2020 20th International Conference on Control, Automation and Systems (ICCAS 2020)
Oct. 13~16, 2020; BEXCO, Busan, Korea

Object Detection-based Semantic Map Building for A Semantic Visual SLAM System

Phuc H. Truong[1] Sujeong You[1] and Sanghoon Ji[1*]

[1] Department of Robot, Korea Institute of Industrial Technology,
Ansan, Gyeongi, 13391, Korea ({phtruong, sjyou21, robot91}@kitech.re.kr)

Abstract: In this paper, we propose a novel semantic visual simultaneous localization and mapping (SLAM) system which can provide rich semantic information of landmarks in the map. The system is created based on a visual SLAM system using a RGBD camera. We integrate an object detection stage on the top of the SLAM system to obtain semantic landmarks of the scene. The semantic object-based landmarks are continuously saved into a semantic map tc not only describe context information, but also update the semantic map in dynamical environments where objects can be moved passively. The experimental results demonstrated the validity of the proposed system.

Keywords: SLAM, VSLAM, Semantic SLAM, Object Detection.

1. INTRODUCTION

Simultaneous localization and mapping (SLAM) is a emerging topic in research community. SLAM systems deals with navigation of robot and generation of local map in parallel using different sensors [1, 2]. One of the appealing approach in SLAM is using off-the-shore cameras for observation due to its lost-cost and portable characteristic [2, 3]. These visual SLAM systems use different vision techniques to detect key points so that the systems can improve navigation accuracy and navigation mapping [4, 5]. In [4], authors detected Harris-Laplace corners to detection keypoints to track for navigation. In [5], SIFT descriptor is applied for keypoints detection. The disadvantage of these features is computation time is so long. In ORB-SLAM system [6, 7], the ORB features are utilized to improve the runtime.

Recently, the development of deep learning techniques has been boosted the accuracy and context description of SLAM systems. Utilizing object detection results [8, 9,10], SLAM systems can improve its navigation accuracy by determining the semantic landmarks in the environment. Authors in [11] used segmentation network to identify static and dynamic objects to improve the process of tracking in navigation.

In this paper, we present a SLAM system that can navigate with RGBD camera and use object detection to generate semantic landmark for the map. The system i s built on the top of the ORB-SLAM2 system as the main navigation and mapping functions. Besides, we add a semantic map building based on results of object detection to give semantic visual landmarks for the systems. Experiment results show the promising application of the proposed system.

2. SEMANTIC VISUAL SLAM SYSTEM

The proposed visual slam system is a combination of a visual SLAM with an object detector and a semantic map to navigate and describe the scene semantically. We build our system on the top of ORB-SLAM2 system which can navigate with different camera options. Integrating a object detector, our system can incorporates semantic information for map building process and visual navigation using object-based landmarks. We have defined static objects as landmarks for the system and applied the YOLO [8, 9] neural network to detect these objects when robot moving in the environment. The semantic mapping process is an important part to incorporating visual landmark to the system and to update the visual change of the environment. The proposed system can be depicted as in Figure 1.

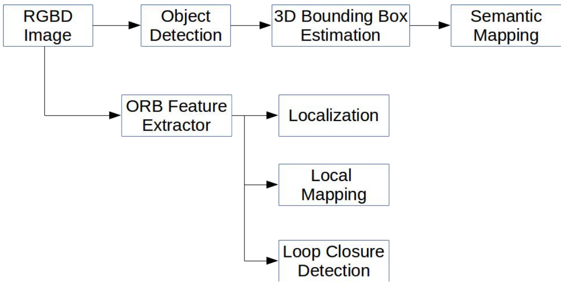


Fig. 1 The architecture of the proposed semantic SLAM. system.

2.1 Simultaneously Localization and Mapping

Because the object detection is conducted on the RGB image and the 3D bounding box is constructed by combing 2D box with depth information, our approach is generic to any RGBD-based SLAM system. However, to demonstrate the validity of the system, we build our system on the top of the ORB-SLAM2 which has been demonstrated efficiently with different camera options [7].

The ORB-SLAM2 system includes an ORB feature extractor and three main threads for localization, local mapping and loop closure detection, respectively. The ORB extractor draws out key points from each frame as the base for further processes. Based on these key points, localization thread estimates the pose of the current

978-89-93215-19-9/20/$31.00 ICROS

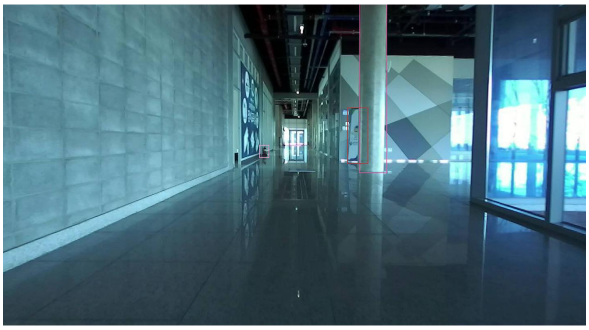


Fig. 2 Landmark objects used to obtain semantic information.

frame, tracks the local map, and adds key frames. The local mapping thread generates map based on the key frames. The key points are also used for loop closure in the navigation.

2.2 Object Detection

The object detection is the critical stage to semantically building map for the system. It provides the visual information in human cognition level to describe the environment. We basically detect landmark objects and calculate their position in the global map to add into the semantic map. These information is an important reference for further processes related to the interaction with the environment. To simplify the semantic landmarks, we pick up some ubiquitous objects in indoor environments to use as the landmark objects to detect. The object list can be described on the Table 1.

Table 1 List of landmark objects.

Fire Extinguisher	Trash Bin	Elevator Door	Chair
Table	Bench	Floor Sign	Hinged Door
Room Sing	Fireplug	Column	Drinking Fountain
Information Sign	Speed Gate	Standing Signboard	

In Figure 2, we show a sample of an indoor environment with landmark objects marked in the scene. The object detector should be quickly and accurately detect these objects and calculate their 3D bounding boxes to create semantic description of the objects.

We use the YOLO detector, which is created based on convolutional neural network (CNN) layers, to detect objects due to its reputation of balancing between accuracy and real-time detection ability. The YOLO is an one-stage detection algorithm which can predict bounding boxes and boxes' classes directly from input images in one computation pass. The YOLO divides an input image into NxN blocks, and each block takes responsibility to detect all objects centering in its region. YOLO predefined 9 anchor boxes for each block which were generated using K-Means clustering to be reference for the predictive boxes. These anchor boxes make the prediction more accurate and robust. A neural network composed of 53 CNN layers organizing in residual blocks, named Darknet53, is used to extract features from the input image. Several layers are added to predict objects' boxes, classes and confident scores from feature maps extracted by the base network. A spatial pyramid technique which locate objects at different feature maps is used to improve the detection at different scales of features. In inference stage, the non-maximum suppression technique is used to remove all boxes of low confident scores and high overlapping with other same-class boxes. This make the algorithm provide more accurate bounding boxes.

Applying the YOLO algorithm, we recalculate the anchor boxes for new object list and retrain the model for the new dataset. The training is conducted on a GPU device. Thus, the model can well detect the objects in different frames. To accelerate the SLAM system, we only apply the detection on the key frames of the navigation.

2.3 Semantic mapping

The results of detection step are processed to obtain the 3D coordinate of the bounding boxes. These are the important information to locate the object in navigation, provide semantic data from the map and possibly promote the loop closure detection of the SLAM system.

Beside, the system can be able to update its object states in the semantic map. If the objects are moved compared with the original position checked by the semantic map, the map will update this change of the corresponding objects by their IDs and corresponding attributes. We used the TOSM semantic database [12], which supports different attributes and is able to quickly query and edit the map itself, to leverage the system. Figure 3 visualizes an example of building semantic map based on results of static object detection.

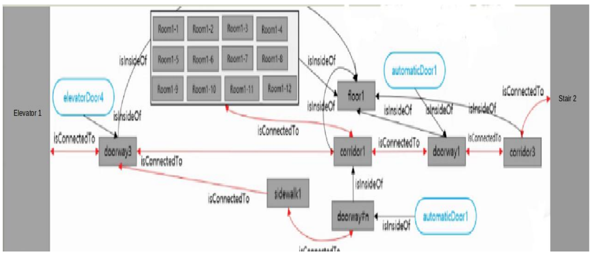


Fig. 3 An example of building semantic map from object detection results.

3. EXPERIMENTAL RESUTLS

We validate our system on the ORB-SLAM2 and YOLO frameworks for navigation and object detection, respectively as described in Section 3. We conducted our experiments on different data and indoor environment to check the quality of navigation, object detection, and semantic mapping of the system. Specifically, we trained the network on a PC with a

GPU of Nvidia Titan XP and then, use this trained model to detect in navigation.

Figure 4 shows the resultant map built by the SLAM system in our navigation data. This map mostly created based on the ORB-SLAM2 with ORB features used to determine the key points. In this figure, the white and red points are global key points which have been saved in to the map and local key points which detected in current frame, respectively, whereas, the green part indicates the key frame in the trajectory.

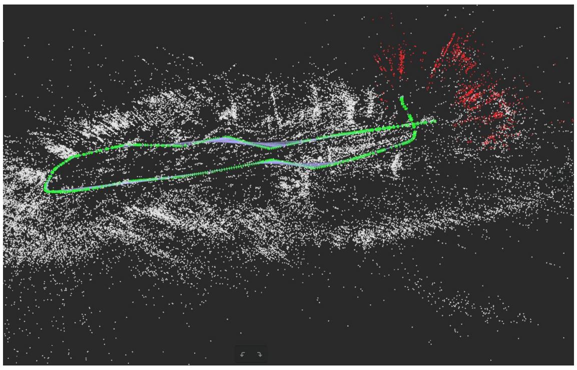


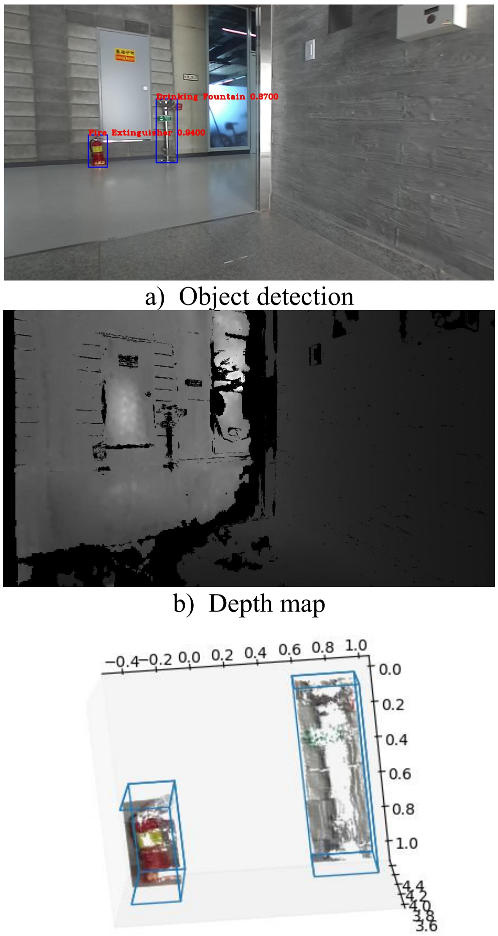
Fig. 4 Trajectory with key points created by the SLAM system.

At each key frame, we apply the object detector to find landmark objects and their positions to analyze. Figure 5a shows the detection results on a key frame in navigation. The detected boxes is converted into 3d boxes using depth information and camera parameters. The 3d box information is compared with information on the semantic map to update the positions of the detected objects in the semantic map. Figure 5b and 5c depict the depth data and point cloud with 3d boxes calculated from the detected bounding box. Figure 6 shows a generated map with landmark objects inserted.

The bounding boxes are indexed in the global map. These information are continuously update into the semantic map. Thus, the map can not only follow the movement, but also can monitor the change of the environment by tracking the landmark objects in each key frames. These results provide a promising resource for further process of the environment.

4. CONCLUSION

In this paper, we presented a semantic SLAM system which can navigate and generate semantic map based object detection. The system was built on the top of the ORB-SLAM2 which works well with navigation using different cameras. Our system used RBG-D camera to navigation and detect landmark objects for semantic mapping.These map provided a better understanding of the environment. Moreover, the landmark information from the semantic map can be possibly used to improve the loop closure function of the SLAM system.



a) Object detection

b) Depth map

c) 3D bounding boxes

Fig. 5 Results of the SLAM system in object detection and 3D bounding box conversion.

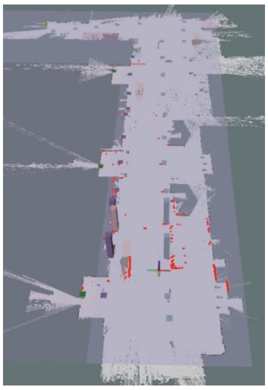


Fig. 6 Semantic map with detected objects.

ACKNOWLEDGMENT

This research was supported by the R&D Program of Ministry of Trade, Industry and Energy (No. 20000683).

REFERENCES

[1] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. F. Durrant-Whyte, and M. Csorba, “A solution to the simultaneous localization and map building (SLAM) problem,” IEEE Trans. Robot. Automat., vol. 17, no. 3, pp. 229–241, 2001.

[2] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, and J. Lenoard, “Simultaneous localization and ampping: Present, future, and the robust perception age,”, arXiv: 1606.05830, 2016.

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

[4] K. Mikolajczyk and C. Schmid, “Indexing based on scale invariant interest points,” in Proc. of Int’l Conf. on Computer Vision (ICCV), 2001.

[5] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int’l J. of Computer Vision (IJCV), vol. 60, no. 2, pp. 91–110, 2004.

[6] R. Mur-Artal, J. Montiel and J. Tardos, “ORB -SLAM: A Versatile and Accurate Monocular SLAM System,” *IEEE Transactions on Robotic s*, Vol. 31, No. 5, pp. 1147-1163, 2015.

[7] R. Mur-Artal and J. Tardos, “ORB-SLAM2: A n Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras,” *IEEE Transacti ons on Robotics*, Vol. 33, No. 5, pp. 1255-126 2, 2017.

[8] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 779-788, 2016.

[9] J. Redmon, A. Farhadi, “*YOLOv3: An Incremental Improvement*,” ArXiv ID: 1804.02767, 2018.

[10] Badrinarayanan V, Kendall A, Cipolla R. “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE transactions on pattern analysis and machine intelligence,* Vol. 39, no. 12, pp. 2481-95, 2017.

[11] V. Murali, H. P. Chiu,S. Samarasekera and R.T. Kumar, “Utilizing semantic visual landmarks for precise vehicle navigation,” *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1-8, 2017.

[12] SH Joo, S Manzoor, YG Rocha, SH Bae, KH Lee, “Autonomous Navigation Framework for Intelligent Robots Based on a Semantic Environment Modeling,” *Applied Sciences*, Vol. 10, No. 9, pp 3219, 2020.