Stereo SLAM in Dynamic Environments using Semantic Segmentation

Yongbao Ai , Qianchong Sun , Zhipeng Xi , Na Li , Jianmeng Dong , Xiang Wang *

Posted Date: 30 May 2023

doi: 10.20944/preprints202305.2072.v1

Keywords: stereo SLAM; semantic segmentation; moving object detection; dynamic scenarios

Article

Stereo SLAM in Dynamic Environments Using Semantic Segmentation

Yongbao Ai, Qianchong Sun, Zhipeng Xi, Na Li, Jianmeng Dong and Xiang Wang *

National Innovation Institute of Defense Technology, Beijing 100071, China
* Correspondence: wang_xiang927@163.com

Abstract: As we all know, many more dynamic objects appear almost continuously in real world, which are immensely able to impair the performance of the majority of vision-based SLAM systems that are based on the static-world assumption. In order to improve the robustness and accuracy of visual SLAM in the high dynamic environment, a real-time and robust stereo SLAM system for dynamic scenes was proposed. To weaken influences of dynamic content, the moving object detection method was put forward in our visual odometry, then the semantic segmentation network was combined into our stereo SLAM to extract pixel-level contours of dynamic objects. Then influences of dynamic objects were extremely weaken and the performance of our system was increased markedly in dynamic, complex and crowed city spaces. Experiment with both on KITTI Odometry dataset and in a real-life scene, the results show that our method can dramatically decrease the tracking error or drift, improve the robustness and stability of our stereo SLAM in high dynamic outdoor scenarios.

Keywords: stereo SLAM; semantic segmentation; moving object detection; dynamic scenarios

1. Introduction

In the 21st century, with the computer vision technology, artificial intelligence and sensor technology increasingly mature, it promotes the gradual transformation of robots from traditional industrial robots to intelligent robots capable of perception, analysis, learning and decision making. The intelligent robots that have characteristics of interconnectivity, virtuality and reality combination and man-machine fusion will play a more and more important role in industry, agriculture, transportation, aerospace, national defense, military, public security and other fields. Nevertheless, how to realize efficient and accurate environment perception, autonomous localization and navigation of intelligent robots in the unstructured environment composed of static obstacles and high dynamic objects is a hot topic in the field, where the key technologies include practical Simultaneous Localization and Mapping (SLAM). It refers to simultaneously estimating locations of newly perceived landmarks and the location of the external environment sensing sensor itself while incrementally mapping in an unknown environment. Moreover, SLAM is considered as the critical technology of visual navigation for several applications, for example, exploring areas where the human lacks access or its manoeuvre is hard, such as in scenarios like underwater, inside of mines or other narrow spaces.

More common in life, the camera offers the superiorities of full 3-D, cheaper than lasers, convenience as well as visual perception like humans. More importantly, the camera can provide seemingly endless potential for the extraction of detail, both geometric and photometric, and for semantic and other higher-level scene understanding, as continually elucidated by the prosperous computer vision research community. Hence, visual SLAM systems are pivotal strategy for the development of the next navigation techniques. The last two decades have seen a significant surge to visual SLAM [1–3] from range sensor-based systems. And many well-performing SLAM systems have been developed, such as SVO [4], LSD-SLAM [5] and ORB-SLAM1/2/3 [6–8]. Nevertheless, to simplify the problem formulation, a majority of SLAM systems typically assume a static world, where

 doi:10.20944/preprints202305.2072.v1

only rigid and non-moving objects are involved. In the real world, it is common to have objects in motion relative to the stationary environment around them, especially in the field of public transportation, where vehicles and pedestrians constantly move back and forth. Although a fraction of moving objects in classical systems can be dealt with by regarding them as noise, the great mass of dynamic objects violate the assumption of a static world, and the accidental appearance of these dynamic objects in front of camera lenses can also cause a decrease in the accuracy of the visual SLAM system, and even the phenomenon of image blur caused by fast moving objects, seriously weakening the robustness and stability of the visual SLAM system, which leads to some available visual SLAM systems are limited for real-world applications. So it requires modelling and tracking them to ensure they do not interfere with the location and map generation of the visual SLAM system. Then, it has proposed some solutions, such as motion target tracking, spatiotemporal consistency, dynamic object segmentation and modelling, and other technologies, to deal with the influence of dynamic objects on the system. Just as Panchpor et al. elucidate in survey [9], since many problems in SLAM of automatic robots in dynamic environments cannot have a robust solution, there's a large amount of scope for further research and development in this domain.

In this paper, we reveal the advantages of accommodating both the semantic segmentation convolutional neural network and the proposed moving object detection algorithm in a visual SLAM system using just a stereo camera. After conducting experiments on the KITTI Odometry datasets [10] and real-world scenarios, the performance of our system has significantly improved in high dynamic scenes compared to the original ORB-SLAM2 [7] system. It also outperforms the DynaSLAM [11] system which is specifically designed for dynamic scenes. The novelty of our paper is summed up as below:

- A new stereo SLAM system based on the ORB-SLAM2 framework combined with a deep learning method is put forward to decrease the impact of dynamic objects on the camera pose and trajectory estimation. The approach of semantic segmentation network plays a role in the data preprocessing stage to filter out features expression of moving objects.
- A novel motion object detection method is presented to reduce influences of moving targets on the camera pose and trajectory estimation, which calculates the likelihood of each keyframe points belonging to the dynamic content and distinguishes between dynamic and static goals in scenarios.
- To the best of our knowledge, the semantic segmentation neural network ENet [12] that is appropriate for city spaces is first utilized to enhance the performance of the visual SLAM system, which makes our system become more robust and practical in high dynamic and complex city streets, and it has practical engineering applications to a certain extent.

The surplus of the paper is organized as follow: Section 2 presents relevant work. The architecture of the proposed stereo SLAM system and the details of our method to solve visual SLAM in dynamic scenes are provided at length in section 3. Whereafter, we show the qualitative and quantitative experiment results in section 4. At last, both a conclusion and a future research are given in section 5.

2. Related Work

The Visual SLAM with a stereo camera. Since Davison and Murray describe the first traditional stereo-SLAM framework that is a real-time EKF-based system [13–15], there are many scholars make their contribution towards this research field. Iocchi et al. combine range mapping and image alignment to reconstruct a planar environment map with just a stereo camera [16]. Se et al. describe a stereo-based mobile robot SLAM algorithm utilizing SIFT features in a small lab scene [17], which is not adapted to large environments or working in challenging outdoor scenes. In [18] and [19], authors demonstrate an autonomous low altitude aero-craft system using the EKF SLAM algorithm for terrain mapping with a stereo sensor. Saez et al. present a full 6-DOF SLAM with a wearable stereo device, which is based on both the ego-motion and the global rectification algorithm [20]. A dense visual SLAM system utilizing Rao-Blackwellized particle filters and SIFT features is demonstrated in

 doi:10.20944/preprints202305.2072.v1

[21,22]. The system can also work in stereo mode, with a camera fixed on a robot moving in 2D space. In [23] a six degrees of freedom SLAM with a stereo handheld camera is presented, which is utilized to construct large-scale indoor or outdoor environments on the basis of the conditionally independent divide and conquer algorithm. The aforementioned many SLAM approaches focus on operating in static environments, it is a strong mathematical modeling assumption, so it restricts the system a lot in practical applications. Actually, there are several dynamic objects in real environments, where the moving objects can generate errors in visual SLAM performing in outdoor or indoor dynamic scenes, it is because dynamic features cause a bad pose estimation and erroneous data association. For this reason, there are a few SLAM systems that specifically address dynamic content, trying to split dynamic and static regions within sequences of images of a dynamic environment. In [24] Lin and Wang present a stereo-based SLAMMOT method which makes use of the inverse depth parameterization and performs EKF to overcome the observability issue. Kawewong et al. [25] use position-invariant robust features (PIRFs) to simultaneously localize and mapping in high dynamic environments. Alcatnarilla et al. [26] detect dynamic objects employing a scene flow representation using stereo cameras for visual odometry [27] in order to improve robustness of the visual SLAM system [28]. But the algorithm of dense optical flow may detect many pixels in an image as dynamic objects, whereas these pixels remain with static content in fact on account of inevitable measurement noises or optical flow inherent problems. In [29] Zou and Tan describe a collaborative vision-based SLAM system with multiple cameras, which differs from now available SLAM systems with just one stereo camera fixed on a single platform. Their system utilizes images from each camera to construct a global map and can run robustly in high dynamic scenes. Their method to deal with dynamic objects is a novel try, but there is a high hardware cost due to the use of multiple cameras. Fan et al. [30] propose an image fusion algorithm to get rid of the influence of dynamic content in a stereo-based SLAM. But their system cannot cope with moving objects with the slower speed of movement or large size.

Deep learning models are used in the visual SLAM system, which has proved its superiority in SLAM systems. There are many outstanding studies that employ deep learning models to replace some non-geometric modules in traditional SLAM systems. B. Bescos et al. [11] present the DynaSLAM that uses Mask R-CNN [31] to segment dynamic objects and background inpainting to synthesize a realistic image without dynamic objects. However, since the network speed of Mask R-CNN is slow and its segmentation algorithm architecture is not a multi-threaded operation mode, their system cannot perform in real time. In the paper [32] a weakly-supervised semantic segmentation neural network is used in the system, which reduces annotations for training. Kang et al. propose a DF-SLAM system which uses a shallow neural network [33] to extract local descriptors [34]. In the literature [35], a two-dimensional detection and classification CNN is used to provide semantic 3D box inferences, then a robust semantic SLAM system in which camera ego-motion and 3D semantic objects in high dynamic environments are being tracked is presented just with a stereo camera. In our previous work [36], a robust RGB-D SLAM with a semantic segmentation neural network is put forward. Since the sensor of our system is an RGB-D camera which is subject to the effects of illumination changes, it is just suitable for indoor dynamic scenes. In this work, we present a robust stereo SLAM that is primarily appropriate for high and complex outdoor dynamic traffic scenes.

3. System Description

The proposed SLAM system will be described at length in this section. Firstly, the framework of our stereo SLAM is presented. After that, a brief presentation of the real-time semantic segmentation method utilized in our system is given. Then the moving object detection method that is proposed to calculate the motion corresponding likelihoods among sequential stereo frames is expounded in detail. Eventually, the means of filtering out outliers in the proposed SLAM system is presented.

3.1. Overview

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 May 2023 doi:10.20944/preprints202305.2072.v1

Nowadays, how to accurately recognize and localize dynamic objects in real scenes, and eliminate their impact on camera pose and trajectory estimation and 3D point cloud map construction is the crucial task of visual SLAM. As we all know, ORB-SLAM2 has an outstanding performance in a variety of scenarios from a handheld camera in the indoor environment, to unmanned aerial vehicles flying in outdoor scenes and pilotless automobiles driving around in a road. Accordingly, our method is integrated with the ORB-SLAM2 system so as to achieve the purpose of improving the stability, robustness and reliability of its performance in large-scale scenes of dynamic and complex urban traffic roads.

Figure 1 shows a flow chart of the stereo SLAM system. It consists of four prime parallel threads: tracking, semantic segmentation, local mapping and loop closing. The raw stereo RGB images obtained by a ZED camera are dealt with in the tracking and semantic segmentation thread at the same time. At the beginning, ORB feature points of frames are extracted in the tracking thread, after that dynamic points are filtered out by the proposed moving object detection (MOD) algorithm. Afterwards, it is waiting for the pixel-wise segmentation mask acquired in the semantic segmentation thread. To incorporate the MOD method with semantic segmentation, we are able to discard feature points which indeed belong to dynamic objects. The reason why is that the inherent problem of semantic prior categories, it maybe mistake static content for dynamic, e.g., parked cars or pedestrians waiting for traffic lights. Originally, these two categories of objects are pre-defined as dynamic objects in the priori category classification of the semantic segmentation algorithm. However, in a real environment, they exist in a static state for a short or long time. If only the semantic segmentation algorithm is used to process such objects, it is easy to misclassify. Therefore, the designed stereo SLAM system is also combined with the MOD method to detect the real dynamic areas in the frame, and treat the ORB feature points belonging to these areas as outliers. Then we obtain the transformation matrix through matching the rest of stable ORB feature points. At last, the fifth thread is launched to perform full BA (Bundle Adjustment) after the local mapping and loop closing thread implemented and optimize all camera poses and landmark points of the whole map to eliminate accumulated errors during system operation.

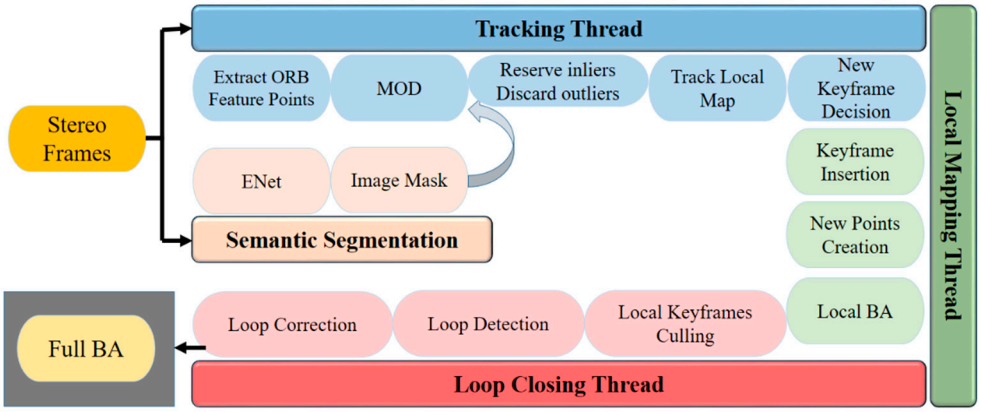


Figure 1. The overview of our VSLAM, main modules: tracking, semantic segmentation, mapping and global optimization modules.

3.2. Semantic Segmentation

Semantic segmentation plays an important role in understanding the content of images and finding target objects, which is useful and helpful in some practical applications, such as unmanned driving and augmented reality. In our stereo SLAM system, ENet is adopted to provide pixel-wise semantic segmentation masks based on the Caffe implementation by TimoSaemann[1] in real time. And it is designed to achieve faster, fewer parameters and more accurate than SegNet that is utilized

[1] https://github.com/TimoSaemann/ENet

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 May 2023 doi:10.20944/preprints202305.2072.v1

in the paper [37]. The ENet is adequate for analyzing applications in the urban street scene [12]. The segmentation network trained on Cityscapes dataset [38] could segment 19 classes in total.

The ENet neural network takes a color image as input, and outputs the corresponding semantic segmentation mask that labels every pixel in the image with one of the several predefined movable categories, for instance, vehicles, persons and riders. These segmentation masks are easy to use in our SLAM system to accurately separate the dynamic object region and the static background area. Whereafter, the binary masks are input to the tracking thread and the detail is clarified in the section 3.4.

3.3. Moving Object Detection

The moving object detection (MOD) method is to set apart moving objects from the background of video sequence images. In many applications of the computer vision, moving object detection is one of the key technologies in the image processing [39], which is utilized to find out dynamic objects accurately in the image and keep from bringing on erroneous data association in the SLAM system running in dynamic crowded real-world environments. Based on classical LK (Lucas Kanada) optical flow method [40], our MOD method leverages Shi-Tomas corner detection [41] to extract sub-pixel corners in the previous frame, then the pyramid LK optical flow algorithm is utilized to track the motion and obtain the matched feature points of the current frame. Next, screen preliminarily feature points according to the following rules: 1) on condition that the matched pairs are close to the edge of the frame; or 2) the pixel disparity of the 3 × 3 image block centered on matched pairs is too large, the matched pairs will be ignored. Afterwards, we find the fundamental matrix through the RANSAC algorithm with a majority of matched feature points, which describes a relationship between any two images of the same scene that restrains where the projection of points from the scene can appear in both images. The p_0, p_1 is denoted as the matched points in the previous frame and current frame separately, and P_0, P_1 is their homogeneous coordinate form.

$$\begin{aligned} p_0 &= [u_0, v_0] & p_1 &= [u_1, v_1] \\ P_0 &= [u_0, v_0, 1] & P_1 &= [u_1, v_1, 1] \end{aligned} \tag{1}$$

where u, v are the pixel coordinates of the frame. Then L denotes the epipolar line that maps the feature points from a previous frame to the correspondences search domain of a current frame, which can be solved as follows:

$$L = \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = FP_0 = F \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} \tag{2}$$

where X, Y, Z denote line vector and F represents fundamental matrix. And the distance from the point P_1 of the current frame to its corresponding epipolar line L is computed by the equation (3):

$$D = \frac{|P_1^T F P_0|}{\sqrt{\|X\|^2 + \|Y\|^2}} \tag{3}$$

where D stands for the distance. If the value of D is higher than the predetermined threshold, matched points will be decided to be dynamic. Figure 2 describes an instance of the four images that can be utilized to compute a sparse optical flow representation of the scene. The original stereo RGB images used in it were collected using a ZED camera in a dynamic real urban road scene. It can be seen that the proposed MOD method can detect moving points in the current frame. However, because the movement of the camera itself interferes with the judgment of the MOD method, some moving points are also detected by mistake on the static buildings in the picture. Therefore, in order to find the real dynamic area in the image, the result obtained by the MOD method is combined with the segmentation mask obtained by real-time semantic segmentation in section 3.2 to accurately distinguish the dynamic object region and the static background area in frames. The specific implementation process will be discussed in detail in the next section 3.4.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 May 2023 doi:10.20944/preprints202305.2072.v1

Figure 2. The schematic diagram of moving object detection based on Sparse Optical Flow method.

3.4. Outliers Removal

One of core tasks for visual SLAM in dynamic environments is the rejection of landmarks which in fact belong to dynamic objects. However, if the semantic segmentation method is only used to differentiate static and dynamic regions in frames, the SLAM system will fail to estimate lifelong models when predetermined moving objects keep static for a short time, for instance, sitting people or parked cars. What's worse, in an extremely challenging environment where dynamic objects may occupy almost the whole image view, there may be many residual correspondences declared as inliers, but they actually belong to moving objects. This leads to large errors in the trajectory estimation and mapping of the SLAM system. Hence, we propose to integrate semantic segmentation with the MOD method to filter out outliers successfully.

The urban street scenarios comprise of stationary and moving objects. The main error obtained in the measures is the presence of dynamic objects, since a total features fraction are localized on the moving objects. Therefore, it is important to avoid them during the system process. Then we will illustrate how to identify regions of frames that belong to dynamic content through combining semantic segmentation with the MOD method. With the MOD algorithm explained in the section 3.3, we can derive motion likelihoods which can be utilized to differentiate dynamic objects, aiding the stereo SLAM performing in crowded and high dynamic environments. Once we have computed the dynamic feature points in the current frame, it is necessary to take into account the image mask of semantic segmentation using ENet network in order to filter out moving objects. According to the experience from multiple tests, we should set a fixed threshold $N=5$ so as to merely detect moving objects in a reliable way. If the dynamic feature points falling into the region of predefined semantic masks are more than N, the segmentation area belongs to moving objects, conversely, static content. Afterwards, in this way, we sweep away ORB feature points located on the dynamic areas of frames from the SLAM process yielding more robust and accurate camera pose and trajectory results. The overall process of outliers removal is depicted in Figure 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 May 2023 doi:10.20944/preprints202305.2072.v1

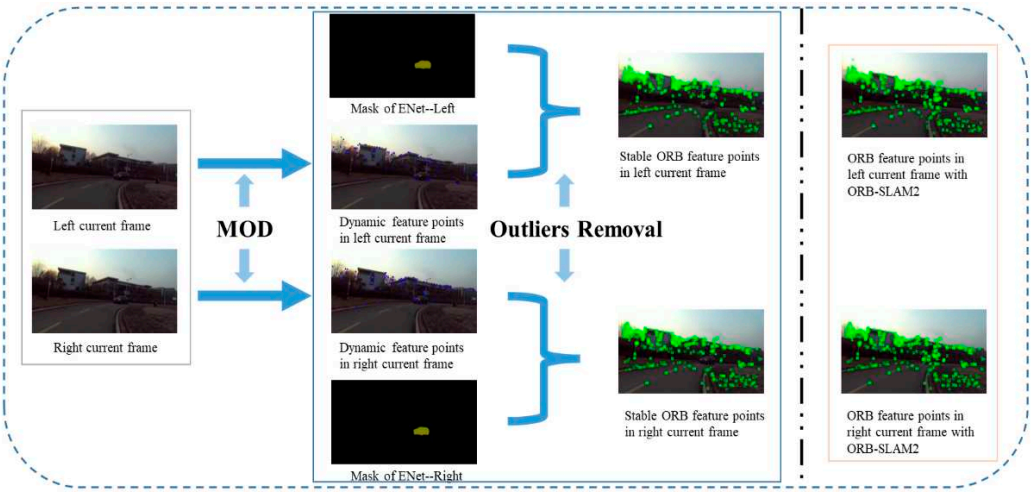


Figure 3. A couple of images are captured by the ZED Stereo Camera. On the left side of demarcation line the process of outliers removal is depicted. Its right side is the ORB feature points result of frames without our approach. As it can be seen that the ORB feature points of a dynamic car is not discarded in the ORB-SLAM2 system, so trajectory errors come out.

4. Experiments

To analyze the effectiveness of our method quantitatively, the performance of the proposed stereo SLAM system in dynamic scenarios is assessed with the KITTI Odometry dataset at first. Then the time required for tracking in our SLAM is counted to test its real-time performance. Furthermore, the stereo SLAM is integrated with ROS system and we qualitatively test it on a physical robot in the dynamic urban traffic scene to evaluate its accuracy, robustness and practicability. Most experiments are executed on a computer with Intel i7 CPU, NVIDIA GeForce GTX TITAN X GPU, and 12 GB memory. The physical robot is a TurtleBot3, and stereo image sequences are captured by the ZED camera, which provides accurate camera calibration parameters and stereo rectification.

4.1. Evaluation using KITTI Benchmark Dataset

The KITTI Odometry dataset [10] is one of the largest computer vision algorithm evaluation datasets suitable for autonomous driving scenarios, and it is captured by a set of vehicle equipment driving around a middle-sized city, in the countryside and on the freeway, which provides many sequences in dynamic scenarios with accurate ground truth trajectories directly attained by the output of the GPS/IMU localization unit projected into the coordinate system of the left camera after rectification. The dataset contains 1240 × 376 stereo color and grayscale images captured at 10 *Hz*. And the 01 sequence is collected on the freeway, and the 04 sequence is in a city road. Both of them are primarily applied to our experiments because they are high dynamic scenes.

To verify the effectivity of our method in this section, besides compared with the ORB-SLAM2 system, the OpenVSLAM is also selected to be analysed, which is based on an indirect SLAM algorithm with sparse features, such as ProSLAM and UcoSLAM [42]. The OpenVSLAM framework can apply sequence images or videos collected by different types of cameras (monocular, stereo, and so on) to locate the current position of the camera in real time and reconstruct the surrounding environment three-dimensional space. It has advantages over the ORB-SLAM2 system in the speed and performance of the algorithm and can quickly locate the newly captured image based on the pre-built map. However, the OpenVSLAM system is also constructed on the basis of the static world assumption. Theoretically, the performance of our stereo SLAM system is more robust and stable than it in high dynamic scenarios. In addition, we also compared with the performance of DynaSLAM (stereo) which is also suitable for dynamic scenes on the KITTI Odometry dataset.

The metric of absolute pose error (APE) is used to measure the performance of a visual SLAM system. And the metric of relative pose error (RPE) is suitable for measuring the drift of a visual

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 May 2023 doi:10.20944/preprints202305.2072.v1

odometry. Therefore, the metrics APE and RPE are computed for the quantitative evaluation analysis. And the values of Root-mean-square Error (RMSE) and Standard Deviation (STD) can preferably give evidence of the robustness and stability of the system. Therefore they are chosen as evaluation indexes. Figure 4 and Figure 5 respectively show APEs and RPEs on the 11 sequences of KITTI dataset. Furthermore compared with the original ORB-SLAM2 system, the RMSE of APE improved by up to 80%, and the STD metric improved by up to 78%. The RMSE of RPE improved by up to 50%, and the STD metric improved by up to 60%. Then it has comparable performance to OpenVSLAM with respect to tracking accuracy in most static sequences (00, 02, 03, 05–10). These fully prove that the proposed stereo SLAM system has much more stability, accuracy and robustness no matter in a high dynamic environment or a low dynamic or static environment.

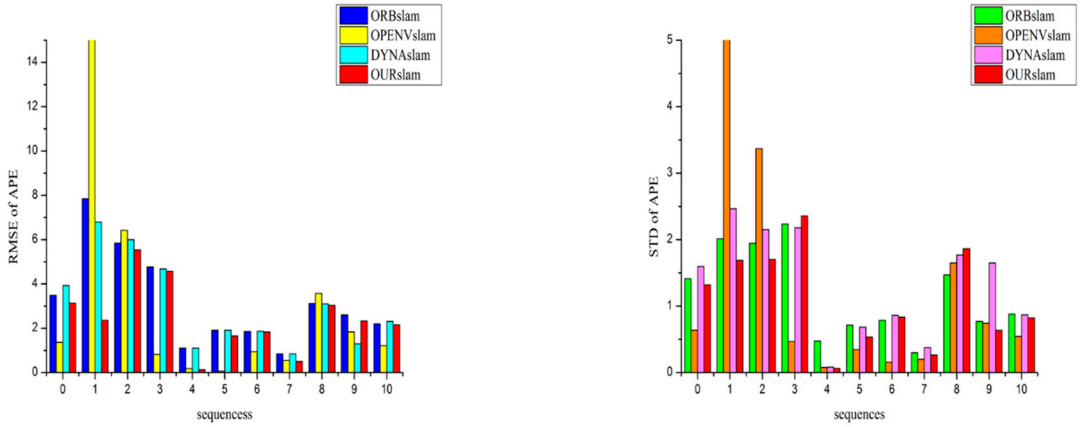


Figure 4. Absolute pose errors on the 11 sequences in KITTI Odometry dataset. Lower is better.

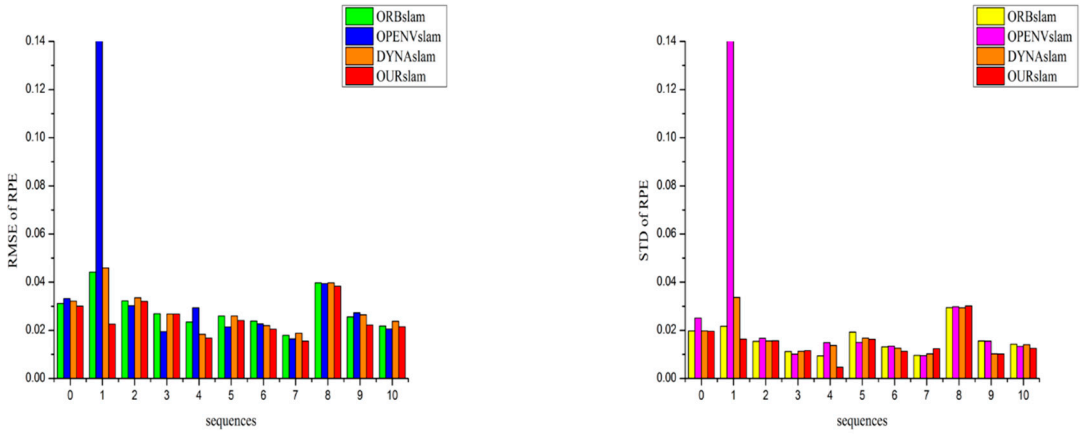
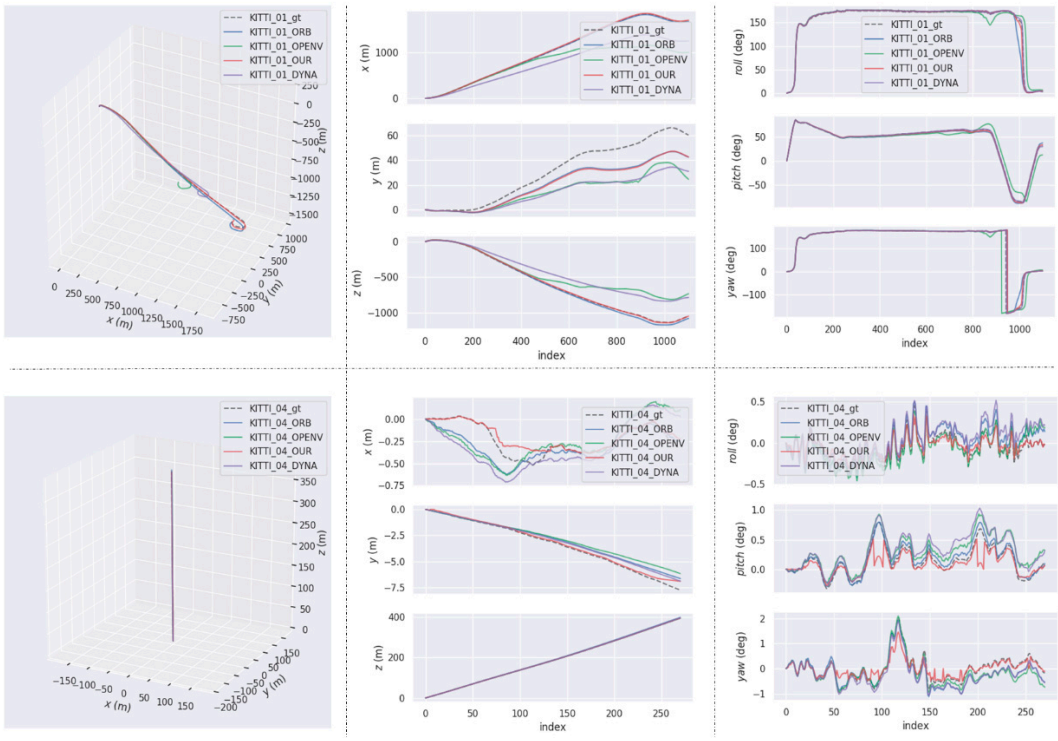


Figure 5. Relative pose errors on the 11 sequences in KITTI Odometry dataset. Lower is better.

The trajectories of four VSLAM systems on KITTI high dynamic sequences are described in the Figure 6. It is found that our trajectories are most close to ground-truth trajectories in dynamic scenes, which is clearly show that our stereo SLAM system is the most accurate and robust among four vision-based SLAM systems in the high dynamic sequences of 01 and 04. In addition, compared with the trajectory of 04 sequence, the 01 sequence is more intricate. It can be inferred from this that our SLAM system will be more stable and robust than others in more challenging dynamic environments.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 May 2023 doi:10.20944/preprints202305.2072.v1



(a) Trajectories in a XYZ Coordinate System (b) Trajectories in the x/y/z plane (c) Trajectories in the r/p/y orientation

Figure 6. The trajectories of four VSLAM systems on KITTI high dynamic sequences. The top row is the result of 01 sequence and the bottom row is the result of 04 sequence.

4.2. Time analysis

In most practical applications of visual SLAM, such as, intelligent robots, virtual and augmented reality on mobile devices, the real-time performance is crucial for them. Then we try out the tracking times of our SLAM with the 05 sequence of KITTI benchmark dataset. The configurations of the test platform are as introduced at the beginning. The DynaSLAM (stereo) is not compared here because it is not a real-time system. Table 1 shows test results. It is obvious that the run time of our SLAM is numerically close with ORB-SLAM2, which is sufficiently short for practical applications although it is a little slower than OpenVSLAM.

Table 1. The mean and median tracking time of three frameworks.

	ORB-SLAM2	OpenVSLAM	OurSLAM
Mean[ms/frame]	65.35	55.46	68.56
Median[ms/frame]	66.43	56.25	69.21

4.3. Evaluation Test in Real Environment

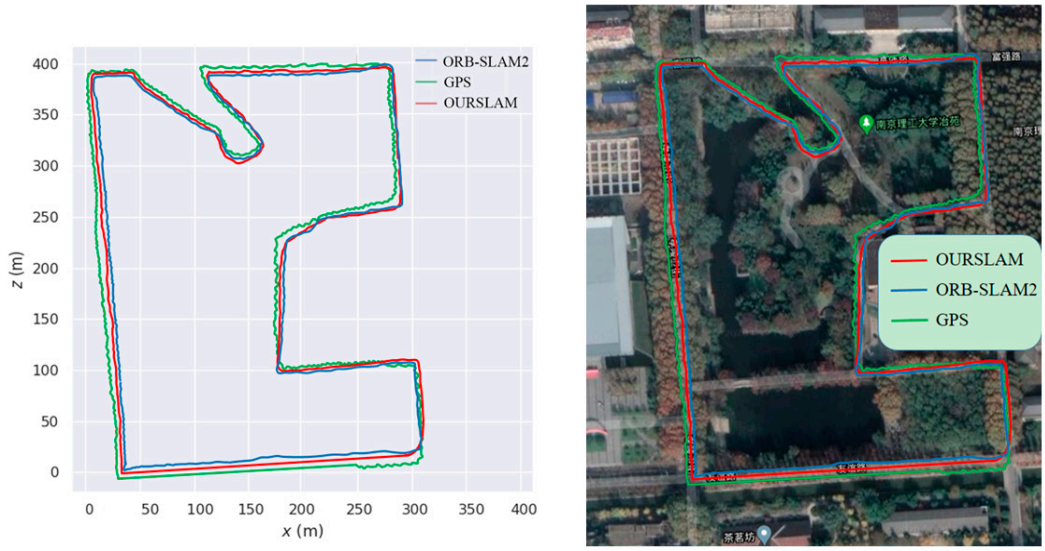
In order to illustrate the stability and availability of our system, we carried out a large-scale visual SLAM experiment in cluttered urban outdoor scenarios, with some independently moving objects for example, pedestrians, riders and cars. We took advantage of a TurtleBot3 robot to perform our SLAM in outdoor scenes. It had to come back to the same place of the route so as to close the loop and correct the accumulated drift during the SLAM process. Images of resolution 320 × 240 captured by a ZED camera at 30 *Hz* were processed on an NVIDIA Jetson TX2 platform. The outdoor sequence was composed of 12,034 stereo pairs gathered in a public square of our city. And the experiment lasted about 15 minutes. The trajectory was about 978 *m* long from the initial position. The quantitative comparison results are shown in Table 2, which turns out the proposed stereo SLAM is more accurate and robust than ORB-SLAM2 in dynamic real city spaces.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 30 May 2023 doi:10.20944/preprints202305.2072.v1

Table 2. The ATE and RPE results of two SLAM systems.

Item	ORB-SLAM2		OurSLAM		Improvements	
	RMSE	STD	RMSE	STD	RMSE	STD
ATE	20.387	10.021	4.632	2.987	77.28%	70.19%
RPE	0.203	0.324	0.085	0.039	58.13%	87.96%

Figure 7a shows the trajectory performed by the robot platform in the city space scenario, considering our stereo SLAM algorithm (in red) and the ORB-SLAM2 system (in blue), as well as the estimated trajectory using a commercial GPS (in green). As it can be observed, the ORB-SLAM2 could not acquire an exact camera trajectory on the major road of the map, since there are many dynamic objects (e.g., cars or riders) impairing its stability. In contrast, our stereo SLAM system can cope with moving objects befittingly, the calculated trajectory is in correspondence with the real camera trajectory. And the trajectory estimated by GPS is close to the one acquired with our SLAM in most cases, however, there are some places in the sequence where the measurement errors are large. Because these situations are the areas where there are many tall buildings or dense woods. In these areas, GPS is liable to failure because of low satellite visibility conditions. Figure 7b describes the same comparison but displaying results onto an aerial image view of the sequence.



(a) Trajectories in a XZ Coordinate System (b) Trajectories in a real environment

Figure 7. Comparison of VSLAM and GPS estimated camera trajectories in urban dynamic environments. (a) Our SLAM, ORB-SLAM2 and GPS (b) Aerial image view of the sequence.

5. Conclusion

A real-time stereo visual odometry method that can cope with dynamic scenes with some independent moving objects has been proposed in this paper. The semantic segmentation and moving object detection method are incorporated into ORB-SLAM2 (stereo) system, which makes some significant performance improvements in urban outdoor high dynamic scenarios. We are capable of dealing with dynamic objects and improving considerably the performance of the visual odometry and consequently, more robust and accurate localization and mapping results are obtained in crowded and high dynamic environments. In the end, we carry out experimental results to illustrate the improved accuracy of our proposed models and the efficiency and availability of our implementation.

Whereas, there is still room for amendment. In our stereo SLAM system, the deep neural network utilized in semantic segmentation thread is the supervised method. Namely, the model may scarcely predict right results when big differences turn up between training environments and actual

scenarios. In the future, we could employ self-supervised or unsupervised deep learning means so as to deal with it.

Author Contributions: Conceptualization, Ai, Y.; methodology, Wang, X.; software, Ai, Y.; validation, Ai, Y.; formal analysis, Ai, Y.; investigation, Sun, Q.; resources, Wang, X.; data curation, Xi, Z.; writing—original draft preparation, Ai, Y.; writing—review and editing, Li, N.; visualization, Dong, J.; supervision, Wang, X.; funding acquisition, Ai, Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Military Scientific Research Project, grant number XXX.

Data Availability Statement: The data presented in this study are available in the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Davison, A.; Reid, I.; Molton, N.; Stasse, O. MonoSLAM: Realtime single camera SLAM. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1052–1067,2007.
2. Neira, J.; Davison, A.; Leonard, J. Guest editorial, special issue in visual slam. IEEE Transactions on Robotics, vol. 24, no. 5, pp.929–931, 2008.
3. Klein, G.; Murray, D. Parallel tracking and mapping for small AR workspaces. 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007.
4. Forster, C.; Pizzoli, M.; Scaramuzza, D. SVO: Fast semi-direct monocular visual odometry. IEEE International Conference on Robotics and Automation, 2014.
5. Engel, J.; Schops, T.; Cremers, D. Lsd-slam: Large-scale direct monocular slam. European conference on computer vision. Springer, pp. 834–849, 2014.
6. Mur-Artal, R.; Montiel, J. M. M.; Tardos, J. D. ORB-SLAM: A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-11631, 2015.
7. Mur-Artal, R.; Tardos, J. D. ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras. IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255–1262, 2017.
8. Campos, C.; Elvira, R.; Rodriguez, J.; Montiel, J.; Tardós, J. ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM. IEEE Transactions on Robotics, vol. 37, no. 6, 2021.
9. Panchpor, A. A.; Shue, S.; Conrad, J. M. A survey of methods for mobile robot localization and mapping in dynamic indoor environments. 2018 Conference on Signal Processing and Communication Engineering Systems, 2018.
10. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets Robotics: The KITTI Dataset. The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231-1237, 2013.
11. Bescos, B.; Facil, J. M.; Civera, J.; Neira, J. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. IEEE Robotics and Automation Letters, vol. 3, no. 4, pp. 4076–4083, 2018.
12. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. 2016.
13. Davison, A. Mobile robot navigation using active vision. Ph.D. dissertation, University Oxford, U.K., 1998.
14. Davison, A. J.; Murray, D. W. Simultaneous localization and map building using active vision. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 865–880, 2002.
15. Davison, A.; Kita, N. 3-D simultaneous localisation and map building using active vision for a robot moving on undulating terrain. IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 384–391, 2001.
16. Iocchi, L.; Konolige, K.; Bajracharya, M. Visually realistic mapping of a planar environment with stereo. International Symposium on Experimental Robotics, pp. 521–532.
17. Se, S.; Lowe, D.; Little, J. Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks. International Journal of Robotics Research, vol. 21, no. 8, pp. 735–758, 2002.
18. Jung, I.; Lacroix, S. High resolution terrain mapping using low altitude aerial stereo imagery. the 9th International Conference on Computer Vision, Nice, France, vol. 2, pp. 946–951, 2003.
19. Hygounenc, E.; Jung, I.; Soueres, P.; Lacroix, S. The autonomous blimp project of LAAS-CNRS: Achievements in flight control and terrain mapping. International Journal of Robotics Research, vol. 23, no. 4, pp. 473–511, 2004.

20. Saez, J.; Escolano, F.; Penalver, A. First Steps towards Stereobased 6DOF SLAM for the visually impaired. IEEE Conference on Computer Vision and Pattern Recognition-Workshops, Washington, vol. 3, pp. 23–23, 2005.
21. Sim, R.; Elinas, P.; Griffin, M.; Little, J. Vision-based SLAM using the Rao–Blackwellised particle filter. International Joint Conference on Artificial Intelligence-Workshop Reason, Uncertainty Robot. Edinburgh, U.K., pp. 9–16, 2005.
22. Sim, R.; Elinas, P.; Little, J. A study of the Rao–Blackwellised particle filter for efficient and accurate vision-based SLAM. International Journal of Computer Vision, vol. 74, no. 3, pp. 303–318, 2007.
23. Paz, L. M.; PiniÉs, P.; TardÓs, J. D.; Neira, J. Large-Scale 6-DOF SLAM With Stereo-in-Hand. IEEE Transactions on Robotics, , vol. 24, no. 5, pp.946-957, 2008.
24. Lin, K. H.; Wang, C. C. Stereo-based simultaneous localization, mapping and moving object tracking. 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, Taipei, Taiwan, 2010.
25. Kawewong, A.; Tongprasit, N.; Tangruamsub, S.; Hasegawa, O. Online and Incremental Appearance-based SLAM in Highly Dynamic Environments. The International Journal of Robotics Research, vol. 30, no. 1, pp.33-55, 2011.
26. Alcantarilla, P. F.; Yebes, J. J.; Almazán, J.; Bergasa, L.M. On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments. IEEE International Conference on Robotics and Automation, pp. 1290-1297, 2012.
27. Kaess, M.; Ni, K.; Dellaert., F. Flow separation for fast and robust stereo odometry. IEEE International Conference on Robotics and Automation, pp. 3539-3544, 2009.
28. Karlsson, N.; Di Bernardo, E.; Ostrowski, J.; Goncalves, L.; Pirjanian, P.; Munich, M. E. The vSLAM algorithm for robust localization and mapping. IEEE International Conference on Robotics and Automation, pp. 24-29, 2005.
29. Zou, D.; Tan. P. CoSLAM: Collaborative Visual SLAM in Dynamic Environments. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 2, pp.354-366, 2013.
30. Fan, Y.; Han, H.; Tang, Y.; Zhi, T. Dynamic objects elimination in SLAM based on image fusion. Pattern Recognition Letters, vol. 127, no. 1, pp. 191-201, 2019.
31. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask r-cnn. IEEE International Conference on Computer Vision, pp. 2961–2969, 2017.
32. Sun, T.; Sun, Y.; Liu, M.; Yeung, D. Y. Movable-Object-Aware Visual SLAM via Weakly Supervised Semantic Segmentation. 2019.
33. Balntas,V.; Riba, E.; Ponsa, D.; Mikolajczyk, K. Learning local feature descriptors with triplets and shallow convolutional neural networks. British Machine Vision Conference, vol. 1, pp. 3, 2016.
34. Kang, R.; Shi, J.; Li, X.; Liu, Y. DF-SLAM: A Deep-Learning Enhanced Visual SLAM System based on Deep Local Features. 2019.
35. Li, P.; Qin, T.; Shen, S. Stereo Vision-based Semantic 3D Object and Ego-motion Tracking for Autonomous Driving. European Conference on Computer Vision, 2018.
36. Ai, Y.; Rui, T.; Lu, M.; Fu, L.; Liu S.; Wang, S. DDL-SLAM: A robust RGB-D SLAM in dynamic environments combined with Deep Learning. IEEE Access, vol. 8, pp. 162335-162342, 2020.
37. Yu, C.; Liu, Z. X.; Liu, X. J.; Xie, F. G.; Yang, Y.; Wei, Q.; Qiao, F. DS-SLAM: A Semantic Visual SLAM towards Dynamic Environments. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2018.
38. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The Cityscapes Dataset for Semantic Urban Scene Understanding. IEEE Conference on Computer Vision and Pattern Recognition, 2016.
39. Zhao, Y.; Shi, H.; Chen, X.; Li, X.; Wang, C. An overview of object detection and tracking. 2015 IEEE International Conference on Information and Automation, 2015.
40. Lucas, B. D.; Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision. the 7*th* International Joint Conference on Artificial Intelligence, 1997.
41. Shi, J.; Tomasi, C. Good Features to Track. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 600, 2000.
42. Sumikura, S.; Shibuya, M.; Sakurada, K. OpenVSLAM: A Versatile Visual SLAM Framework. the 27th ACM International Conference on Multimedia, pp.2292-2295, 2019.

 doi:10.20944/preprints202305.2072.v1