RGB-D SLAM based on semantic information and geometric constraints in indoor dynamic scenes

Tao Liu, Hailong Zhao, Yiqun Liu* and Xuanxia Fan

School of Automotive Engineering, Harbin Institute of Technology, Weihai, 264209, PR China

*Corresponding author: Yiqun Liu (e-mail: lyq.new@163.com)

Abstract. In order to solve the shortcomings of traditional simultaneous localization and mapping in dynamic environment, which is interfered by moving objects, resulting in low accuracy and poor robustness, a visual simultaneous localization and mapping algorithm combining semantic information for motion detection was proposed. First, the SegNet deep neural network is used to extract the semantic information of the environment, and the prior knowledge is used to determine the static attribute objects and dynamic attribute objects. In the motion detection module, the feature points on the dynamic attribute objects are used to perform motion detection using geometric constraint relationships. Then the building module uses semantic information to build a semantic octo-tree map. In order to analyse the effect of motion detection, a control experiment with a motion detection module removed was set up. Finally, experiments were conducted using TUM datasets, and the experimental results of the two schemes were compared and analysed.

1. Introduction

Most of the current SLAM (Simultaneous Localization and Mapping) algorithms, such as RGBD-SLAM-V2 [1], RTAB-MAP [2], ORB-SLAM [3] are based on the assumption of static environment, do not consider moving objects, mainly use RANSAC (Random Sampling Consistency) [4] algorithm to reduce the impact of outliers. This method can easily introduce errors in a highly dynamic environment, causing serious drift in pose estimation.

At present, motion detection mainly uses traditional methods, but these methods are more restrictive. Both the background difference method [5] and the inter-frame difference method [6] can only be used for scenes with a fixed background; the optical flow method [7] is too complicated to calculate, and the calculation is prone to large errors; the geometric constraint method requires accurate pose, usually still use RANSAC method, some of them cannot deal with the movement under degenerate conditions.

In recent years, with the rapid development of deep learning in the field of image processing, many scholars have merged the semantic information of image into the SLAM algorithm to improve the accuracy and robustness of the SLAM algorithm in dynamic scenes. DS-SLAM [8] combined semantic information and epipolar constraints to eliminate the influence of motion feature points on pose estimation and establish a semantic octo-tree map with humans removed. [9] proposed a visual SLAM system based on ORB-SLAM, which adds motion segmentation methods to make it robust in the dynamic environment of monocular, binocular and RGB-D cameras, and can be created a static reusable map.

Combining with the current research results of visual SLAM, motion detection, and semantic segmentation, this paper based on ORB-SLAM2 researched the RGB-D visual semantic SLAM system
of indoor dynamic scenes. This paper combines semantic information and reprojection error to detect motion points, and compares and analyses the scheme using only semantic information for motion detection.

2. Algorithm framework and implementation

2.1. Algorithm framework

In order to improve the robustness and accuracy of the visual SLAM system in an indoor dynamic environment, the algorithm in this paper (named SEM-MD) is improved on the basis of ORB-SLAM2, adding a semantic segmentation module, a motion detection module and a semantic mapping module. The main modules are shown in Figure 1. The semantic segmentation module uses SegNet [10] to segment the image to obtain a semantic image. Then, according to the a priori knowledge and the label value, the feature points are divided into two types: static feature points and dynamic attribute feature points. In this paper, the feature points of the human body region on the image are regarded as dynamic attribute feature points. In the motion detection module, the transformation matrix $T$ is calculated first. Then the reprojection error of each dynamic attribute point is calculated according to the matrix $T$, and the motion point is detected according to the size of the error and the motion point is eliminated. The semantic mapping module uses key frames, depth images and semantic color images to build a semantic octo-tree map. In order to compare and analyse the effect of the motion detection module, this paper sets up a control scheme (named SEM). The SEM removes the motion detection module on the basis of SEM-MD, that is, the tracking thread directly removes the dynamic attribute points after obtaining the semantic information, and only uses the static feature points for tracking to realize the motion detection function.

2.2. Semantic information extraction

SLAM systems need to process data in real time, so we cannot sacrifice real-time performance in pursuit of high accuracy for segmentation. In order to balance the accuracy and real-time, this paper uses SegNet as the semantic segmentation network of this paper, and uses the PASCAL VOC [11] dataset for training.

![Figure 2. SegNet network structure.](image)
data dimension through the Upsampling layer, and finally outputs the maximum value of different classifications through the Softmax layer.

2.3. Transformation matrix solution

In order to obtain the transformation matrix between the two frames of RGB-D images, first use RANSAC to estimate the essential matrix $E$, and decompose the rotation matrix $R$ from $E$. Then use the two sets of matched 3D feature points estimated by $E$ to construct the ICP problem. The SVD method can be used to solve ICP problems. First calculate the centroid of two sets of points:

$$p = \frac{1}{n} \sum_{i=1}^{n} (p_i), \quad p' = \frac{1}{n} \sum_{i=1}^{n} (p'_i)$$ (1)

The feature point $p$ belongs to the current image, and the feature point $p'$ belongs to the reference image. Use the rotation matrix $R$ that has been solved before to calculate the translation matrix $t$ with scale information:

$$t = p - Rp'$$ (2)

The transformation matrix $T$ composed of the rotation matrix $R$ and the translation matrix $t$ obtained above can be represented by equation (3), and its Lie algebra is expressed as $\xi$.

$$T = \begin{bmatrix} R & t \\ 0^T & 1 \end{bmatrix}$$ (3)

2.4. Motion points detection

After the transformation $T$ is obtained, the dynamic attribute feature points are processed next. Suppose that the coordinate of a dynamic attribute feature point $P_d$ is $P_d=[X, Y, Z]^T$. The coordinate of this point in the reference frame image is $P_d_1=[u_1, v_1]^T$, and the depth is $s$. The projection coordinate of $P_d$ in the current image can be calculated according to the transformation matrix $T$ between the two images $P_d_2=[u_2', v_2']^T$, the specific expression is:

$$P_d_2 = \frac{1}{s} K \exp(\xi \cdot \cdot \cdot) P_d_1$$ (4)

The matching point $P_d_2'$ corresponding to $P_d_1$, the pixel coordinate is expressed as $P_d_2'=[u_2', v_2']^T$, which is regarded as the projection of the same 3D point $P_d$, and the reprojection between the projection point $P_d_2$ and the matching point $P_d_2'$ The error $e$ is calculated as follows:

$$e = P_d_2' - \frac{1}{s} K \exp(\xi \cdot \cdot \cdot) P_d = [\Delta u, \Delta v]$$ (5)

The threshold for determining whether the dynamic attribute point is moving or static directly adopts the 2 DOF threshold 5.99 and the 3 DOF threshold 7.81 used in ORB-SLAM2. The calculated projection error value of each pair of matching points is compared with the threshold. The point that are smaller than the threshold is judged to be static point and will be involved in subsequent calculations. The point that are larger than the threshold is moving point and directly eliminated.

2.5. Semantic octo-tree map creation

The semantic mapping module processes the key frames selected by the visual odometer. First calculate the 3D coordinate of each pixel of the current image using the depth value, pixel coordinate and camera internal parameters of the depth image according to equation (6). Then add the colors on the semantic color map to the corresponding 3D points to form a preliminary color point cloud.
\[ z = d, \quad x = \frac{z(y - c_y)}{f_x}, \quad y = \frac{z(u - c_x)}{f_y} \]  

Where \( d \) is the depth value of the pixel, \( (x, y, z) \) is the 3D coordinate of the 3D point, \( (u, v) \) is the pixel coordinate of the projection point, \( (f_x, f_y) \) is the camera focal length, and \( (c_x, c_y) \) is the optical projection center coordinate. Then, perform voxel filtering on the point cloud, use the center of gravity of the points in the voxel to approximate display other points in the voxel, implement downsampling, filter out redundant information, and obtain the filtered point cloud after processing.

Then perform statistical filtering on the point cloud, and for each point, calculate its average distance to all neighboring points. Points whose average distance is outside the standard range (defined by the global distance average and variance) can be defined as outliers and removed from the data. Finally, these point clouds are converted into messages in ROS, and the point cloud map is converted into a semantic octo-tree map using the Octomap library.

3. Experiments and analysis

This paper uses the TUM dataset for experiments. The laptop running the algorithm is configured as: Intel I5-8300H CPU, clocked at 2.3GHz, memory is 16G, graphics card is NVIDIA GeForce GTX 1060 6G video memory, system is Ubuntu16.04.

During the experiment, each sequence was used as input to the SEM, SEM-MD, and ORB-SLAM2 five times, that is, each algorithm can obtain 5 trajectory data files per sequence. In this paper, the translation part of the ATE (Absolute Trajectory Error) is used as an evaluation index to calculate the RMSE (Root Mean Squared Error) of ATE of all trajectory files, and the trajectory is aligned using SE(3). The median of five trajectory data evaluation indicators is selected as the final result for comparison. The translation results of ATE are shown in Table 1.

| Sequence    | ORB-SLAM2 RMSE(m/f) | SEM RMSE(m/f) | Imp. (%) | SEM-MD RMSE(m/f) | Imp. (%) |
|-------------|---------------------|---------------|----------|------------------|---------|
| Low dynamic |                     |               |          |                  |         |
| sitting_static | 0.00777             | 0.00707       | 8.99     | 0.00687          | 11.59   |
| sitting_xyz | 0.00892             | 0.01128       | -26.44   | 0.00976          | -9.44   |
| sitting_half | 0.03108             | 0.02811       | 9.56     | 0.01796          | 42.18   |
| sitting_rpy | 0.02019             | 0.02455       | -21.60   | 0.02068          | -2.45   |
| Mean        | 0.01699             | 0.01775       | -4.48    | 0.01381          | 18.67   |
| High dynamic |                     |               |          |                  |         |
| walking_static | 0.27730             | 0.00784       | 97.17    | 0.00735          | 97.35   |
| walking_xyz | 0.58484             | 0.04203       | 92.81    | 0.02350          | 95.98   |
| walking_half | 0.43994             | 0.03645       | 91.71    | 0.02508          | 94.29   |
| walking_rpy | 0.72585             | 0.39706       | 45.29    | 0.35563          | 51.00   |
| Mean        | 0.50698             | 0.12084       | 76.16    | 0.10289          | 79.70   |

From the data in Table 1, it can be seen that, in order to enable the system to run robustly in a highly dynamic scene, both the SEM and the SEM-MD deal with dynamic attribute objects. In highly dynamic sequences, the feature points of the moving human body area are basically eliminated. It can be seen from Figure. 3 that the motion points of the SEM-MD have been eliminated (the SEM is similar to the SEM-MD), and there are still some motion points on ORB-SLAM2, which seriously affect the pose estimation results. Therefore, the evaluation indicators of the two schemes have been significantly improved relative to ORB-SLAM2.

In the low-dynamic sequence, overall, the effect of the SEM is slightly worse than that of ORB-SLAM2, and the SEM-MD is slightly better than that of ORB-SLAM2. This is mainly because the RANSAC algorithm used by ORB-SLAM2 can handle a small number of motion points in a low dynamic environment. The SEM directly removes the feature points on the dynamic attribute objects, which may cause a part of the static points to be removed, thereby negatively affecting the pose estimation. In the SEM-MD, a motion detection module is added to detect dynamic attribute points, as
far as possible, the points of real motion are eliminated, and the static points are retained. It can be seen from Figure 4 that the SEM-MD can retain the static point on the dynamic attribute object in a low dynamic environment. Therefore, the SEM-MD can achieve better results than the SEM in both low dynamic environment and high dynamic environment.

Figure 3. High dynamic sequence walking_static image processing results: semantic color image (a), ORB-SLAM2 (b) and SEM-MD (c) map points obtained on the same image.

Figure 4. Low dynamic sequence sitting_static image processing results: semantic color image (a), SEM (b) and SEM-MD (c) map points obtained on the same image.

Figure 5 shows the ORB-SLAM2, SEM and SEM-MD error change diagrams and error distribution diagrams of highly dynamic sequence walking_halfsphere. The error change chart intuitively reflects the error change of the three algorithms and the comparison of the error size. From the figure, it can be seen that the error value of the SEM and the SEM-MD is much smaller than the error value of ORB-SLAM2. At the same time, from the error distribution diagram, it can be seen that the concentration of the error distribution of ORB-SLAM2, the SEM and the SEM-MD increases in turn.

Figure 5. Error change graph (a) and error distribution graph (b).
In the semantic mapping module, there are about 200,000 effective 3D points recovered from the depth map, and about 4,000 after voxel filtering, which reduces the number of point clouds by 50 times. The semantic map created after only voxel filtering is shown in Figure 6 (a). It can be seen that there are isolated noises on the map. Figure 6 (b) is an image created after voxel filtering and statistical filtering. It can be seen that the isolated points in Figure 6 (a) are basically filtered.

![Figure 6. (a) is voxel filtering result, (b) is voxel filtering combined with statistical filtering result.](image)

The average time for the odometer to process each frame is 56.16ms, and the semantic segmentation takes 46.75ms. For semantic mapping, the average time for restoring 3D points is 25.20ms, the average time for voxel filtering is 7.34ms, the average time for statistical filtering is 27.05ms, and the release time is 0.15ms. Therefore, the processing speed of the visual odometer part of the algorithm in this paper can reach 17FPS, and the processing speed of the semantic mapping part is 16FPS, which can run in real time.

4. Conclusions
In order to improve the robustness and accuracy of the SLAM system in dynamic scenes, this paper proposes a complete real-time robust semantic SLAM system. In this paper, the SegNet network is used to obtain semantic information in real time, and reprojection errors are used to detect motion of dynamic attribute feature points, which improves the accuracy and robustness of the system in dynamic scenarios. TUM data set experiments prove that the accuracy and robustness of SEM-MD in the dynamic environment is better than ORB-SLAM2, which can effectively reduce the impact of dynamic objects on pose estimation. At the same time, by comparing the experimental results of the SEM-MD and the SEM, it is verified that the combination of semantic information and reprojection error can retain the static point, which is better than the simple pose estimation results obtained by using semantic information for motion detection. The paper finally established a semantic octo-tree map that can be used for navigation.

Acknowledgments
This work was supported by National Natural Science Foundation of China (no. 51705097), Postdoctoral Science Foundation of China (no. 2017M621258), Self-Planned Task of State Key Laboratory of Robotics and System (HIT) (no. SKLRS201803C), and the Fundamental Research Funds for the Central Universities (no. HIT.NSRIF.2020089).

References
[1] Endres F, Hess J, Sturm J, Cremers D and Burgard W 2014 J. IEEE Transactions on Robotics 30 177-187
[2] Labbé M and Michaud F 2014 Int. Conf. on Intelligent Robots and Systems 2661-2666
[3] Mur-Artal R, Montiel JMM and Tardos JD 2015 IEEE Transactions on Robotics 31 1147-1163
[4] Kitt B, Geiger A, Lategahn H 2010 IEEE Intelligent Vehicles Symp. 486-492
[5] Xu Y, Zhou C, Xu S and Xing C 2014 IEEE Workshop on Electronics, Computer and Applications 518-521
[6] Gao F and Lu Y 2019 Int. Conf. on Artificial Intelligence and Advanced Manufacturing 76-81
[7] Wang Z and Yang X 2018 IEEE 3rd Int. Conf. on Image, Vision and Computing 66-69
[8] Yu C, Liu Z, Liu X J, Xie F G, Yang Y, Wei Q and Fei Q 2018 *Int. Conf. on Intelligent Robots and Systems* 1168-1174

[9] Bescos B, Fácil J M, Civera J and Neira J 2018 *J. IEEE Robotics and Automation Letters* 3 4076-4083

[10] Badrinarayanan V, Kendall A and Cipolla R 2017 *J. IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 2481-2495

[11] Everingham M, Van L, Williams C K, Winn J and Zisserman A 2010 *J. Int. journal of computer vision* 88 303-338