IOP Conference Series: Materials Science and Engineering

PAPER • OPEN ACCESS

# Autonomous Mobile Robot Visual SLAM Based on Improved CNN Method

To cite this article: Xuanbo Wang 2018 *IOP Conf. Ser.: Mater. Sci. Eng.* **466** 012114

# Autonomous Mobile Robot Visual SLAM Based on Improved CNN Method

**Xuanbo Wang**

Xi`an Gaoxin No.1 High School, Xi`an, China
kaqi136@163.com

**Abstract.** Although the research on autonomous mobile robot SLAM has received extensive research, the current mobile robot still exists in practical applications: it may move under the condition of disordered and irregular obstacle placement; the shape of the obstacle and the position of the obstacle change; and indoor and outdoor scene switching occurs at different times and other issues. Autonomous mobile robots need to continuously optimize SLAM during motion and obtain real-time information from multiple sensors for real-time identification and rapid response to the surrounding environment. We have improved the CNN-based VSLAM system by replacing the original single convolutional layer with parallelism and reducing the number of model parameters. This approach can reduce the problem of system gradient disappearance and continue to train deeper networks. Finally, a global map is generated with the fully connected layer and passed to the robot's navigation. The experimental results based on RGB-D SLAM Dataset and Benchmark database dataset show that the proposed VSLAM system based on CNN is superior to the traditional CNN VSLAM system in both ATE and RPE indicators.

## 1. Introduction

Significant advances in sensing and mobility of robotic systems have enabled us to address important information gathering issues such as environmental monitoring, search and rescue, source seeking, active sensing and active Simultaneous Localization and Mapping (SLAM). In all these scenarios, the robot is deployed to gather information about the physical phenomena of interest [1].

SLAM is a challenging problem for autonomous navigation of mobile robots: The mobile robot must be able to estimate its position and orientation (posture) within the map of its navigation environment. However, in many practically relevant applications (such as exploration missions or operations in harsh environments), maps are unavailable or highly uncertain. Therefore, in this case, the robot must use the measurements provided by its sensing device to estimate the environmental map while positioning itself within the map. The environment map is constructed by the information mobile robot extraction of the specific target, the position of the feature object is provided by the observation model, and the position of the specific target and the posture of the robot are stored in the variable state vector. As the robot moves, the sensor continuously extracts environmental information and updates the posture state of the robot itself through the motion model [2, 3].

The literature [4] introduced the application of Extended Kalman Filter (EKF) to mobile robot navigation problems in known environments. By developing a model-based positioning algorithm that relies on the concept of geometric beacons (a naturally occurring environmental feature that can be reliably observed in continuous sensor measurements) and can be accurately refined by simple geometric parametrization description. The algorithm is based on EKF and utilizes the match between the observed geometric beacon and the a priori mapping of the beacon position.

The paper [5] made three major contributions to solving the SLAM problem. First, it demonstrates three key convergence characteristics of a complete SLAM filter. Second, it clarifies the true structure of the SLAM problem and shows how it can be used to develop a consistent SLAM algorithm. Finally, it demonstrates and evaluates the implementation of a complete SLAM algorithm in an outdoor environment using millimeter-wave (MMW) radar sensors.

Fast-SLAM is a Rao-Blackwellised particle filter that is simultaneously positioned and mapped [6]. It shows that the algorithm degrades over time, regardless of the number of particles used or the density of the markers in the environment, and always produces an optimistic estimate of long-term uncertainty. Fast-SLAM behaves like a non-optimal local search algorithm; in the short term, it may produce consistent uncertainty estimates, but in the long run, it cannot fully explore the state space is a reasonable shell. Yates estimates. However, the number of particles and landmarks does affect the accuracy of the estimated mean, and Fast-SLAM can produce good non-random estimates in practice given.

In the literature [7], as a review, it introduced the research situation of Visual SLAM. Visual SLAM refers to the problem of using images as the sole source of external information, to simultaneously establish the position of robots, vehicles or mobile cameras in the environment, and construct the exploration area at the same time. Representation. SLAM is an important task for robot autonomy. Nowadays, when using a range sensor such as a laser or sonar to construct a 2D map of a small static environment, the problem of SLAM is considered to have been solved. However, SLAM is an active research area for dynamic, complex and large-scale environments, using vision as the only external sensor. Computer vision techniques used in visual SLAM, such as detection, description and matching of salient features, image recognition and retrieval, remain to be improved.

Another study introduced an autonomous on-site 3D spatial data acquisition and sensing method for mobile robots, specifically for construction sites with many spatial uncertainties [8]. The proposed synchronous positioning and mapping (SLAM) based navigation and object recognition method is implemented and tested by a custom designed mobile robot platform, Ground Robotic Infrastructure (GRoMI), which uses multiple laser scanners and cameras to perceive and building a 3D environment map. Since SLAM did not detect uneven surface conditions and spatial-temporal objects on the ground, an obstacle detection algorithm was developed to identify and avoid obstacles and highly uneven terrain in real time. Given a 3D real-time scan generated by a 3D laser scanner, a path search algorithm was developed for autonomous navigation in an unknown environment with obstacles. Overall, the three-dimensional color map point cloud generated by GRoMI on the construction site is of enough quality to be used in many construction management applications such as construction schedule monitoring, safety hazard identification and defect detection.

Despite all the aspects of mobile robot SLAM research have received attention, the current autonomous mobile robots still have some problems in practical applications. For example, robots are likely to move under messy and irregular obstacle placement conditions; The shape of the obstacle and the position of the obstacle are changed; and the indoor and outdoor scene switching occurs at different times. The robot needs to constantly optimize the motion trajectory during the movement process, and obtain real-time information from multiple sensors for real-time recognition and rapid response to the surrounding environment.

This paper presents an improved visual SLAM indoor robot navigation strategy. This strategy is based on Microsoft Kinect sensors to obtain RGB and depth image information, based on convolutional neural network to process this information, and based on the map quality and effective coverage area to select the optimization objective function to minimize uncertainty; while considering real-time mapping and navigation scenarios, by calculating the maximum entropy to maintain tradeoff between optimization and efficiency.

The rest of this paper is organized as follows: In Section 2, we present the features of Kinect device and some studies related to visual SLAM navigation. Our proposed SLAM is given in Section 3. Then the experimental results and analysis are presented. Finally, the conclusions are given in the Section 5.

## 2. Kinect and Visual SLAM Technology

### *2.1. Kinect Device*

Since the advent of Microsoft Kinect, people have been able to use their bodies for human interaction and entertainment in a natural way. The key technology of human-computer interaction is the understanding of human behavior. The machine needs to know the response of the user's body posture, movements, etc. In the past, computers used machine vision methods for research, but it was very difficult. The Kinect device allows the computer to directly sense the depth information of the user and the environment, as well as the user's speech and direction of travel and posture, and can interpret the user's actions and translate them into a format that developers can handle.

The Kinect, which is shown in Fig.1, includes an array of display infrared (IR) projectors, color cameras and infrared cameras. The depth sensor consists of an infrared projector and an infrared camera, which is a monochrome complementary metal oxide semiconductor (CMOS) sensor. The depth sensing technique is based on the principle of structured light. An infrared projector is an infrared laser that passes through a diffraction grating and becomes a set of infrared points [9].

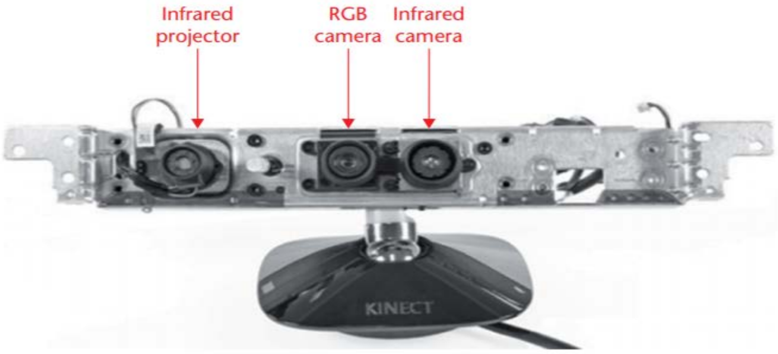

**Figure 1.** Kinect sensor: infrared (IR) projector, IR camera, and RGB camera [9].

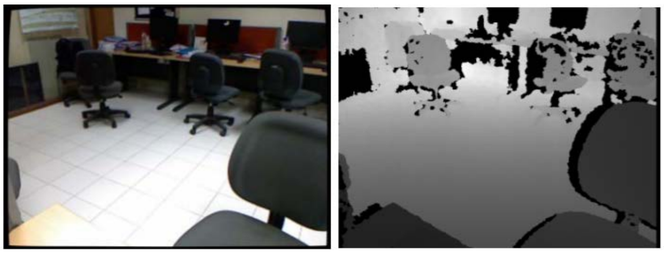

**Figure 2.** RGB image vs. image by Kinect [10].

The depth sensor consists of an infrared laser projector and a monochrome CMOS sensor that captures 3D video data in any ambient light conditions. Figure 2 shows the RGB camera image and the depth image of the same scene captured by the Kinect, where the color gradient from white to dark is used to visualize the depth map. The monochrome depth-sensing video stream uses VGA resolution with a sensitivity of 2,048 [10].

### *2.2. SLAM for Mobile Robots Navigation*

Simultaneous Positioning and Mapping (SLAM) is one of the fundamental challenges of robotics, dealing with the requirement to build an environment map while determining the location of the robot in the map. SLAM is a process by which a mobile robot can construct an environmental map and use this map to infer its location. Initially, the location of the map and the robot in the map is unknown, and the robot motion model is known. The most commonly used sensors for SLAM are laser-based, sonar-based and vision-based systems. Additional sources of sensing are used to better sense robot status information and the outside world, such as compass, infrared technology and Global Positioning

System (GPS). However, all these sensors carry some errors, often referred to as measurement noise, and there are some range limitations [11].

A typical case of SLAM is Visual SLAM (VSLAM) [12], which limits the used sensors to passive vision-based cameras. The use of cameras allows the development of precise automated systems at the same time, reducing costs and overall energy consumption. The first aspect to consider is the number of available views of the scene included one monocular camera, stereo equipment or camera array. Using multiple cameras can have an impact on post processing time, although this effect is not always negative in terms of computation time. Sometimes, getting information from multiple cameras helps simplify the algorithm because more constraints can be built from the extra information that is captured. An example is using a stereo camera to speed up feature extraction.

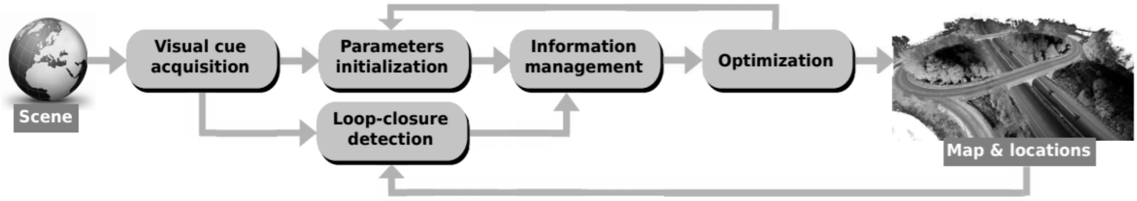

**Figure 3.** A standard VSLAM pipeline for driverless cars [12].

The goal of the VSLAM system is to fuse image and ranging data in a way that enables robust map construction and localization. Since sensor data acquired from mobile robots contains a large amount of noise that is difficult to simulate, "robustness" is the key. The ranging data is incremental, so it will accumulate errors over time. The first thing to understand is that the odometer sensor will never be perfectly calibrated. However, because the robot may slip or may be lifted and moved, the rangefinder will also be exposed to dramatic errors and discrete events. Image data is difficult to process due to image blur, occlusion, limited image resolution, imperfect camera calibration, variable lighting conditions, limited processing power, and the like. For example, if the robot is operating in a densely populated environment, the camera's field of view is often filled with unstable objects, such as moving objects, and cannot be used as a reference point in the map. In addition, lighting conditions can range from very dark to bright, even saturating the captured image [13].

To improve accuracy and efficiency when drawing large areas, two or more robots are usually required to participate in this task. This process is called multi-robot SLAM or cooperative SLAM (CSLAM). While improving efficiency, the complexity of SLAM increases as robots collaborate to build a single joint map of the areas they explore. This problem becomes particularly challenging when the coordinate transformation between the initial poses of the robot is unknown [14].

Recently, many CV-SLAM algorithms have appeared in robotics research. From a traditional front view system, ceiling vision has an advantage in mobile robot positioning because it involves only rotational and affine transformations without changing the scale. Therefore, it is more convenient to utilize the ceiling view feature than the feature detected in the 3D environment. Despite the simple ceiling view, there are still many features that can be used as landmarks, such as corners, key points, lines, circles, and more. As humans, we are experts in quickly and accurately identifying the most visually visible foreground objects in the scene and adaptively focus our attention on these import areas. Therefore, we can select the best features by measuring the significant intensity of the features [15].

For those point cloud matching methods, accuracy and efficiency depend on the type of data. For 2D lidar sensors, plane information can only be detected in one scan, which affects the robustness of the matching method, especially in the absence of certain points. 3D lidars can provide multiple scan lines, and they typically require more scan time and are often costly. Due to the high cost and low scanning speed of 3D lidar, some researchers use 2D lidar to measure the 3D environment and propose many methods. For example, using distance measurements from rotating 2D lidar, real-time algorithms for low-drift ranging and mapping; using 3D point lasers to obtain 3D point clouds, where 6-DOF sensor trajectories are recovered to correct point clouds based on 3D scan matching alignment. 2D Lidar is used to perform full 3D environmental detection and build coherent 3D maps. For mobile robot navigation in indoor environments, accurate 2D maps must be built. When using traditional 2D lidar to map the entire 3D environment, the accuracy of frame-to-frame matching is improved, while mapping will cost more computing resources and work at low frequencies. Literature [16] designed a 2.5D laser radar device with 2.5D point data to improve the robustness and accuracy of the scan matching process. Finally, the map is constructed by compressing 2.5D points to reduce computational complexity and memory consumption.

## 3. An Improved Visual SLAM Indoor Robot Navigation Strategy

### *3.1. Convolutional Neural Network Structure*

Convolutional neural network (CNN) is a deep feedforward artificial neural network in which the neural network preserves the hierarchy by learning internal feature representations and generalizes features in common image problems into object recognition and other computer vision problems. It is not limited to images; it also achieves the most advanced results in natural language processing problems and speech recognition.

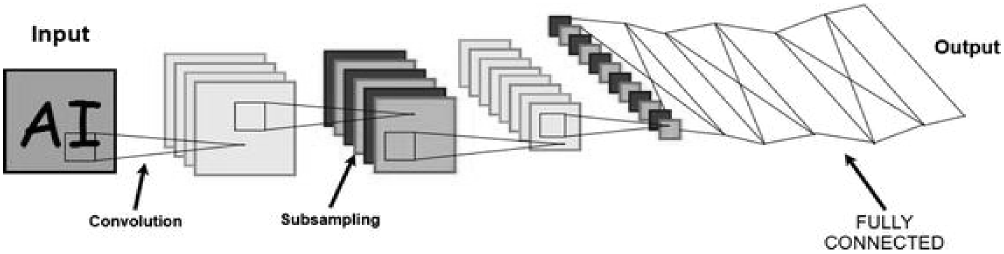

**Figure 4.** Layers in a convolution neural network [17]

In Fig.4, the convolutional layer consists of a filter and an image map. Consider the size of the grayscale input image as 5 x 5, which is a matrix of 25-pixel values. The image data is represented as a three-dimensional matrix of width * height * channels. Convolution is intended to extract features from the input image, thus maintaining spatial relationships between pixels by learning image features using small square input data. Rotation invariance, translation invariance, and scale invariance can be expected. For example, due to the convolution step, the CNN can easily identify a rotated cat image or a rescaled cat image. the filter is slide over the original image and calculate element-wise multiplication at each given position and add a multiply output to get the output of the final integer of the matrix element [17].

Training the convolutional layer means the local receptive field of the training layer, which has a single deviation and a small amount of weight. In this respect, it is like a small logistic regression, which is why the convolutional network is trained quickly: they only need to learn a few parameters. The main structural difference between logistic regression and local receptive fields is that we can use any activation function in the local receptive domain, and we should use logical functions in logistic regression. The most common activation function is a rectifying linear unit or ReLU. ReLU's x is only the maximum of 0 and x, which means 0 if the input is negative, otherwise it returns the original input.

$$\rho(x) = \max(x, 0) \quad \text{(Eq.1)}$$

The idea behind max-pooling is that important information in the image is rarely contained in adjacent pixels, it is usually contained in darker pixels. You may notice immediately that this is a very powerful assumption and usually may not work. It must be said that max-pooling is rarely used for the images themselves, but for the feature maps that are learned, they are images, but they are very strange images. You can try modifying the code in the following section to print out the feature map from the convolutional layer. You can consider the maximum number of pools that reduce the screen resolution. In general, if you identify a dog on a 1200 x 1600 image, you might recognize it on a 600 x 800 image [18].

### *3.2. Improved Visual SLAM Indoor Robot Navigation based on CNN*

In the actual VSLAM indoor navigation system, there are many difficulties, such as the robot may move under the condition of disordered and irregular obstacles; the shape of the obstacle and the position of the obstacle change; and the indoor and outdoor scenes are switched differently. Time has happened. Robots need to continuously optimize their trajectories during motion and get real-time information from multiple sensors. In order to realize the real-time recognition of the robot and the rapid response to the surrounding environment, we improved the VSLAM based on CNN, replaced the original convolution filter with parallel multi-filter, and further reduced the model parameter volume by the parallel pooling method. Such an approach can reduce the likelihood that the system will fall into a gradient and continue to train deeper networks. Then use the fully connected layer to process the output after the parallel convolution, and finally the global map is produced and transferred into the navigation of robots.

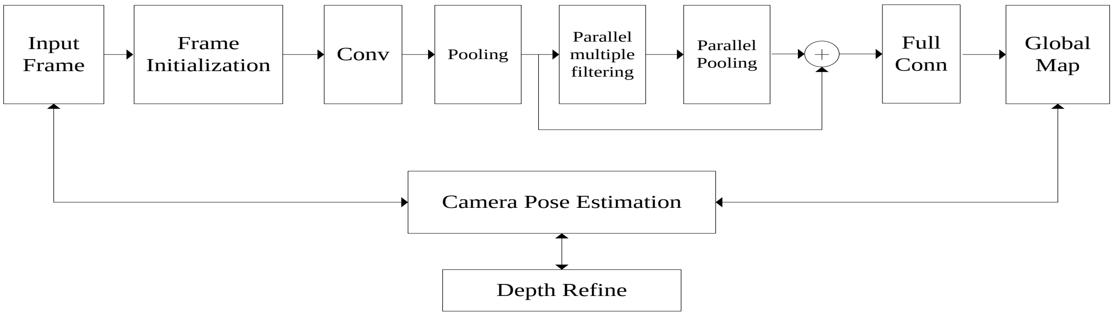

**Figure 5.** Improved visual SLAM indoor robot navigation structure based on improved CNN

## 4. Experiments and Test Results

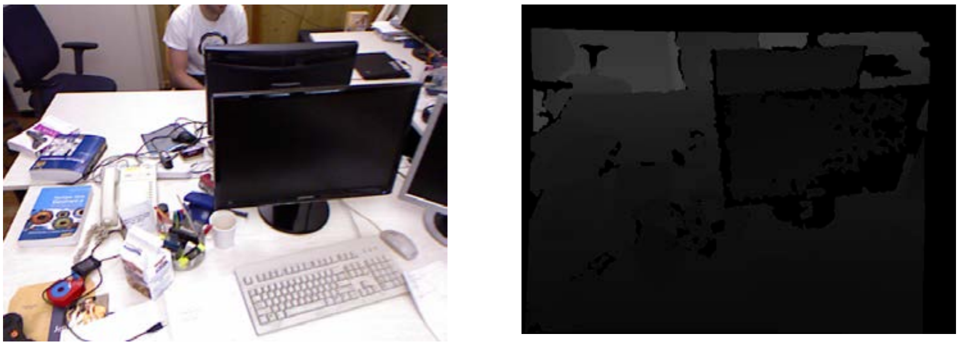

**Figure 6.** The example of Benchmark database [19].

To evaluate the performance of the above methods, we evaluated and tested in the RGB-D SLAM Dataset and Benchmark database [19]. The dataset provides a large data set containing RGB-D data and ground truth data to establish a new benchmark for visual ranging and visual SLAM system evaluation. The data set contains the color and depth images of the Microsoft Kinect sensor along the ground truth track of the sensor. Data was recorded at full frame rate (30Hz) and sensor resolution (640*480). The ground truth trajectory is obtained from a high precision motion capture system with eight high speed tracking cameras (100Hz). Kinect's accelerometer data is also available.

The absolute trajectory error (ATE) and relative pose error (RPE) are used for a quantitative evaluation to compare the absolute distance between the trajectory of the estimated frame and the real trajectory frame. Root mean squared error (RMSE) represents the root mean squared error of the entire trajectory.

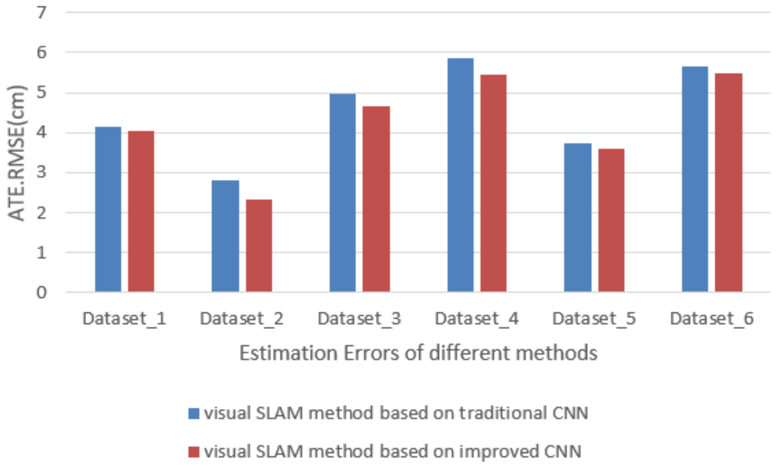

**Figure 7.** Root mean squared error of absolute trajectory error on different method.

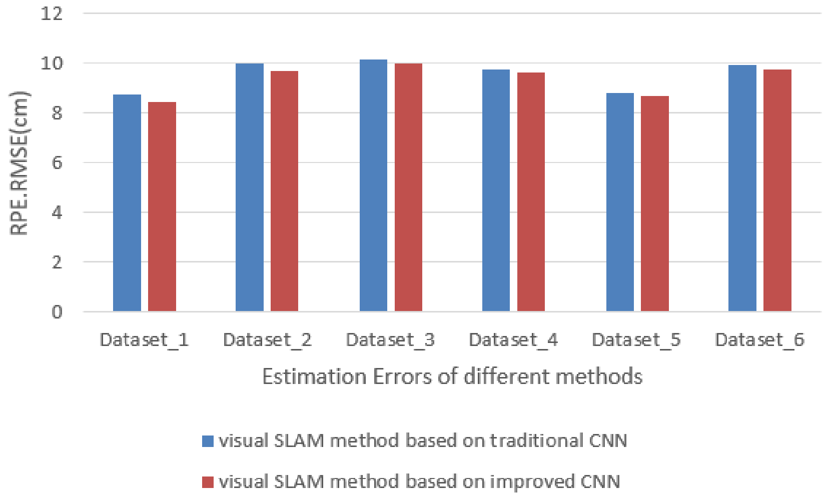

**Figure 8.** Root mean squared error of relative pose error on different method.

We randomly divided the data sets into 6 groups, tested them, and presented the results in Figures 7 and 8. Figure 7 shows the ATE test results of the VSLAM under the traditional CNN method and our CNN method. Figure 8 shows the RPE test results of the VSLAM under the traditional CNN method and our CNN method. According to the data in the figure, in most cases, our method has some error reduction values on the ATE and RPE indicators than the traditional CNN method. This reduction is significantly higher than the RPE index in the ATE index. Therefore, it can be considered that the improved CNN-based VSLAM has better navigation and tracking characteristics than the traditional CNN method.

## 5. Conclusions

In the mobile robot indoor navigation system, the robot moves under indoor and outdoor conditions of disordered and irregular obstacles and the shape and position of the obstacle therein may change. On the other hand, indoor and outdoor scenes need to be switched at any time. To enable mobile robots to obtain real-time information from multiple sensors and continuously optimize their trajectories during travel and navigation to achieve rapid response to the surrounding environment, we improved the CNN-based VSLAM system. The method is to replace the original with paralleling single convolutional layer and reducing the number of model parameters. This approach can reduce the problem of system gradient disappearance and continue to train deeper networks. Finally, a global map is generated with the fully connected layer and passed to the robot's navigation.

## 6. References

[1] Atanasov N, Le Ny J, Daniilidis K, et al. Decentralized active information acquisition: Theory and application to multi-robot SLAM[C]. Robotics and Automation (ICRA), 2015 IEEE International Conference on. IEEE, 2015: 4775-4782

[2] Li X, Huang R, Zhao Y, et al. Mobile robot SLAM algorithm for transformer internal detection and location[C]. 2018 International Conference on Electronics Technology (ICET). IEEE, 2018.

[3] Garulli A, Giannitrapani A, Rossi A, et al. Mobile robot SLAM for line-based environment representation[C]. IEEE Conference on Decision and Control. IEEE; 1998, 2005, 44(2): 2041.

[4] Leonard J J, Durrant-Whyte H F. Mobile robot localization by tracking geometric beacons[J]. IEEE Transactions on robotics and Automation, 1991, 7(3): 376-382.

[5] Dissanayake M W M G, Newman P, Clark S, et al. A solution to the simultaneous localization and map building (SLAM) problem[J]. IEEE Transactions on robotics and automation, 2001, 17(3): 229-241.

[6] Bailey T, Nieto J, Nebot E. Consistency of the FastSLAM algorithm[C]. Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on. IEEE, 2006: 424-429.

[7] Fuentes-Pacheco J, Ruiz-Ascencio J, Rendón-Mancha J M. Visual simultaneous localization and mapping: a survey[J]. Artificial Intelligence Review, 2015, 43(1): 55-81.

[8] Kim P, Chen J, Kim J, et al. SLAM-Driven Intelligent Autonomous Mobile Robot Navigation for Construction Applications[C]. Workshop of the European Group for Intelligent Computing in Engineering. Springer, Cham, 2018: 254-269.

[9] Zhang Z. Microsoft kinect sensor and its effect[J]. IEEE multimedia, 2012, 19(2): 4-10.

[10] Biswas K K, Basu S K. Gesture recognition using microsoft kinect[C]. Automation, Robotics and Applications (ICARA), 2011 5th International Conference on. IEEE, 2011: 100-103.

[11] Aulinas J, Petillot Y R, Salvi J, et al. The SLAM problem: a survey[J]. CCIA, 2008, 184(1): 363-371..

[12] Ros G, Sappa A, Ponsa D, et al. Visual slam for driverless cars: A brief survey[C]. Intelligent Vehicles Symposium (IV) Workshops. 2012, 2.

[13] Karlsson N, Di Bernardo E, Ostrowski J, et al. The vSLAM algorithm for robust localization and mapping[C]. Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on. IEEE, 2005: 24-29.

[14] Zhou X S, Roumeliotis S I. Multi-robot SLAM with unknown initial correspondence: The robot rendezvous case[C]. Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on. IEEE, 2006: 1785-1792.

[15] Li L, Kim H R, Jiang S, et al. Feature saliency based SLAM of mobile robot[C]. 2018 International Conference on Electronics, Information, and Communication (ICEIC). IEEE, 2018: 1-3.

[16] Yang Y, Yang G, Tian Y, et al. A robust and accurate SLAM algorithm for omni-directional mobile robots based on a novel 2.5 D lidar device[C]. 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA). IEEE, 2018: 2123-2127.

[17] Manaswi N K. Convolutional Neural Networks[M]. Deep Learning with Applications Using Python. Apress, Berkeley, CA, 2018: 91-96.

[18] Skansi S. Convolutional Neural Networks[M]. Introduction to Deep Learning. Springer, Cham, 2018: 121-133.

[19] https://vision.in.tum.de/data/datasets/rgbd-dataset