Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Visual Topological Mapping and Navigation for Mobile Robot in Large-Scale Environment*

1st Song Xu, 3rd Huaidong Zhou
School of Mechanical Engineering and Automation
Beihang University
Beijing, China
1st *keithxs@buaa.edu.cn*
3rd *hdzhou@buaa.edu.cn*

2nd Wusheng Chou
School of Mechanical Engineering and Automation
State Key Laboratory of Virtual Reality Technology and Systems
Beihang University
Beijing, China
wschou@buaa.edu.cn

***Abstract*— Autonomous navigation is a basic prerequisite for mobile robot to realize environmental exploration. Current navigation methods are mainly based on metric maps, which require precise geometric coordinates and lack the capability to efficiently store semantic information of the environment. In this paper, we present a visual topological mapping and navigation method for mobile robot in large-scale environment, which is similar to the human navigation system. Topological map represents the environment as a topology diagram with nodes and edges in which the topological nodes record local semantic information of the environment, such as visual features, robot pose and scene properties. In the topological navigation stage, an image-based Monte Carlo localization is proposed to estimate the semantic pose of robot which can help robot judge whether it has reached the target location more flexibility. Experiments are conducted in real world environments and results indicate that the proposed system exhibits great performance in robustness of navigation.**

***Index Terms*— Autonomous Navigation; Topological Map; Semantic Pose**

I. Introduction

Autonomous navigation is a basic prerequisite for mobile robot to perform various tasks such as exploration and homing. A robot moving in an unknown environment first needs to map the environment for navigation. Its basic methodology is to constantly perceive the state of robot and the environment information such as lines, corner and obstacles through sensors during the movement process, so as to guide the robot to reach the target position quickly and safely [1].

In general, robot navigation consists of three processes: mapping, localization and path planning. In the mapping and localization stage, SLAM (Simultaneous Localization and Mapping) technology is often adopted to construct metric maps of the environment based on the information from laser range finder or RGBD camera, which is convenient in small areas, but inefficient in large scale environments [2]. More specifically, the metric map depicts the environment with precise global coordinates by means of grid-based representations which is pure geometric representations. It can provide rich geometric details of the environment, but lack of semantic information of local region such as visual features, robot pose and scene properties. As a result, this may cause some trouble for robot navigation. For example, when human sends an instruction to the robot to go into the conference room and gives the target coordinates which are occupied by obstacle, the robot will lost the target and continuously search around the target coordinates although it has already entered into the conference room.

In contrast, the topological map can store the related information of the local scene or the pose of the robot into topological nodes which is convenient for robot to navigate. The topological map represents the environment as a topological model which focuses on the connectivity and reachability between each local region, without constraining the specific motion trajectory of the robot in the navigation process. The core of the topological navigation model is the construction of the topological map, in which the main scene of the environment is modeled as topological nodes. The topological nodes are usually linked by the topological edges. The topological edges are abstractions of the scene paths, representing the passage relationship between two connected nodes. The advantage of the topological navigation is that the environment awareness and navigation strategy can be decoupled into two independent problems. It is not necessary to store each pixel in the grid map, which takes up less memory and has the capability to navigate in large-scale environment.

In this paper, we proposed a visual topological mapping and navigation method for mobile robot in large-scale environment, in which the topological map is responsible for constructing semantic information of local region in the environment. Global information is not important when the robot is working in a local region of the environment. Just as a person perceives the environment, he is not likely to remember every information about the environment, but

*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)

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

only store the typical local information that is conductive to navigate even in complex environment. In the topological navigation stage, an image-based Monte Carlo localization which integrate image retrieval mechanism into particle filter is proposed to estimate the semantic pose of robot. According to our proposed method, the robot can take semantic pose in consideration to estimate its location which makes the process of localization more flexible and more robust. Our mapping and navigation system has been tested and demonstrated in a real indoor environment. The experimental results indicate that the proposed method exhibits better performance in robustness of navigation.

II. Related Work

A. Topological Mapping

The topological map represents the environment as a graph model which only focus on the connectivity and reachability between each local region in the form of nodes and edges, without constraining the specific motion trajectory of the robot in the navigation process. The topological map can store the related information of the local scene into topological nodes.

Extensive works have been done on topological mapping. [3] proposed a topological map construction method for indoor environment in which the normalized graph-cuts algorithm is employed to segment the topological graph of the environment to obtain a sub-graph corresponding to the convex region. [4] constructed the topological map of environment in the light of neighborhood associations and particle filtering algorithm. In [5], the authors utilized SIFT features to match a series of images. They calculated the transition relationship between local regions based on the SIFT features that can be successfully matched between consecutive frames.

There are also other works focuses on working in environment with similar geometric structure and layouts [6] employed the wide-baseline feature matching to realize the topological map representation which focuses on the creation of a hierarchical topological model. Ulrich et al. [7] proposed an appearance-based topological mapping method in which the color histograms are exploited to realize place recognition and localization.

B. Topological Localization

The core of the topological localization is to how to determine the node corresponding to the actual location of the robot without specific coordinate of this location. In some vision-based methods, the topological localization is cast as a matching process of image-to-node in which the image retrieval technique is employed to estimate the possible node locations of the query image. Towards this end, a similarity measurement between the query image and nodes requires to be defined. For instance, Leandro et al. [8] presented a topological method via classification with a reject option based on moments, texture and keypoint descriptor. Cheng et al. [9] proposed a novel topological map-building-based localization method via multiple observation fusion in which the corridors and the intersections are considered as vertices and branches to construct the topological map for robot localization.

Meanwhile, other researchers focus on the multiple-view geometry validation step to estimate the accurate pose of robot after the process of image retrieval [10]. Suman et al. [11] proposed an image-based navigation via the common line segments and feature points between the current captured image and the surrounding keyframe images in which the environment is represented as a set of key images.

All the above works realize the topological global localization via maximum likelihood scheme which matches the current view of robot with reference views in environment. This strategies may be affected by perceptual aliasing in environment with similar geometric structure and layouts, resulting in localization errors. To this end, other researches utilize Bayesian filtering to generate a maximum a posteriori that integrates the information between past estimates and current observation. For instance, Liu et al. [12] presented a lightweight scene recognition approach utilizing adaptive descriptors constructed by geometric features based on omnidirectional camera. [13] proposed a system to construct a precise topological map in large-scale and complex environment in which the maximum a posteriori scheme is employed to calculate the probability of the position of the query image. In this works, each node in the environment is assigned the same probability in the initial localization phase. After that, the probability of the robot location is calculated by iterative update process based on the observation model of robot.

III. Proposed Mapping and Navigation Method

The overview of the proposed navigation system for mobile robot is shown in Figure1. The proposed navigation process consists of two stages: global navigation based on topological map and local path planner based on grid map. The topological map is utilized to estimate the semantic pose of the robot. In the global navigation process, we employ an image-based Monte Carlo localization to obtain the semantic pose of robot in which an image retrieval network is integrated into particle filter to realize the matching process between the current views of robot and reference views in the environment. The semantic pose of robot can help robot judge whether it has reached the target location more flexibility. The similarity values of image matching are considered as the weights of particles distribution in image-based Monte Carlo localization. In the topological global path planning phase, Dijkstra algorithm is utilized to search the shortest topological nodes path from current node to the target node. Also, the local path plan based on grid map will be activated

rapidly if the global navigation based the topological map cannot assist the robot to reach the target location.

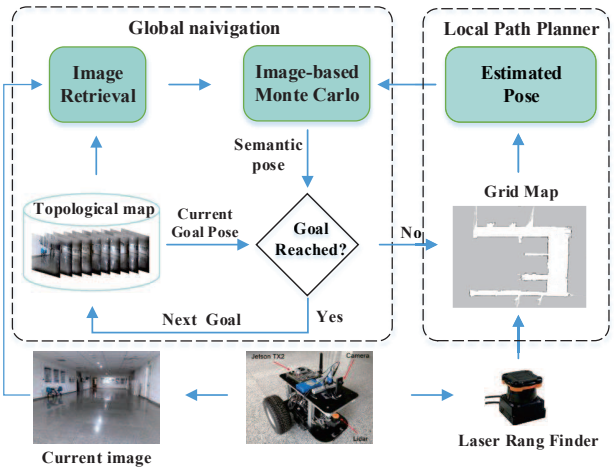

Fig. 1. Overview of the proposed mapping and navigation system.

A. Topological Mapping

In the proposed mapping stage, a metric map and a topological map are constructed simultaneously. The topological map is employed to record available path. In the topological mapping process, the topological node is constructed for every constant distance of 2 meters based on odometry data of mobile robot. In each node, we stores images of different wide-angles in the moveable direction of the robot. The images in topological node can provide semantic information for robot such as visual features and properties of local scene. Additionally, the coordinates based on the metric map is also stored in the corresponding topological node which is responsible for local path planning in navigation stage if necessary as shown in Figure2. Moreover, an edge is added to the topological map between the current node and the previous node. The edge represents that it is passable between adjacent nodes. Since the robot is completely guided by human during the mapping process, the edges in the topological map are regarded as the reference for path planning in the topological navigation stage.

After the above steps, a topological graph is built. However, since the above topological nodes are constructed with a constant distance, the topological map may contain too much redundant information. The environment characteristics of some adjacent nodes may exhibit high resemblance especially in the corridor area. Therefore, we merge two adjacent nodes when frames in current node exhibit too much similarities with frames in previous node which can save the memory of map. But, we stipulate that the distance between adjacent nodes cannot exceed a threshold in case the robot navigation fails. In this paper, we empirically set this threshold as six meters. When the current view image of the robot is similar to the reference images in topological node, we judge that the robot reaches this topological node.

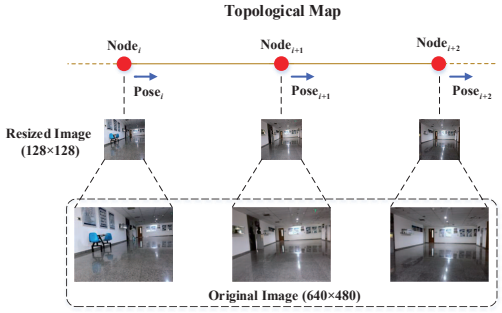

Fig. 2. Topological map. Each topological nodes store 128×128 images, coordinate corresponding to the grid map and the robot pose when constructing the node.

B. Image Retrieval

In recent years, it has been proved that fine-tuning based pre-trained model can obtain excellent performance. To get better feature representations in image retrieval, an effectively way is to explicitly learn weights suited for specific retrieval task based on pre-trained models. Inspired by [14], we employ the pre-trained model of Faster R-CNN to extract global and local features of images for image retrieval. Researches indicate that features extracted from different layers exhibit different performance. According to [15], performing image retrieval using features extracted from top semantic layers is not a wise choice because of the loss of spatial relationship of the object. Conversely, the activations from the convolutional layers show strong discriminability and excellent generalization which can construct competitive compact image representations.

In this paper, we adopt the coarse-to-fine mechanism to realize the image retrieval process which contains two stages: coarse image retrieval based on global feature representations and fine image retrieval based on local feature representations, as shown in Figure3.

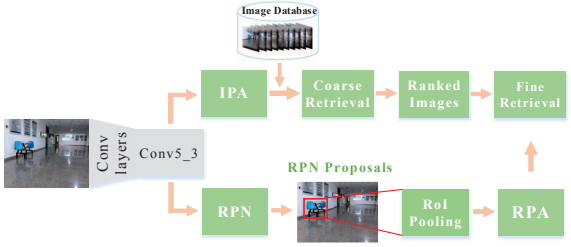

Fig. 3. Architecture of the proposed image retrieval scheme.

In the coarse image retrieval stage, we employ the $\mathrm{conv5\_3}$ layer of VGG16 to extract the global features of images. After feature extraction layer, global image descriptors are built by Image-wise Pooling of Activations (IPA) based on sum pooling. In image retrieval process, we search the images in database with the highest similarity to the current view image of robot based on similarity measurement. Here, the Euclidean distance is adopted to measure the similarity of descriptors between the query image and the database images. After that, we can obtain top-N candidate images corresponding to different topological nodes. However, perceptual aliasing may occur in environment with similar geometric structure and layouts. To this end, spatial re-ranking is necessary. Then, the top-N candidate images is input into the fine retrieval stage.

In the fine image retrieval stage, spatial re-ranking is performed in which Region-wise Pooling of Activations (RPA) with max pooling is employed to build the local region descriptors of images by aggregating the convolutional activations for each object proposals from the RoI pooling layer of Faster R-CNN. After RoI pooling, the feature descriptors are l2-normalized and then whitening is applied to increase its robustness against noise. Local feature representations show strong discriminability and excellent generalization which is conducive to perform image retrieval in complex environment. We do image comparison and generate a similarity between the local descriptors of the bounding box in current view image of robot and the region-wise descriptors of all RPN proposals in the Top-N candidate images based on Euclidean distance. For convenience, the bounding box is resize to the same as the feature maps in the last convolutional layer. The similarity value generated by spatial re-ranking is then adopt to calculate the weights of particle distributions in the next image-based Monte Carlo localization stage.

C. Image-based Monte Carlo Localization

Traditional methods realize the topological global localization via maximum likelihood scheme which matches the current view of robot with reference views in environment. This strategies may be affected by perceptual aliasing in environment with similar geometric structure and layouts, resulting in localization errors. To this end, in this paper, we present an image-based Monte Carlo localization that utilize Bayesian filtering to generate a maximum a posteriori.

According to Bayesian filtering theory, the localization problem is described as estimating the state of the robot x_t at time t in dynamic system based on the initial state distribution of the robot $p(x_0)$, the control sequence of the robot u_t and the observation sequence of the environment z_t which is achieved by estimating the posterior probability distribution $p(x_t|\, z_t, u_t)$. In Monte Carlo localization, the posterior probability of the estimated state is represented by a set of weighted particles. The localization process can be performed as following steps:

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 generated around possible topological nodes in the topological map obtained by image retrieval at time t. The motion model $p\left(x_t \,|x_{t-1}\,, u_{t-1}\right)$ is adopt to predict the current state of the robot in the form of a predictive probability density function (PDF) $p\left(x_t \,|z_t\,\right)$.

2) Update phase: In the update phase, samples is updated through the process of sampling, importance weight calculation and resampling. The state of the robot at time t is sampled based on the importance function and the importance weight of each particle is calculated. The importance weight denotes the probability that robot is in state x_t which is closely related to the observation model $p\left(z_t \mid x_t^{(i)}\right)$. In our image-based Monte Carlo localization, the observation model is determined by the perception data of camera which is a vector composed of feature descriptors extracted from the current view image of robot. When the robot acquires visual information, the importance weights $\omega_t^i = \eta p(\, z_t|\, x_t)$ of all samples are updated according to the camera perception model which is calculated by the similarity of image retrieval. For the current view image of the robot I_t, the importance weight is updated as follows:

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

where $d(I_t, I_j)$ represents the similarity between the current observation image of the robot and the reference image in node j; D denotes the metric distance between the farthest node and the location of the particle i; $dist(I_i, I_j)$ represents the distance between the particle i and the node j.

Finally, the importance weights are normalized and the particle set is resampled according to the importance weights. After resampling, the particle set gradually convergences to the real pose of the robot. Thus, we can estimate the pose distribution of the robot by calculating the expectation of the particle distribution.

IV. Experiments

The proposed system was tested in the new main building of Beihang University and implemented on a real mobile robot, as shown in Figure4. 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 640×480 pixels with 30Hz frame. 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 which can support the neural network. The experimental environment is about $800\mathrm{m}^2$. We steer the robot to construct a hybrid metric-topological map and build the topological node at a distance of every 2 meters based on the data of odometry.

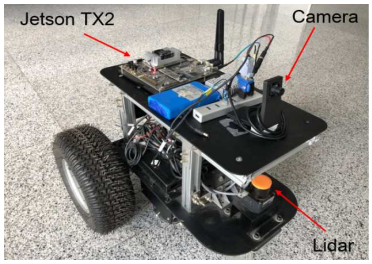

Fig. 4. The mobile robot used to validate the proposed approach.

The plan of the building is shown in Figure5. The red thick line represents the path where the robot constructs the map of environment. Also, we show some images from the experimental environment. As can be seen, the images in the scene exhibit a high degree of similarity in layout and structure which brings great challenges for image-based localization. In this experiment, we empirically judge that the robot reaches the target topological node when the similarity value between two images is greater than the threshold of 0.8.

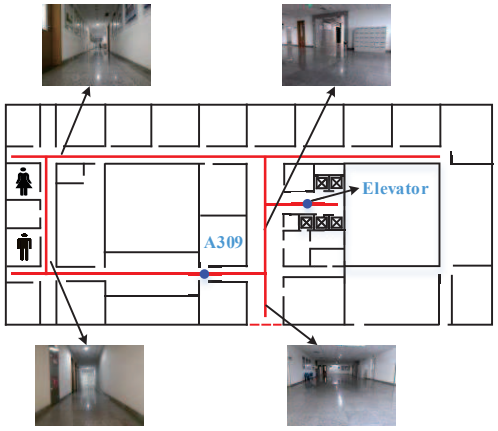

Fig. 5. The plan of the experimental environment.

A. Image Retrieval Experiment

The image retrieval algorithm is verified by the Almere database and our indoor database constructed in this experiment. Due to the high sampling frequency of the Almere database, the robot displacement between two adjacent frames is small. Therefore, we extract two frames of images as matching pairs every 10 frames in which one frame is put into the reference set and other frame is put into the test set. During the test, the coarse-to-fine image retrieval algorithm is used to search the images in reference set that is most similar to the image in the test set. When the matching result is the same as the original match pair, the match result is considered correct. Our indoor database is collected in the environment as shown in Figure5. Since the real coordinates of the collected nodes are known, the matching distance between the matching images can be directly judged. In this experiment, the match result is considered to be correct when the distance of the two image is less than 1m.

TABLE I
ACCURACY OF IMAGE RETRIEVAL.

	Almere(%)	**Our database**(%)
VW	86.3	77.6
FT	70.9	60.5
Proposed Method	90.2	85.7

To quantitatively evaluate the performance, we compare the proposed method with the Fourier Transform (FT) method used in [16] and the Visual Word (VW) method [17]. The image retrieval results of the three methods are shown in Table1. The experiment results show that the proposed method is better than the VW method and the FT method in accuracy of image retrieval. Especially in our indoor database that exhibits a high degree of similarity in geometric structure and layouts, our proposed image retrieval method perform significantly better.

B. Image-based Monte Carlo Localization Experiment

To quantitatively evaluate the performance of proposed image-based Monte Carlo localization, we present the localization error versus execution step of the proposed method in Figure6. As can be seen, the proposed image-based Monte Carlo method provides the robot with the capability to estimate the location of robot efficiently. Additionally, the localization error of the proposed image-based Monte Carlo methods converges quickly after a period of fluctuating localization. The error fluctuation in initial localization stage is mainly because the images in experimental environment present a high degree of perceptual aliasing which causes an incorrect matching of image-to-node in image retrieval. Specially, the localization error gradually converges to less than 50cm after a few iterations which is enough to meet the requirements for global localization.

C. Topological Navigation Experiment

In this section, we conduct an experiment to verify the navigation capability of the proposed system for mobile robot. In the navigation process, the robot start from the door of room A309 and move towards the elevator region. The robot collects the images of environmental in real time for image-based Monte Carlo localization and keep moving until the robot reaches the next topological node. As shown

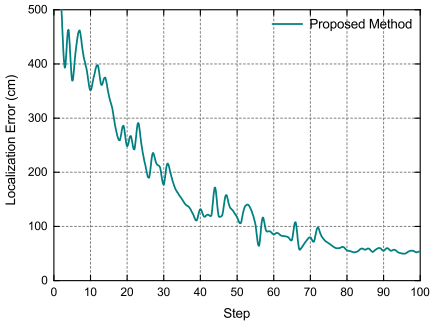

Fig. 6. The localization error of the proposed method.

in Figure7, the blue line represents the ground-truth path, and the red thick line represents the navigation path of the mobile robot recorded by the odometry data. When the current view image of the robot is similar to the images stored in the topological node sufficiently, the sematic pose of the robot transfers to the target node and the robot judges that it has reached the target node. As can be seen, the real navigation path of the mobile robot recorded by the odometry data and the ground-truth path are basically close although the robot has some flaws at the turn of the scene. This is because when the robot transitions from the current scene node to the next scene node at turn, the image features change too fast and the robot temporarily loses its position. Overall, the navigation process of the robot is robust and sufficient to meet the needs of navigation for mobile robot.

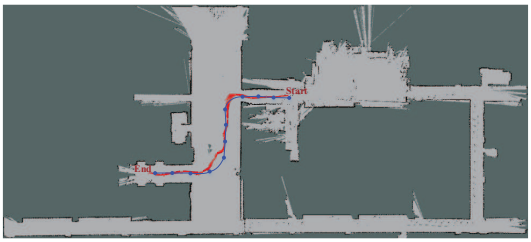

Fig. 7. The topological navigation path of the robot

V. Conclusion

Traditional navigation method depicts the environment with precise global coordinates by means of grid-based representations which is lack of semantic information of local region such as visual features, robot pose and scene properties. In this paper, we proposed a visual topological mapping and navigation method for mobile robot in large-scale environment. The topological map store scene images, coordinate corresponding to the grid map and the robot pose in the topological node which can provide information with more semantic meaning. In the topological navigation stage, an image-based Monte Carlo localization is proposed to estimate the semantic pose of robot which can help robot judge whether it has reached the target position with more flexibility. Experiment results indicate that our proposed system exhibits great performance in robustness of navigation.

References

[1] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*. MIT press, 2005.

[2] R. C. Luo and W. Shih, "Topological map generation for intrinsic visual navigation of an intelligent service robot," in *2019 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2019, pp. 1–6.

[3] Z. Zivkovic, O. Booij, and B. Kröse, "From images to rooms," *Robotics and Autonomous Systems*, vol. 55, no. 5, pp. 411–418, 2007.

[4] F. Werner, F. Maire, J. Sitte, H. Choset, S. Tully, and G. Kantor, "Topological slam using neighbourhood information of places," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2009, pp. 4937–4942.

[5] J. Košecká, F. Li, and X. Yang, "Global localization and relative positioning based on scale-invariant keypoints," *Robotics and Autonomous Systems*, vol. 52, no. 1, pp. 27–38, 2005.

[6] T. Goedemé, T. Tuytelaars, and L. Van Gool, "Visual topological map building in self-similar environments," in *Informatics in Control Automation and Robotics*. Springer, 2008, pp. 195–205.

[7] I. Ulrich and I. Nourbakhsh, "Appearance-based place recognition for topological localization," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, vol. 2. Ieee, 2000, pp. 1023–1029.

[8] L. B. Marinho, J. S. Almeida, J. W. M. Souza, V. H. C. Albuquerque, and P. P. Rebouças Filho, "A novel mobile robot localization approach based on topological maps using classification with reject option in omnidirectional images," *Expert Systems with Applications*, vol. 72, pp. 1–17, 2017.

[9] H. Cheng, H. Chen, and Y. Liu, "Topological indoor localization and navigation for autonomous mobile robot," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 2, pp. 729–738, 2014.

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

[11] S. R. Bista, P. R. Giordano, and F. Chaumette, "Combining line segments and points for appearance-based indoor navigation by image based visual servoing," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 2960–2967.

[12] M. Liu and R. Siegwart, "Topological mapping and scene recognition with lightweight color descriptors for an omnidirectional camera," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 310–324, 2013.

[13] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. Van Gool, "Omnidirectional vision based topological navigation," *International Journal of Computer Vision*, vol. 74, no. 3, pp. 219–236, 2007.

[14] A. Salvador, X. Giró-i Nieto, F. Marqués, and S. Satoh, "Faster r-cnn features for instance search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 9–16.

[15] J. Hao, J. Dong, W. Wang, and T. Tan, "What is the best practice for cnns applied to visual instance retrieval?" *arXiv preprint arXiv:1611.01640*, 2016.

[16] L. Payá, L. Fernández, A. Gil, and O. Reinoso, "Map building and monte carlo localization using global appearance of omnidirectional images," *Sensors*, vol. 10, no. 12, pp. 11 468–11 497, 2010.

[17] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.