Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/15/20230830

Semantic information based solution for visual SLAM in dynamic environment

Chuqi Shao

College of Mathematics and informatics College of Software Engineering, South China Agricultural University, Guangzhou 510642, China

chuqi@stu.scau.edu.cn

Abstract. In recent years, the utilization of visual SLAM with a camera as a sensor has become increasingly widespread, particularly in the context of rapidly developing artificial intelligence such as mobile robots, VR, and AR. This approach is favored due to its affordability, lightweight design, ability to capture comprehensive information, and other advantages. The traditional slam technology has achieved a very mature effect with the static scene as the assumption condition, and the classic representatives are LSD-SLAM and ORB-SLAM which will be introduced in the following. However, since dynamic scenarios are unavoidable in the real world, overcoming the influence of dynamic objects becomes a challenge if researchers want to move forward with more applications. At present, under the dynamic environment of the visual slam algorithm faces positioning problems of low accuracy and poor robustness. To address the challenges posed by dynamic objects in a scene, many researchers are incorporating deep learning techniques and exploring the use of reference semantic information to collaboratively resolve the issue. This paper reviews this and summarizes the development process and important algorithms.

Keywords: visual SLAM, semantic segmentation, dynamic environment.

1. Introduction

SLAM, which stands for Simultaneous Localization and Mapping, is a crucial and essential function of robotic applications. Building a map of an unexplored area based on data from the sensors is a key objective of SLAM, which aims to accomplish this task while minimizing the system's weight. SLAM is mainly used in robot navigation, autopilot, Augmented Reality(AR), and other areas of the widely. Depending on different types of sensors, SLAM can be split in two different directions. One is laser SLAM based on Lidar, and the other is visual SLAM implemented by cameras. Visual SLAM has a wider application than laser sensors since the camera can capture more environmental information for the system. In other words, its core is to obtain RGB and depth information.

Most visual SLAM methods are basically assumed to be in a static environment with no dynamic objects so as to facilitate the implementation of the algorithm and excellent performance has been achieved in this way. Such ideal environments limit most applications of visual SLAM systems under dynamic environments. Currently, accurately and reliably providing real-time information about the position or whereabouts of objects in physical settings is a significant challenge for nearly all existing visual SLAM systems due to their inability to handle dynamic targets in dynamic environments.

DOI: 10.54254/2755-2721/15/20230830

Moving objects cause errors in the calculation of camera motion in dynamic environments, which ultimately leads to low localization accuracy and poor robustness in the system. In addition, the emergence of dynamic objects will also affect the ability to efficiently and accurately process sensor data and provide updated estimates of the camera pose and map, increasing the computing cost and delay.

To solve the above problems, the key lies in the correct detection of dynamic targets in the visual field, and how to deal with the dynamic target subsequently. Currently, several solutions have been proposed for visual SLAM techniques to deal with scenes that contain moving objects or changing lighting conditions. Facing the additional challenge of operating in a state of flux, Cheng J et al. proposed DM-SLAM, which is a combination of optical flow information and instance segmentation network [1]. Tete Ji proposed that RGB-D SLAM uses semantic information to identify dynamic objects and eliminate their effects, and only maintains static maps containing camera tracking [2]. Based on the characteristics of point and line dynamic scene of PLD - SLAM calculates camera position, the robustness and position precision are also improved [3]. In addition, RDS-SLAM uses semantic segmentation results to detect dynamic targets and remove any abnormal or irrelevant data points, while maintaining the whole process in real-time [4].

In short, the challenges visual SLAM faces in dynamic environments are constantly being proposed, although it is facing great challenges. In addition to these methods, several papers have proposed new frameworks for both localization and semantic segmentation, improving their performance through the intermediate results of the two modules. These new methods and frameworks are developing and improving constantly, which provides more ideas and directions for solving the challenges.

The rest of the paper is structured as follows: Sect. 2 describes the structure of the visual slam and the details of each part. In Sect 3: LSD-SLAM and three generations of OGR-SLAM are introduced in detail, which are visual SLAM in a static environment represented by the direct method and feature method respectively. Section 4 is to summarize the current visual slam algorithm combined with semantic information in the dynamic environment. The final section provides conclusions and a discussion about the challenges that still need to be addressed in the future.

2. Visual SLAM

The basic framework structure of the visual SLAM algorithm can be seen in figure1. Visual SLAM has the capability of initiating from an unknown location in an unfamiliar environment based on mobile devices, such as robots, drones, and mobile phones, observing and locating its own position and posture through the camera in the process of movement, and then build incremental maps according to its posture. The ultimate goal is to achieve positioning and map construction at the same time.

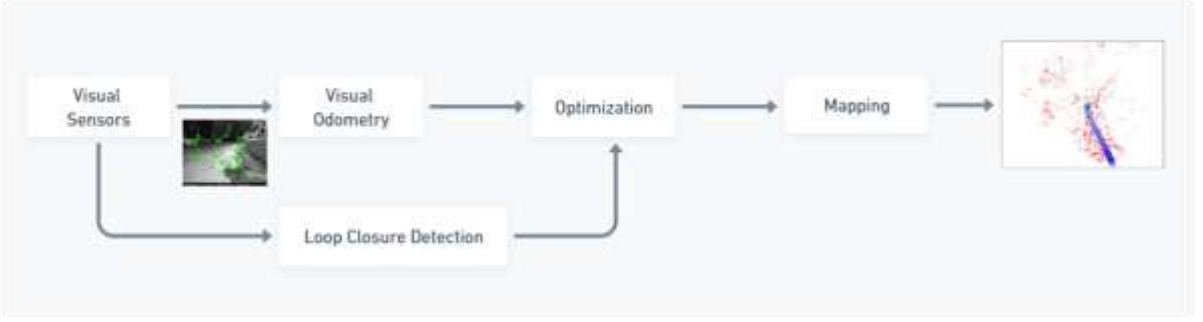

Figure 1. The structure of visual slam.

2.1. Visual sensors

Visual SLAM technology relies on a camera as its primary sensor to perceive and gather information about the surrounding environment, allowing for the simultaneous determination of the robot, vehicle,

DOI: 10.54254/2755-2721/15/20230830

or camera's position in real time. This approach is cheap, lightweight, and versatile, making it an attractive option for a wide range of applications. At present, there have been various types of cameras used in practice and research, including but not limited to monoculars, stereo, RGB-D, pinhole, and fisheye cameras. Visual SLAM systems can even run in micro PCS and embedded devices, and it can be implemented with only a single monocular camera, which is the most cost-effective and smallest sensor device. The mainstream representative single-purpose Visual SLAM methods can be categorized into three groups: Mono-SLAM [5] which uses a filter, PTAM [6] based on keyframe BA and LSD-SLAM [7] based on direct tracking. However, the biggest drawback of a monocular camera is that depth information (distance) cannot be obtained from a single image, and there will be scale drift. The emergence of a binocular camera just solves this problem. It uses the parallax of left and right objects to calculate the distance of pixels, to realize its positioning. RGB-D cameras are preferable to monocular cameras as they can offer more accurate depth information, which enhances the robustness and accuracy of visual SLAM.

2.2. Visual odometry

The visual odometer technique is a process in computer vision where the camera's movement and position are approximated through several multiple consecutive images. Due to the different types of sensors, Visual Odometry can be divided into three categories: direct method, feature point method, and RGB method. The Feature-based approach is the dominant method of the visual odometer, which involves the following three steps: identifying and extracting significant features from consecutive images, matching the features between the images to track their motion, and estimating the camera's movement by optimizing the feature point correspondences. Feature point extraction and matching are often implemented using descriptors and key points, such as ORB-SLAM, and PTAM. Completely different from the feature points, the direct method bypasses the feature point extraction step and instead directly uses the pixel value of consecutive images, specifically the gray-level intensity values of pixels, to estimate the camera's motion. It is considered a complementary approach to the feature points, such as LSD-SLAM. In addition, more scene information can be provided based on the depth information and color(RGB) provided by the RGB-D sensor. However, due to the high cost and high condition requirements, the application is relatively rare. Some algorithms fuse the feature method with the direct method, such as SVO [8].

2.3. Loop closure

Loop closure also referred to as closed-loop detection, is a critical component of Visual SLAM which is utilized to distinguish whether the current observation is from a repeated arrival position or an unexplored location. Davison et al [5]. found the problem of cumulative errors in the MonoSLAM algorithm during the experiment, which caused the deviation and inaccuracy of the SLAM trajectory. This is later referred to as the cumulative error. At present, the popular loop closure method is Bag of Words (BoW) which is generally applied for feature points such as ORB-SLAM. However, this method uses extracted features to judge whether the situation is the same or not, while the direct method does not extract features. If loop closure is needed, the model based on the direct method may need to extract additional features. There are also cases like DSO [9] that are not a complete slam because it lacks loopback detection, resulting in it not eliminating the cumulative error, albeit small.

2.4. Optimization

Optimization in the way of implementation is mainly two kinds of filter method and nonlinear optimization method, also known as the back-end. This module is optimized for receiving information about camera poses from visual odometry measurements at different times, as well as loop detection, resulting in a globally consistent map. The filter optimization method includes the Kalman filter and extended Kalman filter, while the nonlinear optimization method includes Bundle Adjustment(BA), PoseGraph optimization, and factor graph optimization. Due to the existence of frame loss, this sparsity, reflected in the matrix operation can be solved by mathematical techniques such as

DOI: 10.54254/2755-2721/15/20230830

elimination, which makes the nonlinear optimization method can be applied to a real-time SLAM system, called graph optimization. PTAM, ORB-SLAM, LSD-SLAM, and so on are all graph-optimized backends.

3. Classic approaches

Traditional methods can be broadly classified into two categories: direct method and feature-based method, both of which use image information to process problems. Next, the direct approach leading to semi-dense and dense constructs in the case of LSD-SLAM, and the feature-based approach leading to sparse constructs using the three generations of ORB-SLAM as an example, are presented respectively.

3.1. LSD-SLAM

A direct method-based monocular SLAM algorithm, represented by LSD-SLAM, can construct large-scale and consistent semi-dense maps. It exploits the characteristic of monocular slam, namely scale-ambiguity, to seamlessly switch between environments of different scales. LSD-SLAM is based on the direct method to match image feature points, that is, to perform direct, scale-shift-aware image alignment on sim (3), avoiding the problem of scale drift. Because the direct method is to use the pixel value of the image to match, independent of the feature points [7]. The traditional feature method can extract feature points and calculate descriptors. This method can ensure the accuracy of matching to a certain extent, but it also leads to scaling uncertainty.

The algorithm typically comprises three primary components: tracking, depth map estimation, and map optimization. After obtaining the camera pose and map point position in the initialization phase, LSD-SLAM utilizes the direct method for tracking, which involves calculating the camera's motion through the displacement information between the current image captured by the sensor and the previous frame. During the depth map estimation process, the depth of map points is calculated by the triangulation method, and the estimated value of the depth map is updated by optimizing the position and orientation of the camera in the environment and the map point position. In the map optimization process, LSD-SLAM uses direct rendering for map optimization. The whole process uses the local luminosity error as the matching measure to calculate the data error. This new direct monocular slam algorithm shows more functionality, robustness, and flexibility.

3.2. ORB-SLAM

Visual SLAM can be executed with just one monocular camera, which is the most affordable and compact, but also versatile enough to function using a wide range of settings. ORB-SLAM improves on PTAM's algorithmic framework and contains three threads running in parallel: tracking, local mapping, and loop closure. ORB-SLAM [10] is a monocular complete SLAM system that is solely based on sparse feature points, and its fundamental principle is to employ Oriented FAST and BRIEF(ORB) as the key feature of the entire visual SLAM, which makes the system simpler and more robust. The ORB is extremely fast to calculate and match, with a good point-of-view invariance.

ORB-SLAM2 [11] added binocular stereo vision and RGB-D based on ORB-SLAM1 single purpose, which is a set of support for monocular, binocular, and RGB-D complete program, with three main parallel threads: map reuse, loop closing, relocation. An advanced version of ORB-SLAM2 and ORBSLAM-VI is ORB-SLAM3, which is a visual SLAM system that supports vision, vision-plus navigation, and hybrid maps, and can be operated on a variety of cameras [12]. The first major innovation refers to a tightly integrated Visual-Inertial SLAM system that can real-time closed loop and favors the map's sensors over already mapped areas. ORB-SLAM3 was the first visual system and visual inertia system that can accurately match current sensor data with previously mapped data at different time scales, across multiple maps, which was also the key to accuracy. In summary, Compared with other methods of the most advanced monocular SLAM, ORB-SLAM achieved unprecedented performance.

DOI: 10.54254/2755-2721/15/20230830

4. Deep learning

Most traditional slam algorithms, including those mentioned above, assume static scenarios, while dynamic objects cannot be avoided in real scenes, especially to the slam is applied to more scenes. With the rapid development of deep learning, if deep learning technology can be used to deal with dynamic object problems, there will be better development and research direction. In recent years, there are some research on feature extraction and motion estimation, especially on semantic information and depth information. The existing algorithm research combined with semantic information visual SLAM algorithm in recent years is summarized in Table 1. The fr3_walking_xyz and fr3_walking_static in the TUM RGB-D dataset and the Average Displacement (RMSE) error in the KITTI dataset were used as the basis for judging the excellence of the algorithm.

Table 1. Comparison of slam algorithms based on semantic information.

Year	SLAM	Author	Character	KITTI	fr3_walking_xyz	fr3_walking_static
2020	DM-SLAM [1]	Junhao Cheng	Feature-based, Support for monocular, stereo, and RGB-D sensors	2.190	0.0148	0.0079
2020	PSPNet-SLAM[13]	Zhihong Xi	Pyramid scene analysis network, Dynamic scene based on semantic segmentation	-	0.016	0.008
2020	SaD-SLAM [14]	Xun Yuan, Song Chen	Semantic and deep information, Based on ORB-SLAM2	-	0.0167	0.0166
2021	PLD-SLAM [3]	Chengyang Zhang	Point and line features, RGB-D dynamic SLAM method	-	0.0144	0.0065
2021	RDS-SLAM [4]	Yubao Liu	Semantic segmentation method, ORB-SLAM3 real-time visual SLAM	-	0.0269	0.0221
2022	STDC-SLAM [15]	Zgfang Hu	Real-time Semantic SLAM system, ORB-SLAM3, and Qtree-ORB algorithm framework, STDC network	-	0.018	-
2022	STDyn-SLAM [16]	Daniela Esparza	Stereo vision, Dynamic outdoor environment, Semantic segmentation	1.382	-	-

There are some novel visual SLAM techniques that achieve excellent performance in highly dynamic environments, regardless of the sensor used, such as DM-SLAM which incorporates optical flow and semantic masks. It leverages semantic segmentation, self-motion estimation, and dynamic

DOI: 10.54254/2755-2721/15/20230830

point detection in conjunction with a feature-based SLAM framework to mitigate the impact of dynamic objects [1]. Similarly, RDS-SLAM [4] uses dependable feature points to estimate the state of the camera, which is based on the feature points in the static state of the movable object. Moreover, it also uses Mask_Rcnn to obtain semantic information and depth information from the RGB-D camera to discard moving feature points. Chenyang Zhang [3] also proposed that in the framework of RGB-D SLAM, the point-and-line features are used to calculate the posture of the camera in dynamic SLAM, and the semantic segmentation network, named MobileNet, and K-Means algorithm are combined to remove the dynamic features in the scene, which has improved the effect to some extent. Based on ORB-SLAM2, PSPNet-SLAM refers to a parallel semantic thread PSPNet for semantic segmentation at the pixel level. Through pyramidal network structure, it is more effective to obtain more context information and the relation between objects in pixels than the detection based on dynamic feature points. The algorithm also proposes a reverse ant colony search strategy that utilizes dynamic point community distribution to identify and use the most relevant and informative points, which enhances the robustness and achieves accurate and responsive visual SLAM in real-world applications [13].

Such methods typically have an architecture that requires waiting for semantic results in the tracing thread, and processing times that depend on the segmentation method used are not very friendly to demanding applications, such as tasks that need to be completed in real-time. Zgfang Hu et al. [15] proposed an approach that involves incorporating a semantic tracking thread and a semantic-based optimization thread into ORB-SLAM3 for improved performance. The key frame selection strategy of this design is adopted to obtain the latest semantic information to the maximum extent, and the processing of the segmentation method of different speeds, to ensure that the new thread and tracking thread run in parallel and maintain the real-time effect. STDCyn-SLAM [16] is also a real-time system and uses STDC as a semantic segmentation network for semantic thread analysis. Then the dynamic object segmentation graph is obtained. A refinement module is designed to improve semantic segmentation mapping by utilizing image depth information, which is superior to PSPNet-SLAM in localization accuracy and processing speed. Although using a semantic segmentation network can improve the precision and correctness of the system, the segmentation accuracy needs to be improved, as it is easy to lead to tracking faults where the surroundings are rapidly and constantly changing.

5. Conclusion

Nowadays Visual SLAM combined with deep learning technology has become one of the hot research fields. This article introduces the classic framework of visual SLAM, including sensors, the visual odometer, back-end optimization, and loopback detection. Two classical models, LSD-SLAM based on the direct method and ORG-SLAM based on the feature point method, are also introduced. For dynamic environments, there are research and discussions based on deep learning technology to solve dynamic objects. By using semantic information, pixels in the image are divided into different categories, such as sky, vehicle, and floor, which can better account for the presence of moving subjects and improve the precision and robustness of positioning and built figure. In addition, deep learning can also be used to improve feature extraction and motion estimation. For example, feature extraction based on Convolutional Neural Networks(CNN) can better deal with visual SLAM in dynamic scenes, and Recurrent Neural Networks (RNN) can be used to predict the trajectory of dynamic objects.

To some extent, using semantic information to eliminate dynamic objects improves robustness and accuracy in dynamic scenes. However, breakthroughs are still needed in the following aspects in the future. Firstly, the discernable types of moving objects are limited to some extent. Secondly, many static potentials moving objects such as parked vehicles are not well processed. Then, the accuracy of the semantic segmentation network needs to be enhanced, especially in a highly dynamic environment, which is easy to lead to tracking faults. Finally, the running speed is relatively slow, and improving the real-time performance would enable the system to process and analyze data more quickly, allowing for faster adaptation to changes in the environment.

DOI: 10.54254/2755-2721/15/20230830

References

[1] Cheng J, Wang Z, Zhou H, Li L and Yao J 2020 DM-SLAM: A Feature-Based SLAM System for Rigid Dynamic Scenes. ISPRS International Journal of Geo-Information; 9(4):202

[2] Ji T, Wang C and Xie L 2021 "Towards Real-time Semantic RGB-D SLAM in Dynamic Environments," 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, pp. 11175-11181

[3] Zhang C, Huang T, Zhang R and Yi X 2021 "PLD-SLAM: A New RGB-D SLAM Method with Point and Line Features for Indoor Dynamic Scene," ISPRS International Journal of Geo-Information. 2021; 10(3):163

[4] Liu Y and Miura J 2021 "RDS-SLAM: Real-Time Dynamic SLAM Using Semantic Segmentation Methods," IEEE Access, 9, 23772-23785

[5] A. J. Davison, I. D. Reid, N. D. Molton and O. Stasse 2007 "MonoSLAM: Real-Time Single Camera SLAM," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1052-1067

[6] Klein G and Murray D 2007 "Parallel Tracking and Mapping for Small AR Workspaces," 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan, pp. 225-234

[7] ngel, J., Schöps, T., and Cremers, D. 2014 "LSD-SLAM: Large-Scale Direct Monocular SLAM," In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8690. Springer, Cham.

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

[9] J. Engel, V. Koltun and D. Cremers 2018 "Direct Sparse Odometry," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 611-625, 1

[10] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós 2015 "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," in IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163

[11] R. Mur-Artal and J. D. Tardós 2017 "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," in IEEE Transactions on Robotics, vol. 33, no. 5, pp. 1255-1262

[12] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel and J. D. Tardós 2021 "ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM," in IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874-1890

[13] X. Long, W. Zhang and B. Zhao 2020 "PSPNet-SLAM: A Semantic SLAM Detect Dynamic Object by Pyramid Scene Parsing Network," in IEEE Access, vol. 8, pp. 214685-214695

[14] X. Yuan and S. Chen 2020 "SaD-SLAM: A Visual SLAM Based on Semantic and Depth Information," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, pp. 4930-4935

[15] Z. Hu, J. Chen, Y. Luo and Y. Zhang, 2022 "STDC-SLAM: A Real-Time Semantic SLAM Detect Object by Short-Term Dense Concatenate Network," in IEEE Access, vol. 10, pp. 129419-129428

[16] Esparza D and Flores G 2022 "The STDyn-SLAM: A Stereo Vision and Semantic Segmentation Approach for VSLAM in Dynamic Outdoor Environments," in IEEE Access, vol. 10, pp. 18201-18209