RESEARCH ON SEMANTIC-ASSISTED SLAM IN COMPLEX DYNAMIC INDOOR ENVIRONMENT

C. Li [1], Z. Kang [1, *], J. Yang[1], F. Li[1], Y. Wang[1]

[1] Department of Remote Sensing and Geo-Information Engineering, School of Land Science and Technology, China University of Geosciences, Xueyuan Road, Beijing, 100083 CN – 724675560@qq.com, zzkang@cugb.edu.cn, jtyang66@126.com, foudar@163.com, wangyao@cugb.edu.cn

Commission IV, WG IV/5

KEY WORDS: Simultaneous localization and mapping, Semantic segmentation, Dynamic object recognition, Robot indoor navigation, Scene understanding

ABSTRACT:

Visual Simultaneous Localization and Mapping (SLAM) systems have been widely investigated in response to requirements, since the traditional positioning technology, such as Global Navigation Satellite System (GNSS), cannot accomplish tasks in restricted environments. However, traditional SLAM methods which are mostly based on point feature tracking, usually fail in harsh environments. Previous works have proven that insufficient feature points caused by missing textures, feature mismatches caused by too fast camera movements, and abrupt illumination changes will eventually cause state estimation to fail. And meanwhile, pedestrians are unavoidable, which introduces fake feature associations, thus violating the strict assumption that the unknown environment is static in SLAM. In order to ensure how our system copes with the huge challenges brought by these factors in a complex indoor environment, this paper proposes a semantic-assisted Visual Inertial Odometer (VIO) system towards low-textured scenes and highly dynamic environments. The trained U-net will be used to detect moving objects. Then all feature points in the dynamic object area need to be eliminated, so as to avoid moving objects to participate in the pose solution process and improve robustness in dynamic environments. Finally, the constraints of inertial measurement unit (IMU) are added for low-textured environments. To evaluate the performance of the proposed method, experiments were conducted on the EuRoC and TUM public dataset, and the results demonstrate that the performance of our approach is robust in complex indoor environments.

1. INTRODUCTION

As the basis of intelligent services for robots, autonomous navigation and positioning are the frontiers of the robotics research field. Its applications are increasingly embedded in people's daily lives including autonomous parking, advanced home services, medical services, path planning and obstacle avoidance. Traditional positioning technology, such as Global Navigation Satellite System (GNSS), enables accurate outdoor positioning. However, positioning and navigation tasks cannot be accomplished in GNSS-restricted environments, especially indoor environments due to complex indoor structures and the failure to receive satellite signals.

Over the past decades, visual Simultaneous Localization and Mapping (SLAM) systems have been widely investigated in response to requirements. It provides autonomous navigation and positioning in an unknown environment since rich image information obtained by a vision sensor can be used to estimate its own motion. In an ideal environment, many visual SLAM systems have achieved high location accuracy. MonoSLAM has been developed to generate 3D trajectory of unknown scene quickly through monocular camera (Andrew et al., 2007). However, the computing efficiency is limited by the size of the scene since extended Kalman filtering is used to optimize camera pose (Hauke et al., 2012). PTAM is the first system to propose parallel computing of tracking and mapping (Klein, Murray, 2007). It is worth mentioning that nonlinear optimization instead of filters is used as backend in this system. Based on ORB SLAM (Mur-Artal et al., 2015), ORB SLAM2 system (Mur-Artal, Tardós, 2017) was developed for monocular, stereo and RGB-D camera. The structure includes feature tracking of front-end modules, optimization of back-end modules, loop closing modules to identify known locations and mapping modules. It adopts Oriented FAST and Rotated BRIEF (ORB) as the feature point detection algorithm (Rublee et al., 2011). ORB SLAM2 is the representative of the feature point methods and the association between points is obtained by feature matching. The camera poses and map points position will be calculated by minimizing reprojection errors. LSD-SLAM (Engel, Cremers, 2014), as one of the direct method SLAM systems, optimizes the camera pose and 3D points coordinates by constructing a photometric error function, without the correspondence between points. It achieves a semi-dense scene reproduction on the CPU. The emergence of DSO (Engel et al., 2018) makes the direct method more mature, which uses fully direct method. It proposes photometric calibration to solve the effect of light on the direct method, Photometric parameters of uncalibrated cameras will be dynamically estimated. These make the direct method more robust.

In recent years, visual SLAM has gradually merged with various sensors (Heng et al., 2018). The integration of inertial measurement unit (IMU) measurements and visual SLAM can overcome the shortcomings of the arbitrary scale of the monocular system (Mur-Artal, Tardós, 2017). The sliding window strategy has been used in most Visual Inertial Odometer (VIO) systems (Dong-Si, Mourikis, 2012). VINS (Li et al., 2017) is a system based on semi-direct method, which uses optical flow for front-end tracking, but the back-end is still optimized for reprojection error.

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIII-B4-2020-353-2020

The above SLAM systems have achieved high location accuracy under ideal circumstances. However, traditional methods which are mostly based on point feature tracking, usually fail in harsh environments. Previous works have proven that insufficient feature points caused by missing textures, feature mismatches caused by too fast camera movements, and abrupt illumination changes will eventually cause state estimation to fail. And meanwhile, pedestrians are unavoidable, which introduces fake feature associations, thus violating the strict assumption that the unknown environment is static in SLAM. With the development of deep learning technology, researchers can accurately obtain the semantic information and geometric information of the scene at the same time, which is helpful for us to recognize the dynamic objects in the scene and eliminate their effect (Bescos et al., 2018).

In recent years, the semantic SLAM, which has the ability of scene understanding, assists mapping and positioning through accurate understanding of object targets in the environment (Bowman et al., 2017). The results of semantic segmentation and object detection can provide SLAM with higher-level information (Zhi, 2019). It provides an understanding of the surrounding environment. Semantic SLAM focuses on the incremental storage of image sequences and the update of semantic information (Civera, et al., 2011), as well as the fusion of multi-view semantic labels. The CNN-SLAM (Tateno et al., 2017) predicts both depth and labels at the same time, realizes label fusion and integrates the predicted depth information into SLAM, so that it can restore the true scale and obtain a semantically consistent map. At present, researchers mostly use the probabilistic model for data association, and correctly match the target object detected in the image to the existing 3D object of this category in the map data.

In order to ensure how our system copes with the huge challenges brought by these factors in a complex indoor environment, this paper proposes a semantic-assisted VIO system towards low-textured scenes and highly dynamic environments. Our approach is committed to the improvement of the system from the following two aspects: (1) The constraints of IMU are added for low-textured environments and (2) the geometric consistency of the system is improved by semantic information provided by deep learning technology in the highly dynamic environments.

2. METHODOLOGY

As shown in Figure 1, we propose a semantic-assisted VIO system, which contains the following two aspects:
(1) Firstly, feature extraction and semantic segmentation are performed. Semantic segmentation is used to identify dynamic objects and eliminate the effects of outliers, and finally obtain reliable feature points. (2) After eliminating outliers, the joint optimization of IMU measurement errors and reprojection errors ensures the system to acquire good pose calculation results.

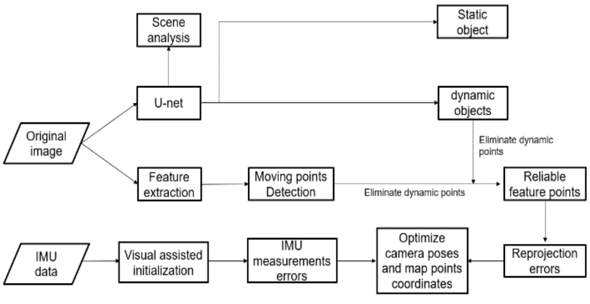


Figure 1. The flow charts

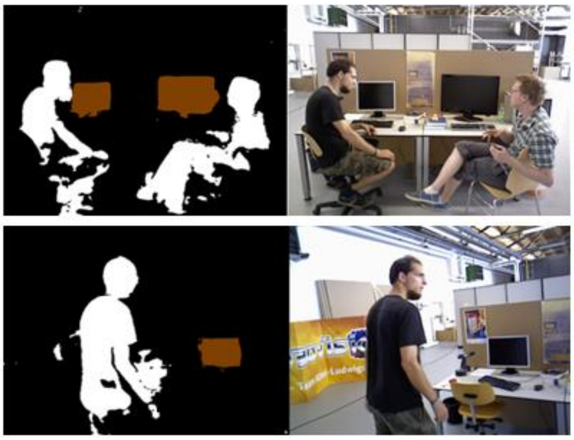


Figure 2. Semantic segmentation results

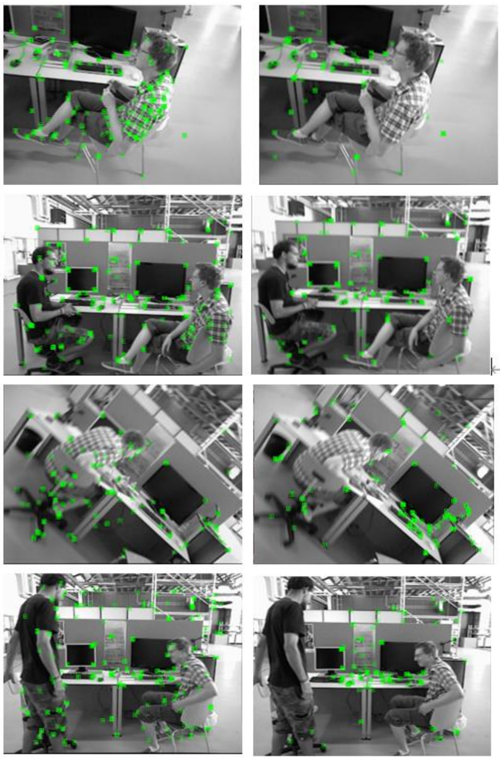


Figure 3. Feature extraction of ORB SLAM2 and our system

2.1 Semantic SLAM

In this section, we will introduce the combination of traditional SLAM system and semantic segmentation module. The feature extraction module and outlier filtering module of the traditional SLAM system will be used. In addition, we introduced the semantic module, which will assist the system to identify dynamic objects and finally filter outliers based on the object level.

2.1.1 Semantic segmentation: Our system adopts U-net based on PyTorch to provide pixel-wise semantic segmentation. The U-net trained on PASCAL VOC dataset could segment each image. In real life, people are most likely to be dynamic objects, so we assume that feature points extracted from people are most likely to be outliers. Figure 2 shows segmentation results of sitting people and walking people. The state of the person reflects the level of the dynamic environment.

2.1.2 Feature extraction: The feature detection algorithm in ORB SLAM2 will be used, with Features from Accelerated Segments Test (FAST) corners as features and Binary Robust Independent Elementary Features (BRIET) as descriptors. The left side of figure 3 shows the results of traditional SLAM feature extraction.

The matched feature points will be used to calculate the camera poses. However, we can know that people in motion will also be extracted to the feature points. In order to obtain reliable poses, it is necessary to ensure that the feature points involved in the pose calculation have accurate data associations. RANSAC method tests the data for consistency by random sampling and iterate continuously. Although it can filter out some outliers, it does not consider the connection between samples in a multi-object motion scene, that is, the points on the same rigid body have the same motion state. But the semantic segmentation has the advantage of identifying whether points are on the same object.

2.1.3 Outlier elimination: The result of semantic segmentation provides semantic level information for visual SLAM. People are regarded as potential dynamic objects, and feature points existing on potential dynamic objects will be eliminated to reduce the impact of mismatch. Descriptors are only calculated for the remaining key points to perform the feature matching process. The right side of Figure 3 shows the results of our system feature extraction.

2.2 VIO

As mentioned in the introduction, low-texture environments and abrupt illumination changes will be inevitable in the real world. However, such scenes are either insufficient in feature points or mismatched, which ultimately leads to the failure of pose estimation. Figure 4 shows the complex environment in a room. This shows the disadvantage of the camera: it is greatly affected by the external environment. Compared with the camera, IMU has the characteristics of small error accumulation in a short time and is not affected by the external environment. Therefore, multi-sensor fusion has become the trend of research.

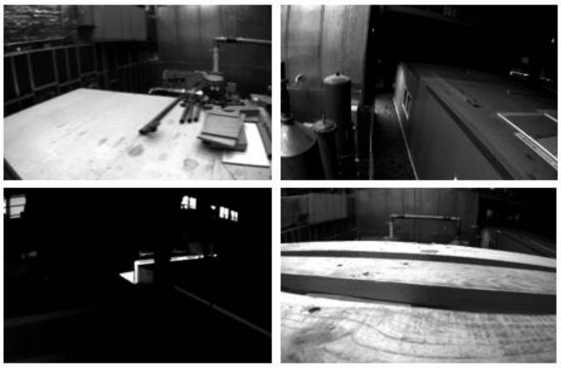


Figure 4. Complex environment

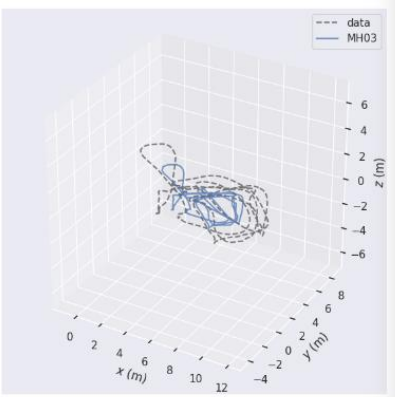


Figure 5. Trajectory of visual SLAM system.

When the IMU performs navigation alone, it needs to remain stationary for a period of time for initialization. In fact, the SLAM system does not provide enough initialization time for IMU. Therefore, the visual information will be used to assist IMU initialization as most VIO systems. The initialization process will be described in detail later.

The scale obtained by traditional monocular SLAM is arbitrary. Figure 5 shows the trajectory graph predicted by ORB SLAM2 system. The IMU measurements can assist the monocular system to obtain the true scale. In this section, we introduce the VIO system in detail.

2.2.1 IMU initialization: The IMU data will be processed concurrently with the image data. The method of obtaining IMU measures with visual assistance is mainly performed by calculating the camera poses of the first few frames. That is, first, the gyro deviation can be obtained from the relative direction between the two frames. Then the scale factor and gravity approximation will be calculated from the gyro deviation if accelerometer deviation is ignored. After that, the acceleration is taken into consideration to get the accelerometer deviation, the corrected scale factor and gravity. This completes the initialization process.

2.2.2 Tracking: Reliable camera pose will be predicted with IMU. The map points in the local map are projected and matched with reliable feature points from the previous section. Then the IMU error terms and reprojection errors of all matched static feature points are jointly optimized to calculate the camera pose. When there is a map update from the Local Mapping or Loop Closing thread, these two errors are calculated as follows for the current frame j and last key frame i:

$$\theta = \{\mathbf{R}_{\mathbf{WB}}^{j}, \mathbf{wp}_{\mathbf{B}}^{j}, \mathbf{wv}_{\mathbf{B}}^{j}, \mathbf{b}_{\mathrm{g}}^{j}, \mathbf{b}_{\mathrm{a}}^{j}\} \tag{1}$$

$$\theta^{*} = \underset{\theta}{\operatorname{argmin}}(\sum_{\mathrm{k}} \mathbf{e}_{\mathrm{proj}}(k,j) + \mathbf{e}_{\mathrm{IMU}}(i,j)) \tag{2}$$

where $\mathbf{R}_{\mathbf{WB}}^{j}, \mathbf{wp}_{\mathbf{B}}^{j}, \mathbf{wv}_{\mathbf{B}}^{j}$ are the orientation, position and velocity of IMU respectively, and $\mathbf{b}_{\mathrm{g}}^{j}, \mathbf{b}_{\mathrm{a}}^{j}$ are biases of the accelerometer and gyroscope respectively.

$$\mathbf{e}_{\mathrm{proj}}(k,j) = \rho(\left(\mathbf{x}^{k} - \pi(\mathbf{X}_{\mathrm{c}}^{k})\right)^{\mathrm{T}} \mathbf{\Sigma}_{\mathrm{k}}(\mathbf{x}^{k} - \pi(\mathbf{X}_{\mathrm{c}}^{k}))) \tag{3}$$

$$\begin{aligned} \mathbf{e}_{IMU}(i,j) = \rho\big([\mathbf{e}_R^T\mathbf{e}_v^T\mathbf{e}_p^T]\mathbf{\Sigma}_I[\mathbf{e}_R^T\mathbf{e}_v^T\mathbf{e}_p^T]^T\big) \\ + \rho(\mathbf{e}^T\mathbf{\Sigma}_R\mathbf{e}_b) \end{aligned} \tag{4}$$

Where

$$\mathbf{X}_{\mathrm{c}}^{\mathrm{k}} = \mathbf{R}_{\mathrm{CB}}\mathbf{R}_{\mathrm{BW}}^{j}\left(\mathbf{X}_{\mathrm{w}}^{k} - \mathbf{wp}_{\mathrm{B}}^{j}\right) + \mathrm{c}\mathbf{p}_{\mathrm{B}} \tag{5}$$

$$\mathbf{e}_R = \log\left((\Delta\mathbf{R}_{ij}\mathrm{Exp}(\mathbf{J}_{\Delta R}^{g}\mathbf{b}_g^{j})^T\mathbf{R}_{\mathbf{BW}}^{i}\mathbf{R}_{\mathbf{WB}}^{j})\right) \tag{6}$$

$$\begin{aligned} \mathbf{e}_v = \mathbf{R}_{\mathbf{BW}}^{\mathrm{i}}\left(\mathbf{wv}_{\mathbf{B}}^{j} - \mathbf{wv}_{\mathbf{B}}^{i} - \mathbf{g}_{\mathbf{w}}\Delta \mathrm{t}_{ij}\right) - (\Delta\mathbf{v}_{\mathrm{ij}} \\ + \mathbf{J}_{\Delta\mathrm{v}}^{g}\mathbf{b}_g^{j} + \mathbf{J}_{\Delta v}^{a}\mathbf{b}_a^{j}) \end{aligned} \tag{7}$$

$$\begin{aligned} \mathbf{e}_p = \mathbf{R}_{\mathbf{BW}}^{i}\left(\mathbf{wp}_{\mathbf{B}}^{j} - \mathbf{wp}_{\mathbf{B}}^{i} - \mathbf{wv}_{\mathbf{B}}^{i}\Delta\mathrm{t}_{\mathrm{ij}} - \frac{1}{2}\mathbf{g}_{\mathbf{w}}\Delta\mathrm{t}_{ij}^{2}\right) \\ - (\Delta\mathbf{p}_{ij} + \mathbf{J}_{\Delta p}^{\mathrm{g}}\mathbf{b}_g^{\mathrm{j}} + \mathbf{J}_{\Delta p}^{a} \\ + \mathbf{J}_{\Delta p}^{a}\mathbf{b}_a^{j}) \end{aligned} \tag{8}$$

$$\mathbf{e}_b = \mathbf{b}^{j} - \mathbf{b}^{i} \tag{9}$$

When there is not a map update, for frame j + 1:

$$\theta = \{\mathbf{R}_{\mathbf{WB}}^{j}, \mathbf{wp}_{\mathbf{B}}^{j}, \mathbf{wv}_{\mathbf{B}}^{j}, \mathbf{b}_{g}^{j}, \mathbf{b}_{a}^{j}, \mathbf{R}_{\mathbf{WB}}^{j+1}, \mathbf{wp}_{\mathbf{B}}^{j+1}, \mathbf{wv}_{\mathbf{B}}^{j+1}, \mathbf{b}_{g}^{j+1}, \mathbf{b}_{a}^{j+1}\} \tag{10}$$

$$\theta^{*} = \underset{\theta}{\operatorname{argmin}}(\sum_{\mathrm{k}} \mathbf{e}_{\mathrm{proj}}(k, j+1) + \mathbf{e}_{\mathrm{IMU}}(j, j + 1) + \mathbf{e}_{\mathrm{proj}}(j)) \tag{11}$$

Where

$$\mathbf{e}_{proj}(j) = \rho\left([\mathbf{e}_{R}^{T}\mathbf{e}_{v}^{T}\mathbf{e}_{p}^{T}\mathbf{e}_{b}^{T}]\mathbf{\Sigma}_{p}[\mathbf{e}_{R}^{T}\mathbf{e}_{v}^{T}\mathbf{e}_{p}^{T}\mathbf{e}_{b}^{T}]^{T}\right) \tag{12}$$

$$\mathbf{e}_{R} = \mathbf{w}\bar{\mathbf{v}}_{\mathbf{B}}^{j} - \mathbf{wv}_{\mathbf{B}}^{j} \tag{13}$$

$$\mathbf{e}_{p} = \mathbf{w}\bar{\mathbf{p}}_{\mathbf{B}}^{j} - \mathbf{wp}_{\mathbf{B}}^{j}\mathbf{e}_{\mathbf{b}} \tag{14}$$

$$\mathbf{e}_{\mathrm{b}} = \bar{\mathbf{b}}^{j} - \mathbf{b}^{j} \tag{15}$$

2.2.3 Local mapping: Compared to traditional visual SLAM, the local window of our system is determined by the last few key frames since the IMU errors are cumulative. Thus, IMU error terms and reprojection error terms are cost functions to further optimize the local window. Old frames that can observe map points only provide visual constraints. In this way, IMU measurements provide constraints for pose calculations to improve performance in low-textured.

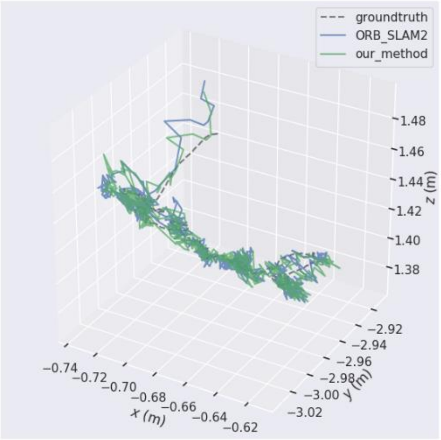


Figure 6. Trajectory graph of fr3_sitting_static predicted by ORB SLAM2 and our system.

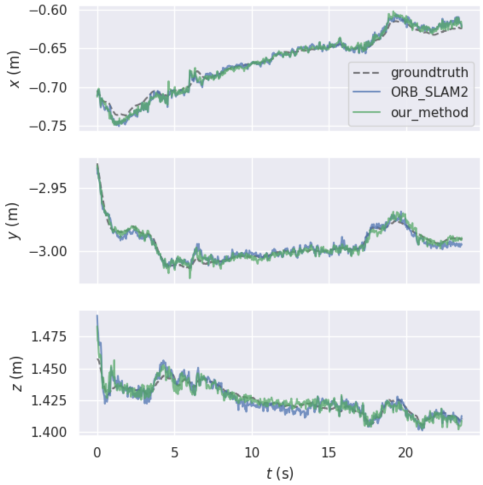


Figure 7. Trajectory of fr3_sitting_static projected in xyz direction

3. EXPERIMENTS AND RESULTS

To evaluate the performance of the proposed method, experiments were conducted on the TUM public dataset and EuRoC public dataset. The low-texture scenes and highly dynamic environments have been encountered.

The fr3 image sequences in the TUM dataset contain sitting and walking people. Datasets are classified into low dynamic environment and high dynamic environment according to the state of people. This can be used to detect the impact of dynamic objects on the positioning result.

The EuRoC dataset consists of three scenes and 11 sequences, including abrupt illumination changes, fast camera movement and low-textured environments. The datasets are divided into three levels of easy, medium and difficult according to the above situation. Compared with the monocular visual system, our system not only obtains the observability scale, but also improves the global positioning accuracy.

The results of the two datasets will be analysed separately. In order to illustrate the resistance of our system to dynamic objects, we selected five sequences on the TUM dataset and ran them on ORB SLAM2 system and our system respectively. Table 1 illustrates the positioning results of sequences. The first two sequences are low dynamic environments and the last three sequences are high dynamic environments. The RMSE reflects the robustness and accuracy of the system, and S.D. reflects the stability of the system.

From the Table 1, it can be seen that whether it is a low dynamic environment or a high dynamic environment, compared with the ORB system, the RMSE and S.D. of our system are reduced. It proves that our system performance has been improved, especially in high dynamic environment.

In order to express the ability of our system to resist dynamic environments, the positioning results will be visualized. We show the predicted trajectory and the real trajectory of fr3_sitting_static and fr3_walking_rpy in the following figures.

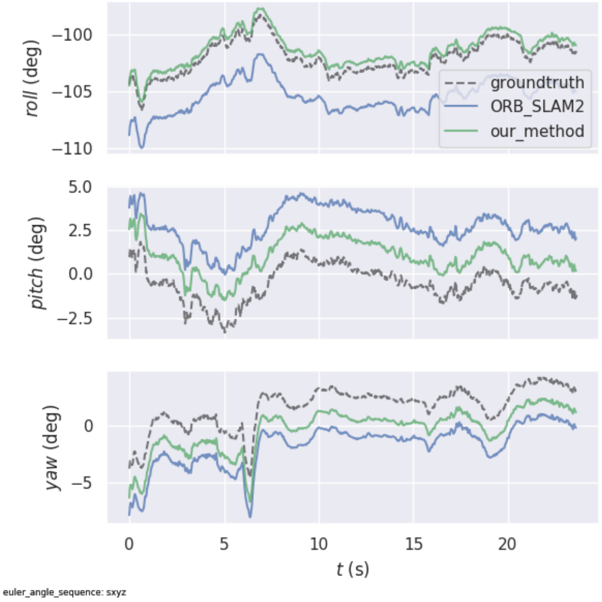


Figure 8. Euler angle of fr3_sitting_static predicted by ORB_SLAM2 and our system.

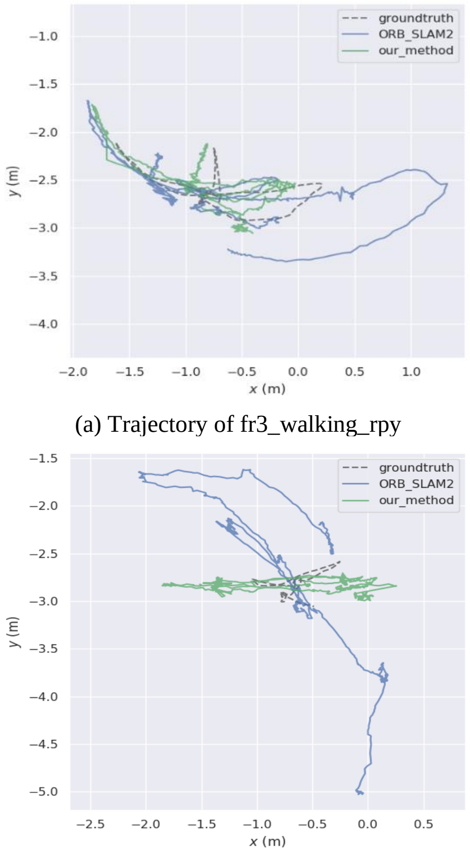
sequence		ORB_SLAM2		our method		improvement	
		RMSE	S.D.	RMSE	S.D.	RMSE	S.D.
Low dynamic scene	Sitting half	0.0242	0.0129	0.0226	0.0118	6.81%	8.38%
Low dynamic scene	Sitting static	0.0084	0.0039	0.0077	0.0038	8.46%	0.98%
High dynamic scene	Walking half	0.4175	0.2160	0.3838	0.1324	8.08%	38.67%
High dynamic scene	Walking rpy	1.0034	0.5387	0.5539	0.1819	44.80%	66.23%
High dynamic scene	Walkng static	0.4201	0.1710	0.3292	0.0999	21.66%	41.56%

Table 1. Absolute trajectory error (ATE) of TUM dataset (m)

sequence		ORB_SLAM2		Our method		Improvement	
		RMSE	S.D.	RMSE	S.D.	RMSE	S.D.
easy	MH01	0.0434	0.0189	0.0258	0.0154	40.55%	18.64%
medium	MH02	0.0359	0.0186	0.0210	0.0099	41.49%	46.58%
medium	MH03	0.0392	0.0171	0.0279	0.0118	28.79%	31.08%
difficult	MH04	0.0594	0.0257	0.0640	0.0244	-7.76%	4.91%
difficult	MH05	0.0737	0.0400	0.0497	0.0225	32.57%	43.69%

Table 2. Absolute trajectory error (ATE) of EuRoC datasets(m)

They are representatives of low dynamic environment and high dynamic environment respectively.



(a) Trajectory of fr3_walking_rpy

(b) Trajectory of fr3_walking_half

Figure 9. Trajectory predicted by ORB_SLAM2 and our system.

Figure 6 shows the trajectory of fr3_sitting_static predicted by ORB SLAM2 and our method. It can be seen from the Table 1 and Figure 6 that the performance in the low dynamic environment has improved even if it is small. Figure 7 and Figure 8 are used to clearly express the difference in system positioning accuracy.

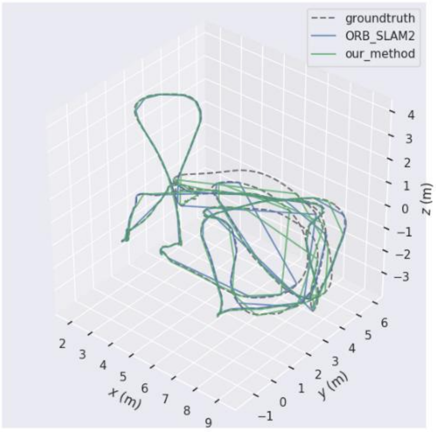


Figure 10. Trajectory graph of MH03 predicted by ORB_SLAM2 and our system.

Figure 7 shows the trajectory of fr3_sitting_static projected in xyz direction. It can be seen from the Figure 7 that although the errors of the two systems are similar in low dynamic environment, the trajectory predicted by our system is closer to the true trajectory.

Figure 8 shows the predicted Euler angle. Obviously, the performance of our system is better than ORB SLAM2.

Figure 9 shows the trajectory of fr3_walking_rpy and fr3_walking_half projected onto the plane in high dynamic scenes. As can be seen from the figure, the positioning error of the ORB SLAM2 is large in the high dynamic environment, which is quite different from the real trajectory. It proves that the positioning result is disturbed by dynamic objects in the traditional SLAM system. However, as can be seen from Figure 9, our system can greatly reduce the interference of dynamic objects. It is consistent with the results in Table 1 that the improvement of positioning performance in high dynamic scene is very obvious.

Table 2 shows the results of experiments conducted on the EuRoc dataset. The environment of the EuRoc dataset is shown in Figure 4. Five sequences were selected for experiments to illustrate the resistance of our system to complex environments.

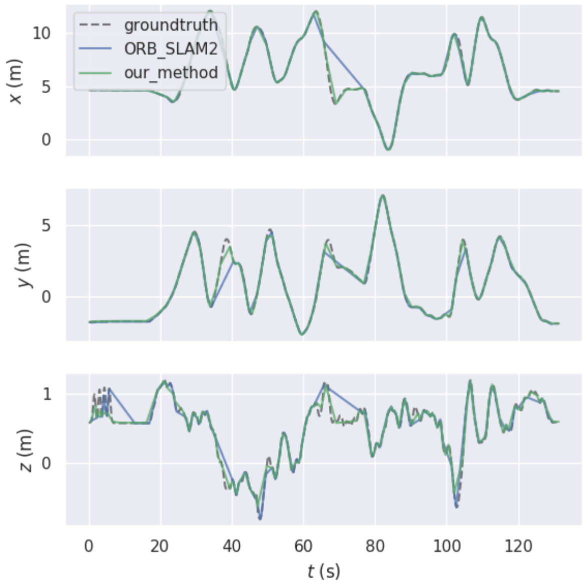


Figure 11. Trajectory of MH03 projected in xyz direction

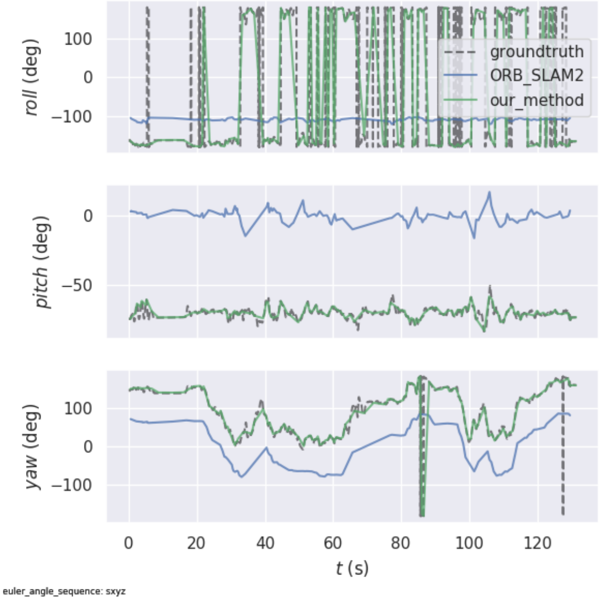


Figure 12. Euler angle of MH03 predicted by ORB_SLAM2 and our system.

As can be seen from the Table 2, in the five sequences, the RMSE and S.D. of our system are mostly lower than the ORB SLAM2 system, which proves that the stability and robustness of our system is better than the ORB SLAM2 system. Figure 10 plots the trajectory of MH03 sequence predicted by the ORB SLAM2 system and our system respectively.

For the MH03, the maximum ATE of ORB SLAM2 is 0.101m, and that of our method is 0.059m. The trajectory predicted by our system is closer to the real trajectory than ORB SLAM2 system. In order to show the performance of the system more clearly, the trajectory error graph is displayed in Figure 11 and 12.

4. CONCLUSION

In our paper, a VIO system based on semantic assistance is proposed. Compared with the traditional visual SLAM, it has a module for semantic recognition of dynamic objects, and its performance is improved in dynamic environment by removing feature points on dynamic objects. In addition, the joint optimization of IMU measurement errors and reprojection errors ensures the system to acquire good pose calculation results under low-textured environment. Experiments prove that the performance of our system has been improved in a complex indoor environment. In the future research, we will introduce semantic data association, that is, the fusion of semantic labels of static objects to the proposed system, and a semantic consistency map will eventually be established.

REFERENCES

Berta B, Facil J M, Javier C, et al. DynaSLAM: Tracking, Mapping and Inpainting in Dynamic Scenes. *IEEE Robotics & Automation Letters*, 2018:1-1.

Bowman S L, Atanasov N, Daniilidis K, et al. Probabilistic data association for semantic slam. *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1722-1729.

Davison A J, Reid I D, Molton N D, et al. MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2007, 29: pp.1052-1067.

Dong-Si T C, Mourikis A I. Consistency analysis for sliding-window visual odometry. *IEEE International Conference on Robotics and Automation*, 2012, pp. 5202-5209.

Engel J., Koltun V. and Cremers D., Direct Sparse Odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, pp. 611-625.

Engel, J., Schöps, T and Cremers, D. LSD-SLAM: Large-Scale Direct Monocular SLAM, DC, USA (2014), 13–16, pp. 1–10

Hauke Strasdat, J.M.M. Montiel, Andrew J. Davison. Visual SLAM: Why filter? *Image & Vision Computing*, 2012, 30(2):65-77.

Heng D, Usman A, Qiang F, et al. Visual–inertial estimation of velocity for multicopters based on vision motion constraint. *Robotics and Autonomous Systems*, 2018, pp. 262-279.

J. Civera, D. Gálvez-López, L. Riazuelo, J. D. Tardós and J. M. M. Montiel, Towards semantic SLAM using a monocular camera, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1277-1284.

Klein, G., and Murray, D., 2007. Parallel tracking and mapping for small AR workspaces (PTAM). *IEEE and ACM International Symposium on Mixed and Augmented Reality, Washington, DC, USA*, 13–16, pp. 1–10

Li, P. L., Qin, T., Hu, B. T., Zhu, F. Y., and Shen, S. J., 2017. Monocular Visual-Inertial State Estimation for Mobile Augmented Reality. *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Nantes, 2017, pp. 11-21.

Li S P, Zhang T, Gao X, et al. Semi-direct monocular visual and visual-inertial SLAM with loop closure detection. *Robotics and Autonomous Systems*, 2018.

Mur-Artal, R., J, M. M. Montiel., and J. D. Tardós., 2015. ORBSLAM: a Versatile and Accurate Monocular SLAM System, *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163.

Mur-Artal, R., and Tardós, J. D., 2017. ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras. *IEEE Trans. Robot*, vol. 33, pp.1255–1262.

Raúl Mur-Artal, Juan D. Tardós. Visual-Inertial Monocular SLAM With Map Reuse. *IEEE Robotics and Automation Letters*, 2017, pp. 796-803.

Renato, F. S. M., Richard, A.N., Hauke, S., Paul, H. J. K., and Andrew, J. D., 2013. SLAM++: Simultaneous global location and mapping at the level of objects. *IEEE International Conference on Computer Vision and Pattern Recognition*, pp. 1352-1359.

Rublee, E., Rabaud, V., and Konolige, K., 2011. ORB: An efficient alternative to SIFT or SURF. *IEEE International Conference on Computer Vision (ICCV)*, 6–13, pp. 2564–2571.

Zhi S, Bloesch M, Leutenegger S, et al. SceneCode: Monocular Dense Semantic Reconstruction using Learned Encoded Scene Representations. 2019, pp. 11776-11785