sensors

MDPI

Article

A Robust Indoor Localization System Integrating Visual Localization Aided by CNN-Based Image Retrieval with Monte Carlo Localization

Song Xu [1,*], Wusheng Chou [1,2] and Hongyi Dong [1]

[1] School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China; wschou@buaa.edu.cn (W.C.); donghylucky@buaa.edu.cn (H.D.)
[2] State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China
* Correspondence: keithxs@buaa.edu.cn

Received: 6 November 2018; Accepted: 7 January 2019; Published: 10 January 2019

Abstract: This paper proposes a novel multi-sensor-based indoor global localization system integrating visual localization aided by CNN-based image retrieval with a probabilistic localization approach. The global localization system consists of three parts: coarse place recognition, fine localization and re-localization from kidnapping. Coarse place recognition exploits a monocular camera to realize the initial localization based on image retrieval, in which off-the-shelf features extracted from a pre-trained Convolutional Neural Network (CNN) are adopted to determine the candidate locations of the robot. In the fine localization, a laser range finder is equipped to estimate the accurate pose of a mobile robot by means of an adaptive Monte Carlo localization, in which the candidate locations obtained by image retrieval are considered as seeds for initial random sampling. Additionally, to address the problem of robot kidnapping, we present a closed-loop localization mechanism to monitor the state of the robot in real time and make adaptive adjustments when the robot is kidnapped. The closed-loop mechanism effectively exploits the correlation of image sequences to realize the re-localization based on Long-Short Term Memory (LSTM) network. Extensive experiments were conducted and the results indicate that the proposed method not only exhibits great improvement on accuracy and speed, but also can recover from localization failures compared to two conventional localization methods.

Keywords: indoor global localization; monocular camera; laser range finder; image retrieval; convolutional neural network; kidnapping

1. Introduction

Global localization is a basic prerequisite for mobile robot navigation and control. The goal of mobile robot localization is to estimate the exact pose of mobile robot using only current sensor data based on a previously learned map. Accurate localization makes sense to many tasks such as motion control, path planning and target tracking [1–3].

In the past several years, extensive research has been conducted on indoor global localization, which could be divided into three categories from the perspective of the sensor: Wireless Local Area Network (WLAN)-based localization, laser-based localization and vision-based localization. WLAN-based localization realizes the localization process by combining the experience test and the signal propagation model based on the information of each network node, resulting in low cost but extremely susceptible to signal fluctuations and environmental interference [4,5]. In contrast, laser-based localization is more robust, in which the Bayesian filtering is leveraged to transform the mobile robot localization into the probability distribution estimation problem based on grid maps [6–9].

It not only achieves precise localization, but also enables continuous pose tracking of mobile robot. However, due to the presence of dynamic targets, frequent corrupted observations are prone to result in localization failures. With respect to vision-based localization, images can provide rich visual information such as geometric features and color textures [10–14], which is usually cast as image retrieval that finds the closest image in the geo-tagged database to the query image by means of feature matching. The retrieved geo-tagged image is likely to present the position where the query image is taken. Vision-based localization is widely applied to indoor localization due to its low cost, utmost convenience in use and rich information of geometric feature [15,16]. It is suitable for robot operation particularly in populated environments and does not suffer from the interferences often observed in light- or sound-based localization methods. However, it is hard to only use a camera for indoor localization because of the lack of distance information. Beside, its localization process is sensitive to the illumination and angle of objects.

The indoor environment tends to be highly similar in structure and layout. Images in different locations may contain the same object or repeated structural elements, resulting in global ambiguities, which is challenging for indoor localization if only through a single sensor. In this case, it is prone to cause mismatch if the localization is only performed by a single sensor. Therefore, it is necessary to fuse multi-sensor information for indoor localization [17–19].

In this paper, we propose a novel multi-sensor-based indoor global localization system. The proposed global localization system employs the coarse-to-fine mechanism to realize robust and efficient localization, which consists of coarse localization and fine localization. The coarse localization, which is cast as a multiclass place recognition problem, determines possible candidate locations where the robot may be located based on an image-based localization method, which comprises offline stage and online stage. In offline stage, we use a monocular camera to capture images by steering the mobile robot through the experiment environment, building a 2D geo-tagged image database applied for image retrieval. Each image in the database is labeled with ground truth pose information. Then, in online stage, the query image captured during the motion of robot is fed to the image retrieval system to determine the possible locations of robot. In the fine localization, a 2D laser range finder developed by HOKUYO in Japan is adopted to estimate the accurate pose of mobile robot by means of Monte Carlo localization based on the results of coarse localization. In addition, to address the "robot kidnapping" problem, a closed-loop mechanism is introduced to monitor the state of the robot in real time. When the mobile robot is kidnapped, the closed-loop mechanism will be activated rapidly and make adaptive adjustments to recognition the real location of robot based on the LSTM network with image sequences.

Our global localization system was tested and demonstrated in a real indoor environment. The experimental results show that the proposed system can perform better in different environments than the conventional localization method in terms of speed and accuracy for mobile robot. Besides, we further conducted robot-kidnapping experiments to verify that the proposed system provides the robot with the ability to recover from localization failure.

The remainder of this paper is organized as follows. Section 2 presents the related work. In Section 3, after describing the proposed indoor localization system, we elaborate the details of the coarse place recognition, fine localization and image sequence correlation with LSTM. Then, to verify the effectiveness and robustness of the proposed method, the results of various experiments are presented in Section 4. Finally, we conclude this paper in Section 5.

2. Related Work

2.1. Image Retrieval

Vision-based localization is usually cast as image retrieval that find the closest image in the geo-tagged database to the query image via feature matching [20–24]. Traditional image retrieval is mainly divided into text-based and content-based methods [25–28]. Text-based image retrieval [29,30]

adopts text annotation to describe the content of images, which could be formed with keywords, such as objects, scenes, etc. When performing image retrieval, the system searches the images marked with the corresponding keywords according to the query keywords provided by the user. The text-based image retrieval method is only suitable for small-scale image data, requiring manual intervention in the labeling process. For precise queries, it is sometimes difficult for users to describe the images they want to obtain with short keywords. Besides, the manual annotation process is inevitably affected by the level of cognition, verbal use, and subjective judgment of the tagger, which will cause the difference in image description. Content-Based Image Retrieval (CBIR) was then proposed, establishing the image feature vector description based on low-level visual features of images such as color, shape and texture, which can be divided into global features and local features. The global features extracted from the visual content of the entire image are not suitable for scenes with similar layouts and are susceptible to the interference of occlusion and illumination. In contrast, the local features extracted from regions of interest exhibit greater discrimination and robustness. After more than a decade of development, many CBIR methods have been presented. The work in [31] proposes Hierarchical k-means for scalable recognition with a vocabulary tree. Zhu et al. [32] later adopted the Hierarchical k-means to build a large vocabulary tree for quantization and represent each video clip by a visual Bag-of-Words for instance search that outperformed all submissions on the instance search dataset TRECVID 2011. The authors believed that the use of larger vocabularies is the reason for better performance. In [33], Jegou et al. proposed Hamming embedding which could maintain the discriminative power of descriptor in the cluster. Then, Vector of Locally Aggregated Descriptors (VLAD) [34] was proposed by Jegou, which has good performance in large-scale image retrieval. However, the global descriptor extensively exploited in CBIR has difficulty achieving the desired performance in the case of illumination, deformation and occlusion, which compromise the retrieval accuracy.

Recently, CNN-based image retrieval [35–38] has gained popularity and gradually overtaken conventional image retrieval methods. CNNs trained with massive volume of data have been shown the superior ability to learn feature representations. In 2014, Babenko et al. [39] first proposed a CNN-based method for image retrieval based on fine-turning model. The work in [40] introduces a method to extract convolutional features from each layer of the networks, which are encoded into a vector of image features by VLAD. In [41], a regional maximum activation of convolution is built to aggregate several image regions into a compact feature vector, which is robust to scale and translation. Salvador et al. [42] extracted the features of images based on pre-trained Faster-RCNN and constructed the image descriptors via Image-wise pooling of activations and Region-wise pooling of activations for spatial re-ranking.

2.2. Monte Carlo Localization

Over the past few decades, extensive research has been conducted on the laser-based localization. blue Dellaert et al. [6] proposed a Monte Carlo Localization (MCL) with the particle filter algorithm based on laser rang scan that utilizes a discrete particle set to represent the posterior probability distribution, in which the correlation of the observation information is not considered, resulting in the degradation of the particle set. In response to the problem, many scholars have proposed different strategies for improvement. blue Merwe et al. [43] proposed to adopt the unscented Kalman filter to generate the particle filter importance density function, which does not need to calculate the Jacobian matrix, and the estimation accuracy is better. However, Unscented Kalman filter is susceptible to system noise and has poor adaptive characteristics. Kim et al. [44] proposed to update the Gaussian mixture model by incremental method and introduce a new weight calculation method to make the localization algorithm more robust. To reduce the computational burden, Zhang et al. [45] proposed a self-adaptive Monte Carlo localization for mobile robots by employing a pre-caching technique. The work in [9] proposes a global localization approach that merge SVM-based place recognition and particle filter for indoor environment with high similarity based on 2D range scan data.

In this paper, we adopt an adaptive Monte Carlo localization method to estimate the accurate pose of robot that dynamically adjusts the number of particles based on the observation information. It not only effectively avoids the particle degradation in global localization, but also prevents robots from being kidnapped.

3. Proposed Multi-Sensor-Based Indoor Localization System

The overview of the proposed localization system for mobile robot is shown in Figure 1. The global localization process consists of two stages: coarse place recognition and fine localization. In the coarse place recognition, the monocular camera is utilized to capture the image during the motion of the mobile robot and build a 2D geo-tagged image database applied for image retrieval. In the image retrieval, a pre-trained CNN for object detection is exploited to extract the high dimensional convolution features of images, which are adopted to search for images in the 2D geo-tagged image database most similar to the query image. The retrieved images are likely to present the possible locations of mobile robot, which are taken as the input to the next stage. In the fine localization, a 2D laser range is equipped to estimate the accurate pose of mobile robot by means of Monte Carlo localization, in which the candidate locations obtained from coarse place recognition are considered as seeds for initial random sampling. A set of sampled particles is injected into the candidate locations in a timely manner to track the real location of robot. In this process, the CNN-based image retrieval not only accelerates the convergence of samples in Monte Carlo localization, but also makes the global localization more robust in indoor environments with highly similar style and layout. Furthermore, to address the "robot kidnapping" problem, a closed-loop mechanism is introduced to monitor the state of the robot in real time. When the mobile robot is kidnapped, the closed-loop mechanism will be activated rapidly and make adaptive adjustments, which effectively exploits the correlation of image sequences via LSTM network. The LSTM network provides the localization system with a memory function ability to model sequence image data, which provides great discrimination for localization system in highly similar indoor environments.

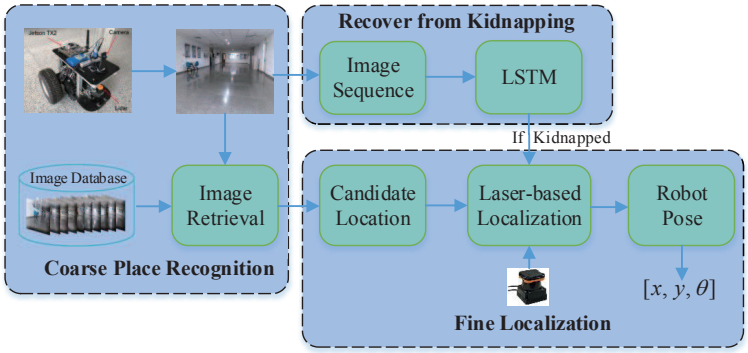


Figure 1. Overview of the proposed localization system.

3.1. Image-Based Localization

Image-based localization is usually cast as image retrieval process. There mainly exist two image-based localization methods: 2D image-based localization and 3D image-based localization [46]. The 2D image-based localization can be a component of 3D image-based localization to estimate where the query image might be taken. The 2D image-based localization searches the closest image in the geo-tagged database that presents the approximation position of mobile robot to the query image by means of feature matching. Then, the 3D image-based localization based on SFM is adopted to establish a set of 2D–3D matches to estimate the 6 DOF camera pose. However, building the 3D models required by structure-based techniques is not a trivial work, especially for large scale complex environment. Besides, the retrieval database image cannot provide valid information of local SFM

model resulting in occasional localization failure. By contrast, 2D image-based localization methods only needs a geo-tagged images database, which is easy to construct.

In this paper, we adopt a 2D image-based localization methods to estimate the coarse location of mobile robot, which are effective in terms of computational efficiency and stability based on CNN-based image retrieval. As one of the transfer learning methods, fine-tuning based on the pre-trained model can transform the general features into special features, thus enabling the transferred features to adapt the tasks. To get better feature representation, an effective way is to explicitly learn weights suited for specific retrieval task based on pre-trained models. The work of Babenko [39] represents that models pre-trained on ImageNet for object classification could be improved by fine-tuning them on an external set of Landmarks images, even when using a classification loss. Inspired by Gordo [42], the pre-trained model of Faster R-CNN is used to extract the high dimensional convolution features of images, which searches the Top-N images with high similarity to the query image in the geo-tagged image database. The proposed image retrieval method can be divided into offline phase and online phase.

3.1.1. Offline Phase

The dataset for image retrieval comprises 3550 images for training and 1032 images for testing with geo-referenced pose information, which is resized to 224 $\times$ 224 pixels. To generate ground truth pose information for each image, we steered the mobile robot through the indoor environment and captured the pictures from different positions. The positions of images were spaced roughly 2 m apart. At each position, we presented images at six different wide-angles covering full 360°. blue The dataset contains eight classes of common indoor objects, which can be used to construct local region descriptors for image retrieval.

During the training stage, we modified the output layers of the Faster RCNN to return nine class probabilities (eight classes in the dataset and an extra class for the background) and their corresponding regressed bounding box coordinates. Additionally, to obtain better feature representations for our image database, we chose to fine tune the model of Faster RCNN, which is considered as a feature extractor for image retrieval. In our experiments, we chose to update the weights of all layers after the first two convolutional layers to adapt to our query tasks. For the training details, we used a learning rate of 0.001 for 6000 iterations and 0.0001 for the next 2000 iterations. Weight decay and momentum were, respectively, set to 0.0005 and 0.9.

3.1.2. Online Phase

Figure 2 shows the architecture of image retrieval. CNNs trained with massive volume of data have been demonstrated the superior ability to learn feature representations. During image retrieval, features extracted from different layers perform differently. According to Hao [47], the features extracted from top semantic layers are not conducive to object retrieval, which loses the object's spatial information. In contrast, the activations of the convolutional layers may exhibit higher generalization ability and produce highly competitive compact image representations for object or scene retrieval. In fact, the fully connected layers can be removed in the process of exploiting intermediate layer activations as feature vectors and can remove the constraint of the input image size. In this paper, the conv5_3 layer of VGG16 is adopted to extract the features and to derive the representations for images. The extracted feature vector that represents the image to be localized after FC8 layer is one thousand dimensional. The process of image retrieval can be divided into two stages: initial filter stage and spatial re-ranking stage.

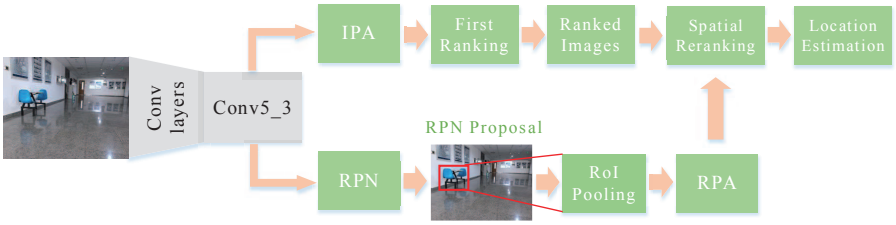


Figure 2. Architecture of the proposed image retrieval scheme.

In the initial filter stage, Image-wise Pooling of Activations (IPA) strategy is used to construct global image descriptors for both query and database images, in which sum pooling is adopted for feature representation after the layer of feature extraction. Exploiting sum pooling to aggregate features from activations of the last convolutional layer has proven to have highly competitive performance [40]. In the test phase, we search for the Top-N images in the geo-tagged image database that are most similar to the query image based on the similarity measurement for first ranking. In this paper, we use the Euclidean distance to compute the distances between the descriptors of the query image and the descriptors of database images. In traditional image-based localization system, the candidate image with highest ranking in similarity is considered as the best match, and the position where the candidate image is located is returned to achieve the localization. However, for indoor environments, mismatching may occur in scenes with similar structures and layouts since the spatial relationship of feature vectors is neglected. Therefore, spatial re-ranking is necessary in image retrieval.

Post-processing with spatial verification has proven to be an effective way to improve the performance of image retrieval. After first ranking, the Top-N ranked images are verified for spatial re-ranking. In the re-ranking stage, Region-wise Pooling of Activations (RPA) with max pooling is adopted to construct the local region descriptors of images by aggregating the convolutional activations for each of the object proposals from the RoI pooling layer of Faster R-CNN. Then, we compare the local region descriptors of the query bounding box with the region-wise descriptors for all RPN proposals of the Top-N ranked images based on the Euclidean distance. For the high-level or complex scene, local feature representation exhibits superior generalization ability and strong distinction in image retrieval. For the convenience of comparison, we warp the bounding box to the size of the feature maps in the last convolutional layer. The region with maximum similarity for the Top-N ranked images provides the object localization and its score is kept for re-ranking. After the RoI pooling, the feature descriptors are l2-normalized and then whitening is applied to increase its robustness against noise.

3.2. Laser-Based Localization

In this section, we focus on the proposed fine localization method, which is based on an adaptive Monte Carlo localization. The adaptive Monte Carlo localization method approximates the posterior probability of the estimated state through a set of discrete weighted particles and gradually implements filtering through state prediction, update weight and resampling. The robot cannot recover from the localization failure when the pose suddenly changes. The traditional Monte Carlo localization method addresses this problem by adding random particles. Its localization accuracy is related to the number of particles. The more particles there are, the better the robustness is. However, the computational complexity caused by the increase of particles will affect the real-time performance of global localization. In this paper, our adaptive Monte Carlo localization method combines the augmented Monte Carlo method and the KLD-sampling, which not only solves the kidnapping problem, but also adopts KLD-sampling to dynamically adjust the particle number to reduce the computational complexity and improve localization efficiency. Its localization process can be performed as the following steps:

3.2.1. Prediction Phase

In the prediction phase, a set of particles $S_t = \left\{ \left(x_t^{(i)}, \omega_t^{(i)} \right), i = 1, 2, \cdots, N \right\}$ is randomly injected into all possible positions of the robot obtained by image-based localization method. The motion model $p(x_t | x_{t-1}, u_{t-1})$ is adopted to predict the current state of the robot in the form of a predictive Probability Density Function (PDF) $p(x_t | z_t)$ where z_t represents the observation information at time t.

3.2.2. Update Phase

In the update phase, the importance weight ω_t of each particle is obtained by the proposal distribution $\pi\left(x_t^{(i)} \middle| m, x_{1:t-1}^{(i)}, z_{1:t}, u_{1:t-1} \right)$, which is considered as the motion model $p(x_t | x_{t-1}, u_{t-1})$.

$$\omega_t^{(i)} = \omega_{t-1}^{(i)} \frac{p\left(z_t \mid x_t^{(i)}, m \right) p\left(x_t^{(i)} \middle| x_{t-1}^{(i)}, u_{1:t-1} \right)}{\pi\left(x_t^{(i)} \middle| m, x_{1:t-1}^{(i)}, z_{1:t}, u_{1:t-1} \right)} = \omega_{t-1}^{(i)} p\left(z_t \mid x_t^{(i)}, m \right) \tag{1}$$

where $p\left(z_t \mid x_t^{(i)}, m \right)$ represents the observation model; m represents the a priori grid map; and $u_t - 1$ denotes the control information at time $t - 1$. The importance weights of each particle are updated using the observation model $p\left(z_t \mid x_t^{(i)}, m \right)$. In most cases, the sampled particles will gradually converge to a certain position. If the robot suddenly is kidnapped, the observation of the new particles is inaccurate. Additionally, the weight is normalized by the observation model and the difference is still obvious, which will lead to the localization failure. In this case, we inject random particles into the map when the observation of the particles is not very accurate based on augmented Monte Carlo method and assign the weight according to the observation model of the random particles. The quality of the current observation information is evaluated by comparing the long-term average weight of the particle set ω_{slow} with the short-term average weight of the particle set ω_{fast}. They are calculated as follows

$$\omega_{avg,\,t} = \frac{1}{N} \sum_{i=1}^{N} \omega_t^{(i)} \tag{2}$$

$$\omega_{slow,\,t} = \alpha_{slow} \omega_{avg,\,t} + (1 - \alpha_{slow}) \omega_{slow,\,t-1} \tag{3}$$

$$\omega_{fast,\,t} = \alpha_{fast} \omega_{avg,\,t} + (1 - \alpha_{fast}) \omega_{fast,\,t-1} \tag{4}$$

where $\omega_{a\,v\,g,\,t}$ is average weight of the particles at the current moment. Since the weight of sampled particles update is based on the observation model, the average weight can represent the quality of the current observation. The coefficients α_{slow} and α_{fast} are used to assign the ratio of the current average weight $\omega_{a\,v\,g,\,t}$, which is generally set to $0 < \alpha_{slow} \leq \alpha_{fast} < 1$. Then, the probability of adding particles during resampling is $\max\left\{ 0,\ 1 - \frac{\alpha_{fast}}{\alpha_{slow}} \right\}$. Therefore, when the long-term average weight is larger than the short-term average weight, the current average weight of the particles is small, the observation of the particle set is inaccurate and the probability of adding the random particles is higher.

After that, the particle set $\left\{ x_t^{(i)}, \omega_t^{(i)} \right\}$ is resampled. In the process of resampling, the weight of some particles is prone to be very small and almost can be ignored after several iterations, which is called particle degradation. Thus, many computations are wasted on particles that have little effect on estimating the state of sampled particles. Hence, we adopt KLD-sampling to solve the particle degradation. In the KLD-sampling, the Kullback–Leiber distance is used to measure the error between the true probability distribution and the approximation distribution. When uncertainty of the distribution is relatively high, that is, when the error between the true probability distribution and the approximation distribution is relatively large, we increase the number of sampled particles, which can ensure the robustness of the localization system, and vice versa. The KLD-sampling

method dynamically adjusts the number of particles, ensuring that the distribution of current particles approximates the target distribution with a minimum number of particles. The minimum number of samples N_{kld} is

$$N_{kld} = \frac{N_b - 1}{2\varepsilon}\left\{1 - \frac{2}{9\,(N_b - 1)} + \sqrt{\frac{2}{9\,(N_b - 1)}}z_{1-\delta}\right\}^3 \tag{5}$$

where N_b denotes the number of subspaces occupied by sample; ε is the maximum value of the target distribution error; and $1 - \delta$ represents the probability that the error is less than ε. When the number of generated particles is larger than N_{kld}, the process of sampling can stop.

The process of the adaptive Monte Carlo localization is shown in Algorithm 1.

Algorithm 1 Adaptive Monte Carlo localization

Input: observation information z_t, control information u_t;

Output: S_t;

Initialization: $N_r = 0$, $N_m = 0$, $N_{kld} = 0$, $N_b = 0$, $\omega_{avg} = 0$;
for all $b \in H$ **do** $b = 0$;
end for
if $rand() < max\{0, 1 - \omega_{fast}/\omega_{slow}\}$ **then**
    add random pose to S_{t-1};
    $N_r = N_r + 1$;
else
    draw i with probability $\propto \omega_{t-1}^{(i)}$;
    $N_m = N_m + 1$;
    $x_t^{(N_m)}$ = motion_model $(u_{t-1}, x_t^{(i-1)})$;
    $\omega_t^{(N_m)}$ = observation_model $(z_t, x_t^{(N_m)}, m)$;
    $S_t = S_t \cup x_t^{(i)}, w_t^{(i)}$;
    $\omega_{avg} = \omega_{avg} + \omega_t^{(i)}$;
    if $b(x_t^{(N_m)}) = 0$ **then**
        $b(x_t^{(N_m)}) = 1$;
        $N_b = N_b + 1$;
        if $N_b > 1$ **then**
            $N_{kld} = \frac{N_b - 1}{2\varepsilon}\left\{1 - \frac{2}{9(N_b-1)} + \sqrt{\frac{2}{9(N_b-1)}}z_{1-\delta}\right\}^3$;
        end if
    end if
end if
while $(N_m < N_{kld}$ && $N_r + N_m < N_{max})$ or $N_r + N_m < N_{min}$ **do**
    $\omega_{slow} = \alpha_{slow}\omega_{avg}/N_m + (1 - \alpha_{slow})\omega_{slow}$;
    $\omega_{fast} = \alpha_{fast}\omega_{avg}/N_m + (1 - \alpha_{fast})\omega_{fast}$;
end while
return S_t;

3.3. Image Sequence Correlation with LSTM

LSTM is a recurrent neural network (RNN) with memory function that can process temporal sequence, allowing the network to choose to forget the previous hidden state or update the hidden state during the learning process. Compared with traditional RNN, gradient explosion and gradient disappearance can be effectively addressed in LSTM. It has been successfully applied to many tasks, such as action recognition [48] and language modeling [49] and translation [50]. Recently, integrating LSTM and CNN has become a common means of capturing long-term temporal dependencies in the computer vision community. The work in [51] applied spatial LSTM to human re-identification, analyzing the bounding box of pedestrian detection to learn the relevance of embedded feature

space. LRCNs [52] present a novel end-to-end optimizable mapping from pixels to sentence-level natural language descriptions for video activity recognition by means of parsing the spatial and temporal dependencies.

In this paper, LSTM and CNN are intelligently combined to better learn the contextual information of image sequences and to enhance the discriminative capability of the local features, which can effectively avoid mis-retrieval caused by insufficient feature information of single image. Due to the high similarity in structure and layout of the indoor environment, the robot are prone to be kidnapped, resulting in the failure of localization. To address the above problem, we embed the LSTM network in the global localization system, which is a parallel process to the previous coarse place recognition and fine localization. Once the robot is kidnapped, the LSTM network will be activated to restore the pose of robot via image sequence correlation. We determine whether the robot is kidnapped by measuring the probabilities of samples. If the maximum of probabilities of samples is less than a threshold, we assume that the robot has been kidnapped. The re-localization based on LSTM can determine the accurate position of the robot, which is then input into the adaptive Monte Carlo localization to recover and track the pose of robot. Its architecture is shown in Figure 3. After the robot is kidnapped, we use multiple frames of time continuous images captured during the motion of the robot to realize the re-localization. The pre-trained VGG16 is used to extract the feature of each query image. VGG16 has five convolution layers, five pooling layers and three fully connected layers. The extracted features vector that represents the image to be localized after FC8 layer is one thousand dimensional. The features of each frame are considered as one chunk for one input of LSTM. In this paper, the length of the sequence in LSTM is empirically set to six frames in order to achieve a balance between the accuracy and speed of localization. After the LSTM layer, a mean pooling is adopted to construct feature descriptors for feature representation. Since the dimensionality of the feature descriptor is relatively high, which is not conductive to direct comparison with the images in database, we apply PCA to reduce the dimensionality of the feature descriptors to improve retrieval efficiency and then whitening to enhance its robustness against noise. The Adam is adopted for optimization with NVIDIA Tian X. In the training stage, we use a learning rate of 0.001 for cost minimization and a batch size of 75.

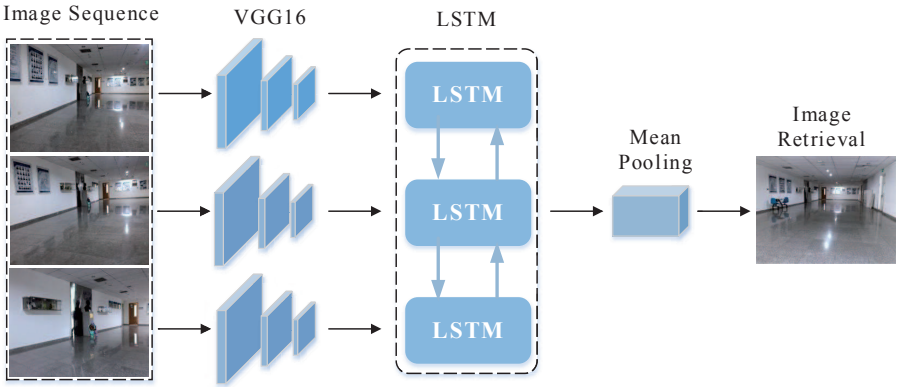


Figure 3. Architecture of the proposed LSTM network.

4. Experiments

The above localization method was verified in the new main building of Beihang University and tested on a real mobile robot, as shown in Figure 4. The mobile robot is equipped with a monocular camera and a 2D laser range finder. The monocular camera used to capture the images in the front of the robot is 640 $\times$ 480 pixels in resolution. The laser developed by HOKUYO can cover 30 m and 270°. The localization method was processed on an Nvidia Jetson TX2 with 256 core-Pascal GPUs.

The experimental environment is about 80 m $\times$ 80 m in size. The testing dataset consists of 1032 images with 2D geo-tagged information blue(X, Y, θ) for each image. They were obtained by steering the mobile robot through the indoor environment and capturing pictures from different positions in

the environment. The position of images are spaced roughly 2 m apart. At each position, we presented images at six different wide-angles covering 360°.

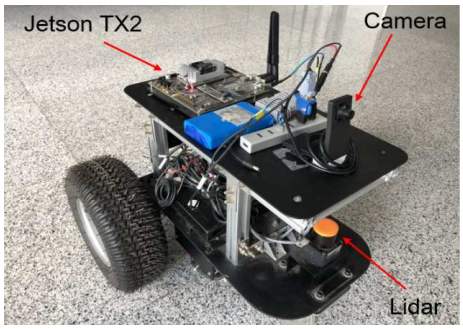


Figure 4. The mobile robot used to validate the proposed localization approach.

4.1. Experimental Setup

To obtain the feature representations for image retrieval in coarse place recognition, we chose to fine-tune the architecture of Faster RCNN to output the regressed bounding box coordinates and the class scores for each geo-tagged image in the tested database. The following eight object categories were selected for object detection to construct local feature descriptors in image retrieval, as shown in Figure 5. The selected eight classes of objects are most common in indoor environments, facilitating image-based localization. Moreover, they contain rich and significant texture information, which is conductive to detection.

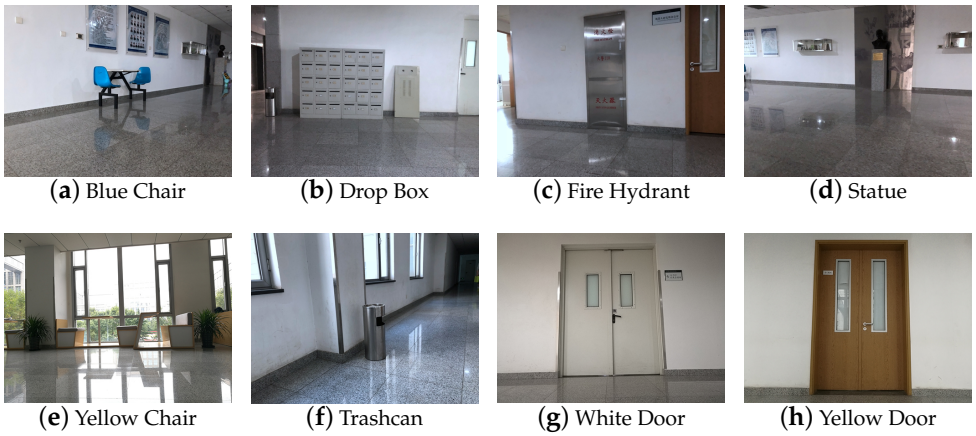


Figure 5. Classes for object detection in dataset.

In the image-based localization, we searched the Top-50 images that are similar to the query image in the geo-tagged image database based on the Euclidean distance between global image descriptors constructed by IPA for first ranking. After first ranking, the ranked Top-50 images were verified for spatial re-ranking, in which RPA with max pooling was adopted to construct the local region descriptors of images. In the traditional image-based localization system, the candidate image with the highest ranking in the similarity is considered as the best match, and the position where the Top-N candidate image is located is returned to achieve the localization. However, for indoor environments, mismatching may occur in scenes with similar styles and layouts since the spatial relationship of feature vectors is neglected in image retrieval. In this experiment, the locations of the Top-4 candidate images of voting in order were all considered to be the possible locations of the mobile robot, and used as seeds for initial random sampling of Adaptive Monte Carlo localization. Then, the particles were

randomly distributed at all possible initial location and the 2D fine pose of mobile robot was estimated based on probabilistic scan matching in conjunction with laser range scan and coarse pose information.

4.2. Coarse Place Recognition

The experiment was carried out to evaluate the performance of image-based localization in detail. Figure 6 shows the grid map constructed by mobile robot on the third floor of the new main building of Beihang University. The experimental environment is mainly composed of indoor corridors, which are quasi-symmetrical, so positioning in such an environment is fairly challenging. There are many repetitive objects in the experimental environment such as chairs, doors and windows. The numerical size of the grid-cell is 5 cm $\times$ 5 cm. In this established grid map, the level of gray color denotes the occupancy probability. The darker is the color, the higher is the occupancy probability. The blue arrows represent the location and the direction of taking images in the indoor environment. Additionally, five images captured from different nodes are shown.

Figure 7 presents some image retrieval results in the stage of coarse localization. For each row, the first image represents the query image taken by the mobile robot and the other four images are the re-ranked images. We adopted the image captured from node 1 as the query image for the localization experiment. At the testing stage, the MAP of the image retrieval method was about 0.742 for the Top-4 re-ranked images. For the query image in the first row of Figure 7, which corresponds to the image of node 1 in Figure 6, its Top-4 retrieved images correspond to four locations in the experimental environment. The first two candidate images located near node 1 are fairly accurate retrieval results, but the third and fourth candidate image taken from nodes 4 and 5 (in Figure 6) are mis-retrievals. Due to the structural similarity of the indoor environment and the interference of illumination and viewpoint, the occurrence of mis-retrieval is inevitable. In principle, the higher is the rank of the candidate image, the more similar it is to the query image. Therefore, the importance of the Top-4 re-ranked images is different. We defined the importance weight of the four retrieval images correspond to four candidate locations as 4:3:2:1.

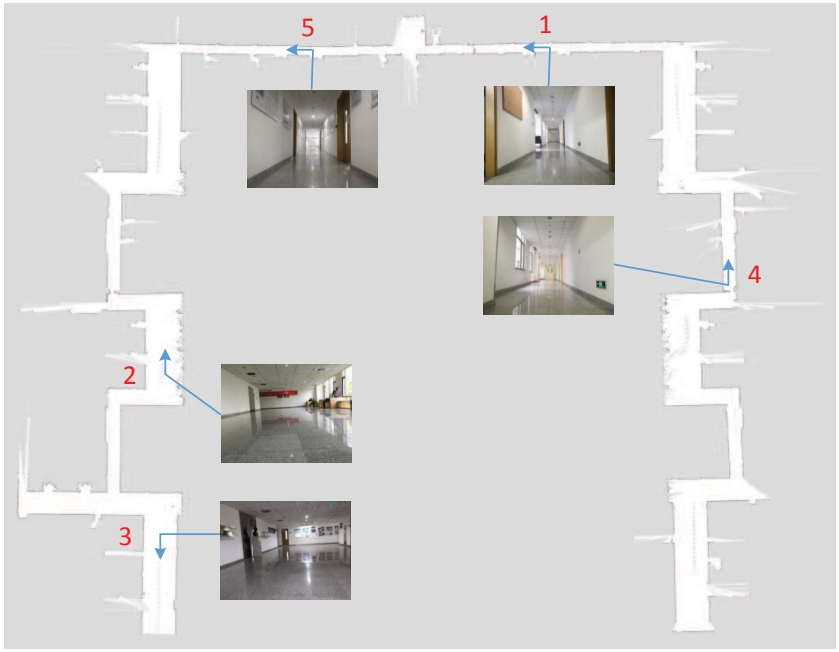


Figure 6. The grid map in the experimental site.

Figure 7. Three image retrieval results in the stage of coarse place recognition. The first column is the query image with a blue contour and the others are retrieved images. The images with red contour are mismatching.

4.3. Accurate Localization

4.3.1. Accurate Localization Estimation

Based on the importance weight obtained by the coarse place recognition, a set of particles was randomly injected around these locations and output to the initialization process of Adaptive Monte Carlo localization.

We focused on verifying the accuracy of the localization system to estimate the pose of the mobile robot. In Table 1, the 2D pose (X, Y, θ) of node 1, node 4 and node 5 estimated by coarse place recognition was selected as the candidate pose for random sampling. After several iterations, most particles tend to converge to the real position of the robot and estimate the fine pose of robot. As shown in Table 1, candidate node 1 was determined as the correct node. Node 4 and node 5 were candidate nodes similar to node 1 in structure and layout. The fine pose of the robot estimated by the proposed localization method was $(X, Y, \theta) = (1572.23\text{ cm},\ 3723.51\text{ cm},\ -179.64^\circ)$, and the ground truth pose of the robot was $(X, Y, \theta) = (1561.45\text{ cm},\ 3730.72\text{ cm},\ -173.42^\circ)$. It can be seen that the system could accurately estimate the pose of the robot and reliably keep track of it afterwards.

Table 1. Results of pose estimation relative to the query image from node 1.

	X (cm)	Y (cm)	θ (°)
Candidate node 1	1592.87	3720.63	−172.11
Candidate node 4	3681.13	304.76	94.25
Candidate node 5	−1548.34	3728.05	−169.68
Fine pose	1572.23	3723.51	−179.64
Ground truth	1561.45	3730.72	−173.42
Error	−10.78	7.21	6.22

4.3.2. Impact of the Image-Based Localization

The coarse place recognition results obtained by image retrieval can provide the candidate locations of robot for Monte Carlo localization, which not only speeds up the convergence of sampled particles but also significantly improves the robustness of the localization system, especially in complex environments with high similarities. All sampled particles are randomly generated in arbitrary areas of the grid map with the same weight when the image-based localization is removed. To evaluate the contribution of the image retrieval system, the Cumulative Distribution Function (CDF) of the localization error with and without image-based localization is shown in Figure 8. The case of not performing the image-based localization reflects running Adaptive Monte Carlo localization without

any prior information. As shown in Figure 8, the 90% localization error of the proposed method without image-based localization is less than 6 m. In contrast, the proposed localization with image-based localization can realize the localization process effectively. When the image retrieval is performed, the localization accuracy is fairly high and the 90% localization error is less than 2 m.

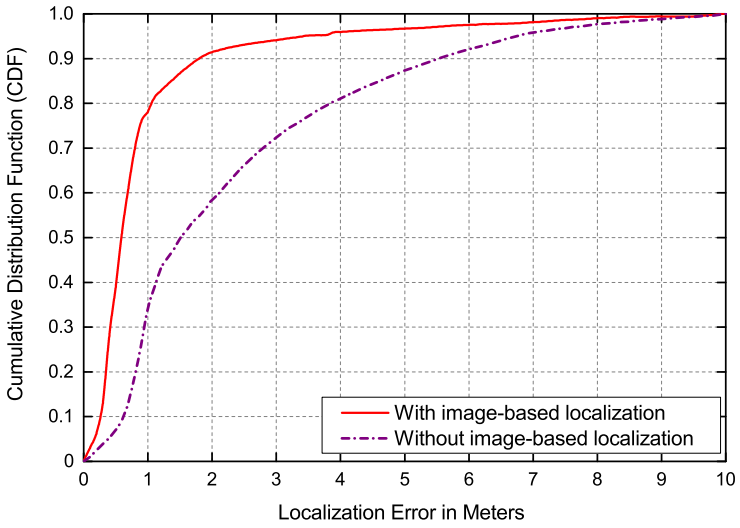
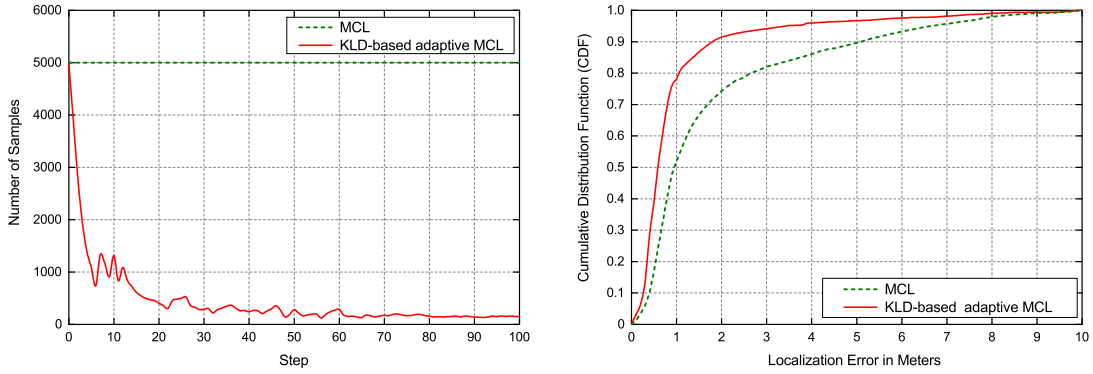


Figure 8. CDF of localization accuracy with and without image-based localization.

4.3.3. Impact of the Laser-Based Localization

To verify the performance of the laser-based localization proposed in this paper, comparative experiments among Monte Carlo Localization (MCL) and KLD-based adaptive MCL were performed, as shown in Figure 9. After determining the coarse location of the robot obtained by image-based localization, we adopted MCL and KLD-based adaptive MCL for fine localization to estimate the accurate pose of robot. The initial number of samples in global localization was uniformly set to 5000. Figure 9a presents the typical evolution of the number of samples of the two methods in global localization. As can be seen, the MCL method utilizes fixed number of samples for importance sampling during the entire state estimate process while the KLD-based adaptive MCL dynamically adjusts the number of samples according to the underlying state uncertainty, which can significantly improve the efficiency and reduce the computational complexity. As expected, with the improvement of localization accuracy, the samples decrease gradually in the KLD-based adaptive MCL. Figure 9b shows the CDF of localization accuracy of two methods in global localization. Obviously, the performance of KLD-based adaptive MCL outperforms MCL in global localization accuracy. The 50% localization error of KLD-based adaptive MCL is less than 0.6 m, while 50% localization error of MCL is less than 0.95 m.



(**a**) (**b**)

Figure 9. (**a**) Typical evolution of number of samples of two methods in global localization. (**b**) CDF of localization accuracy of two methods in global localization.

4.3.4. Evaluation of Proposed Localization System

To quantitatively evaluate the performance of the proposed global localization method, we chose mixture Monte Carlo localization (MCL) [3], image-based localization [20] and the proposed global localization method for comparative experiments. The maximum number of particles in the Monte Carlo localization was uniformly set to 5000. In the process of Monte Carlo localization, the mean of the sampled particles was used to estimate the pose of robot. The localization error versus execution step of three localization methods is depicted in Figure 10. As can be seen, the proposed localization method could estimate the pose of the robot efficiently. Moreover, in the initial stage localization, compared with the other two localization methods, the localization error of the proposed method converges significantly more quickly, because the coarse localization results of image retrieval provides reliable initial pose information robot for the next accurate localization stage, which can accelerate the convergence of particles. In addition, the proposed method is overall more robust in localization process compared with the other two localization methods. The adaptive Monte Carlo localization method in this paper adjusts the distribution of particles according to the observation model of system. Specifically, the average localization error of the proposed method is less than 40 cm, which is enough to meet the requirements for indoor localization.

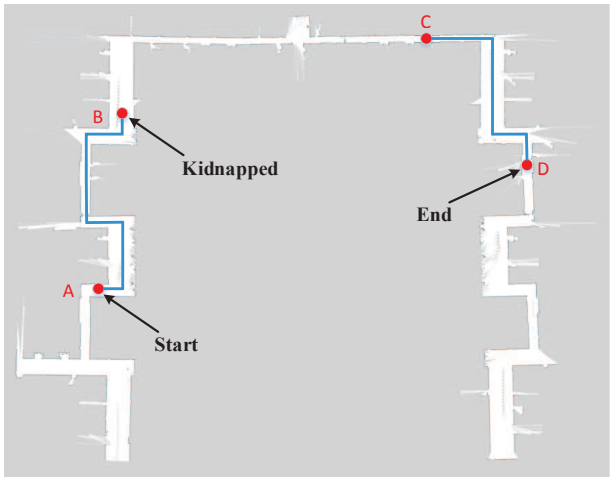


Figure 10. Localization error of three methods in global localization.

4.4. Re-Localization from Kidnapping

We focused on the ability of proposed localization system to deal with robot kidnapping which is a fairly tricky problem. To provide the robot with the ability to escape from kidnapping, we integrated the LSTM unit into CNN, so that the network can learn the feature correlation between image sequences. The embedding of the LSTM network will slow down the test speed. To reduce the computational cost of training and testing, we empirically limited the length of the sequence to six frames in this experiment, which not only ensures the recognition speed of the network, but also allows the captured image sequence to fall into the same area as much as possible.

The parameters and procedure of this experiment were consistent with the global localization experiment presented above. As shown in Figure 11, the motion path of robot represented by blue lines started from Point A to Point B. When the robot moved to Point B, its pose was already determined. Then, the robot was kidnapped to Point C. In this case, the system needed to re-localize the robot. When kidnapped to Point C, the robot moved towards Point D. In the relocation stage, 1000 particles were randomly injected into the grid map at each execution step based on the results of image sequence processing with LSTM. Figure 12 presents the typical localization error of three localization methods when the robot was kidnapped. The robot was kidnapped after its accurate pose was already determined. As can be seen, the MCL could not effectively restore the location of robot. Although the image-based localization method enables the robot to escape from the kidnapping, it takes a long time to recover because its image-based re-localization is not accurate enough. In contrast, exploiting the correlation between image sequences, the proposed method could relocate the robot rapidly based on its powerful temporal sequence analysis capabilities. After the robot is kidnapped, the proposed method needs less than 40 steps to recover the location of robot and its localization error is less than 40 cm. Moreover, the robustness of proposed method is much higher than the other two localization methods. The experiment results demonstrate that the proposed localization method are indeed a promising avenue to tackle kidnapping problem in repetitive structures or similar layouts, which are predominant in modern indoor environment and are a tricky issue for traditional global localization methods.

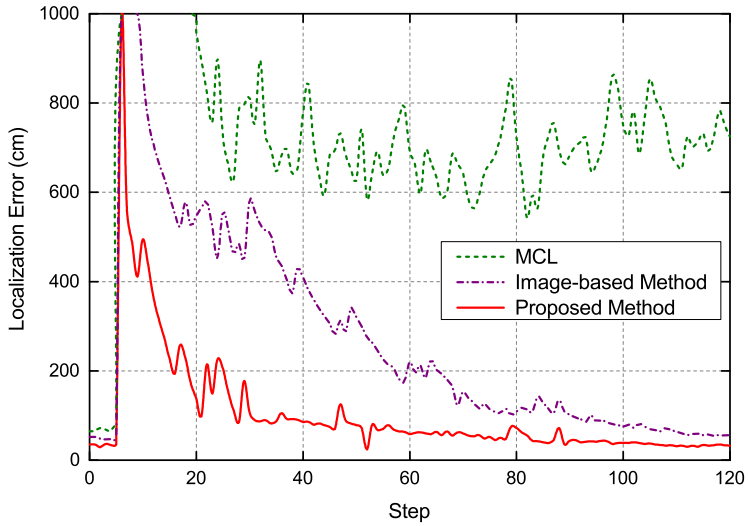


Figure 11. The motion path of the robot when it is kidnapped.

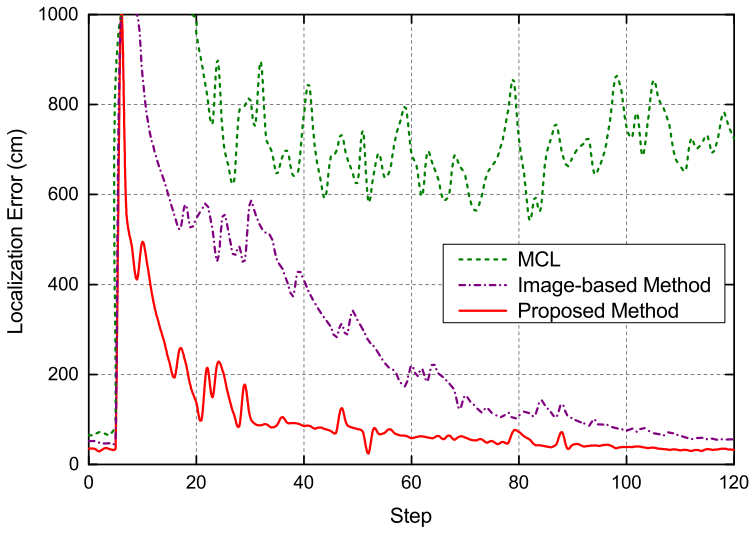


Figure 12. Typical localization error of three methods when the robot is kidnapped.

5. Conclusions

In this work, we propose a novel multi-sensor-based indoor global localization system integrating visual localization aided by CNN-based image retrieval with a probabilistic approach denoted as Monte Carlo localization. The image retrieval performed with a pre-trained Faster RCNN model seeks to estimate the coarse locations of a mobile robot, which is taken as the input for the next stage of Monte Carlo localization. Besides, exploiting the correlation between image sequences, a closed-loop mechanism is introduced to deal with the robot kidnapping problem. When the robot is kidnapped, the LSTM network will be activated rapidly and make adaptive adjustments to provide a guidance for global localization. The integration of both techniques constructs a robust and efficient localization system. Systematic experiments were conducted on a real mobile robot. The results indicate that the proposed localization method exhibits great improvement on the speed and robustness of indoor localization compared to conventional localization methods. In addition, the proposed localization system enables the robot with the ability to recover from localization failure.

In the future, we plan to validate the localization system in different indoor environments such as low illumination and occlusion, which bring great disruption for the image retrieval. Moreover, we plan to further optimize the performance of image retrieval to suit various indoor environments.

Author Contributions: S.X. and W.C. designed the study; S.X. and H.D. conducted the experiments; S.X. implemented the system and wrote the paper under the supervision of W.C.

Funding: This work was supported by the National Key R&D Program of China (Grant No. 2017YFB1302503) and the National Natural Science Foundation of China (Grant No. 61633002).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Thrun, S.; Burgard, W.; Fox, D. *Probabilistic Robotics*; MIT Press: Cambridge, MA, USA 2005.
2. Thrun, S.; Fox, D.; Burgard, W.; Dallaert, F. Robust Monte Carlo localization for mobile robots. *Artif. Intell.* **2001**, *128*, 99–141. [CrossRef]
3. Thrun, S.; Fox, D.; Burgard, W. Monte Carlo Localization with Mixture Proposal Distribution. In Proceedings of the National Conference on Artificial Intelligence (AAAI), Austin, TX, USA, 30 July–3 August 2000; pp. 859–865.
4. Liu, H.; Darabi, H.; Banerjee, P.; Liu, J. Survey of Wireless Indoor Positioning Techniques and Systems. *IEEE Trans. Syst. Man Cybern. C.* **2007**, *37*, 1067–1080. [CrossRef]

5. Jiang, Y.; Pan, X.; Li, K.; Lv, Q.; Dick, RP.; Hannigan, M.; Shang, L. ARIEL: Automatic wi-fi based room fingerprinting for indoor localization. In Proceedings of the ACM International Conference on Ubiquitous Computing (UbiComp), Pittsburgh, PA, USA, 5–8 September 2012; pp. 441–450.
6. Dellaert, F.; Fox, D.; Burgard, W.; Thrun, S. Monte Carlo localization for mobile robots. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Detroit, MI, USA, 10–15 May 1999; pp. 1322–1328.
7. Fox, D.; Burgard, W.; Thrun, S. Markov localization for mobile robots in dynamic environments. *J. Artif. Intell. Res.* **1999**, *2*, 327–391. [CrossRef]
8. Roumeliotis, S.I.; Bekey, G.A.; Burgard, W.; Thrun, S. Bayesian estimation and Kalman filtering: A unified framework for mobile robot localization. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), San Francisco, CA, USA, 24–28 April 2000; pp. 2985–2992.
9. Park, S.; Roh, K.S. Coarse-to-Fine Localization for a Mobile Robot Based on Place Learning With a 2-D Range Scan. *IEEE Trans. Robot.* **2016**, *32*, 528–544. [CrossRef]
10. Wang, J.; Zha, H.; Cipolla, R. Coarse-to-fine vision-based localization by indexing scale-Invariant features. *IEEE Trans. Syst. Man Cybern. B.* **2006**, *36*, 413–422. [CrossRef]
11. Sattler, T.; Leibe, B.; Kobbelt, L. Image retrieval for image-based localization revisited. In Proceedings of the British Machine Vision Conference (BMVC), Guildford, Surrey, UK, 3–7 September 2012; pp. 1–12.
12. Sattler, T.; Havlena, M.; Scjindler, K.; Pollefeys, M. Large-Scale Location Recognition and the Geometric Burstiness Problem. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1582–1590.
13. Sunderhaulf, N.; Dayoub, F.; McMahon, S.; Talbot, B.; Schulz, R.; Corke, P.; Wyeth, G.; Upcroft, B.; Milford, M. Place categorization and semantic mapping on a mobile robot. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 5729–5736.
14. Zeisl, B.; Sattler, T.; Pollefeys, M. Camera Pose Voting for Large-Scale Image-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 2704–2712.
15. Zamir, A.R.; Shah, M. Image Geo-Localization Based on MultipleNearest Neighbor Feature Matching UsingGeneralized Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1546–1558. [CrossRef] [PubMed]
16. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Minneapolis, MN, USA, 18–23 June 2007; pp. 1–8.
17. Biswas, J.; Veloso, M. Multi-sensor Mobile Robot Localization for Diverse Environments. In Proceedings of the Robot Soccer World Cup (RobotCup), Eindhoven, The Netherlands, 24 June–1 July 2013; pp. 468–479.
18. Srinivasan, K.; Gu, J. Multiple Sensor Fusion in Mobile Robot Localization. In Proceedings of the Canadian Conference on Electrical and Computer Engineering (CCECE), Vancouver, BC, Canada, 22–26 April 2007; pp. 1207–1210.
19. Duan, P.; Tian, G.; Wu, H. A multi-sensor-based mobile robot localization framework. In Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO), Zhuhai, China, 6–9 December 2015; pp. 642–647.
20. Wolf, J.; Burgard, W.; Burkhardt, H. Robust vision-based localization by combining an image-retrieval system with Monte Carlo localization. *IEEE Trans. Robot.* **2005**, *21*, 208–216. [CrossRef]
21. Irschara, A.; Zach, C.; Frahm, J.M.; Bischof, H. From structure-from-motion point clouds to fast location recognition. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Kyoto, Japan, 27 September–4 October 2009; pp. 2599–2606.
22. Sattler, T.; Leibe, B.; Kobbelt, L. Efficient & Effective Prioritized Matching for Large-Scale Image-Based Localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1744–1756. [PubMed]
23. Zamir, A.R.; Shah, M. Accurate Image Localization Based on Google Maps Street View. In Proceedings of the IEEE International Conference on Computer Vision (ECCV), Heraklion, Crete, Greece, 5–11 September 2010; pp. 255–268.
24. Kim, H.; Lee, D.; Oh, T.; Myung, H. A Probabilistic Feature Map-Based Localization System Using a Monocular Camera. *Sensors* **2015**, *15*, 21636–21659. [CrossRef] [PubMed]
25. Smeulders, A.W.M.; Worring, M.; Santini, S.; Gupta, A.; Jain, R. Content-Based Image Retrieval at the End of the Early Years. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1349–1380. [CrossRef]

26. Jing, F.; Li, M.; Zhang, H.J.; Zhang, B. A unified framework for image retrieval using keyword and visual features. *IEEE Trans. Image Process.* **2005**, *14*, 979–989. [CrossRef] [PubMed]
27. Zhou, X.S.; Huang, T.S. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst.* **2003**, *8*, 536–544. [CrossRef]
28. Smith, J.R.; Chang, S.F. VisualSEEk: A fully automated content-based image query system. In Proceedings of the Acm International Conference on Multimedia (ACMMM), Boston, MA, USA, 18–22 November 1996; pp. 87–98.
29. Wu, V.; Manmatha, R.; Riseman, E.M. TextFinder: An Automatic System to Detect and Recognize Text In Images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 1224–1229. [CrossRef]
30. Jung, K.; Kim, K.I.; Jain, A.K. Text information extraction in images and video: A survey. *Pattern Recognit.* **2004**, *37*, 977–997. [CrossRef]
31. Er, N.; Stewennius, H. Scalable recognition with a vocabulary tree. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, 17–22 June 2006; pp. 2161–2168.
32. Zhu, C.Z.; Satoh, S. Large vocabulary quantization for searching instances from videos. In Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR), Hong Kong, China, 5–8 June 2012; pp. 1–8.
33. Jegou, H.; Douze, M.; Schmid, C. Hamming Embedding and Weak Geometry Consistency for Large Scale Image Search. In Proceedings of the IEEE International Conference on Computer Vision (ECCV), Marseille, France, 12–18 October 2008; pp. 304–317.
34. Jegou, H.; Douze, M.; Schmid, C.; Perez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
35. Gong, Y.; Wang, L.; Guo, R.; Lazebnik, S. Multi-scale Orderless Pooling of Deep Convolutional Activation Features. In Proceedings of the IEEE International Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 392–407.
36. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
37. Radenović, F.; Tolias, G.; Chum, O. CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples. In Proceedings of the IEEE International Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 3–20.
38. Liu, H.; Tian, Y.; Wang, Y.; Pang, L.; Huang, T. Deep Relative Distance Learning: Tell the Difference between Similar Vehicles. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2167–2175.
39. Babenko, A.; Slesarev, A.; Chigorin, A.; Lempitsky, V. Neural Codes for Image Retrieval. In Proceedings of the IEEE International Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; pp. 584–599.
40. Ng, J.Y.H.; Yang, F.; Davis, L.S. Exploiting local features from deep networks for image retrieval. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA, 7–12 June 2015; pp. 53–61.
41. Gordo, A.; Almazán, J.; Revaud, J.; Larlus, D. Deep Image Retrieval: Learning Global Representations for Image Search. In Proceedings of the IEEE International Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 241–257.
42. Salvador, A.; Giroinieto, X.; Marques, F.; Satoh, S. Faster R-CNN Features for Instance Search. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 26 June–1 July 2016; pp. 394–401.
43. Merwe, R.V.D.; Doucet, A.; Freitas, N.D. The unscented particle filter. In Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS), Denver, CO, USA, 1–2 December 2000; pp. 563–569.
44. Kim, J.; Lin, Z.; Kweon, I.S. Rao-Blackwellized particle filtering with Gaussian mixture models for robust visual tracking. *Comput. Vis. Image Underst.* **2014**, *125*, 128–137. [CrossRef]
45. Zhang, L.; Zapata, R.; Lépinay, P. Self-adaptive Monte Carlo localization for mobile robots using range finders. *Robot* **2012**, *30*, 229–244. [CrossRef]

46. Sattler, T.; Torii, A.; Sivic, J.; Pollefeys, M.; Taira, H.; Okutomi, M.; Pajdla, T. Are Large-Scale 3D Models Really Necessary for Accurate Visual Localization?. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1637–1646.
47. Hao, J.D.; Dong, J.; Wang, W.; Tan, T.N. What Is the Best Practice for CNNs Applied to Visual Instance Retrieval?. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017; pp. 1–13.
48. Ullah, A.; Ahmad, J.; Muhammad, K.; Sajjad, M.; Baik, S.W. Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features. *IEEE Access* **2018**, *6*, 1155–1166. [CrossRef]
49. Zaremba, W.; Sutskever, I.; Vinyals, O. Recurrent Neural Network Regularization. *arXiv* **2014**, arXiv:1409.2329.
50. Luong, M.T.; Sutskever, I.; Le, Q.V.; Vinyals, O.; Zaremba, W. Addressing the Rare Word Problem in Neural Machine Translation. *arXiv* **2014**, arXiv:1410.8206.
51. Varior, R.R.; Shuai, B.; Lu, J.; Xu, D.; Wang, G. A Siamese Long Short-Term Memory Architecture for Human Re-identification. In Proceedings of the IEEE International Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 135–153.
52. Donahue, J.; Hendricks, L.A.; Marcus, R.; Venugopalan, S.; Guadarrama, S.; Saenko, K.; Darrell, T. Long-term Recurrent Convolutional Networks for Visual Recognition and Description. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 2625–2634.