Semantic Segmentation based Dense RGB-D SLAM in Dynamic Environments

Jianbo Zhang¹, Yanjie Liu*¹, Junguo Chen², Liulong Ma¹, Dong Jin¹ and Jiao Chen¹

¹State Key Laboratory of Robotics and System (HIT)
²China Petroleum Materials Company Limited

E-mail: 17S008048@stu.hit.edu.cn

Abstract. Visual Simultaneous Location and Mapping (SLAM) based on RGB-D has developed as a fundamental capability for intelligent mobile robot. However, most of existing SLAM algorithms assume that the environment is static and not suitable for dynamic environments. This is because moving objects in dynamic environments can interfere with camera pose tracking, cause undesired objects to be integrated into the map. In this paper, we modify the existing framework for RGB-D SLAM in dynamic environments, which reduces the influence of moving objects and reconstructs the background. The method starts by semantic segmentation and motion points detection, then removing feature points on moving objects. Meanwhile, a clean and accurate semantic map is produced, which contains semantic information maintenance part. Quantitative experiments using TUM RGB-D dataset are conducted. The results show that the absolute trajectory accuracy and real-time performance in dynamic scenes can be improved.

1. Introduction
For navigation purposes, visual SLAM is widely used in many robotic applications. In the past years, RGB-D based visual SLAM is extensively researched due to the emergence of economical depth cameras such as Kinect2. In order to estimate the camera’s ego-motion and perceive environmental information, the SLAM system based on RGB-D can directly use the color and depth information provided by the camera. Lots of advanced SLAM algorithms have been developed and achieve good results.

In order to simplify the problem, most of current visual SLAM systems assume that the environment is static. It means that the change of image information depends only on the motion of the camera itself. However, in the actual environment, there are inevitably dynamic objects, such as walking people, moving tables and chairs, etc. Some algorithms (such as random sampling consistency, RANSAC) can be used to handle the outliers introduced by small moving objects. However, when the dynamic object is relatively large in the image, it will introduce errors into the SLAM system. Then, the system gets the wrong pose information, which affects the accurate construction of the map. Besides, most of the semantic maps built by the system are less accurate and lack of effective semantic information maintenance [1].

Improving the performance of the SLAM system in dynamic environments is very important for mobile robots. SLAM uses the continuous image sequence output by the camera to estimate the pose. Dynamic objects in the scene can lead to mismatches in feature matching, affecting pose estimation.
accuracy. Therefore, Dynamic objects in the scene should be removed. Some methods are adopted to reduce the impact of moving objects, for instance, moving objects can be inspected and removed using scene flow in order to avoid mismatches in the alignment step [2]. But this problem has not been solved very well.

Semantic map can provide rich environmental information for robot navigation. The traditional semantic map construction mainly uses the support vector, CRF and other machine learning methods to detect and segment the scene, but the performance and accuracy are limited. At present, the deep learning method provides a new way of constructing a semantic map, which can achieve better performance.

In this paper, we are concerned about real-time detection of dynamic objects to reduce the impact of moving objects. Meanwhile, the system should provide a semantic octo-tree map, which contains accurate semantic information.

The contribution of this paper are as follows:
- We propose to use ENet to obtain pixel-level semantic segmentation of the image. Then, moving objects can be detected by combining semantic segmentation with moving consistency check method, so that the SLAM algorithm will not extract features on them. Compared with ORB, the accuracy and robustness of the system are improved. We achieve a balance of system accuracy and real-time performance.
- We build a semantic octo-tree map in the map creation thread. In order to avoid fusing moving objects into map, the dense semantic octo-tree map uses semantic information and log-odds score method to filter out them [1]. Meanwhile, this thread combines the semantic information of different viewpoints to achieve multi-view consistent semantic segmentation.

The paper is organized as follows. We review state-of-art visual SLAM in dynamic environments as well as several approaches about semantic SLAM in Section 2. Then Section 3 presents the overall structure of the whole SLAM system, that how to detect moving objects on images and update semantic information. Section 4 provide our evaluation and results of performance of the system against dynamic environments, where we compare the method against related works. Section 5 concludes our work and future works for research.

2. Related work

2.1. Moving objects detection

Most SLAM systems are based on the assumption of static environment. However, there are inevitably dynamic objects in the actual environment, which will affect the accuracy and robustness of SALM system.

To address this problem, special measure should be taken to enhance its robustness and accuracy in dynamic environments. In most SLAM systems, dynamic objects are classified as noise, separated out before doing SLAM, therefore neither included in the map nor used for pose estimation. There are several SLAM systems that can detect dynamic objects more specifically. P.Ochs et al. [3] detect dynamic object by long term point trajectory analysis without priori model. The method is robust, but the time cost is very expensive. Jens et al. [4] combine optical flow and depth information to detect moving objects. However, optimization methods need to be used to reduce processing complexity and minimize the high computational time. Besides, Wangsiripitak and Murray [5] deal with dynamic objects by tracking known 3D objects in the scene. The method can detect those a priori dynamic objects, but it can’t detect changes made by static objects, such as a chair pushed by a person, or a ball thrown by someone.

2.2. Semantic SLAM

In most SLAM algorithms, the position of the robot and the map point information are only dense or sparse geometric points in the space. By estimating the position of these spatial points, we can obtain relatively accurate position information, but can’t get higher-level semantic information. Thus, the rich
semantic information in the environment can’t be used in complex task. Semantic information can help SLAM better understand the scene and get more information about the scene, thus improving the accuracy of positioning and map construction.

So semantic map has become a popular research domain in recent years. Galindo et al. [6] divide the map into spatial map and conceptual map, organized in a hierarchical manner, and merge the spatial map nodes with the conceptual map nodes. With the development of visual target detection algorithms based on deep neural networks, we can use more accurate semantic information to construct semantic maps. J.McCormack et al. [7] propose a dense 3D semantic mapping method using convolutional neural networks by combining CNN with ElasticFusion. Their work focus on building maps using semantic information, while semantic information is not used in other parts. Besides, many methods have been proposed to fuse semantic information from different perspectives to perform multi-view consistent semantic information. Li et al. [8] propose to fuse the semantics information by using bayesian fusion on key frames from their SLAM backend.

In this paper, semantic information is not only used to generate semantic octree maps, but also utilized to filter out moving object in the process of tracking in dynamic environments [1]. And we will fuse semantic information from different perspectives.

3. System description

In this section, first, the framework of the system is presented. Second, a robust moving object segmentation method will be introduced in details, which enhance the robustness and accuracy of dense SLAM in dynamic environments. We use a real-time semantic segmentation network named ENet and optical flow to detect moving objects. Finally, we present the method to create semantic octo-tree map, which includes semantic information maintenance.

3.1. Framework of the system

ORB-SLAM2 performs well in most practical scenarios, so the system is improved based on ORB-SLAM2 to provide global feature-based SLAM in dynamic environments. we modify the existing framework [1] for RGB-D SLAM in dynamic environments, and the overview of the system is shown in Figure 1.

![Figure 1. The framework of system](image)

Five threads run in parallel in the system: tracking, semantic segmentation, local mapping, loop closing, and map creation. Local mapping and loop closing are the same as ORB-SLAM2. The RGB image captured by Kinect2 is utilized to semantic segmentation thread and moving consistency check simultaneously. Moving objects are detected by combining the potential outliers from moving consistency check and pixel-wise semantic label from semantic segmentation thread, then the ORB feature points located in moving objects will be filtered out. Finally, the system calculates the transformation matrix by matching the stable ORB feature points [1].

The system is based on ROS (Robot Operating System). ROS is an information delivery system in which many parts can work together by communicating output information.
3.2. Moving objects detection
This system combines semantic segmentation with moving consistency check to detect dynamic objects, using the contours extracted by semantic segmentation thread and motion information provided by moving consistency check thread.

To extract the contours of dynamic objects, we use a CNN to obtain a pixel-level semantic segmentation of the image. Considering the application of the system in the actual scene, the system should achieve a balance between accuracy and real-time. So we use ENet to process images in our experiments, which is one of the most advanced semantic segmentation networks. ENet consists of a main branch and an additional branch with a convolution kernel, and finally performs pixel-level addition and fusion. ENet is 18 times faster, has 79 less parameters, requires 75 less FLOPs, provides similar or better accuracy to other existing models [9]. The input to ENet is the RGB raw image and the idea is to segment those classes that may be dynamic or moveable, such as people, bicycles, cars, motorcycles etc.

Dataset plays an important role in the training process of semantic segmentation network, which will affect the accuracy of segmentation. This article combines the original pascal-voc2012 dataset with the voc_aug dataset provided by B. Hariharan et al. for our semantic segmentation task. The Adam optimization algorithm is used to train the network, which allows ENet to quickly converge using GTX 1080Ti GPU. The training is performed in two phases: First, we only train the encoder to classify the downsampled regions of the input image, then add the decoder and train the network for upsampling and pixel classification. Learning rate of 5e-3 and L2 weight decay of 2e-4, along with batch size of 8 consistently provided the best results. Training result is shown in Figure 2, Figure 3.

Figure 2. RGB image  Figure 3. The semantic segmentation

The test result demonstrates that ENet has similar accuracy, but better real-time performance compared with Segnet. The inference time of ENet per frame is 23.86ms, while Segnet is 40ms.

We can detect moving points by moving consistency check based on sparse optical flow method. Firstly, optical flow pyramid is calculated to get the matched points in the current frames. Secondly, the fundamental matrix is calculated by using RANSAC with the most inliers. Then, the epipolar line in the current frame can be found by using the fundamental matrix. Finally, the key points will be determined to be moving if the distance from a matched point to its epipolar line is more than a threshold [2]. Therefore, we can detect the dynamic points in the current frame.

This system combines semantic segmentation with moving consistency check to detect dynamic objects. If the number of dynamic points in the contours of a segmented object is more than a certain threshold, this object is regard as a dynamic object. Then remove the feature points located in dynamic regions. The experimental results are shown in the Figure 4, Figure 5.

Figure 4. ORB feature points  Figure 5. Removing outliers
3.3. Dense semantic map building

The oct-tree map saves a lot of storage space compared to the point cloud map. In addition, the oct-tree map can be used to query occupancy information and used for navigation. Therefore, the system creates a separate thread to build a dense 3D oct-tree map using new key frames from tracking and segmentation result. In this process, semantic information is not only used to filter out dynamic objects but also to create semantic map [1]. When we look at a scene from a different point of view, we may get different semantic results. We combine the semantic information of different viewpoints to achieve multi-view consistent semantic segmentation. The system uses max fusion to fuse semantics information of different measurements. When max fusion is performed, the voxel only includes the semantic color with the highest confidence generated by the ENet and its highest confidence. When inserting the point cloud into an oct-tree map, if the voxel has new semantic result, we fuse two semantics in this way [10]. The pseudo code is shown in Algorithm 1.

Algorithm 1 max fusion

1. function max fusion(sem1,sem2)
2. if sem1.color == sem2.color then semfusion.color ← sem1.color
3. semfusion.confidence ← (sem1.confidence + sem2.confidence)/2
4. else
5. if sem1.confidence > sem2.confidence then semfusion > sem1
6. else semfusion > sem2
7. semfusion.confidence ← semfusion.confidence × 0.9
8. return semfusion

4. Result and evaluation

In this section, to evaluate the effectiveness of the system, we tested the real-time performance and the accuracy of the system in dynamic environments using public TUM RGB-D dataset. The dataset has a ground truth of the camera trajectory, which is used to measure both relative and absolute drift. All experiments are performed on a computer with Intel I7 CPU, GTX 1080Ti GPU, and 16GB memory.

The system is based on ORB-SLAM2, so we compared Metrics Absolute Trajectory Error(ATE) that stands for the global quality of the trajectory. We evaluate the values of RMSE, Mean, Error, and Standard Deviation (S.D.) of the ATE in the result. And the improvement values are calculated as shown: improvement=(m – n)/m × 100%, where m is the value of ORB-SLAM2, n is the value of the system. The comparison results are shown in Table 1. The improvement of the system can reach 78% on average. Besides, the system is compared with previous methods to demonstrate the advantages of system. The results in comparison to previous SLAM are shown in Table 2. As we can see from the results, the system performs well in dynamic environments.

Table 1. Result of Absolute trajectory error (ATE)

| sequence           | ORB-SLAM2(cm) | The system(cm) | improvement |
|--------------------|---------------|----------------|-------------|
|                    | RMSE  | Mean  | S.D.  | RMSE  | Mean  | S.D.  | RMSE  | Mean  | S.D.  |
| fr3_walking_xyz    | 73.8  | 64.5  | 38.1  | 32.6  | 28.9  | 15.1  | 55%   | 55%   | 54%   |
| fr3_walking_half   | 76.8  | 69.9  | 31.7  | 10.9  | 5.6   | 9.3   | 85%   | 92%   | 71%   |
| fr3_walking_static | 39.3  | 36.1  | 15.7  | 2.1   | 1.4   | 1.6   | 94%   | 96%   | 90%   |
Real-time performance is an important indicator to evaluate SLAM system. The time required of each module of the system is compared with a work which uses Segnet to segment the image. The result is shown in Table 3. Because the real-time semantic segmentation network ENet is used, the real-time performance of the system is relatively improved. The average computation time per frame of the system is 60ms when the resolution of the image is 360×360, which is less than SLAM system adopting other semantic segmentation network. Compared to the system, the work [2] requires 85ms per frame. Compared with no real-time dynamic SLAM system, the system has considerable advantages in real-time performance.

### Table 3. Time analysis (image resolution: 360*360)

| Module                        | ENet | Segnet |
|-------------------------------|------|--------|
| ORB extraction                | 9.4ms| 9.1ms  |
| Moving consistency check      | 14.9ms| 16ms  |
| Semantic segmentation         | 37ms | 60ms   |

### 5. Conclusion

In this paper, we modify a real-time SLAM system that can handle dynamic environments. The system combines a real-time semantic segmentation network ENet with moving consistency check to handle dynamic objects in scene, then the feature points located in dynamic regions will be filtered out to improve the accuracy of the system. Besides, we build a semantic octo-tree map in the map creation thread. Meanwhile, this thread combines the semantic information of different viewpoints to achieve multi-view consistent semantic segmentation. The real-time performance and accuracy of the system are improved considerably, especially the real-time performance. Besides, we get a map with more accurate semantic information.

### Acknowledgment

This work was supported by National Key R&D Program of China, Provincial Funding for National Key R&D Program (Task), Support by Self-Planned Task (NO. SKLRS201813B) of State Key Laboratory of Robotics and System (HIT).

### References

[1] Y. C, Liu Z, Liu X, Xie F, Yang Y and Wei Q 2018 Ds-slam: a semantic visual slam towards dynamic environments *IEEE/RSJ International Conference on Intelligent Robots and Systems*

[2] Bakkay M, Majdi A and Zagrouba E 2015 Dense 3D SLAM in dynamic scenes using kinect *7th Iberian Conference IbPRIA, Pattern Recognition and Image Analysis, Lecture Notes in Computer Science*

[3] Ochs P, Malik J and Brox T 2014 Segmentation of moving objects by long term video analysis *IEEE Transactions on Pattern Analysis & Machine Intelligence* 36(6) 1187-1200

[4] Klappstein J, Vaudrey T, Rabe C, Wedel A and Klette R 2008 Moving object segmentation using optical flow and depth information *Lecture Notes in Computer Science* 5414 611-623

[5] Wangsiripitak S and Murray D 2009 Avoiding moving outliers in visual SLAM by tracking moving objects *IEEE International Conference on Robotics & Automation*
[6] Galindo C, Saffiotti A, Coradeschi S, Buschka P, Fernandez-Madrigal J and Gonzalez J 2005 Multi-hierarchical semantic maps for mobile robotics *IEEE/RSJ International Conference on Intelligent Robots & Systems*

[7] McCormac J, Handa A, Davison A and Leutenegger S 2016 Semanticfusion: dense 3d semantic mapping with convolutional neural networks.

[8] Li X and Belaroussi R 2016 Semi-dense 3d semantic mapping from monocular slam

[9] Paszke A, Chaurasia A, Kim S and Culurciello E 2016 Enet: a deep neural network architecture for real-time semantic segmentation

[10] Zhang X 2018 Real-time voxel based 3D semantic mapping with a hand held RGB-D camera

[11] Jaimez M, Kerl C, Gonzalez-Jimenez J and Cremers D 2017 Fast odometry and scene flow from RGB-D cameras based on geometric clustering *IEEE International Conference on Robotics & Automation.*

[12] Whelan T, Salas-Moreno R, Glocker B, Davison A and Leutenegger S 2016 Elasticfusion. *International Journal of Robotics Research*

[13] Rünz M and Agapito L 2017 Co-fusion: real-time segmentation, tracking and fusion of multiple objects

[14] Sun Y, Ming L and Meng Q 2017 Improving rgb-d slam in dynamic environments: a motion removal approach *Robotics & Autonomous Systems* 89*(Complete)* 110–122