Article

ERF-IMCS: An Efficient and Robust Framework with Image-Based Monte Carlo Scheme for Indoor Topological Navigation

Song Xu ^{1,*} , Huaidong Zhou ¹ and Wusheng Chou ^{1,2}¹ School of Mechanical Engineering and Automation, Beihang University, Beijing 100191, China; hdzhou@buaa.edu.cn (H.Z.); wschou@buaa.edu.cn (W.C.)² State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China

* Correspondence: keithxs@buaa.edu.cn

Received: 4 September 2020; Accepted: 25 September 2020; Published: 29 September 2020

Abstract: Conventional approaches to global localization and navigation mainly rely on metric maps to provide precise geometric coordinates, which may cause the problem of large-scale structural ambiguity and lack semantic information of the environment. This paper presents a scalable vision-based topological mapping and navigation method for a mobile robot to work robustly and flexibly in large-scale environment. In the vision-based topological navigation, an image-based Monte Carlo localization method is presented to realize global topological localization based on image retrieval, in which fine-tuned local region features from an object detection convolutional neural network (CNN) are adopted to perform image matching. The combination of image retrieval and Monte Carlo provide the robot with the ability to effectively avoid perceptual aliasing. Additionally, we propose an effective visual localization method, simultaneously employing the global and local CNN features of images to construct discriminative representation for environment, which makes the navigation system more robust to the interference of occlusion, translation, and illumination. Extensive experimental results demonstrate that ERF-IMCS exhibits great performance in the robustness and efficiency of navigation.

Keywords: vision-based topological mapping and navigation; image-based monte carlo localization; image retrieval; CNN features

1. Introduction

Conventional approaches to robot navigation mainly rely on metric maps to provide rich geometric information of the environment. Metric maps are easy to maintain and can provide precise navigation strategies for the mobile robot [1,2]. However, this kind of map will cause the problem of large-scale structural ambiguity and lack semantic information of the environment, which is not conducive to human-robot interaction [3,4].

In recent years, computer vision has been extensively applied to robot navigation [5–7]. Since it can provide rich features of environment which is conducive to eliminate the ambiguity of objects and places categories compared with traditional range sensors. As a highly abstracted graph-based representation of the environment, the vision-based topological map does not require precise location information of the robot and compactly describes the environment as topological diagram with nodes and edges, in which each node corresponds to a scenario with distinctive features and each edge encodes adjacency relationships between nodes [8,9]. Global consistent information is not essential when the robot navigates in a local region. Just as a person perceives a novel environment, he is unlikely to take notice of all the details of the environment, but only record the typical local landmarks

that is conducive to navigate even in a complex environment, such as corners and doors. Compared to a metric map, the topological map takes up less storage memory and, consequently, is suitable for long distance qualitative navigation [10]. More specifically, vision-based topological mapping has drawn extensive attention due to its low cost and rich informative representation of the environment. In vision-based topological mapping, local feature descriptors are generally adopted to search visual keyframes for estimating relative pose of the robot, such as binary robust independent element feature (BRIEF) [11] and scale-invariant feature transform (SIFT) [12]. In practice, such feature descriptions perform well for providing certain degree of invariance in camera rotations and image aliasing, but they are short of ability to cope well with some challenging visual scenarios in indoor navigation, such as large illumination variance and lack of texture, because these methods only extract the local gradient feature description and do not capitalize on other discriminative global information encoded in images. In a nutshell, the improvement of robustness and efficiency is still the main issue facing indoor visual navigation [13].

To tackle above challenges, for this paper, we proposed a scalable vision-based topological mapping and navigation method for a mobile robot to work robustly and flexibly in large-scale indoor environment, exploiting the fine-tuned convolutional neural network (CNN) feature as the image representation to perform image matching. Our major contributions are summarized as follows:

- An efficient and robust framework with image-based Monte Carlo scheme for indoor topological navigation, namely ERF-IMCS, is presented which combines the image retrieval and Monte Carlo localization to provide the robot with the capability to avoid perceptual aliasing.
- To the best of our knowledge, ERF-IMCS is the first work to employ hash encoding technology to perform efficient CNN feature indexing and visual localization in topological navigation.
- We proposed a effective visual localization method, simultaneously employing the global and local CNN features of images to construct discriminative representation for environment, which makes the navigation system more robust to the interference of occlusion, translation and illumination. Experiments on a real robot platform show that ERF-IMCS exhibits great performance in terms of robustness and efficiency.

2. Related Work

2.1. Image Retrieval

Vision-based topological localization is generally cast as image retrieval that searches the most similar keyframe in the geo-tagged topological map to the query image via feature descriptions matching [14,15]. To this end, the quality of the topological mapping and localization directly depends on the performance of image retrieval method.

As a long-standing research topic in the computer vision field, image retrieval has been extensively investigated [16,17]. Conventional approaches to image retrieval can be divided into two categories: text-based methods and content-based methods [18–20]. In text-based image retrieval, text annotation is employed to represent the images, forming keywords for each image, such as objects and scenes. Manual intervention is required in the labeling process, so text-based method is only suitable for small-scale image data. Afterwards, content-based was proposed, indexing the images by visual cues, such as color, shape, and texture. Early works for content-based image retrieval are based on global descriptors [21]. However, a well-known problem is that global features are susceptible to the interference of occlusion, translation, and illumination, which restricts its application scope. This issue prompted extensive studies on local feature based image retrieval to emerge, such as SIFT and SURF. Concurrently, encoding techniques that aggregate the features to construct an image descriptor has been extensively studied, such as vector of locally aggregated descriptors (VLAD) [22] and Fisher vector (FV) [23].

Recently, features generated by convolutional neural networks has exhibited superior ability for image representation in various tasks. Extensive works have proved that features extracted from

CNN can achieve state of the art performance in image retrieval [24,25]. The work [26] studied the applicability of using the activation from full connectivity layer of CNN to describe the image for image retrieval. In the work [27], off-the-shelf features extracted from the OverFeat network was adopted as the image representation, which exhibited powerful ability for image retrieval. In addition, there exists many works focus on the region descriptors for image retrieval. The work [28] constructed a regional maximum activation of convolution (R-MAC) by aggregating compact features in a fixed layout spatial region which is robust to scale and translation. In the work [29], off-the-shelf features extracted from a pre-trained object detection CNN are used for initial filtering and region features constructed by region-wise pooling of activations are adopted for spatial re-ranking. Our image retrieval method required in topological localization is inspired by the work [29]. However, complementing previous work, we address the efficiency problem in real-time image retrieval by introducing the hash index technology.

2.2. Topological Mapping and Localization

Generally, the visual topological localization is cast as a matching process of image-to-node, determining the topological node corresponding to the position of the robot [30]. To this end, the measurement between the acquired image of the robot and topological nodes need to be determined. For instance, in Reference [31], a novel topological localization method adopting classification with reject option in omnidirectional images was proposed to increase the accuracy in all cases. The work [13] employed sharpness measure to select high-quality keyframes and used vector field consensus to retrieve the keyframe efficiently and robustly. In Reference [32], a novel topological localization method based on multiple observation fusion was proposed, regarding the corridors and the intersections as vertices and branches correspondingly to build the topological map. Concurrently, many works pay attention to the multiple view geometry validation technology to obtain the accurate pose of the robot for navigation. The work [33] proposed Topomap, a lightweight topological mapping and navigation system designed to extract free space from sparse visual features. This method can deal with noisy and sparse visual observation. The work [10] presented a robust topological navigation method for omnidirectional mobile robot. In the hierarchical localization process, the robot recognizes the coarse location via node recognition algorithm and computes the precise metric pose via five-point relative pose estimation algorithm.

With respect to robot topological localization, the above research works are based on maximum likelihood scheme, matching the current view image of the robot with the reference images in topological node directly. This strategy may suffer from perceptual aliasing in an indoor environment with similar geometric structure and layouts, resulting in localization failure. To tackle above problem, many works adopt Bayesian filtering to generate a maximum posteriori probability, fusing the past estimation and current observation. The work in Reference [34] proposed a lightweight scene recognition approach to realize topological modeling via adaptive descriptors, using a hierarchical Bayesian modeling to perform online inference process of topological localization automatically. Reference [35] described a topological navigation system in large-scale and complex environment, utilizing the maximum posteriori scheme to estimate the probability of the current view image, in which each node in topological map was assigned the same probability in the initial localization phase. Then, the observation of the robot was utilized to estimate the probability of the robot location iteratively for global localization.

3. Main Design of ERF-IMCS

The logical scheme of ERF-IMCS to perform topological navigation for a mobile robot is shown in Figure 1. ERF-IMCS consists of three stages, named topological mapping, global localization, and path planning, respectively. The purpose of the topological mapping stage is to search the keyframes to represent the environment discriminatively. These keyframes described by CNN features represent visually different nodes of the topological map in the environment. In each topological

node, the keyframes provide semantic visual information for the robot to perform navigation more flexible and more robust. In the global localization stage, an image-based Monte Carlo localization is adopted to estimate the robot pose, integrating the image retrieval mechanism into particle filter. The image retrieval is responsible for calculating the importance weight for the resampling of samples. Besides, with respect to the image retrieval, we employ the coarse-to-fine mechanism and hash index technology to perform the process of image matching robustly and efficiently. The final stage is path planning, in which Dijkstra algorithm is employed to search the shortest graph-based path from current topological node to the target topological node. Path planning is performed by traversing a sequence of keyframe images in the topological map.

Figure 1. Architecture of ERF-IMCS.

3.1. Vision-Based Topological Mapping

During the topological mapping, we steer the robot through all parts of the indoor environment, recording the images and odometry information to search keyframes by CNN-based feature representation. A threshold of the similarity is adopted to judge whether the query image matches to the keyframes in database. The query image will be added into the database if it does not match the keyframes in database. And a link between the query image and the previous keyframes will be added to the link graph. In the case of finding a match in the database, the query image will not be added, only the link is constructed. This creates a loop in the topological structure. In the process of topological localization, the robot is located to the closest node by means of comparing the currently captured image with the keyframes in topological nodes. Therefore, the selection of keyframes is the main issue in constructing topological map, which should fulfill following principles: (1) the similarity between the selected keyframes should be greater than a certain threshold to ensure the discriminative ability and representativeness; and, (2) for collected images with few local visual features, more keyframes should be selected nearby to avoid instability during the localization process. To this end, the feature representation of the keyframes is quite crucial. In order to obtain a compact representation of the keyframes, we characterize the images by a set of most relevant features of the environment. In each topological node, an object detection CNN are adopted to extract the features and object categories of keyframes for the robot to perform image retrieval.

The flowchart of the topological mapping is shown in Figure 2. For the current image captured by robot, we compute the similarity between the current image and the existing keyframes in topological map based on the image matching method in Section 3.2. We construct a new topological node when the similarity between the current image and the existing keyframes in topological map is greater than a certain threshold, inserting the current image as a new keyframe. Furthermore, an edge is added between the current node and previous node, which represents the connectivity between different nodes. Since the path of capturing keyframes in a topological mapping is completely guided by a human, the edges in a topological map are regarded as the reference for path planning. After that, a topological map is constructed. As shown in Figure 3, the keyframe images and robot pose are stored in each topological node. The keyframe images in each node are responsible for providing

the visual semantic information of local scene, such as visual features and observed object category for the robot. The visual features of images in the database are transformed to binary hash codes to facilitate efficient image matching, in real-time, for global localization. In this way, it is conducive to perform navigation flexibly in a human-like mechanism for the mobile robot which is not available in conventional metric-based navigation method. For example, when the target coordinates where human commands the robot to navigate is occupied, the robot will continuously search around the target location, even though it has already arrived the goal in metric-based navigation. In contrast, the proposed vision-based topological navigation system provides the robot with the capability to navigation flexibly without requiring specific coordinates of target location.

Figure 2. The flowchart of the topological mapping.

Figure 3. Schematic diagram of the topological map.

3.2. Image Retrieval

Image-based topological localization is usually cast as an image retrieval process, searching for the keyframes closest to the current image captured by the robot in the geo-tagged topological map based on feature matching. These keyframes described by visual features represent different nodes of the topological map in the environment. As one of the transfer learning methods, fine-tuning based on pre-trained model that transforms the general features into special features has been proven to achieve state of the art performance in feature representation. To obtain better feature representation, an effective way is to explicitly learn weights suited for specific image retrieval task based on pre-trained models. Inspired by Reference [29], we adopted a pre-trained object detection CNN to extract the high dimensional convolution features of query image for image retrieval. However, complementing the works in Reference [29], we addressed the efficiency problem in real-time image retrieval by introducing the hash index technology for performing the process of image matching quickly. In the image retrieval, we presented a coarse-to-fine mechanism consisting of two stages, i.e., coarse filtering based on global feature and fine retrieval based on local feature, as depicted in Figure 4.

Figure 4. Architecture of the proposed image retrieval scheme.

(1) Coarse Filtering

According to Reference [36], the features extracted from top semantic layers are not suitable for image retrieval because of lack of spatial information of objects. In contrast, the activations pooled from convolutional layers exhibit powerful ability to generalize images. Additionally, extensive works demonstrate that the activations from convolutional layers significantly outperform the activations from fully connected layers in image retrieval because the fully connected layer is highly related to classification tasks, which is less effective than the convolution layers at image retrieval tasks. To this end, we adopted the conv5 layer of ZF [37] to extract the global feature of images in which the dimension is 256 in the coarse filtering stage. According to the results in Reference [38], ZF architecture, which has 5 shareable conv layers, is significantly more efficient than VGG-16 architectures in real-time image retrieval although the performance of ZF architectures in Faster R-CNN is slightly inferior to VGG16 architectures in accuracy. After feature extraction layer, we adopt Image-wise Pooling of Activations (IPA) strategy to generate the compact global descriptors for both query images and keyframe images in topological map based on sum pooling. Sum pooling has proven to achieve competitive performance in aggregating feature activations [39]. However, quickly searching the top-N keyframe images closest to query image is a non-trivial work in real-time image retrieval, especially for large scale dataset. Therefore, we introduce a hash index technology to accelerate the image matching process. After obtaining the compact global features from the pooling layer, we employ iterative quantization (ITQ) [40] to transform the 256-dimensional CNN features to 64-bit binary hash codes. ITQ is an efficient image matching method that can encode high-dimensional CNN features into a discriminative binary code. This method transforms the features into low-dimensional codes while maintaining the adjacent relationship of data in the original space, thereby minimizing the loss of information. In the coarse retrieval stage, the Hamming distance of the hash code is computed. Then, we search for the top-N keyframes in each topological node closest to the current query image captured by the robot based on the similarity measurement of Hamming distance.

(2) Fine Retrieval

Perceptive aliasing may occur in an indoor environment with similar layouts and geometric structure if the top-N keyframes obtained by the coarse filtering based on global features is taken as the final result of image retrieval. In contrast, local features exhibit superior generalization in environment with similar geometric structure and layouts. Hence, it is imperative to adopt fine retrieval for re-ranking based on local region features of images. In the fine retrieval stage, we employ the Region-wise Pooling of Activations (RPA) to build the local compact descriptors for both query

images and keyframe images in topological map, aggregating the activations for each proposal from the ROI pooling layer. After that, the local region descriptors, which are l2-normalized and whitened, are encoded as hash codes via ITQ for fast image matching as in the coarse filtering stage. Then, the hash code of the query bounding box is compared with all the RPN proposals of the top-N ranked keyframes to generate a similarity based on Hamming distance. For convenience, we warp the bounding box of each proposal to the same size of the feature maps in the last convolutional layer. After that, query expansion which significantly increases the retrieval accuracy at a comparatively low time cost is performed. Finally, the similarity between the query image and the images in database generated by fine retrieval is injected into the next image-based Monte Carlo localization stage to compute the weights of the possible location of the robot. Based on the importance weights, a set of particles was randomly injected around these locations and output to the initialization process of image-based Monte Carlo localization.

(3) Iterative Quantization

In this work, the ITQ hash encoding method [40] is introduced to transform the CNN features obtained in coarse filtering and fine retrieval into 64-bit binary hash codes while maintaining the adjacent relationship of data in the original space, which minimizes the loss of information. Assuming that $X = \{x_1, x_2, \dots, x_n\}$, $x_i \in R^d$ is a set of feature data and are zero-centered. The goal of ITQ is to learn a binary code matrix $B \in \{-1, 1\}^{n \times c}$, where c is the code length. The binary encoding function is defined as $B = \text{sgn}(XW)$, where W is the matrix with column vector of hyperplane coefficients w_k . In order to make the hash codes of different data perpendicular to each other and maximize the variance between each bit, the following objective function can be constructed:

$$\begin{aligned} \tilde{F}(W) &= \sum_k E(\|xw_k\|_2^2) = \frac{1}{n} \sum_k w_k^T X^T X w_k \\ &= \frac{1}{n} \text{tr}(W^T X^T X W), \quad W^T W = I \end{aligned} \quad (1)$$

where x is original feature vector, and w_k is a column vector of hyperplane coefficients. The constraint $W^T W = I$ requires the hashing hyperplanes to be orthogonal to each other. As can be seen, if W is an optional solution, there is $\tilde{W} = WR$ for any orthogonal matrix R . So, the projected data $V = XW$ can be orthogonally transformed to minimize the quantization loss:

$$Q(B, R) = \|B - VR\|_F^2. \quad (2)$$

The process of ITQ is to search a local minimum of the quantization loss, which can be divided into two steps.

First step: fix R and update B .

For a fixed R , minimizing (2) is equivalent to maximizing

$$\text{tr}(BR^T V^T) = \sum_{i=1}^n \sum_{j=1}^c B_{ij} \tilde{V}_{ij}, \quad (3)$$

where $\tilde{V} = VR$, updating B through $B = \text{sgn}(VR)$ to maximize (3).

Second step: fix B and update R .

For a fixed B , minimizing (2) is equivalent to a classic Orthogonal Procrustes problem. First, decompose the matrix $B^T V$ into $S\Omega\hat{S}^T$ via SVD algorithm, and then update the matrix R through $R = \hat{S}\hat{S}^T$. The learning process is completed by repeated iteration until convergence.

3.3. Image-Based Monte Carlo Localization

For image-based topological localization, the top retrieved images in above image retrieval process are likely to present the possible locations of the mobile robot. Conventional methods mainly perform

the topological localization by means of maximum likelihood, directly taking the result of image matching in image retrieval as the location of the robot. These methods may be affected by noise or occlusion in indistinctive environment, resulting in perceptual aliasing. To tackle above challenge, an image-based Monte Carlo localization is introduced to estimate the maximum a posterior over the topological space via Bayesian Filtering in this paper. By computing the posterior distribution of sample set, the data association between the current observation, and the corresponding topological node is determined to provide the most general solution for perceptual aliasing.

The place of the topological map is defined as $x \in X$. Z denotes the collection of observed images z_s and $Z(x)$ corresponds to the observation at place x . According to Bayesian Filtering theory, for a new image captured by the robot at time t , the localization process that is responsible for determining the place of the robot x_t is cast as estimating the posterior probability distribution $p(x_t|z_t, u_t)$ based on the control sequence of the robot u_t and the observation z_t . The image-based Monte Carlo localization consists of the following two phases:

(1) Prediction Phase

In prediction phase, a set of particles $s_t = \{(x_t^i, \omega_t^i), i = 1, 2, \dots, N\}$ is randomly injected around most possible topological nodes estimated by image retrieval at time t , where ω_t^i represents the weight of each particle. Each particle contains similarities corresponding to different topological nodes. The motion model $p(x_t|x_{t-1}, u_t)$ is employed to estimate the current state of the robot that exhibits the probability of the transition from place x_{t-1} to place x_t .

(2) Update Phase

In update phase, the probability distribution of the robot in the environment is represented by a series of samples randomly selected based on posterior probability $\text{bel}(x_{0:t})$, where $x_{0:t}$ denotes the distribution of x from time 0 to time t . The samples are updated through the steps of sampling, importance weight calculation, and resampling, iteratively estimating the pose probability of the robot as follows:

$$\begin{aligned} \text{bel}(x_{0:t}) &= p(x_{0:t}|z_{1:t}, u_{1:t}) = \\ &\eta p(z_t|x_t) \int p(x_t|x_{t-1}, u_t) p(x_{0:t-1}|z_{1:t-1}, u_{1:t-1}) dx_{t-1} , \\ &= \eta p(z_t|x_t) \int p(x_t|x_{t-1}, u_t) \text{bel}(x_{0:t-1}) dx_{t-1} \end{aligned} \quad (4)$$

where η is the normalized coefficient. At each iteration, samples are taken from the posterior probability distribution of the robot according to proposal distribution $p(x_t|x_{t-1}, u_t)$. Then, the importance weight $\omega_t^i = \eta p(z_t|x_t)$, representing the probability of the robot at the particle position, is estimated for each particle based on the observation model, where $p(z_t|x_t)$ represents the observation model of the robot. In the image-based Monte Carlo localization, the observation model $p(z_t|x_t)$ is related to the similarity $S(I_t, I_j)$ between the current view image I_t and the keyframe image I_j in topological node j . The similarity $S(I_t, I_j)$ is generated by means of comparing the current view image of the robot with all the keyframe images in topological nodes, as shown in Figure 5. Then, the similarity $S(I_t, I_j)$ is fed into the process of calculating the importance weight. For the current view image I_t of the robot, the importance weight ω_t^i is defined as:

$$\omega_t^i = \frac{1}{N} \sum_{j=1}^N S(I_t, I_j) (D - \text{dist}(I_i, I_j)), \quad (5)$$

where D is the metric distance between the particle i and the farthest node, and $\text{dist}(I_i, I_j)$ is the distance between the particle i and the node j .

Figure 5. The definition of the similarity between the current view and topological nodes.

(3) Adaptive Genetic Algorithm

However, the particle set is prone to diverge and fall into kidnapping when the samples do not lie in the vicinity of ground-truth location, if the weighted particles generated by the above model are directly adopt to approximate the nonlinear posterior distribution. Then, a large number of particles are required to inject into the map and approximate the posterior distribution, resulting in computational burden. To this end, we introduce the adaptive evolution strategy of genetic algorithm to increase the diversity of samples and reduce the computational burden, which requires relatively few particles to approximate the posterior distribution rapidly and precisely.

The adaptive genetic algorithm simulates the particle convergence problem as a biological evolution process which performs the iteratively updates through operations of crossover and mutation to gradually eliminate samples with low fitness function values and increase samples with high fitness function values. Its process is as follows:

Crossover: selecting two parent particles $\{x_t^m, x_t^n\}$ for s_t randomly and generating two new particles as defined below:

$$\tilde{x}_t^m = \alpha x_t^m + (1 - \alpha)x_t^n + \eta, \quad (6)$$

$$\tilde{x}_t^n = \alpha x_t^n + (1 - \alpha)x_t^m + \eta, \quad (7)$$

where $\eta \sim N(0, \Sigma)$, $\alpha \sim U(0, 1)$, $\eta \sim N(0, \Sigma)$ is normal distribution, and $U(0, 1)$ is uniform distribution. $\tilde{x}_t^m$ is accepted if $p(z_t | \tilde{x}_t^m) > \max\{p(z_t | x_t^m), p(z_t | x_t^n)\}$. Otherwise, accept with the probability of $p(z_t | \tilde{x}_t^m) / \max\{p(z_t | x_t^m), p(z_t | x_t^n)\}$. The same operation for $\tilde{x}_t^n$.

Mutation: selecting a parent particles. Then, the mutation operation is performed to generate a new particle according to the following formula:

$$\tilde{x}_k^j = x_k^j + \eta, \quad \eta \sim N(0, \Sigma). \quad (8)$$

$\tilde{x}_t^j$ is accepted if $p(z_t | \tilde{x}_t^j) > p(z_t | x_t^k)$. Otherwise, accept with the probability of $p(z_t | \tilde{x}_t^j) / p(z_t | x_t^k)$.

In basic genetic algorithm, crossover probability p_c and mutation probability p_m generally take fixed values which may lead to problems in the state estimation of complex system, such as low accuracy and premature convergence of particles. Therefore, an adaptive adjustment strategy is

introduced to adjust the crossover probability and mutation probability by means of individual fitness value. The implementation is as follows:

$$p_c = \begin{cases} \frac{k_1(f_{\max} - f')}{f_{\max} - f_{\text{avg}}}, & f' \geq f_{\text{avg}} \\ k_3, & f' < f_{\text{avg}} \end{cases} \quad (9)$$

$$p_m = \begin{cases} \frac{k_2(f_{\max} - f)}{f_{\max} - f_{\text{avg}}}, & f \geq f_{\text{avg}} \\ k_4, & f < f_{\text{avg}} \end{cases} \quad (10)$$

where $f_{\max}$ is maximum fitness value, f_{avg} is average fitness value, f' represents the larger fitness value of the two crossover samples, and f represents the fitness value of the mutation sample. By setting k_1 , k_2 , k_3 , and k_4 , the adaptive adjustment of p_c and p_m can be achieved. .

The detailed process of the image-based Monte Carlo localization is shown in Algorithm 1.

Algorithm 1 Image-based Monte Carlo localization

Input: observation information z_t , control information u_t ;

Output: the state of robot s_t ;

- 1: **Initialization:** Generate the initial particles set $\{x_0^i\}_{i=1}^N$ based on prior distribution $p(x_0)$, $t = 1$;
 - 2: **for** $t = 1, 2, \dots$ **do**
 - 3: Motion model = $p(x_t | x_{t-1}, u_t)$;
 - 4: Observation model = $p(z_t | x_t)$;
 - 5: Calculate ω_t^i according to Equation (5);
 - 6: **Crossover:** set the weight ω_t^i as the fitness value;
 - 7: Calculate p_c with low fitness value according to Equation (9);
 - 8: Perform crossover according to Equation (6) and (7);
 - 9: **Mutation:** calculate p_m according to Equation (10);
 - 10: Perform Mutation according to Equation (8);
 - 11: Normalization: $\bar{\omega}_t^i = \omega_t^i / \sum_{i=1}^N \omega_t^i$
 - 12: **end for**
 - 13: **return** The state of robot at time t: $s_t = \{x_t^i, \bar{\omega}_t^i\}$;
-

4. Experiments

To verify the performance of ERF-IMCS, the experiment has been implemented on a real mobile robot and tested intensively in Beihang University, as shown in Figure 6. The maximum linear speed of the mobile robot is approximate 0.5 m/s. The robot is equipped with a monocular camera which can provide the RGB image at the resolution of 640×480 . The vision-based topological navigation system is processed on an Nvidia Jetson TX2 with 256 core-Pascal GPUs.

Figure 6. The mobile robot used to perform ERF-IMCS experiment.

The plan of the experimental environment is shown in Figure 7. The experimental environment is about 800 m² in size. During the testing stage, we firstly steer the robot through the indoor experimental environment and record the image frames and odometry information in which the keyframes is extracted to construct the topological nodes via CNN-based image matching. In the map building, the travel distance of the robot between each keyframe is set to be relatively small. In this way, the topological nodes and the links between them in indoor environments can be determined more robustly through visual feature matching without the need for geometry information of the environment. For the topological map, we build a database containing 826 images that are uniformly resized to 224 × 224 pixels. The path length traversed by the robot in topological mapping is about 120 m. In Figure 7, a resulting topological map is presented in which the blue dots represent the nodes with distinctive visual features in topological map and the links exhibit topological relationships between nodes. Additionally, several keyframe images from the experimental environment are displayed.

Figure 7. The plan of the experimental environment.

4.1. Image Retrieval Experiment

In this section, we perform experiments on the Oxford Building dataset [41] and Paris Building dataset [42] to validate the performance of our coarse-to-fine image retrieval method. The Oxford Building dataset contains 5063 images with 11 different buildings in Oxford, in which the bounding boxes surrounding the objects is also provided. The Paris building dataset contains 6412 images with 11 different buildings. With respect to the training details, the learning rate is set to 0.001, and weight decay and momentum are set to 0.0005 and 0.9 respectively. In coarse filtering stage, we adopt the conv5 layer of ZF to build the global descriptors of images for searching the top-50 images in the database closest to the query image. After that, the ranked Top-50 images were verified for spatial re-ranking based on local descriptors. In order to get better compact representations for image retrieval, we fine-tune the pre-trained network to return the regressed bounding box coordinates and the class scores for each image, which is conducive to obtain the local features of the interest region. In this way, the feature descriptors exhibit strong discrimination in the environment with a similar layout.

To quantitatively evaluate the retrieval performance, we compare the proposed methods with several image retrieval works in the literature in terms of the accuracy and time cost. We use the widely-used Mean Average Precision (MAP) to evaluate the retrieval accuracy of image retrieval. To make the comparison fair, we fine tune the network model of each method on the auxiliary dataset

and performed query expansion with top-5 images. Table 1 shows the results of image retrieval. As can be seen, in terms of MAP, our result is significantly better than J'egou et al. [43] and Babenko et al. [44] on the Oxford 5k dataset. This may be the reason that our method simultaneously considers global and local features of the image, which is conducive to extract discriminative descriptors from images with similar layouts and geometric structure. Comparing with the well-developed methods [29,45], our method also obtains comparable performance on the Oxford 5k dataset and Pair 6k dataset. However, in terms of time consumption, they are both relative expensive. The works [29,45] require a costly spatial verification in test time, comparing the query feature with each of the features in the dataset. In this way, it will be very time-consuming when the dimension of the query vector is high. By contrast, our method only takes 0.17 s, which is about seventeen times faster than Reference [45] and six times faster than Reference [29]. This is mainly because our method transforms the high dimensional CNN features into a more efficient binary code with almost negligible loss of accuracy.

Table 1. Performance comparison with existing methods.

	MAP		Times(s)	
	Oxford 5k	Pair 6k	Oxford 5k	Pair 6k
Jégou et al. Reference [43]	56.0	—	—	—
Babenko et al. Reference [44]	65.7	—	—	—
Tolias et al. Reference [45]	77.3	86.5	2.9	3.52
Salvador et al. Reference [29]	78.6	84.2	1.08	1.36
Ours	78.1	83.8	0.17	0.21

Figure 8 shows the top retrieved results of several query examples in the Oxford Building dataset and Paris Building dataset. The first image depicted with a blue contour at each row is the query image. The others are the top retrieved images, while positive and negative images are depicted with green and red contour, respectively. The black number below is the ranking of the image and the red number in brackets is the ranking of the image before spatial reranking. It can be seen that our spatial reranking strategy based on local features can effectively improve the performance of image retrieval, especially for images that contain salient local object.

Figure 8. Examples of the proposed image retrieval method. The first column is the query image with blue contour, and the others with green contour are retrieved images.

4.2. Image-Based Monte Carlo Localization

To verify the robustness of our topological navigation method, we performed the experiment on the topological map shown in Figure 7. In image-based Monte Carlo localization, a series of similarities between the currently view image and the keyframe images in topological nodes generated by image retrieval are adopted to compute the importance weights of the possible location of the robot. Generally speaking, the higher the ranking of the retrieved image, and the more similar to the query image, the greater the weight of the samples assigned to that location. For an indoor mobile robot, the similarity between two images decreases as the captured distance increases. In an actual navigation experiment, in order to judge whether the robot reaches a certain target position, we steered the robot to collect a series of image in different indoor environments and randomly select frames to analyze the

relationship between the similarity and spatial distance of two images. In this way, the approximate distribution of the captured distance under a certain degree of similarity can be obtained through experimental statistics. The result is shown in Figure 9. As can be seen, the captured distance between two images is within 1 m when the similarity is greater than 0.9 and the captured distance between two images is within 0.5 m when the similarity is greater than 0.85. Therefore, we set the threshold of similarity to 0.85 in this experiment, which means that we judged the robot to reach the target position when the similarity between the currently view image and the keyframes in target nodes was greater than 0.85.

Figure 9. The relationship between image similarity and collection distance.

In this section, to quantitatively evaluate the performance of localization, comparative experiments among the proposed image-based Monte Carlo localization method and a well-known metric-based localization method [46] were performed, as shown in Figure 10. The metric-based navigation method which adopts a particle set of fixed size to simulate the posterior probability is performed by providing the specific geometric coordinates. Figure 10a depicts the number of samples versus execution step of the two localization method. Figure 10b shows the CDF of the localization error versus execution step of two localization methods. It can be seen that the performance of our image-based localization method obviously outperforms metric-based method in terms of localization accuracy. The localization error of the proposed localization method converges rapidly after a short period of fluctuation. Such fluctuation of localization error generally occur in the initial stage of global localization, which is mainly due to the incorrect matching of image-to-node in topological localization, namely perceptual aliasing. But the coarse-to-fine mechanism in our image-based Monte Carlo localization provides the robot with the capability to restore from the perceptual aliasing and converge to the real location rapidly. Additionally, as depicted in Figure 10, the proposed visual localization method is overall robust, and the average localization error is less than 50 cm when the localization process is stable after a few iterations, which is enough to meet the requirements for global localization. In a nutshell, the experiment results indicate that the proposed localization method is indeed a promising avenue to tackle the visual localization problem in an indoor environment with repetitive structures or similar layouts.

Figure 10. (a) A typical evolution of number of samples in global localization. (b) CDF of localization accuracy.

4.3. Vision-Based Topological Navigation

In this section, we adopt the topological map in Figure 11 to perform the final navigation experiment for the mobile robot. For robot navigation, it is challenging because the indoor experimental environment contains many repetitive structures and similar layouts. The topological navigation is performed by means of traversing a sequence of way-point keyframes in topological nodes, which is constructed by graph-based path planning. During the path planning, the robot continually matches the current image with the keyframes in the next target node. After reaching each target node, the robot re-estimates its pose and makes adjustments until reaching the final goal location. In order to provide the robot with the capability to cope with the incorrect estimation of relative orientation robustly, we limit the maximum directional change of the mobile robot to 20° and the maximum turning angle to 15° , respectively. In this way, the target nodes is always in the view of camera even if the localization estimation is incorrect, which is conducive to restore the pose for the robot. In the topological navigation experiment, the mobile robot is initialized and then navigates from the starting position P0 to the doorways of different rooms which corresponds to nodes labeled with P1 to P5, as shown in Figure 11. In most cases, the robot pose successfully converges to the target node and reaches the goal location. To quantitatively evaluate the performance of navigation, we compare the proposed topological navigation method with a well-known metric-based navigation method [46]. The result of the two methods in navigation efficiency is shown in Figure 12. Each trial is performed 5 times. Compared with traditional metric-based navigation, our proposed method is slightly better in time consumption. For the path of P0 to P4, the topological navigation method takes more time. This is mainly because there is a warehouse with uneven illumination and lack of obvious visual features in the path of P0 to P4, which makes it non-trivial to resample the particles via image retrieval. In this case, the robot will rely more on the odometry to update the particles. Additionally, during the navigation experiment, ERF-IMCS only requires us to input the label attribute of the navigation goal, such as doorway of conference room, since the labels of local location are stored in each corresponding topological node, in contrast to other systems that adopt sequences of metric poses. In this way, the mobile robot can perform human-like navigation behavior. Overall, ERF-IMCS is completely sufficient to meet the requirements for indoor robot navigation.

Figure 11. The topological navigation path of the robot.

Figure 12. Time cost of each navigation method. Each trial was performed 5 times.

5. Conclusions

In this paper, ERF-IMCS was proposed, which integrates visual topological localization aided by CNN-based image retrieval with a probabilistic approach of Monte Carlo localization. The image retrieval strategy is responsible for calculating the similarity between current view image and keyframes in topological nodes based on a pre-trained faster RCNN model, which is injected into the next image-based Monte Carlo localization stage to compute the weights of the possible location of the robot. Moreover, during the topological mapping stage, we stored visual features and scene properties in topological nodes, which is conducive to judge whether it has reached the goal location with more flexibility for the mobile robot. Systematic experiments were conducted on a real mobile robot with a monocular camera. The results indicate that ERF-IMCS exhibits great performance in efficiency and robustness compared to existing navigation methods.

Author Contributions: S.X. and W.C. designed the study; S.X. and H.Z. conducted the experiments; S.X. implemented the system and wrote the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (Grant No. 2019YFB1310802).

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. 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](#)]
3. Chen, X.; Sun, H.; Zhang, H. A New Method of Simultaneous Localization and Mapping for Mobile Robots Using Acoustic Landmarks. *Appl. Sci.* **2019**, *9*, 1352. [[CrossRef](#)]
4. Bista, S.R.; Giordano, P.R.; Chaumette, F. Combining line segments and points for appearance-based indoor navigation by image based visual servoing. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 2960–2967.
5. Ferro, M.; Paolillo, A.; Cherubini, A.; Vendittelli, M. Vision-Based Navigation of Omnidirectional Mobile Robots. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2691–2698. [[CrossRef](#)]
6. Luo, R.C.; Shih, W. Topological Map Generation for Intrinsic Visual Navigation of an Intelligent Service Robot. In Proceedings of the 2019 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 11–13 January 2019; pp. 1–6.
7. Valiente, D.; Gil, A.; Paya, L.; Sebastian, J.M.; Reinoso, O. Robust Visual Localization with Dynamic Uncertainty Management in Omnidirectional SLAM. *Appl. Sci.* **2017**, *7*, 1294. [[CrossRef](#)]
8. Wang, J.; Zha, H.; Cipolla, R. Coarse-to-fine vision-based localization by indexing scale-Invariant features. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* **2006**, *36*, 413–422. [[CrossRef](#)]
9. Maldonadoramirez, A.; Torresmendez, L. A Collaborative Human-Robot Framework for Visual Topological Mapping of Coral Reefs. *Appl. Sci.* **2019**, *9*, 261. [[CrossRef](#)]
10. Maohai, L.; Han, W.; Lining, S.; Zesu, C. Robust omnidirectional mobile robot topological navigation system using omnidirectional vision. *Eng. Appl. Artif. Intell.* **2013**, *26*, 1942–1952. [[CrossRef](#)]
11. Calonder, M.; Lepetit, V.; Strecha, C.; Fua, P. BRIEF: Binary Robust Independent Elementary Features. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6314, pp. 778–792.
12. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [[CrossRef](#)]
13. Ma, J.; Zhao, J. Robust Topological Navigation via Convolutional Neural Network Feature and Sharpness Measure. *IEEE Access* **2017**, *5*, 20707–20715. [[CrossRef](#)]
14. 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.
15. Kalantidis, Y.; Tolias, G.; Avrithis, Y.; Phinikettos, M.; Spyrou, E.; Mylonas, P.; Kollias, S. VIRaL: Visual Image Retrieval and Localization. *Multimed. Tools Appl.* **2011**, *51*, 555–592. [[CrossRef](#)]
16. Mansourian, L.; Abdullah, M.T.; Abdullah, L.N.; Azman, A.; Mustaffa, M.R. An effective fusion model for image retrieval. *Multimed. Tools Appl.* **2018**, *77*, 16131–16154. [[CrossRef](#)]
17. Arandjelović, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. NetVLAD: CNN Architecture for Weakly Supervised Place Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1437–1451. [[CrossRef](#)] [[PubMed](#)]
18. Preethy Byju, A.; Demir, B.; Bruzzone, L. A Progressive Content-Based Image Retrieval in JPEG 2000 Compressed Remote Sensing Archives. *IEEE Trans. Geosci. Remote. Sens.* **2020**, *58*, 5739–5751. [[CrossRef](#)]
19. 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](#)]
20. 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](#)]
21. Zhou, X.S.; Huang, T.S. Relevance Feedback in Image Retrieval: A Comprehensive Review. *Multimed. Syst.* **2003**, *8*, 536–544. [[CrossRef](#)]
22. Jégou, H.; Douze, M.; Schmid, C.; Pérez, P. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311.
23. Perronnin, F.; Dance, C. Fisher Kernels on Visual Vocabularies for Image Categorization. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.

24. Jose, A.; Lopez, R.D.; Heisterklaus, I.; Wien, M. Pyramid Pooling of Convolutional Feature Maps for Image Retrieval. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 480–484.
25. Liu, P.; Guo, J.; Wu, C.; Cai, D. Fusion of Deep Learning and Compressed Domain Features for Content-Based Image Retrieval. *IEEE Trans. Image Process.* **2017**, *26*, 5706–5717. [[CrossRef](#)]
26. Babenko, A.; Slesarev, A.; Chigorin, A.; Lempitsky, V. Neural codes for image retrieval. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 584–599.
27. Sharif Razavian, A.; Azizpour, H.; Sullivan, J.; Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Columbus, OH, USA, 23–28 June 2014.
28. Gordo, A.; Almazán, J.; Revaud, J.; Larlus, D. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 241–257.
29. Salvador, A.; Giró-i Nieto, X.; Marqués, F.; Satoh, S. Faster r-cnn features for instance search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 9–16.
30. Shalev, O.; Degani, A. Canopy-Based Monte Carlo Localization in Orchards Using Top-View Imagery. *IEEE Robot. Autom. Lett.* **2020**, *5*, 2403–2410. [[CrossRef](#)]
31. Marinho, L.B.; Almeida, J.S.; Souza, J.W.M.; Albuquerque, V.H.C.; Reboucas Filho, P.P. A novel mobile robot localization approach based on topological maps using classification with reject option in omnidirectional images. *Expert Syst. Appl.* **2017**, *72*, 1–17. [[CrossRef](#)]
32. Cheng, H.; Chen, H.; Liu, Y. Topological Indoor Localization and Navigation for Autonomous Mobile Robot. *IEEE Trans. Autom. Eng.* **2015**, *12*, 729–738. [[CrossRef](#)]
33. Blochliger, F.; Fehr, M.; Dymczyk, M.; Schneider, T.; Siegwart, R. Topomap: Topological mapping and navigation based on visual slam maps. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; pp. 1–9.
34. Liu, M.; Siegwart, R. Topological Mapping and Scene Recognition With Lightweight Color Descriptors for an Omnidirectional Camera. *IEEE Trans. Robot.* **2017**, *30*, 310–324. [[CrossRef](#)]
35. Goedemé, T.; Nuttin, M.; Tuytelaars, T.; Gool, L.V. Omnidirectional Vision Based Topological Navigation. *Int. J. Comput. Vis.* **2007**, *74*, 219–236. [[CrossRef](#)]
36. Hao, J.; Dong, J.; Wang, W.; Tan, T. What Is the Best Practice for CNNs Applied to Visual Instance Retrieval? *arXiv* **2016**, arXiv:1611.01640.
37. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 818–833.
38. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)] [[PubMed](#)]
39. Yue-Hei Ng, J.; Yang, F.; Davis, L.S. Exploiting local features from deep networks for image retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 53–61.
40. Gong, Y.; Lazebnik, S.; Gordo, A.; Perronnin, F. Iterative Quantization: A Procrustean Approach to Learning Binary Codes for Large-Scale Image Retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2916–2929. [[CrossRef](#)]
41. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Object retrieval with large vocabularies and fast spatial matching. In Proceedings of the IEEE Computer Society Conference on Computer Vision & Pattern Recognition, Minneapolis, MI, USA, 18–23 June 2007.
42. Philbin, J.; Chum, O.; Isard, M.; Sivic, J.; Zisserman, A. Lost in quantization: Improving particular object retrieval in large scale image databases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
43. Jégou, H.; Zisserman, A. Triangulation embedding and democratic aggregation for image search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 3310–3317.
44. Babenko, A.; Lempitsky, V. Aggregating Local Deep Features for Image Retrieval. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1269–1277.

45. Tolias, G.; Sivic, R.; Jégou, H. Particular Object Retrieval With Integral Max-Pooling of CNN Activations. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016; pp. 1–12.
46. Thrun, S.; Fox, D.; Burgard, W.; Dellaert, F. Robust Monte Carlo localization for mobile robots. *Artif. Intell.* **2001**, *128*, 99–141. [[CrossRef](#)]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).