which is sensitive to object occlusion and image blur. In this paper, the keyframes are selected by two strategies: the fixed-distance strategy and the matching-ratio strategy. A frame sequence is defined as $N = \{n_1, n_2, \ldots, n_l\}$, where n_1 is the start frame captured by the RGB camera and n_l is the current frame. For the fixed-distance strategy, the element in the keyframe set N_1 is selected every p frames from the frame sequence N:

$$N_1 = \left\{ n_{pq+1} \mid n_{pq+1} \in N,\ q = 0, 1, \ldots \right\}. \tag{10}$$

However, in practice, the camera controlled by the operator often occurs in a sudden speed increment of the movement, which will lead to the loss of some essential visual information by selecting the keyframes every p frames. If the essential visual information is missed, the loop closure detection could fail because there is no previous keyframe matched with the current keyframe. With the purpose of

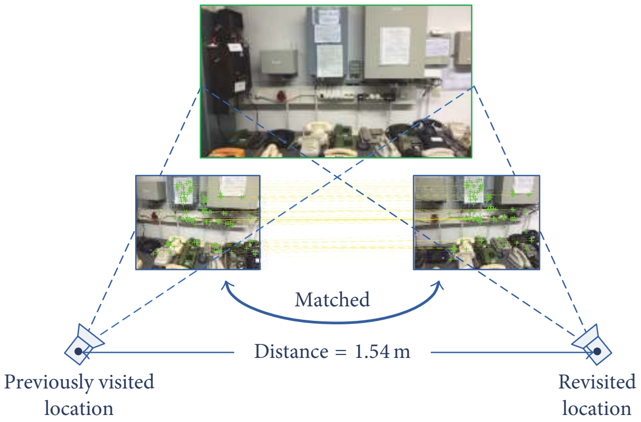


Figure 4: A false detection of loop closure only by feature descriptor matching.

conserving enough visual information to achieve accurate detection of the loop closure, the matching-ratio strategy is introduced to select the keyframe, ensuring that the keyframes contain sufficient visual features to match with the potential loop closure frame.

The matching-ratio selection strategy is proposed on the basis of the fixed-distance strategy to reserve as much visual information between frames as possible. Figure 5 is an illustration of the matching-ratio selection strategy. Between the two successive keyframes $n_{pq} \in N_1$ and $n_{p(q+1)} \in N_1$, there will be p frames that are not selected as the keyframes in the time domain. n_b is a frame that is not selected by the fixed-distance strategy between n_{pq} and $n_{p(q+1)}$. Frame $n_a \in N_2$ denotes a keyframe that has already been selected by the matching-ratio strategy. N_2 is a keyframe set containing the keyframes selected by the matching-ratio strategy. According to the matching-ratio strategy, every frame that does not belong to set N_1 should be separately matched with the nearest previous keyframes selected by the fixed-distance strategy and the matching-ratio strategy. Therefore, frame n_b should be matched with keyframe n_a and keyframe n_{pq}, separately. This process relies on the ORB descriptor matching as described in Section 3.2, and, after that, the MLESAC algorithm is implemented to exclude the outliers in the matching result. Then, the matching ratios α_1 and α_2 can be calculated by

$$\alpha_i = \frac{m_i}{m_{\text{total}}}, \quad (i = 1, 2), \tag{11}$$

where m_1 represents the number of matched inliers between n_b and n_a, m_2 represents the number of matched inliers between n_b and n_{pq}, and m_{total} is the total number of ORB descriptors extracted from n_b. The threshold α_T is set as a threshold value to verify whether n_b should be taken as a keyframe and put in set N_2. The decision rule is that if $\alpha_1 < \alpha_T$ or $\alpha_2 < \alpha_T$, n_b will be regarded as a keyframe and put in N_2. Each frame outside the set N_1 will have an estimation as to whether it should be a keyframe for set N_2, and, after this procedure, the complete keyframe set N_{key} can be obtained by $N_{\text{key}} = N_1 \cup N_2$.

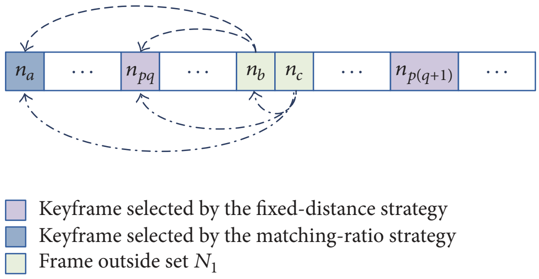

n_a
⋯
n_{pq}
⋯
n_b
n_c
⋯
$n_{p(q+1)}$
⋯
Keyframe selected by the fixed-distance strategy
Keyframe selected by the matching-ratio strategy
Frame outside set N_1

Figure 5: An illustration of the keyframe selection by the matching-ratio strategy.

It is worth noting that a frame outside of set N_1 is always matched with its two nearest keyframes, which are separately selected by the fixed-distance strategy and the matching-ratio strategy. For example, if n_b shown in Figure 5 is selected as a keyframe by the matching-ratio strategy, frame n_c outside of set N_1 next to n_b will be matched with n_b and n_{pq}. Otherwise, n_c should be matched with n_a and n_{pq}. In addition, when selecting the first keyframe in set N_2, a nonkeyframe outside of set N_1 only matches the keyframes in N_1 by (11).

3.2. Loop Closure Detection by the RGB-D Constraint. In general, the keyframe set is large when the camera has a long distance trajectory, and then the keyframe searching is a heavy burden. Therefore, an image retrieval method based on the geographic restriction and bag of features (BOF) [51] is employed in this paper to avoid the global search for the keyframes. In contrast to the general BOF retrieval that searches for similar images in the entire database by the K-Nearest-Neighbor (KNN) algorithm, the proposed method only searches the keyframes that are within a finite distance. The finite distance will be set depending on the scale of the indoor scene. In this manner, the previous keyframes that are similar to the current keyframe are selected as the candidates to detect the loop closure.

For the graph-based global optimization for RGB-D mapping, the performance of the optimization depends on the exact loop closure detection. In the proposed method, the visual and depth data acquired by the RGB-D sensor provide more helpful information to recognize the loop closure more accurately. The main task for the closure detection is to estimate whether the RGB-D camera returns to the position that it has visited before. Therefore, in this paper, the loop closure will be detected by two evaluation criterions: (1) the matching rate of the visual features and (2) the depth consistency of the visual features. The loop closure will be determined only if the current keyframe and the previous keyframe satisfy the two criteria simultaneously.

The matching rate of the visual features is proposed based on the vision similarity between the current keyframe and the previous keyframes. The ORB descriptors extracted from the RGB images in the local optimization phase are used to match the current keyframe and the previous keyframes. To reflect the visual similarity clearly between the frames, the matching rate $\sigma_{\text{mat}} \in [0,1]$ is defined as

$$\sigma_{\text{mat}} = \frac{n_{\text{mat}}}{(1/2)\left(n_{\text{cur}} + n_{\text{key}}\right)}, \tag{12}$$

where n_{mat} is the total number of inliers refined by the MLESAC algorithm. In (12), n_{cur} and n_{key} are the number of descriptors extracted from the current keyframe and the keyframes, respectively.

In some cases, the matching rate can indicate the positions of the camera; however, there are some exceptions in which the camera captures the same object with different poses, as shown in Figure 4. Therefore, it is necessary to further determine the similarity between the current keyframe and the previous keyframes by the other criterion.

In the process of visual feature matching, the ORB descriptors have been extracted and matched between the current keyframe and the previous keyframes. According to the result of the RGB-D camera calibration, each pixel in the RGB image relates to a depth value. For each pair of the matched descriptors in the keyframes, an $r \times r$ pixel patch is defined around the matched descriptor in the keyframes. Then, the depth value d_P for the pixel patch P can be described as

$$d_P = \frac{1}{r^2}\left(\sum_{s,t}\left(d_{11} + d_{12} + \cdots + d_{st}\right)\right), \quad (s = 1,\ldots,r,\ t = 1,\ldots,r), \tag{13}$$

where d_{st} is the depth value corresponding to the pixel (s,t) in the patch. Centered on each matched descriptor, a depth value d_P can be calculated by the depth of the pixel in the patch. If there are m_d matched descriptors in the current keyframe, there will accordingly be m_d matched descriptors in the previous keyframe. To measure the distinction of the distance between the object and the camera, an error metric of the depth is defined as

$$\sigma_{\text{dep}} = \frac{1}{m_d}\sum_{j}\left(\left|v_j^c - v_j^k\right|\right), \quad (j = 1,\ldots,m_d), \tag{14}$$

where v_j^c denotes the depth value of the patch in the current keyframe and v_j^k denotes the depth value of the patch in the previous keyframe. The value of σ_{dep} reflects the distinction of the distance between the camera and objects, which also indicates the position difference that the camera shot at.

The final decision of the loop closure will be drawn from the two criteria: the matching rate σ_{mat} and the depth distinction σ_{dep}. In practical operation, it is necessary to define a set of thresholds $(\sigma_{\text{mat}}^T, \sigma_{\text{dep}}^T)$ to determine the loop closure. Only if the current keyframe and the previous keyframe satisfy (1) $\sigma_{\text{mat}} \geq \sigma_{\text{mat}}^T$ and (2) $\sigma_{\text{dep}} \leq \sigma_{\text{dep}}^T$ can the loop closure be ultimately determined.

3.3. Global Optimization by the Graph-Based Method. Due to the errors caused by the camera pose estimation, the path of the camera usually cannot form a globally consistent trajectory. Therefore, the global optimization methods are used to correct the drift introduced by the iterations of the camera pose estimation. The graph-based method for SLAM global optimization is a typical technology to build a consistent map by estimating the parameters associated with vertices (camera poses) in a graph. The g2o framework [52] as an effective graph-based method is employed in this paper to optimize the trajectory of the RGB-D camera estimated by the MC-ICP algorithm.

The g2o framework presents the global optimization problem as a graph constraint solved with the nonlinear least squares algorithm. In the framework, the poses of the camera are regarded as the vertices, and the edges serve as the constraints in the graph. For the total of n_{pose} camera poses in the graph, the error function can be defined as

$$f_{\text{g2o}} = \sum_{i,j \in n_{\text{pose}}} \mathbf{e}\left(\mathbf{v}_i, \mathbf{v}_j, \mathbf{w}_{ij}\right)^T \boldsymbol{\Omega}_{ij} \mathbf{e}\left(\mathbf{v}_i, \mathbf{v}_j, \mathbf{w}_{ij}\right), \tag{15}$$

where $\mathbf{v}_i$ and $\mathbf{v}_j$ denote the ordered poses of the camera related to the constraint mean matrix $\mathbf{w}_{ij}$ and $\boldsymbol{\Omega}_{ij}$. Vector $\mathbf{e}(\mathbf{v}_i, \mathbf{v}_j, \mathbf{w}_{ij})$ is the error function measuring how well the poses $\mathbf{v}_i$ and $\mathbf{v}_j$ satisfy the constraint $\mathbf{w}_{ij}$. According to (15), the optimal trajectory can be described as

$$\mathbf{V}^* = \arg\min_{\mathbf{V}} f_{\text{g2o}}(\mathbf{V}), \tag{16}$$

where $\mathbf{V}$ contains the original camera poses and $\mathbf{V}^*$ contains the optimized poses.

The process of the global optimization is described in Algorithm 2. After the global optimization of the camera poses, an integrated visual map constructed by the RGB-D sensor is stored as the database for the visual localization, containing (1) a 3D structure and 2D floor plan of the indoor scene, (2) the database images captured by the RGB-D sensor, and (3) the poses of the database camera. The entire process of the visual map construction is conducted in the offline stage at the service of the online localization.

4. Image-Based Localization Using the Visual Map

The image-based localization is an economical and effective method to recognize the position of the user using the

```
$N_1$ ←Select the keyframes by the fixed-distance strategy;
$N_2$ ←Select the keyframes by the matching-ratio strategy;
$N_{\text{key}}$ ←Obtain the keyframe set by $N_1 \cup N_2$;
$t$ ←Set the total number of the keyframe set $N_{\text{key}} = \{n_1, n_2, \ldots, n_t\}$;
while ($t \geq 2$)
    $\sigma_{\text{mat}}$ ←Match the current keyframe $n_t$ with the previous
    keyframes $\{n_1, \ldots, n_{t-1}\}$;
    $\sigma_{\text{dep}}$ ←Compute the depth errors of the pixel patches in the
    current keyframe $n_t$ and the previous keyframes $\{n_1, \ldots, n_{t-1}\}$;
    if $\sigma_{\text{mat}} \geq \sigma_{\text{mat}}^T$ and $\sigma_{\text{dep}} \leq \sigma_{\text{dep}}^T$
      $\mathbf{V}^*$ ←Perform the g2o algorithm with the camera pose $\mathbf{V}$;
      break;
    end
    $t = t - 1$;
end
```

Algorithm 2: Global optimization based on the RGB-D loop closure detection and the g2o algorithm.

query image captured by the smartphone's camera. This part will focus on the online phase, which is the image-based localization. Taking advantage of the completed visual map, the position of the query camera can be estimated according to the relationship constrained by the epipolar constraint between the query image and the database images. To utilize the epipolar constraint, image retrieval is employed to find the candidate database images that contain the same visual features as the query image.

4.1. Retrieval and Matching between Query and Database Images. In the online stage of the localization, the user captures the query image by the camera equipped on the smartphone, and then the query image is uploaded to the server through the wireless network. In addition, if the smartphone has powerful data processing capabilities, the query image can also be preprocessed on the smartphone, such as the feature descriptor extraction, and, in this way, it is only necessary to upload the data of the feature descriptors to the server. Either way, the visual information captured by the user is the necessary data required by the image-based localization.

When the server receives the query image, the ORB descriptors are extracted from the query image. After that, the descriptors are quantized to the visual features, and the occurrences of the features are recorded as a vector to index the query image. For the bag of features, each database image is stored in the bag in the term of the vector related to the occurrence of the visual features. The search engine calculates the similarity between the query vector and the database vector based on the L_2 distance. In this manner, the most similar database images are selected as the candidate database images using the KNN strategy.

The query image retrieval based on the BOF searching is efficient but imprecise and can be regarded as a coarse retrieval. This is because there is no rigid alignment between the query image and the database image using the feature descriptor matching. However, the epipolar constraint acts on the matched descriptors in the images, so it is necessary to refine the candidate database images obtained by the BOF retrieval. The specific method is that the query image matches each candidate database image using the ORB descriptors, and the outliers are rejected by the MLSAC algorithm. Then, some database images in the candidate set S_{can} are excluded, since there are few inliers in the matching between the database images and the query image.

4.2. Query Camera Pose Estimation by the Epipolar Constraint. By the retrieval between the query image and database images, the candidate database images are selected according to their visual similarity. In the process of the visual map construction, the poses of the database camera related to the database images are recorded. Therefore, utilizing the query image, the candidate database images, and the poses of the database camera, the location of the query camera can be estimated based on the epipolar constraint.

In this paper, the database camera is the RGB camera of the RGB-D device that has been calibrated in the offline stage, and the query camera equipped on the user's smartphone should be calibrated before the localization. $\mathbf{K}_{\text{database}}$ and $\mathbf{K}_{\text{query}}$ denote the database camera calibration matrix and the query camera calibration matrix, respectively. The coordinates $\mathbf{X}_{\text{database}}$ and $\mathbf{X}_{\text{query}}$ of the ORB descriptors extracted from the database image and the query image can be normalized by

$$\begin{aligned} \widehat{\mathbf{X}}_{\text{database}} &= \mathbf{K}_{\text{database}}^{-1} \mathbf{X}_{\text{database}}, \\ \widehat{\mathbf{X}}_{\text{query}} &= \mathbf{K}_{\text{query}}^{-1} \mathbf{X}_{\text{query}}. \end{aligned} \tag{17}$$

The relations between the normalized coordinates of the ORB descriptors in the query image and the database image can be described as

$$\widehat{\mathbf{X}}_{\text{database}} \mathbf{E} \widehat{\mathbf{X}}_{\text{query}} = 0, \tag{18}$$

where $\mathbf{E}$ is the essential matrix, which contains the rotation matrix and the translation vector of the camera transformation. The essential matrix is used to present the camera

```
$S_{\text{can}}$ ←Select the candidate database images by the BOF image retrieval.
for $m = 1 : k_{\text{DB}}$
    $\mathbf{E}_m$ ←Compute the essential matrix between the query image and the database image.
    $(\mathbf{R}_E^m, \mathbf{t}_E^m)$ ←Decompose the essential matrix $\mathbf{E}_m$ into the rotation matrix and the translation vector by SVD.
end
$\mathbf{L}_i$ ←Compute the position of the intersection determined by each two connecting lines.
$\mathbf{L}_{\text{query}}$ ←Compute the estimated location of the query camera.
```

Algorithm 3: Image-based localization algorithm.

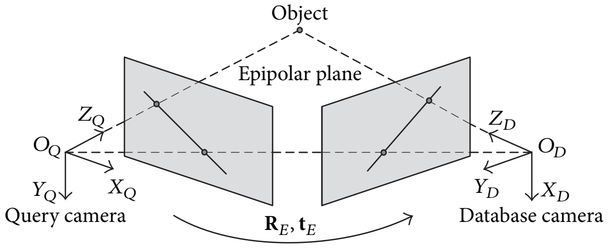


Figure 6: An illustration of the epipolar constraint.

transformation parameters between the database camera and the query camera in the following form:

$$\mathbf{E} \simeq \widehat{\mathbf{t}}_E \mathbf{R}_E, \tag{19}$$

where $\widehat{\mathbf{t}}_E$ denotes the translation vector and $\mathbf{R}_E$ denotes the rotation matrix. For $\mathbf{t}_E = [t_x, t_y, t_z]^{-1}$, $\widehat{\mathbf{t}}_E$ is the skew-symmetric matrix of $\mathbf{t}_E$ and

$$\widehat{\mathbf{t}}_E = \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}. \tag{20}$$

The essential matrix $\mathbf{E}$ is determined by the normalized image coordinates of the matched ORB descriptors in the query image and the database image. One efficient method for computing the essential matrix is the five-point algorithm [53]. Via the essential matrix $\mathbf{E}$, the translation vector $\mathbf{t}_E$ and rotation matrix $\mathbf{R}_E$ can be extracted using Singular Value Decomposition (SVD).

The epipolar constraint reflects the pose relationship between the query camera and the database camera, which is shown in Figure 6. By the translation vector $\mathbf{t}_E$ and the rotation matrix $\mathbf{R}_E$ extracted from the essential matrix $\mathbf{E}$, the 3D position relationship between the query camera and the database camera can be described as

$$\mathbf{X}_q = \mathbf{R}_E \mathbf{X}_d + \mathbf{t}_E, \tag{21}$$

where $\mathbf{X}_q$ denotes the 3D position (without the scale) of the query camera and $\mathbf{X}_d$ denotes the 3D position of the database camera. Under the epipolar constraint, the position of the query camera can be estimated by each pair of the query image and the candidate database image in S_{can}.

However, the estimated 3D position $\mathbf{X}_q$ of the query camera obtained by the epipolar constraint is not the definite location, because $\mathbf{X}_q$ is always a unit vector. Without a relative scale, $\mathbf{X}_q$ cannot indicate the definite distance between the query camera and the database camera, but it contains relative orientation from the database camera coordinate system $O_D X_D Y_D Z_D$ to the query database coordinate system $O_Q X_Q Y_Q Z_Q$. Utilizing a pair of the query image and the candidate database image, the connecting line can be determined based on the known position of the database camera. As shown in Figure 7, for the query image, there are two candidate database images resulting in two connecting lines l_{13} and l_{23}, and the intersection of the connecting lines is the estimated position of the query camera on the ground plane.

Since more than two candidate database images are selected in general, there will be more than two connecting lines that usually cannot intersect one point due to the errors caused by localization. Therefore the average location of intersections is defined as the estimated query camera location $\mathbf{L}_{\text{query}}$:

$$\mathbf{L}_{\text{query}} = \frac{1}{n_{\text{inter}}} \left(\sum_{i=1}^{n_{\text{inter}}} \mathbf{L}_i \right), \quad \left(1 \le n_{\text{inter}} \le \frac{k_{\text{DB}} \left(k_{\text{DB}} - 1 \right)}{2} \right), \tag{22}$$

where n_{inter} represents the total number of intersections and $\mathbf{L}_i$ represents the position of each intersection. The process of the image-based localization is described in Algorithm 3.

5. Simulation Results and Discussion

To evaluate the performance of the proposed visual map construction method, an extensive experiment is conducted in our office area as shown in Figure 8. The experimental area includes five offices, two corridors, a classroom, and an exhibition room. These rooms are typical indoor places that have different visual and structural complexities. According to functions and sizes, these rooms are divided into four scene classes, namely, the small room scene (the size is less than 25 square meters, including Office 1 and Office 4), the medium

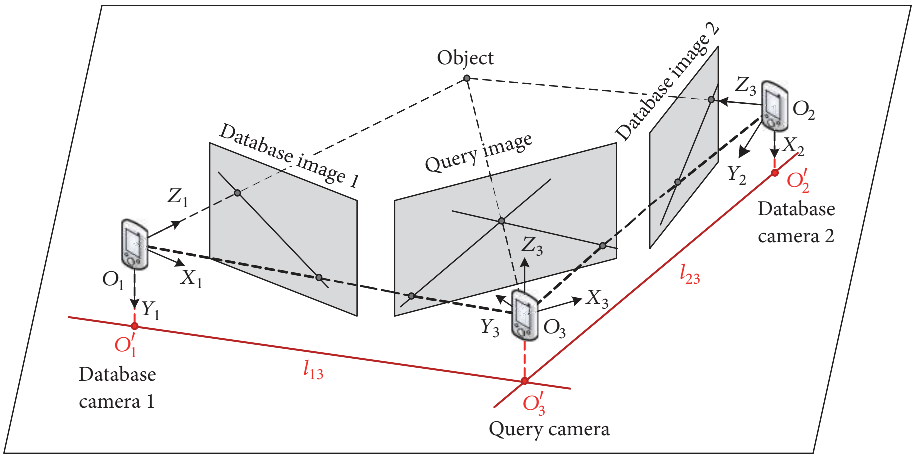


FIGURE 7: The schematic of the image-based localization algorithm.

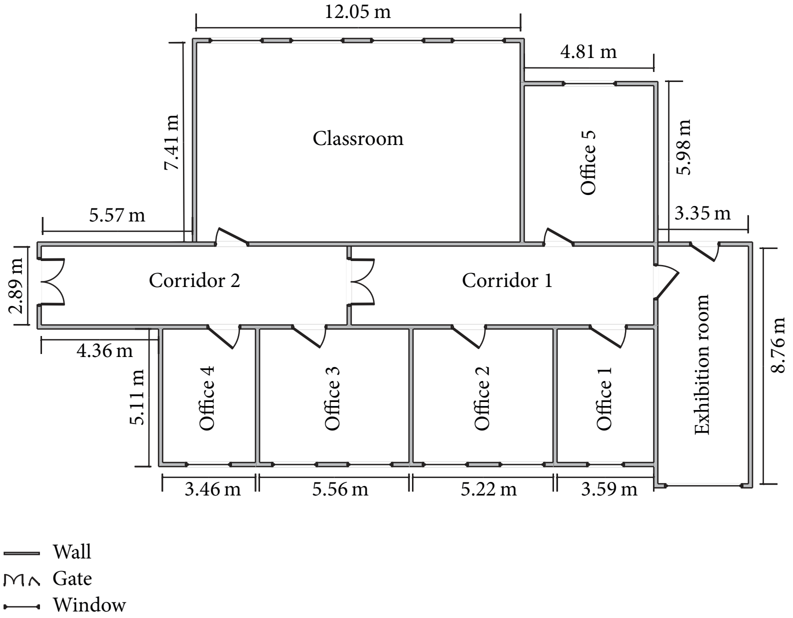
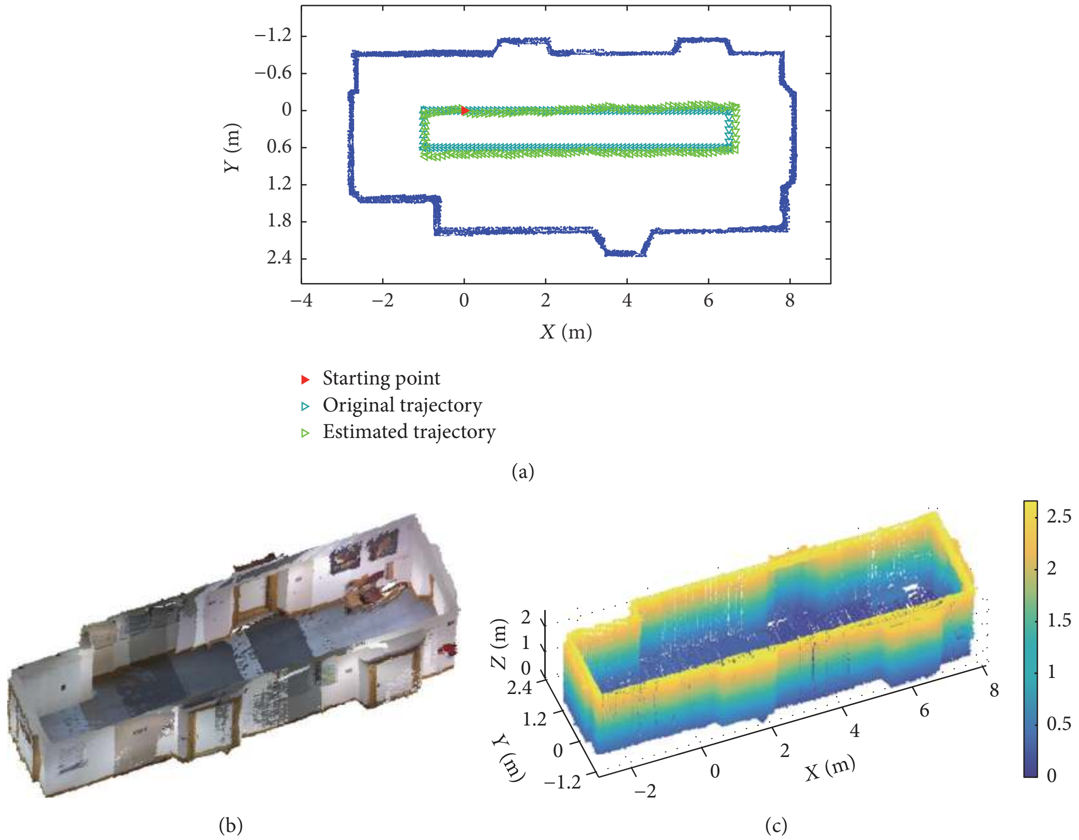


FIGURE 8: The floor plan of the indoor experimental area.

room scene (the size is between 25 and 40 square meters, including Office 2, Office 3, Office 5, and the exhibition room), the large room scene (the size is more than 40 square meters, including the classroom), and the corridor scene (including Corridor 1 and Corridor 2).

Because the implementation of image-based localization relies on the database images and the database camera poses stored in the visual map, the accuracy of the localization is restricted to the precision of the visual map construction, especially for the precision of the estimated database camera positions. To effectively evaluate the precision of the database camera positions, the original trajectory of the RGB-D camera is recorded as the true positions to compare with the estimated positions. For each room or corridor, test points on the camera trajectory are uniformly selected by the step distance of 10 centimeters. The RGB-D camera is moved sequentially between the test points and placed on a tripod at each test point from the starting location and finally back to the starting location. In order to further explore the influence of different light conditions to the visual map construction, the visual map is constructed both in the daytime and at the nighttime. In the daytime, the lamps in the rooms (with windows) are turned off, but the lamps in the corridors (without windows) are turned on. At the nighttime, all the lamps are turned on, and then the light conditions of the rooms and the corridors are well.

The experiment is conducted by Microsoft Kinect v1, and the parameters of Kinect are shown in Table 1. The visual image is captured by the RGB camera on Kinect, and the depth image is simultaneously acquired by the infrared camera on Kinect. All data processing runs on MATLAB 2016B with an Intel Core i5 CPU and an 8 GB RAM.



(a)

(b)

(c)

Figure 9: An example of the visual map construction in Corridor 1.

Table 1: Parameters of Kinect v1.

Parameter	Variable	Value [unit]
Focal length of Kinect RGB camera	f_{RGB}	4.884 [mm]
Pixel size of Kinect RGB camera	p_x	9.3 [μm]
	p_y	9.3 [μm]
Principle point coordinates of Kinect RGB camera	u_0	318.57 [pixel]
	v_0	262.08 [pixel]

Figure 9 is an example of the visual map construction in Corridor 1. Figure 9(a) shows the construction result with a 2D view, which contains the completed floor plan of the corridor and the original trajectory (blue triangle markers) and estimated trajectory (green triangle markers) of the RGB-D camera. In this paper, the localization precision is evaluated in the 2D plane (in the *XOY* coordinate system). Figures 9(b) and 9(c) separately show the 3D visualization map and 3D point clouds of Corridor 1. The 3D visualization map can offer a 3D view to the user during the localization and navigation, and the 3D point clouds can be used to reconstruct the structures of the building. The visual map contains adequate visual and structural information of the building, and, in addition, the database camera poses are also stored in the visual map, all of which are indispensable for the proposed image-based localization.

To completely evaluate the performance of the proposed mapping method, the classical ICP mapping method [32] and the RGB-D ICP mapping method [33] are also implemented under the same conditions. The average lengths of the test trajectories in the small room scene, the medium room scene, the large room scene, and the corridor scene are 7.3 meters, 10.6 meters, 32.8 meters, and 16.9 meters, respectively. Because the area scales of the scenes are different, the average lengths of the trajectories are also different. In most cases, the length of the trajectory meets the proportional relation of the area scale. To clearly analyze the position errors of the test points, the average error of the Euclidean distance in the *XOY* coordinate system is defined as

$$e = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \sqrt{\left(x_i^o - x_i^e\right)^2 + \left(y_i^o - y_i^e\right)^2}, \tag{23}$$

where (x_i^o, y_i^o) is the original position of the camera, (x_i^e, y_i^e) is the estimated position of the camera, and n_{test} is the total number of the test points. In addition, the average errors in the X direction and Y direction are calculated by

$$\begin{aligned} e_X &= \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left(\left|x_i^o - x_i^e\right|\right), \\ e_Y &= \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \left(\left|y_i^o - y_i^e\right|\right). \end{aligned} \tag{24}$$

Table 2: Position errors of the database camera with different mapping methods.

		Average errors (meters)						Accuracy improvement (%)	
Scenes	Methods of modeling	In X direction		In Y direction		Of Euclidean distance			
		Day	Night	Day	Night	Day	Night	Day	Night
Small room	Proposed method	0.0565	0.0440	0.0424	0.0397	0.0750	0.0640	—	—
	RGB-D ICP	0.0739	0.0721	0.0617	0.0566	0.1040	0.0980	27.88	34.69
	Classical ICP	0.0822	0.0823	0.0726	0.0759	0.1186	0.1187	36.76	46.08
Medium room	Proposed method	0.0608	0.0592	0.0418	0.0296	0.0775	0.0695	—	—
	RGB-D ICP	0.0906	0.0874	0.0761	0.0691	0.1237	0.1197	37.35	41.94
	Classical ICP	0.1211	0.1220	0.0902	0.0883	0.1629	0.1632	52.42	57.41
Large room	Proposed method	0.1187	0.1089	0.0975	0.0903	0.1643	0.1525	—	—
	RGB-D ICP	0.2318	0.2374	0.1946	0.1702	0.3247	0.3144	49.40	51.49
	Classical ICP	0.3578	0.3171	0.2241	0.2029	0.4539	0.4505	63.80	66.15
Corridor	Proposed method	0.0800	0.0838	0.0413	0.0411	0.0953	0.0977	—	—
	RGB-D ICP	0.1584	0.1644	0.0537	0.0525	0.1727	0.1787	44.82	45.33
	Classical ICP	0.1970	0.2057	0.1253	0.1216	0.2518	0.2548	62.15	61.66

The average position errors of the database camera in different scenes are shown in Table 2. Because visual maps are constructed both in the daytime and at the nighttime, there are two results of position errors related to each mapping method in different scenes. From the results of the average errors, it can be observed that, in the small room scene, the medium room scene, and the large room scene, the construction precision of the visual map at the nighttime is better than that in the daytime. But either in the daytime or at the nighttime, the construction precision is the same in the corridor scene. The reason is that the lamps in rooms are turned off in the daytime, so, with well lighting, light conditions are better at the nighttime. However, in the corridor scene, lamps are turned on all day. Therefore, the construction precision of the corridor scene is satisfactory either in the daytime or at the nighttime.

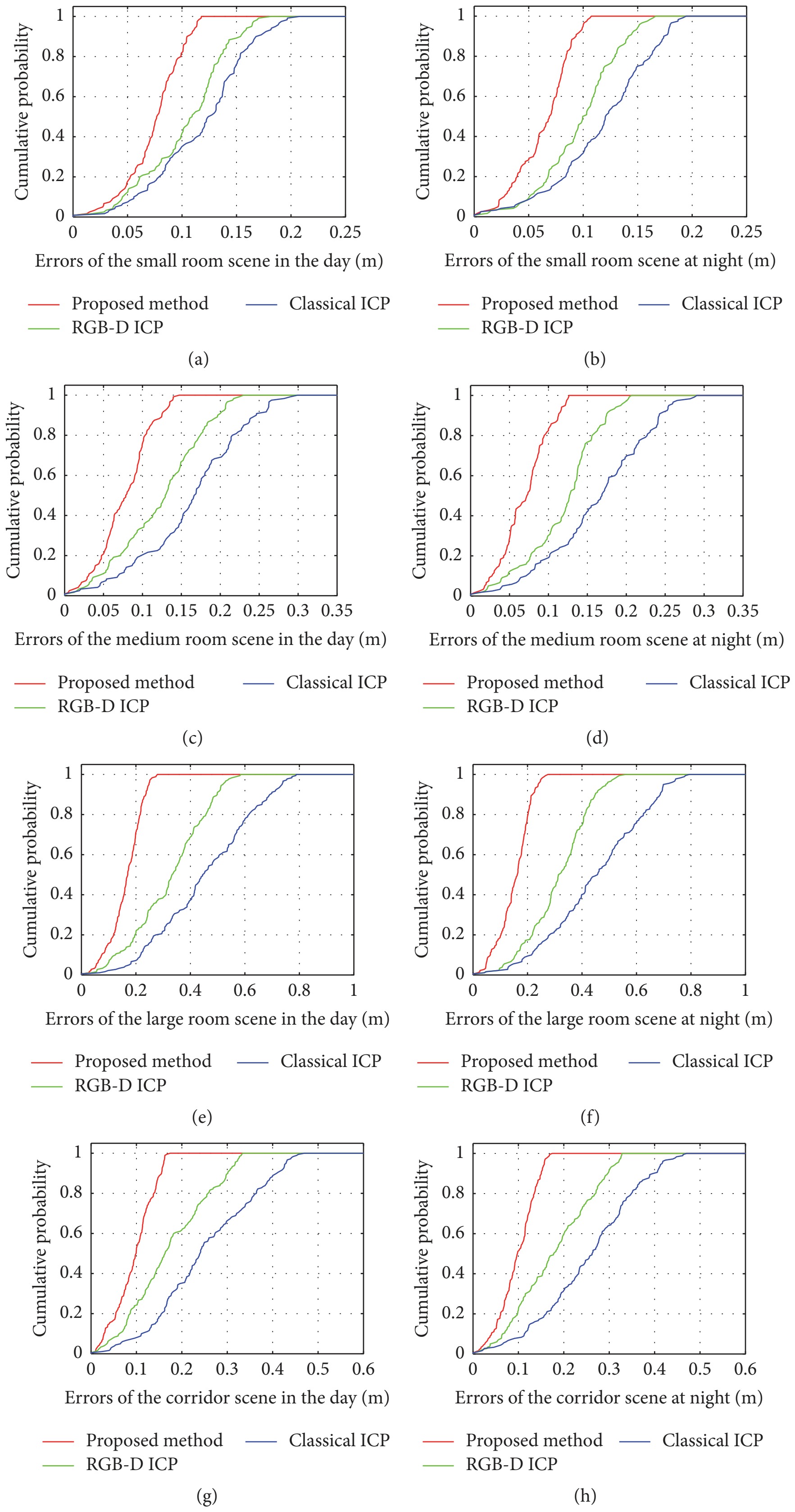
For the average errors of the Euclidean distance, the accuracy improvement percentage i_{im} of the proposed method is calculated by

$$i_{\text{im}} = \frac{\left|e_p - e_c\right|}{e_c} \times 100, \tag{25}$$

where e_p denotes the average error of the proposed method and e_c denotes the average error of the comparative method, such as the classical ICP or the RGB-D ICP mapping method. As shown in Table 2, the accuracy improvement of the proposed method reaches more than 27.88% and 36.76%, respectively, compared with the classical ICP and the RGB-D ICP methods in all scenes. Moreover, if the visual map is constructed at the nighttime, the precision can be improved at last by 34.69% and 46.08%, compared with the other two methods. Therefore, in order to achieve a high precision, it is better to construct the visual map at the nighttime under well light conditions.

According to the result of the position errors, the average errors accumulate along with the increase of the trajectory length. Although the local optimization and global optimization are employed to reduce the errors introduced in the visual map construction, the cumulative error is inevitable and cannot be completely eliminated. Compared with the classical ICP and the RGB-D ICP mapping methods, the performance of the proposed method is evidently better than those of the other two methods, especially in the large room scene. With the increase in the scene size, the proposed method improves the construction precision of the visual map significantly. This is because the proposed method eliminates the errors of the camera transformation as much as possible via multiple constraints that take advantage of RGB-D information.

Figure 10 shows the cumulative probabilities of the Euclidean distance errors with different mapping methods in the daytime and at the nighttime. In the small room scene, the performances of the three mapping methods are satisfactory because the maximum position errors are all limited to within 0.25 meters. As the area of the indoor scene grows, the performances of the three methods degrade due to cumulative errors. However, the performance of the proposed method is still better than those of the other two methods in each scene. The maximum error of the proposed method is limited to within 0.3 meters in the large room scene, but the maximum errors of the RGB-D ICP method and the classical ICP method reach 0.5603 meters and 0.7943 meters, respectively, although the visual map is constructed at night. As for the proposed method, the visual maps constructed in the day or at night achieve different maximum errors. Except for the corridor scene, the maximum errors obtained at the nighttime are less than those obtained in the daytime. The maximum errors of the corridor scene are approximately equivalent because light conditions are well both in the day and at night. Since the accuracy of the image-based localization is significantly influenced by the precision of the estimated positions of the database camera, the proposed method will be beneficial to the image-based localization.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

FIGURE 10: Cumulative probability of the position errors for the database camera.

Table 3: Hardware resource usage and time consumption.

Methods of modeling	CPU usage (%)	R_c (%)	Memory usage (GB)	R_m (%)	Time consumption (second/frame)	R_t (%)
Proposed method	39.08	—	1.42	—	0.2478	—
RGB-D ICP	37.84	3.28	1.37	3.65	0.2236	10.82
Classical ICP	32.61	19.84	1.19	19.33	0.1998	24.02

Table 4: Localization errors of the query camera with different mapping methods.

		Average errors (meters)						Accuracy improvement (%)	
Scenes	Methods of modeling	In X direction		In Y direction		Of Euclidean distance			
		Day	Night	Day	Night	Day	Night	Day	Night
Small room	Proposed method	0.4624	0.4152	0.3915	0.3689	0.6735	0.5940	—	—
	RGB-D ICP	0.5872	0.5490	0.4238	0.4065	0.7775	0.7208	13.38	17.59
	Classical ICP	0.6375	0.6304	0.4917	0.4375	0.8648	0.8182	22.12	27.40
Medium room	Proposed method	0.3956	0.4684	0.3956	0.3518	0.6885	0.6276	—	—
	RGB-D ICP	0.5974	0.5939	0.4376	0.4252	0.8019	0.7757	14.14	19.09
	Classical ICP	0.6951	0.6643	0.4636	0.4587	0.9042	0.8714	23.86	27.98
Large room	Proposed method	0.6656	0.6491	0.4802	0.4256	0.8727	0.8235	—	—
	RGB-D ICP	0.8070	0.7645	0.5468	0.5893	1.0429	1.0358	16.32	20.50
	Classical ICP	0.9372	0.9489	0.7466	0.6824	1.2848	1.2538	32.08	34.32
Corridor	Proposed method	0.5677	0.5890	0.4020	0.3974	0.7527	0.7575	—	—
	RGB-D ICP	0.6953	0.7041	0.4576	0.4539	0.8849	0.8888	14.94	14.77
	Classical ICP	0.7977	0.8084	0.6305	0.5918	1.0779	1.0708	30.17	29.26

In order to fully evaluate the performance of the proposed visual map construction method, the average usage of hardware resource and computational consumption are recorded as shown in the Table 3. The time consumption is the average processing time of one frame during the visual map construction. For the proposed method, R_c, R_m, and R_t represent the increment ratios in aspects of the CPU usage, the memory usage, and the time consumption, compared with the RGB-D ICP and the classical ICP.

Compared with the classical ICP method, the hardware resource usage and the time consumption increase with respect to the RGB-D ICP method and the proposed method. The reason is that the two methods utilize both visual features and depth values to construct the visual map rather than only using point clouds in the classical ICP method. The proposed method needs more hardware resource and running time, because, based on RGB-D information, the multiple constraints are introduced in this method and thereby the algorithm complexity increases. According to the data in Table 3, the maximum increment ratio that appears in time consumption is 24.02%; however, the accuracy improvement ratio of the proposed method is at least 27.88% as shown in Table 2. Moreover, the average processing time for one frame of the proposed method is 0.2478 seconds, which can satisfy the frame rate (3 frames per second) in the visual map construction. Therefore, the proposed method is practicable and well performed in the visual map construction.

On the basis of the accomplished visual map created by different mapping methods, the image-based localization is implemented to evaluate the localization accuracy of the query camera in different scenes. In the small room scene, the medium room scene, the large room scene, and the corridor scene, there are 30, 40, 90, and 60 test points, respectively, which are uniformly selected to evaluate the positioning accuracy of the proposed localization method. For each test point, two query images are captured in different orientations for image-based localization. For each scene, the image-based localization is performed using the visual maps that are constructed in the day and at night. The average localization errors of the query camera with different mapping methods are shown in Table 4.

According to the average localization errors of the query camera, the proposed mapping method improves the performance of the image-based localization. Moreover, the performance of the image-based localization is better when the visual map is constructed at night. In the corridor and the large room scenes, the average errors of the Euclidean distance exceed 1 meter using the visual maps constructed by the classical ICP, which is unsatisfactory for indoor localization. Compared with the RGB-D ICP method, the accuracy improvements of the proposed method can reach at least 13.38% and 14.77%, corresponding to the visual maps constructed in the day and at night. For indoor localization, improvement of precision is challenging, especially when the localization error is reduced to less than 1 meter. Therefore, our proposed mapping method is meaningful for the improvement of image-based indoor localization.

In the process of visual map construction, because the proposed mapping method estimates camera pose transformation utilizing local and global optimizations, the average position errors of the database camera are limited to within 0.2 meters. On the basis of the visual map, image-based localization is conducted in different scenes. By the extensive localization experiments, the localization errors of the query camera can be limited to within 0.9 meters in all scenes, which will satisfy most requirements of indoor location-based service. However, according to the results of the experiments, the position errors of the database camera and the query camera both increase with the area of the indoor scene. Because cumulative errors still exist even though the visual map is constructed by the proposed mapping method, the accuracy of the proposed image-based localization will degrade as the area of the indoor scene increases, which cannot be completely prevented in theory.

6. Conclusions

In this paper, a novel method of visual map construction for indoor environments is proposed to support image-based localization. The main objective of this work is to minimize the position errors of the database camera when constructing the visual map by the RGB-D sensor in the offline stage. In the process of visual map construction, the multiconstraint ICP mapping method is utilized to estimate pose transformation of the database camera. As global optimization, the g2o algorithm is employed based on the RGB-D loop closure detection to achieve a consistent map. On the basis of the visual map, image-based camera localization is conducted via the use of the epipolar constraint. The proposed system that contains the visual map construction and the image-based localization is a novel practical application for indoor 3D dense map.

As illustrated in the simulation results, the proposed method of visual map construction is much more efficient than other mapping methods for image-based localization. Specifically, the position accuracy of the database camera and the query camera is improved by at least 27.88% and 13.38%, respectively, compared with the other two methods in all test scenes. The experimental results also show that the improvement of the position accuracy of the database camera can enhance the performance of the image-based localization.

In the future, the Augmented Reality (AR) technology will be introduced in our image-based localization method to provide users with practical or entertaining information such as site introductions in railway stations or airports and text instructions of exhibits in museums.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper is supported by the National Natural Science Foundation of China (61571162) and Natural Science Foundation of Heilongjiang Province (F2016019).

References

[1] S. Adler, S. Schmitt, K. Wolter, and M. Kyas, "A survey of experimental evaluation in indoor localization research," in *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation, IPIN 2015*, pp. 1–10, October 2015.

[2] P. Davidson and R. Piche, "A survey of selected indoor positioning methods for smartphones," *IEEE Communications Surveys & Tutorials*, no. 99, pp. 1–44, 2016.

[3] S. He and S. G. Chan, "Wi-Fi fingerprint-based indoor positioning: recent advances and comparisons," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 466–490, 2016.

[4] J. Xiao, Z. Zhou, Y. Yi, and L. M. Ni, "A survey on wireless indoor localization from the device perspective," *ACM Computing Surveys*, vol. 49, no. 2, pp. 1–31, 2016.

[5] C. Yang and H.-R. Shao, "WiFi-based indoor positioning," *IEEE Communications Magazine*, vol. 53, no. 3, pp. 150–157, 2015.

[6] R. Faragher and R. Harle, "Location fingerprinting with bluetooth low energy beacons," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2418–2428, 2015.

[7] H. Li, "Low-cost 3D bluetooth indoor positioning with least square," *Wireless Personal Communications*, vol. 78, no. 2, pp. 1331–1344, 2014.

[8] P. Dickinson, G. Cielniak, O. Szymanezyk, and M. Mannion, "Indoor positioning of shoppers using a network of Bluetooth Low Energy beacons," in *Proceedings of the 2016 International Conference on Indoor Positioning and Indoor Navigation, IPIN 2016*, pp. 1–8, October 2016.

[9] J. Z. Liang, N. Corso, E. Turner, and A. Zakhor, "Image based localization in indoor environments," in *Proceedings of the 2013 4th International Conference on Computing for Geospatial Research and Application, COM.Geo 2013*, pp. 70–75, July 2013.

[10] A. Hallquist and A. Zakhor, "Single view pose estimation of mobile devices in urban environments," in *Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision, WACV 2013*, pp. 347–354, January 2013.

[11] H. Sadeghi, S. Valaee, and S. Shirani, "A weighted KNN Epipolar Geometry-based approach for vision-based indoor localization using smartphone cameras," in *Proceedings of the 2014 IEEE 8th Sensor Array and Multichannel Signal Processing Workshop, SAM 2014*, pp. 37–40, June 2014.

[12] H. Liu, H. Li, T. Mei, and J. Luo, "Vision-based fine-grained location estimation," *Multimodal Location Estimation of Videos and Images*, pp. 63–83, 2015.

[13] H. Sadeghi, S. Valaee, and S. Shirani, "Ocrapose: An indoor positioning system using smartphone/tablet cameras and OCR-aided stereo feature matching," in *Proceedings of the 40th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2015*, pp. 1473–1477, April 2014.

[14] H. Sadeghi, S. Valaee, and S. Shirani, "Semi-supervised logo-based indoor localization using smartphone cameras," in *Proceedings of the 2014 25th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communication, IEEE PIMRC 2014*, pp. 2024–2028, September 2014.

[15] R. Huitl, G. Schroth, S. Hilsenbeck, F. Schweiger, and E. Steinbach, "TUMindoor: An extensive image and point cloud dataset for visual indoor localization and mapping," in *Proceedings of the 2012 19th IEEE International Conference on Image Processing, ICIP 2012*, pp. 1773–1776, October 2012.

[16] R. Huitl, G. Schroth, S. Hilsenbeck, F. Schweiger, and E. Steinbach, "Virtual reference view generation for CBIR-based visual pose estimation," in *Proceedings of the 20th ACM International Conference on Multimedia, MM 2012*, pp. 993–996, November 2012.

[17] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, and E. Steinbach, "Mobile visual location recognition," *IEEE Signal Processing Magazine*, vol. 28, no. 4, pp. 77–89, 2011.

[18] J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha, "Visual simultaneous localization and mapping: a survey," *Artificial Intelligence Review*, vol. 43, no. 1, pp. 55–81, 2015.

[19] N. K. Dhiman, D. Deodhare, and D. Khemani, "Where am I? Creating spatial awareness in unmanned ground robots using slam: a survey," *Sādhanā*, vol. 40, no. 5, pp. 1385–1433, 2015.

[20] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i the essential algorithms," *IEEE Robotics Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[21] B. J. Guerreiro, P. Batista, C. Silvestre, and P. Oliveira, "Sensor-based simultaneous localization and mapping - Part II: online inertial map and trajectory estimation," in *Proceedings of the IEEE American Control Conference*, pp. 6334–6339, 2012.

[22] T. Pire, T. Fischer, J. Civera, P. De Cristoforis, and J. J. Berlles, "Stereo parallel tracking and mapping for robot localization," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015*, pp. 1373–1378, October 2015.

[23] F. Bellavia, M. Fanfani, F. Pazzaglia, and C. Colombo, "Robust selective stereo SLAM without loop closure and bundle adjustment," in *Proceedings of the International Conference on Image Analysis and Processing*, pp. 462–471, 2013.

[24] C. Brand, M. J. Schuster, H. Hirschmüller, and M. Suppa, "Stereo-vision based obstacle mapping for indoor/outdoor SLAM," in *Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2014*, pp. 1846–1853, September 2014.

[25] K. Schauwecker and A. Zell, "On-board dual-stereo-vision for the navigation of an autonomous MAV," *Journal of Intelligent & Robotic Systems*, vol. 74, no. 1-2, pp. 1–16, 2014.

[26] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proceedings of the European Conference on Computer Vision*, pp. 834–849, 2014.

[27] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: a versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[28] J. Ventura, C. Arth, G. Reitmayr, and D. Schmalstieg, "Global localization from monocular SLAM on a mobile phone," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 4, pp. 531–539, 2014.

[29] G. Bresson, T. Feraud, R. Aufrere, P. Checchin, and R. Chapuis, "Real-Time Monocular SLAM with Low Memory Requirements," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 1827–1839, 2015.

[30] T. Schops, T. Sattler, C. Hane, and M. Pollefeys, "3D modeling on the go: interactive 3D reconstruction of large-scale scenes on mobile devices," in *Proceedings of the 2015 International Conference on 3D Vision, 3DV 2015*, pp. 291–299, October 2015.

[31] A. Al-Nuaimi, M. Piccolrovazzi, S. Gedikli, E. Steinbach, and G. Schroth, "Indoor location retrieval using shape matching of kinectfusion scans to large-scale indoor point clouds," in *Proceedings of the 2015 Eurographics Workshop on 3D Object Retrieval*, pp. 31–38, 2015.

[32] Y. Takeda, N. Aoyama, T. Tanaami, S. Mizumi, and H. Kamata, "Study on the indoor slam using kinect," *Advanced Methods, Techniques, and Applications in Modeling and Simulation*, pp. 217–225, 2012.

[33] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "RGB-D mapping: Using depth cameras for dense 3D modeling of indoor environments," in *Proceedings of the 12th International Symposium on Experimental Robotics*, pp. 477–491, 2014.

[34] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[35] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, 2012.

[36] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[37] C. Daniel Herrera, J. Kannala, and J. Heikkilä, "Joint depth and color camera calibration with distortion correction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2058–2064, 2012.

[38] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[39] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[40] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: an efficient alternative to SIFT or SURF," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV '11)*, pp. 2564–2571, Barcelona, Spain, November 2011.

[41] P. H. S. Torr and A. Zisserman, "MLESAC: a new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

[42] T. F. Coleman and Y. Li, "An interior trust region approach for nonlinear minimization subject to bounds," *SIAM Journal on Optimization*, vol. 6, no. 2, pp. 418–445, 1996.

[43] A. V. Zykina, "A lexicographic optimization algorithm," *Automation and Remote Control*, vol. 65, no. 3, pp. 363–368, 2004.

[44] H. Zhang, "BoRF: Loop-closure detection with scale invariant visual features," in *Proceedings of the 2011 IEEE International Conference on Robotics and Automation, ICRA 2011*, pp. 3125–3130, May 2011.

[45] J. Wu, H. Zhang, and Y. Guan, "Visual loop closure detection by matching binary visual features using locality sensitive hashing," in *Proceedings of the 2014 11th World Congress on Intelligent Control and Automation, WCICA 2014*, pp. 940–945, July 2014.

[46] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós, "An image-to-map loop closing method for monocular SLAM," in *Proceedings of the 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, pp. 2053–2059, September 2008.

[47] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós, "A comparison of loop closing techniques in monocular

SLAM," *Robotics and Autonomous Systems*, vol. 57, no. 12, pp. 1188–1197, 2009.

[48] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proceedings of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR '07)*, pp. 225–234, Nara, Japan, November 2007.

[49] S. Leutenegger, P. Furgale, V. Rabaud, M. Chli, K. Konolige, and R. Siegwart, "Keyframe-based visualinertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[50] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 2320–2327, November 2011.

[51] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proceedings of the European Conference on Computer Vision*, pp. 490–503, 2006.

[52] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G^2o: a general framework for graph optimization," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA '11)*, pp. 3607–3613, Shanghai, China, May 2011.

[53] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, pp. 756–770, 2004.

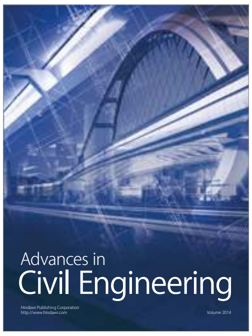
Journal of
Engineering

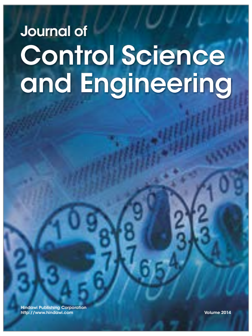
The Scientific
World Journal

International Journal of
Rotating
Machinery

Journal of
Sensors

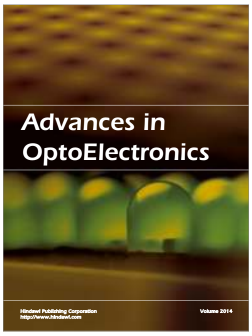
International Journal of
Distributed
Sensor Networks


Advances in
Civil Engineering


Journal of
Control Science
and Engineering

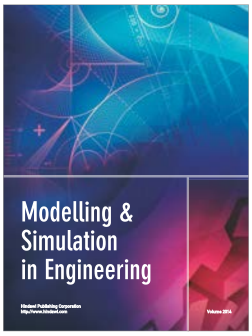
Journal of
Robotics

Journal of
Electrical and Computer
Engineering


Advances in
OptoElectronics

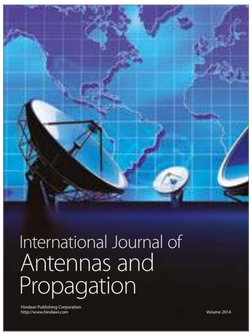
VLSI Design

International Journal of
Navigation and
Observation

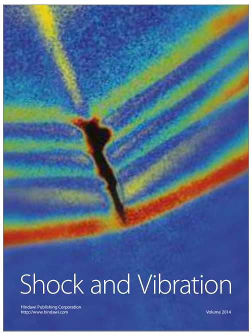

Modelling &
Simulation
in Engineering

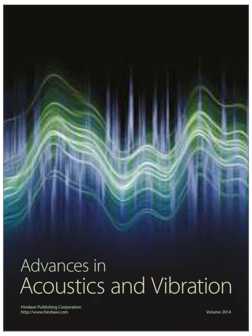
International Journal of
Aerospace
Engineering

International Journal of
Chemical Engineering


International Journal of
Antennas and
Propagation

Active and Passive
Electronic Components


Shock and Vibration


Advances in
Acoustics and Vibration