Dynamic SLAM System Using Histogram-based Outlier Score to Improve Anomaly Detection

Fujun Pei
Faculty of Information Technology,
Beijing University of Technology
Beijing Key Laboratory of Computational Intelligence and Intelligent System
Beijing, China
pfj@bjut.edu.cn

Zhu Miao
Faculty of Information Technology,
Beijing University of Technology
Beijing Key Laboratory of Computational Intelligence and Intelligent System
Beijing, China
miaozhu84@emails.bjut.edu.cn

Jinghui Wang
Faculty of Information Technology,
Beijing University of Technology
Beijing Key Laboratory of Computational Intelligence and Intelligent System
Beijing, China
245387251@qq.com

2021 China Automation Congress (CAC) | 978-1-6654-2647-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/CAC53003.2021.9728124

***Abstract*— It is well known that the traditional visual SLAM systems have performed well in a relatively ideal environment, but for the dynamic environment, there are some problems in the instance segmentation algorithm such as incomplete segmentation or wrong segmentation which will result in outliers. Aiming at this problem, this paper proposed a dynamic SLAM system using Histogram-based Outlier Score (HBOS) to improve the reprojection residual constraint to increase the robustness in removing outliers. In this method, HBOS method is used to process the reprojection residual, and the feature association data obtained after the segmentation mask processing is further filtered to improve the robustness. This method can combine the semantic segmentation network with traditional geometric methods to solve the data association problem caused by the wrong information generated from the dynamic objects. To verify the effectiveness of the proposed method, the experiment is carried out on the public TUM data set and the results show that the proposed method improves the accuracy of SLAM positioning in the dynamic environment.**

Keywords—visual SLAM, Mask R-CNN, dynamic scenes, ORB-SLAM2

I. INTRODUCTION

Simultaneous Localization and Mapping (SLAM) mainly refers to that in an unknown environment, the robot carries the required sensors according to the functions to be completed. Then it moves autonomously and incrementally to get the map of the surrounding environment and estimate the trajectory of its pose. Today, many vision SLAM systems which can run under the robot operating system have been developed[1-5]. These systems make the robot have a good ability of pose estimation and mapping in the ideal static environment. However, according to the requirements of specific scenes, robots often need to plan their navigation path in the environment with dynamic targets or potential dynamic targets. The feature information extracted from dynamic targets often leads to the drift of robot pose estimation trajectory and the increase of accumulated error in the process of mapping, which affects the accuracy and robustness of the visual SLAM system.

When some dynamic objects appear in the indoor scene, usually passers-by walking or carrying things, they will interfere with the positioning accuracy of the indoor robot. The optimization of robot pose estimation in indoor dynamic scenes is discussed in this article, mainly in low-dynamic scenes where pedestrians are relatively sparse. According to the characteristics of the dynamic environment, some dynamic SLAM algorithms are developed under the premise of prior information of dynamic objects. They can keep good robustness and accuracy in the dynamic environment. At the same time, some corresponding 3D reconstruction algorithms in dynamic environment structure from motion (SFM) have been developed [6].

For a low dynamic environment that has a short time and fewer dynamic objects, traditional visual SLAM treats the feature points extracted from dynamic objects as outliers. Through various methods, it tries to distinguish the dynamic feature information from the static feature information. For example, some papers use better feature extraction methods, such as FAST, ORB, etc.[7], to obtain better feature descriptors. Some edge feature points of dynamic objects can be filtered in the following feature matching, which makes the system robust to a low dynamic environment. More algorithms use RANSAC[8] algorithm to filter dynamic outliers by statistical sampling calculation, which makes the algorithm robust to low-time dynamic interference significantly. However, these methods are not specifically for the dynamic environment, most of which take the static environment as the premise, which indirectly improves the robustness of the visual SLAM system dynamic environment.

With the rise of artificial intelligence technology, image semantic segmentation technology in deep learning is gradually applied to dynamic visual SLAM. For example, different semantic segmentation methods are used as the front-end assistance of visual SLAM to eliminate the dynamic target feature information in the image, and then the static environment is simulated for estimation [9, 10]. However, the semantic segmentation network still has instability and cannot recognize the motion state of the object separately So in the dynamic environment, the dynamic SLAM system combines the semantic target detection network and the traditional method to deal with the dynamic outliers. For example, combining the visual SLAM system ORB-SLAM2 with the PSP-Net semantic segmentation network, a PSP-Net SLAM system[10] is proposed. The system uses the optical flow and semantic segmentation to detect and eliminate dynamic points and realize the dynamic scene semantic SLAM. A relatively fast and accurate instance segmentation algorithm MASK R-CNN[11] is used in some dynamic SLAM algorithms to process dynamic objects. For example, in [12], after using MASK R-CNN, the feature points in the reference frame are reprojected to the current frame, and the vector difference between the feature points is used to set the threshold to filter the feature points. And based on the depth of the RGB-D camera, B. Bescos, and J. M. Facil, etc.[13] gets the threshold of experience depth difference through multi-scene indoor experiments regarding the accuracy and recall rate, and then filters on them. In [14], Y. Wang and S. Huang use the semantic network and polar constraint to filter outliers. Similarly, when the polar constraint is used, the threshold of the control range should be set to distinguish feature points.

Semantic segmentation algorithms are usually applied in the above methods. However, due to the instability of the semantic segmentation algorithm caused by the complexity and diversity of the actual environment, the traditional multi-view geometric method is used in conjunction with it. When the traditional multi-view geometry method based on rough pose is used to screen out abnormal points, a robust threshold with good effect needs to be set. When applied to complex real scenes, the robustness is poor and the experiment needs to be recalibrated. To solve this problem, a method to adjust the screening threshold of reprojection residuals adaptively according to the actual situation is discussed. From multiple perspectives such as threshold value and direction, this method can obtain more reliable data association information to optimize attitude estimation, which is mainly aimed at low dynamic scenes with relatively sparse pedestrians in indoor scenes. The positioning accuracy of indoor dynamic SLAM is improved by this system, which consists of semantic segmentation, lightweight tracking, HBOS-based reprojection method, and ORB-SLAM2 framework.

The research status is introduced in chapter 1. In chapter 2, the general research background of the problem is stated and the mathematical principles are explained. In chapter 3, we introduce our system and method from the semantic segmentation module, lightweight tracking module, and multi-view geometry module. In chapter 4, according to our improved algorithm, the indoor verification experiment is carried out and the effectiveness is proved. Finally, the algorithm and content in the paper are summarized.

II. BACKGROUND

A. Mathematical representation of SLAM

In a continuous period, the robot moves in an unknown environment and discretely records the position of the robot and the landmarks it passes during the movement. These landmarks are mainly obtained from the observation data of the robot's sensors. In the robot system, the algorithm is adopted to realizes its positioning and perception of the surrounding environment by obtaining its motion information and observation information.

The mathematical models are as follows：

$$\begin{cases} x_k = f(x_{k-1}, u_k, w_k) \\ z_{k,j} = h(y_j, x_k, v_{k,j}) \end{cases} \tag{1}$$

where $x_k$represents the position information of the robot at k time, u_k is the input of the robot's motion sensor, w_k is the noise, and f is the abstract function. The motion equation is that the robot can estimate its position under the influence of noise when it gets the input information of the sensor when the position information of the previous time is known.

In the observation equation, y_j is the jth landmark, $v_{k,j}$is the noise, and $z_{k,j}$ is the observation information of the kth landmark robot. It is generally assumed that the noise w_k, $v_{k,j}$ obey the Gaussian distribution of zero means.

In the observation equation, the error is expressed as:

$$e_{k,j} = z_{k,j} - h(y_i, x_k) \tag{2}$$

B. Problem Statement

In the dynamic scene, due to the interference of dynamic abnormal information, the sensor input information of the robot often contains error information, which will directly lead to the position estimation error. Therefore, eliminating the interference of dynamic information is the most direct way to improve positioning accuracy.

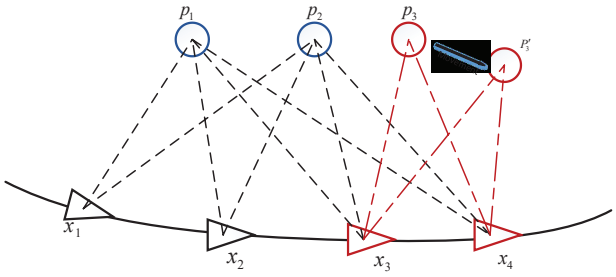


Fig. 1. Graph representation of SLAM in a dynamic environment

When the interference of dynamic object to observation is expressed as $d_{k,j}$, the error in the dynamic environment is expressed as:

$$e_{k,j} = z_{k,j} - h(y_i, x_k) - d_{k,j} \tag{3}$$

The cost function is adjusted accordingly due to the adjustment of the error. The optimization problem usually adopts the nonlinear least square method, and the optimization equation can be expressed as:

$$J(x) = \frac{1}{2}\sum_k^m \sum_j^n e_{k,j}(x)^T Q_{k,j}^{-1} e_{k,j}(x) \tag{4}$$

As shown in Fig. 1, x_k represents the pose of the robot, which is connected to form a trajectory. p_k represents the landmarks, where p_1, p_2 represents the static landmark point and the dynamic landmark point p_3 moves to p_3' .The movement of signpost point p_3 will lead to the incorrect estimation of position point x_3, x_4.

III. SYSTEM INTRODUCTION

Given the above problems, our system and method will be introduced.

As shown in Fig. 2, the system is mainly based on the monocular vision in ORB-SLAM2. After adding the MASK R-CNN instance segmentation module, it is combined with the multi-view geometric method based on the statistical histogram to further remove outliers.

Each frame information is preprocessed by MASK R-CNN to remove most of the dynamic abnormal points, and then the rough pose is obtained through the lightweight tracking module. After obtaining the pose, the reprojection module detects the position and pose, and the square statistics module processes the detection results to screen out the abnormal points. Finally, the processed results are sent to the network framework of ORB-SLAM2.

A. Instance segmentation module

To eliminate the dynamic characteristics of large targets in the dynamic environment, the system adopts the current method that performs better in the direction of detection and image segmentation.

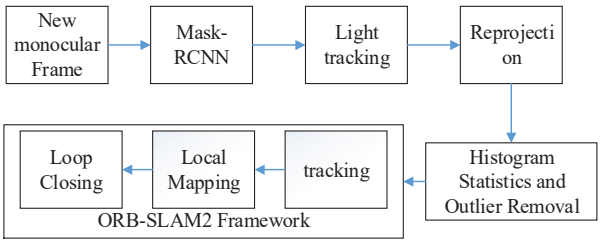


Fig. 2. System Framework

Mask R-CNN [3] is an instance segmentation algorithm, which can be used for "target detection", "target instance segmentation" and "target key point detection". Mask R-CNN is a very flexible framework, which can add different branches to complete different tasks, such as target classification, target detection, semantic segmentation, instance segmentation, human posture recognition, and so on.

B. Light motion estimation

After the instance segmentation processing, most of the influence factors have been removed in the input frame image, the remaining influencing factors include the incomplete edge contour of dynamic target segmentation and human hand objects. As shown in Fig. 3. If you want to get more accurate positioning results, this information needs further processing and filtering.

In ORB-SLAM2, there is a lightweight tracking mode algorithm, which can directly extract and match the feature points, and get the current rough pose information by minimizing the projection error. At the same time, this lightweight pose estimation module has the characteristics of low cost and fast calculation.

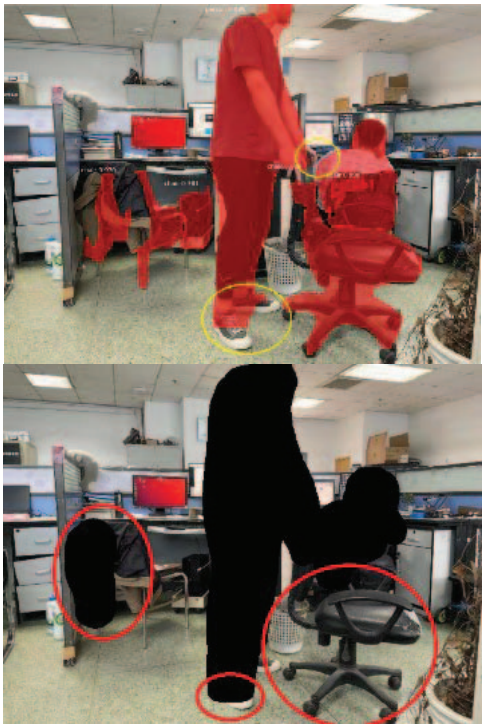


Fig. 3. Segmentation results of Mask R-CNN in an indoor dynamic environment. some of the yellow and red circles in the graph still have the problems of false segmentation and incomplete segmentation

C. Multi-angle geometric removal of dynamic points

The projection error between the two frames is obtained by projecting static and dynamic points from the previous phase plane onto the current phase plane with rough pose information.

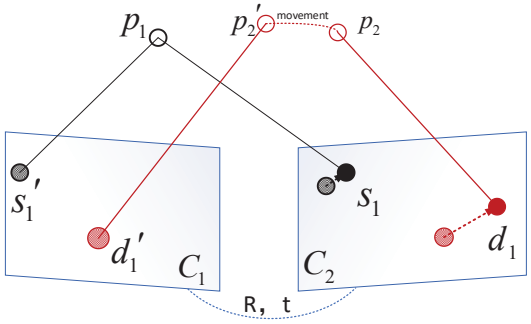


Fig. 4 Reprojection error representation of the dynamic scene.

In Fig. 4, Static and dynamic punctuation mappings are represented in the front and back frames of the camera. C1 and C2 represent the phase planes of the two frames before and after respectively, R and t denote the pose transformation relations between the two frames, namely rotation, and translation. p1 and p2 represent static and dynamic punctuation marks, respectively. The mapping of p1 and p2 in the phase plane of two frames is represented by s_1'-s_1 and d_1'-d_2. s_1''-s_1 and d_1''-d_2 represent the offset vectors of static point p1 and dynamic point p2 in the current frame and the offset can be expressed by a directional arrow line.

In the indoor dynamic environment, the visual distribution of the module length of the reprojection error of all feature points is shown in Fig. 5, in which the green connecting line is the line between the feature points projected from the previous reference frame and the corresponding feature points in the current frame. the reprojection offset error on the dynamic human body is large. After the mask processing, some outliers still be interfering with the positioning accuracy.

When the reprojection error information in the indoor environment is statistically distributed many times, the correct information in the reprojection information is often concentrated, and the less error information is outlier distribution. In addition, the change vector of feature points obtained by reprojection has two information for screening: direction and module length. The algorithm based on direct square statistics needs to count the two separately. Taking the module length as an example, the statistical process based on the histogram is shown in Fig. 6. The module length value is recorded as θ and the angle is recorded as the φ value.

Fig. 5 Comparison of reprojection errors the feature points. The reprojection error between two reference frames is projected onto the current frame and is represented by a connected thin line, where the length of the thin line indicates the size of the error.

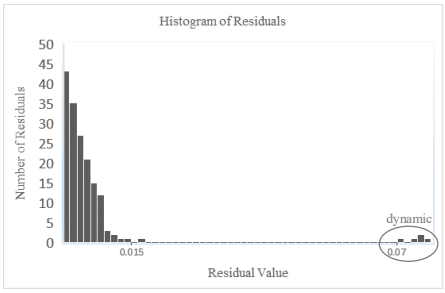


Fig. 6 The square distribution of reprojection residuals

As shown in Fig. 6, the re-projection residuals between two frames in the actual situation are represented by histograms. It can be seen that the moving process of dynamic points often causes these obvious abnormal outliers, which are distributed around 0.07m. However, the projection residuals of most static feature points are less than 0.015m. However, only setting a fixed threshold is difficult to adapt to the more complex actual operating environment. Therefore, the Histogram-based Outlier Score (HBOS) algorithm that removes outliers based on square statistics is used. The higher the score calculated by the algorithm, the more likely it is to become an outlier, which is characterized by high efficiency. The scoring model is as follows:

$$HBOS(p_i) = log(\frac{1}{hist(\varepsilon_i)} * \frac{1}{hist(\theta_i)}) \quad (5)$$

where *HBOS ()* is the score and $\mathrm{hist}_i()$ is the probability density of the part where the point is located, and ε is the residual threshold, θ is the angle residual value. Generally, it is more appropriate to set the interval of the statistical straight side of the threshold to 0.01. HBOS algorithm can score from multiple dimensions, and then add multiple scores to make a comprehensive score. Here, the angle and threshold are used as our scoring dimensions and the sum of the two scores is normalized into a comprehensive result. As shown in Equation 6.

$$\delta = \alpha(HBOS_{max}(P) - HBOS_{min}(P)) + HBOS_{min}(P) \quad (6)$$

In the formula, P is the sample set, and a is the variable that can be adjusted according to our needs. Choose between 0 and 1. The larger the variable is, the looser the screening range is. δ is the screening threshold of the score, and the sample points larger than δ will be discarded, which is generally set to 0.7 according to experience. HBOS score can greatly enhance the ability of threshold screening, which is no longer limited to the weak adaptability of hard threshold screening.

The result of matching is shown in Fig. 7.

IV. EXPERIMENTAL RESULT

In this part, the algorithm will be verified on the public data set of monocular vision, showing the experimental results obtained and comparing with some advanced algorithms in the current field.

A. TUM Dataset

The public Technical University of Munich (TUM) RGB-D data set includes color and depth images and accurate real ground data. It contains many scenes, and the data set used is a low dynamic scene with relatively few dynamic objects.

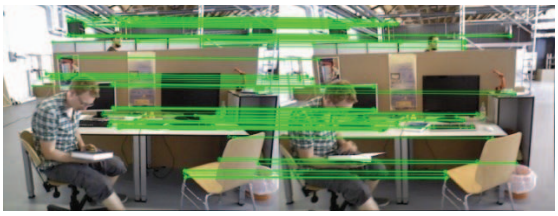


Fig. 7 Schematic diagram of final matching point

Among them, the data set of the dynamic scene mainly shows that two people are walking and communicating in the indoor environment, and the main interference comes from the people wearing plaid shirts and the moving parts that are not recognized and segmented, such as the partially segmented people and the books carried by people. The motion of the camera is (1) *xyz*: the camera mainly moves along the x-y-z axis (2) *rpy:* the camera uses rolling, pitch angle, roll angle rotation (3) static: the camera generally remains stationary.

According to custom, common abbreviations are used to denote the names of these data sets: such as fr, w, v in the sequence name represents Freiburg, walking, and validation. xyz and rpy indicate the movement mode of the camera.

B. Evaluation criterion

1) Absolute Trajectory Error (ATE): It is mainly used to directly compare the attitude error relationship between the estimated results and the reference trajectory, and calculate the statistics of the whole trajectory. The global consistency of trajectory is evaluated by this metric. And in the table, trajectory errors are represented by the abbreviation Traj.

2) Relative Pose Error（RPE）: Unlike ATE, this data is mainly used to compare local motion (attitude increment), that is, local drift, and can be used to calculate the drift of translation and rotation per meter.

3) Root Mean Squared Error (RMSE) of ATE: It is the square root of the ratio of the square of the deviation between the observed value and the true value and the observation times n. The square root error is very sensitive to a group of large or small errors in measurement, so the root means square error can well reflect the precision of measurement.

C. Experimental data

RGB-D data is chosen to sets from TUM, mainly Freiburg3_walking_xyz and Freiburg3_walking_xyz data sets. The main content is indoor common dynamic scenes, and the main dynamic objects are pedestrians. The method is evaluated according to the above indicators.

As shown in Table 1 and Table 2, compared with ORB-SLAM2 which is suitable for general scenes, our method has better robustness in the dynamic environment and can counteract most of the impact from pedestrians. Moreover, compared with DynaSLAM which also uses Mask R-CNN and other geometric methods, our method has a certain improvement. In Figure 8, the trajectory distributions of our SLAM and DynaSLAM algorithms are compared to the real trajectory. The error distribution is mainly distributed in a small interval, and the error distribution is better in our SLAM algorithm from the perspective of range and means square error. From the perspective of the method and the actual effect comparison, our method is more suitable for the indoor dynamic environment and has better improvement than the method in DynaSLAM in the case of monocular

operation. For complex and changeable indoor environments, the threshold can be adjusted adaptively according to the HBOS scoring mechanism, which is more robust.

All the experiments were performed on a computer equipped with an i7 CPU, 16GB RAM, and a GTX1060 graphics card. Table 3 shows the comparison of the average computation time between this algorithm and the DynaSLAM algorithm at different stages. Due to the high running time of the instance splitting algorithm, the hardware requirements of the real-time running algorithm are high.

TABLE I. RESULTS OF METRIC ROTATIONAL DRIFT (ATE)

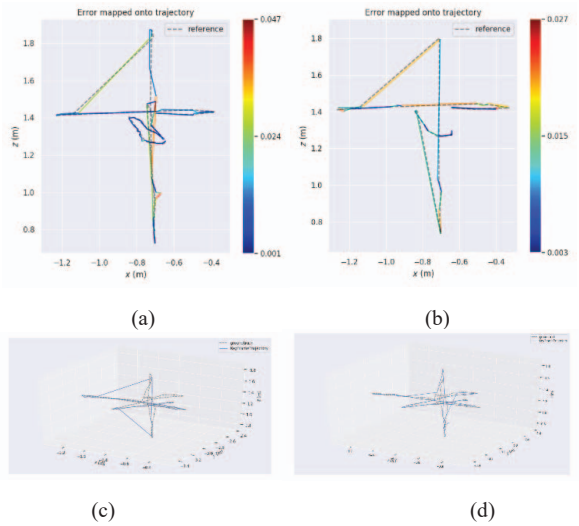
Sequences	ORB-SLAM2		DynaSLAM		Our SLAM	
	RMSE	*Traj (%)*	*RMSE*	*Traj (%)*	*RMSE*	*Traj (%)*
fr3/w/xyz	0.014	85.61	0.014	87.37	0.011	91.88
fr3/w/rpy	0.066	85.82	0.052	85.11	0.054	88.32

TABLE II. RESULTS OF METRIC ROTATIONAL DRIFT (RPE)

Sequences	ORB-SLAM2		DynaSLAM		Our SLAM	
	RMSE	*MEAN*	*RMSE*	*MEAN*	*RMSE*	*MEAN*
fr3/w/xyz	0.4124	0.311	0.0601	0.0514	0.0583	0.0459
fr3/w/rpy	0.4249	0.2825	0.0908	0.0795	0.0957	0.0743

TABLE III. COMPARISON OF THE AVERAGE RUNNING TIME AT DIFFERENT STAGES (UNIT: ms)

Sequences	Instance Segmentation [ms]	Motion Estimation [ms]	Dynamic Point Detection and removal [ms]
DynaSLAM	205.32	3.41	42.01
Our SLAM	202.73	3.38	43.28



(a) (b)

(c) (d)

Fig. 8 (a) and (c) are ATE and Track diagrams from Dyna-SLAM respectively. (b) and (d) are ATE and Track diagrams from our SLAM, respectively.

V. CONCLUSION

In this paper, a visual dynamic SLAM system that combines an instance segmentation module with an outliers removal module based on the HBOS method of reprojection errors is proposed. The HBOS algorithm based on a statistical histogram is developed to filter dynamic outliers with reprojection errors, which can compensate for the instability caused by the semantic segmentation algorithm. By setting the reprojection error, this method increases the robustness of the filtering method, combines the traditional geometric algorithm with the semantic segmentation algorithm better, and improves the positioning accuracy of this dynamic SLAM system. On indoor low dynamic TUM data sets, the system has improved over DynaSLAM in the case of monocular vision, because the failure of semantic segmentation is effectively dealt with and the robustness of the multi-view geometric algorithm is enhanced when outliers are removed. For the characteristics of the dynamic environment, this processing method is mainly applied in front-end data preprocessing and data association, which is suitable for dynamic scenes with richer static texture in the indoor environment.

REFERENCES

[1] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," IEEE Transactions on Robotics, vol. 31; 31, pp. 1147-1163, 2015-01-01; 2015-01-01 2015.

[2] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," IEEE Transactions on Robotics, vol. 33, pp. 1255-1262, 2017-01-01; 2017-01-01 2017.

[3] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual-Inertial and Multi-Map SLAM," 2020-01-01 2020.

[4] C. Forster, M. Pizzoli and D. Scaramuzza, "SVO: Fast semi-direct monocular visual odometry," in 2014 IEEE International Conference on Robotics and Automation (ICRA), 2014, pp. 15-22.

[5] J. Engel, T. Schops and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM," in Lecture Notes in Computer Science. vol. 8690, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. BERLIN: SPRINGER-VERLAG BERLIN, 2014, pp. 834-849.

[6] M. Runz, M. Buffier, and L. Agapito, "MaskFusion: Real-Time Recognition, Tracking and Reconstruction of Multiple Moving Objects," in International Symposium on Mixed and Augmented Reality, NEW YORK, 2018, pp. 10-20.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in 2011 International Conference on Computer Vision 2011, pp. 2564-2571.

[8] R. Raguram, J. M. Frahm and M. Pollefeys, "A Comparative Analysis of RANSAC Techniques Leading to Adaptive Real-Time Random Sample Consensus," in Lecture Notes in Computer Science. vol. 5303, D. Forsyth, P. Torr, and A. Zisserman, Eds. BERLIN: SPRINGER-VERLAG BERLIN, 2008, pp. 500-513.

[9] S. Muthu, R. Tennakoon, T. Rathnayake, R. Hoseinnezhad, D. Suter, and A. Bab-Hadiashar, "Motion Segmentation of RGB-D Sequences: Combining Semantic and Motion Information Using Statistical Inference," IEEE Transactions on Image Processing, vol. 29, pp. 5557-5570, 2020.

[10] S. Han and Z. Xi, "Dynamic Scene Semantics SLAM Based on Semantic Segmentation," IEEE Access, vol. 8, pp. 43563-43570, 2020.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988.

[12] J. Cheng, Z. Wang, H. Zhou, L. Li, and J. Yao, "DM-SLAM: A Feature-Based SLAM System for Rigid Dynamic Scenes," ISPRS International Journal of Geo-Information, vol. 9, p. 202, 2020-03-27 2020.

[13] B. Bescos, J. M. Facil, J. Civera, and J. Neira, "DynaSLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes," IEEE Robotics and Automation Letters, vol. 3, pp. 4076-4083, 2018.

[14] Y. Wang and S. Huang, "Motion segmentation based robust RGB-D SLAM," in Proceeding of the 11th World Congress on Intelligent Control and Automation Proceeding of the 11th World Congress on Intelligent Control and Automation, 2014, pp. 3122-3127.