2023 IEEE International Conference on
Integrated Circuits, Technologies and Applications

2023 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA) | 979-8-3503-4428-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTA60488.2023.10364289

MSeg-SLAM: A Semantic Visual SLAM System for Dynamic Scenes

Peijun Li, Weiyi Zhang, Zeyu Wan, Chun Zhang*

School of Integrated Circuits, Tsinghua University,
Beijing, 100084, China

***Abstract*—Visual Simultaneous Localization and Mapping (VSLAM) technique is crucial for intelligent mobile devices to acquire their current position and pose information. The traditional VSLAM system assumes that all objects in the environment are static by default. But in the physical world, some dynamic objects like humans are unavoidable, which imposes a huge burden on the system. This work proposes a semantic VSLAM system named MSeg-SLAM, which combines the VSLAM system and MSeg semantic segmentation network to reduce the impact of dynamic objects. At the same time, we generate semantic octree maps to optimize the storage space occupied by the dense point cloud map. Compared with the original ORB-SLAM2 system, the absolute trajectory error (ATE) can be reduced by over 93% in high dynamic scenes of the public TUM dataset. The results indicate our system has strong robustness and stability both in datasets and real-world practical applications.**

Keywords—Visual SLAM, semantic segmentation, octree, robotics

I. Introduction

As the key technology in intelligent mobile devices, such as robotics, self-driving cars, and drones, Visual Simultaneous Localization and Mapping (VSLAM) help them locate their current position and perceive the surrounding environments, thus providing information for high-level tasks such as navigation, trajectory planning, and human-computer interaction. The basic assumption of the VSLAM system is all objects in the current environment are static. However, there are lots of dynamic objects like humans and animals in real scenarios which could interfere with the system and even cause it to collapse. In addition, the dense point cloud map constructed by the ORB-SLAM2 system occupies a large amount of storage space and has no practical significance for mobile devices.

In this work, we propose MSeg-SLAM, a semantic VSLAM system for dynamic environments. The system can not only remove dynamic objects but also generate an efficient octree map using semantic information. The results show that the ATE of MSeg-SLAM can achieve up to 98.25% improvements in high dynamic scenes compared with the ORB-SLAM2 system.

II. Proposed Architecture

A. Framework of MSeg-SLAM

The MSeg-SLAM system is developed based on the traditional ORB-SLAM2 system [1], which has outstanding localization accuracy and pose estimation performance in static environments. Fig. 1. illustrates the overall framework of the MSeg-SLAM system. The whole system consists of five main threads: tracking, semantic segmentation, loop mapping, loop closing, and map creation.

Considering semantic segmentation is a time-consuming process, we designed a parallel system that carries out ORB extraction and semantic segmentation simultaneously to deal with each input RGB image captured by the camera, as shown in Fig. 2. The next loop mapping and loop closing threads are similar to the ORB-SLAM2 system. The map creation thread will generate the final semantic octree map according to the existing semantic and keyframe information.

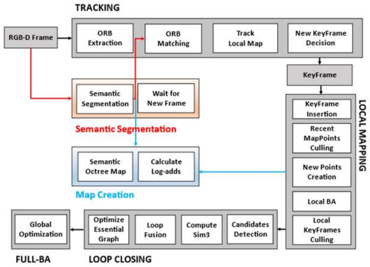

Fig. 1. The framework of MSeg-SLAM

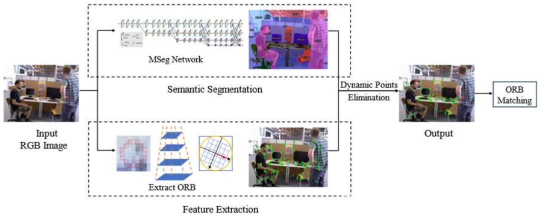

Fig. 2. The parallel tracking thread of MSeg-SLAM

B. Semantic Segmentation Network

The pixel-level segmentation networks used in related work only focus on pursuing high accuracy on a single dataset. The segmentation performance will dramatically decrease when the settings change, which poses challenges in practical applications. Lambert et al. presented a composite semantic segmentation dataset with more than 80,000 images from different domains and proposed the MSeg model, which was trained on the dataset and exhibited competitive performance and strong generalization ability [2]. The MSeg network can recognize 194 categories of common objects and deal with single-scale inference in real-time, meeting the latency requirements of VSLAM systems.

When obtaining the semantic results returned by MSeg, the system will remove the dynamic feature points in the current frame. To make the process more precise, each feature point is extended by i pixels outward to check for potential objects, as shown in Algorithm 1.

Algorithm1: Dynamic Points Elimination Algorithm

```
Input: The set of all feature points, AllFPs; Dynamic objects returned from MSeg, DynamicObjects; Expansion value, i;
Output: The set of static feature points, StaticFPs;
for AllFPs do
    Calculate current feature point coordinates $(x,y)$
    Calculate the expansion area $E$ by $i$
    if DynamicObjects not in $E$:
        add current feature point to StaticFPs
    end if
end for
```

C. Semantic Octree Map

The ORB-SLAM2 system will append all recognized feature points to the dense point cloud map. Thus, the storage requirements will greatly increase when there is a vast amount of feature points in the environment. To solve this problem, we introduced octree types to store the map. An octree can divide a node into eight sub-nodes. Only when all the sub-nodes of a node are occupied or not occupied, will the node not be further divided, thereby effectively compressing storage space.

Assuming $x \in [0,1]$ represents the probability of a voxel being occupied, $y \in R$ is the Log-adds of the probability. The logit transformation and its inverse transformation can be calculated as (1) and (2):

$$y = logit(x) = log\left(\frac{1}{1-x}\right) \tag{1}$$

$$x = logit^{-1}(y) = \frac{\exp(y)}{\exp(y)+1} \tag{2}$$

979-8-3503-4428-8/23/$31.00 © 2023 IEEE

Let Z_t represent the results of a voxel node n, from the start time to time t, the Log-adds can be denoted as $L(n|Z_{1:t})$. The Log-adds at time $t+1$ can be written as:

$$L(n|Z_{1:t+1}) = L(n|Z_{1:t-1}) + L(n|Z_t) \tag{3}$$

If the value of the Log-adds exceeds a pre-set threshold, it is considered that the voxel is occupied. Meanwhile, to achieve a succinct semantic octree map, we divided 194 main classes into 25 categories for display.

III. Experiment Results

The open-source TUM dataset is often used to test VSLAM algorithms [3]. The sequences are named by the action of the human body and the movement method of the camera. Based on the actions of individuals within the sequences, it can be divided into high dynamic sequences (with two individuals walking and exhibiting noticeable motion changes) and low dynamic sequences (with two individuals sitting and displaying only slight limb movements). The camera is moved in four different ways: along the x-y-z axes, on a hemisphere, along roll-pitch-yaw axes, and almost stationary. All experiments were conducted on a computer platform consisting of an Intel Core i7 CPU, 32 GB of memory, Nvidia RTX 3060 GPU, Ubuntu 18.04 operating system, and Intel RealSense D457 camera.

Table I presents a comparison of both the translational (T) and rotational (R) components of the relative pose error (RPE) between the two systems. The accuracy of the MSeg-SLAM system is enhanced by about 90% compared to ORB-SLAM2 in high dynamic sequences, significantly improving drift during system operation. However, in low-dynamic scenarios, humans only exhibit slight limb changes, and the presence of feature points on their bodies assists in localization, resulting in less noticeable improvement. It can be seen in Table II that the improvements of the absolute trajectory error (ATE) in the MSeg-SLAM system are all above 93% in four walking sequences, which is consistent with the RPE results. It is noted that in the w_xyz sequence, the increase in RMSE can reach up to 98.25%.

To further evaluate the robustness and performance of our system, we also compare it with state-of-the-art semantic VSLAM systems. The results demonstrate that the MSeg-SLAM system can achieve relatively higher accuracy in most cases.

Table III shows a comparison of storage space among different map forms in highly dynamic scenes. Compared to traditional dense point cloud map types, octree maps can be compressed to around 1% of the original storage space, which is friendly for resource-constrained edge devices to save map data.

TABLE I. Results of Metric Translational(T) and Rotational(R) Drift (RPE)

	Type	Sequences				
		w_xyz	**w_half**	**w_rpy**	**w_static**	**s_static**
ORB-SLAM2 [1]	T	0.4131	0.3352	0.3424	0.1725	0.0090
	R	7.8706	6.8016	6.4672	2.9782	0.3004
MSeg-SLAM (Ours)	T	**0.0194**	**0.0246**	**0.0405**	**0.0102**	**0.0073**
	R	**0.6001**	**0.8695**	**1.0251**	**0.2947**	**0.2833**

TABLE II. Results of Absolute Trajectory Error (ATE)

	ORB-SLAM2 [1]	**DS-SLAM [4]**	**PSPNet-SLAM [5]**	**DRSO-SLAM [6]**	**MSeg-SLAM (Ours)**
w_xyz	0.8391	0.0247	0.0156	0.0158	**0.0147**
w_half	0.4709	0.0303	0.0256	0.0268	**0.0234**
w_rpy	0.6354	0.4442	0.0334	0.0751	**0.0327**
w_static	0.3068	0.0081	**0.0073**	0.0111	0.0082

TABLE III. Storage Space Comparison

	Semantic Dense Point Cloud Map	**Semantic Octree Map**
w_xyz	10.3 MB	102.5 KB
w_half	31.2 MB	360.7 KB
w_rpy	60.0 MB	543.2 KB
w_static	6.4 MB	72.6 KB

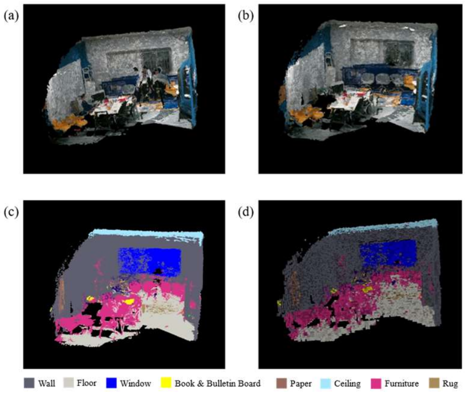

Fig. 3. Different map types in an office setting

We also tested our system in a real-world office setting, with one person sitting and the other sitting and walking around in the room. As shown in Fig. 3, human shadows are presented on the map at the beginning. The MSeg-SLAM system will recognize and filter out those dynamic objects and only preserve static point clouds within the room. After converting them to semantic labels, we can obtain a map with efficient semantic classification information. The storage space of the semantic dense point cloud map and the octree map is 99.8 MB and 697.4 KB, respectively.

IV. Conclusion

In this work, we propose the MSeg-SLAM system to improve the generalization ability and eliminate the impact of dynamic and irrelevant objects. To improve the storage efficiency of the dense point cloud map, we leverage the octree format to save the map, which could only need 1% storage space of the original map. All results on datasets and real scenes show the robustness and usefulness of our system. The MSeg-SLAM system may open new avenues for autonomous devices to run in a real, complex, and multi-interference environment.

References

[1] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," in *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255-1262.

[2] J. Lambert, Z. Liu, O. Sener, et al., "MSeg: A Composite Dataset for Multi-Domain Semantic Segmentation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 796-810.

[3] J. Sturm, N. Engelhard, F. Endres, et al., "A benchmark for the evaluation of RGB-D SLAM systems," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 573-580.

[4] C. Yu, Z. Liu, X. Liu, et al., "DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1168-1174.

[5] S. Han and Z. Xi, "Dynamic Scene Semantics SLAM Based on Semantic Segmentation," in *IEEE Access*, vol. 8, pp. 43563-43570.

[6] N. Yu, M. Gan, H. Yu, et al., "DRSO-SLAM: A Dynamic RGB-D SLAM Algorithm for Indoor Dynamic Scenes," in *2021 33rd Chinese Control and Decision Conference*, pp. 1052-1058.