ISSN: 2710-0588

A Review of SLAM Research on Mobile Robot Vision

Longfei Wei, Hongchang Sun, Guoqing Qiao, Xiangyan Wu, Hang Xu

School of Mechanical Engineering, Tianjin University of Technology and Education, Tianjin 300350, China

Abstract

Simultaneous Localization and Mapping (SLAM) refers to the process of constructing an unknown environmental map while locating the position of an intelligent agent. In recent years, visual sensors have shown significant performance, accuracy, and efficiency in simultaneous localization and mapping systems. The main purpose of this paper is to introduce the progress of visual SLAM systems, introduce representative classic visual SLAM methods and multi-sensor fusion methods, summarize some of the problems existing in the visual SLAM system discussed in the article, and finally explore the hot research directions and development prospects of visual SLAM in the future.

Keywords

Visual SLAM, Feature extraction, Multi-sensor fusion.

1. Introduction

Simultaneous Localization and Mapping (SLAM) refers to the process of constructing an unknown environmental map while locating the position of an intelligent agent[1]. SLAM technology was proposed by R. C. Smith et al[2] at the 1986 IEEE Robotics and Automation Conference and is considered a key technology for achieving autonomous robot movement. SLAM technology enables autonomous movement of robots in the presence of unknown positioning beacons [3]. Simultaneous Localization and Mapping (SLAM) is an important research topic in the fields of robotics and computer vision. Its core task is to estimate the real-time position of intelligent agents in unknown environments and construct environmental maps.

In recent years, with the advancement of sensor technology and the improvement of computing power, vision based SLAM (VSLAM) methods have gradually become a research hotspot. These methods utilize visual information obtained from cameras, which can not only achieve high-precision positioning, but also generate rich environmental maps. Visual SLAM has advantages such as low cost and rich data compared to traditional SLAM methods, but it also faces challenges such as lighting changes, dynamic object interference, and feature extraction. Unlike sensors such as LiDAR, the processing of visual information is often more complex, which requires more refined considerations in algorithm design and optimization for visual SLAM systems. This article will describe the development history and main methods of VSLAM system, covering classic visual SLAM methods as well as fusion schemes with other sensors or algorithms. Finally, the article will look forward to the future research trends and development directions of VSLAM, providing reference for both academia and industry.

2. The Development of Visual SLAM

In 2007, Davison et al[4] first proposed the MonoSLAM visual SLAM system based on monocular cameras, which is implemented using the Extended Kalman Filter (EKF). The core method is to add poses to the EKF to update the mean and variance, enabling the system to operate in real-time and accurately. In the same year, Mourikis et al[5] also fused visual and inertial

ISSN: 2710-0588

measurement unit (IMU) data based on the EKF algorithm and proposed the Multi State Constraint Kalman Filter (MSCKF) system. This method only takes the motion parameters between adjacent frames as state variables and converts the constraints of pose and feature points into pose constraints, greatly reducing computational complexity. Early visual SLAM based on filter based methods only considered the constraints of system states at adjacent moments for estimation, without taking into account the constraints between global states.

With the development of nonlinear optimization and the recognition of the sparsity of incremental equation coefficient matrices, visual SLAM based on nonlinear optimization has gradually become the mainstream research direction. The visual SLAM model based on nonlinear optimization establishes motion and observation equations, treats all camera poses and spatial points as variables, and constructs an error function. The error is minimized by beam adjustment to reduce the accumulation of errors. The PTAM (parallel tracking and mapping) system proposed by Georg Klein et al[6] for the first time divides visual SLAM into two parts: tracking and mapping, and adopts nonlinear optimization based on keyframes to process monocular visual data, which has extremely high real-time performance. This multi-threaded standard has been recognized by many subsequent works, and this article will discuss it. The main idea of their method is to reduce computational costs and apply parallel processing to achieve real-time performance. When the tracking thread estimates camera motion in real-time, the mapping thread predicts the 3D position of feature points. PTAM is also the first method to jointly optimize camera pose and create 3D maps using beam adjustment (BA). It uses the FAST[7] corner detector algorithm for keypoint matching and tracking. Although the performance of this algorithm is better than Mono SLAM, its design is complex, requiring users to manually set it in the first stage, and tracking is prone to loss, making it not suitable for large scenarios.

In 2011, Peter Henry et al[8]. from MIT first proposed a depth camera based SLAM algorithm using Kinect camera as the visual sensor for visual SLAM. This algorithm uses SIFT as the feature extraction part and uses ICP iterative visual feature points to solve the pose and achieve the construction of indoor environment. However, the author did not verify it in a large scene and the real-time performance is poor. Afterwards, the team made improvements to the feature extraction part[9], abandoning the computationally complex SIFT algorithm and using the lower complexity SURF algorithm. They also used the ICP iterative method to solve the pose, and finally used the pose graph optimization method to obtain the global optimal pose, improving the efficiency of the system. Later, the team proposed a block based approach combined with GPU for accelerated processing[10], further improving the speed of 3D scene construction. In the same year, Endres et al[11] proposed the RGBD-SLAM scheme, which uses SURF algorithm and Random Sample Consensus (RANSAC) algorithm for feature extraction and screening. By iterating the nearest point algorithm for inter frame pose estimation, real-time mapping on GPU can be achieved. Newcombe et al. introduced a direct method called Dense Tracking and Mapping (DTAM) in the same year for measuring depth values and motion parameters to construct maps. DTAM is a real-time framework with dense mapping and tracking, which can determine camera pose by aligning the entire frame with a given depth map. The above stages estimate the depth and motion parameters of the scene to form a map. Although this algorithm can provide detailed maps, it requires high computational costs.

In 2013, Endres et al. proposed a method based on RGB-D cameras. This method is executed in real-time and focuses on low-cost embedded systems and small robots, but it is difficult to produce accurate results in featureless or challenging scenarios. In the same year, Salas Moreno et al[12] proposed the first attempt to utilize semantic information in a real-time SLAM framework, called SLAM++. Their system uses RGB-D sensor output and performs 3D camera pose estimation and tracking to form a pose graph. The nodes in the pose graph represent pose estimation and are connected by edges representing the relative poses between nodes with

ISSN: 2710-0588

measurement uncertainty[13]. Then, the predicted pose will be optimized by merging the relative 3D poses obtained from semantic objects in the scene.

As the basic framework of visual SLAM matures, researchers focus on the performance and accuracy of the system. In this regard, Forster et al. proposed a hybrid VO method as part of a visual SLAM system in 2014, called semi direct visual odometry (SVO)[14]. Their approach can combine feature-based methods and direct methods for sensor motion estimation and mapping tasks. SVO can work with monocular and stereo cameras and is equipped with a pose refinement module to minimize reprojection errors. However, the main drawback of SVO is the use of short-term data association and the inability to perform loop detection and global optimization.

In 2014, LSD-SLAM[15] was introduced by Engel et al., which includes tracking, depth map estimation, and map optimization. This method can use its pose estimation module to reconstruct large-scale maps and has global optimization and loop detection functions. The weakness of this method lies in its challenging initialization phase, which requires all points in the plane, making it a computationally intensive approach. In the same year, Endres et al[16] improved RGBD-SLAM using the ORB (Oriented FAST and Rotated BRIEF, ORB) algorithm, enabling it to bypass GPU acceleration and achieve real-time processing on the CPU. At the 2015 IEEE Transactions on Robotics conference, Mur Artal et al[17] proposed the first generation ORB-SLAM system, which comprehensively implemented the keyframe based monocular SLAM algorithm for the first time. The system was innovatively divided into three parallel threads: tracking, map creation, and closed-loop control. The ORB feature extraction algorithm was used for feature extraction and matching, as well as sparse map creation. On the basis of real-time operation, it also had good accuracy.

In 2016, Jakob Engel proposed DSO[18] (direct sparse odometry), which is a sparse direct visual odometry based on monocular cameras. It outperforms LSD-SLAM and ORB-SLAM in terms of robustness, accuracy, and speed. Unlike the feature point method, which requires finding which frames a certain feature point matches, DSO projects each point into all frames, calculates its residuals in each frame, and considers points with residuals within the same reasonable range as projections of the same point, without caring about the correspondence between these points. In general, the system runs quickly and can generate very dense point clouds. The backend is a sliding window composed of multiple keyframes, which requires maintenance of the sliding window and optimization related structures. In order to reduce the impact of lighting on the performance of visual SLAM, DSO proposed photometric calibration, which can reduce the influence of lighting on tracking and effectively improve the robustness of the system. Although DSO has significant advantages in speed compared to previous methods, like SVO, DSO does not have closed-loop detection, making it difficult to reposition after tracking failure. In 2017, Mur Artal et al. developed an upgraded version of ORB-SLAM2[19] based on ORB-SLAM, which achieved compatibility between binocular cameras and RGB-D cameras. In 2020, Campos et al[20] developed the ORB-SLAM3 system based on ORB-SLAM2, which summarized the relevant work of the research group in recent years, added IMU module and Atlas atlas functions, and was compatible with wide-angle fisheye cameras.

3. Multi sensor Fusion Visual SLAM Method

The research on multi-sensor fusion for localization and mapping can be roughly divided into filtering based fusion frameworks and graph optimization based fusion frameworks. Mourikis et al[21] first proposed the multi-sensor fusion framework MSCKF based on filtering methods in 2007. Previous studies mostly added feature points to the state vector and estimated them together with the IMU state. However, MSCKF added camera poses to the state vector and only

ISSN: 2710-0588

maintained a specific number of camera poses, solving the problem of exploding state vector dimensions.

The idea of multi-sensor fusion based on graph optimization originates from SFM[22], which mainly constructs an error function through matching between images, and then optimizes the error function to recover the camera's motion. Zhang et al. proposed a fusion framework V-LOAM for visual odometry and laser odometry. In this system, the visual odometry can handle high-speed motion, while the laser odometry can maintain low drift and robustness of the system under poor lighting conditions. In 2017, Professor Shen Shaojie's team at the Hong Kong University of Science and Technology developed the VINS-MONO visual SLAM system[23]. This algorithm minimizes photometric error matching for feature points, and its backend nonlinear optimization is based on sliding windows to optimize the visual state variables and IMU errors and biases. Accurate pose estimation is achieved through tight fusion optimization[24]. By tightly integrating visual information and IMU data, IMU data can be used to supplement pose estimation when visual features are insufficient, while visual information can correct IMU bias drift and provide repositioning and map reuse functions[25]. This algorithm also provides new ideas for subsequent SLAM multi-sensor fusion research.

R3LIVE is a novel fusion framework of lidar inertial visual sensors, which utilizes measurements from lidar, inertial, and visual sensors to achieve robust and accurate state estimation. R3LIVE is built upon the previous R2LIVE and consists of two subsystems: radar inertial odometer (LIO) and visual inertial odometer (VIO). The LIO subsystem (FAST-LIO) utilizes measurements from LiDAR and inertial sensors to establish the geometric structure of the global map (i.e., the positions of 3D points). The VIO subsystem utilizes data from visual inertial sensors to present texture features of 3D maps (i.e., colors of 3D points). In 2021, Tixiaoshan, the author of liosam, proposed LVIS-SAM[26] under the loose coupling of liosam and Vins mono to achieve multi-sensor fusion under factor graph optimization method. Compared with the fusion framework of filtering, graph optimization based methods have poor real-time performance. However, for indoor or low-speed outdoor scenes that do not require high real-time requirements, graph optimization based fusion methods have better positioning accuracy and higher robustness. Therefore, this project mainly focuses on the research of multi-sensor fusion positioning and mapping based on the idea of graph optimization.

4. The Main Problems of Visual SLAM

In visual SLAM (Simultaneous Localization and Map Building), feature extraction and matching are the core steps, which directly affect the performance and accuracy of the system. However, this field faces multiple technological challenges, and the main issues and their impacts of visual SLAM will be discussed in detail below.

Visual SLAM needs to run in real-time environments, so the computational complexity of the algorithm is crucial. Feature matching typically involves a large number of feature points and descriptor comparisons, which is particularly evident in high-resolution images and large-scale scenes. As the number of features increases, the time cost of the matching process also increases, which may lead to slow system response and require efficient matching algorithms to maintain real-time performance. Therefore, high-performance matching algorithms are an important research topic. In practical applications, feature matching is often affected by occlusion, noise, and other interference factors. For example, when there are obstructions in the scene, some features may not be detected, resulting in matching failure. In addition, noise in the image can affect the quality of feature points, thereby affecting the accuracy of matching. Therefore, improving the robustness of feature matching is the key to ensuring the reliability of visual SLAM systems.

ISSN: 2710-0588

Feature descriptors are used to describe extracted feature points for matching purposes. Different feature descriptors have their own advantages and disadvantages, such as SIFT, SURF, and ORB. In specific application scenarios, a descriptor may perform well, but in other cases it may not be applicable. For example, although high-dimensional descriptors contain more information, they also have higher computational and storage costs. Therefore, selecting appropriate feature descriptors and optimizing them in different scenarios is an important task in feature matching. Establishing the correlation of feature data in multiple frames of images is another major challenge in feature matching. Especially in long running SLAM systems, loop detection (i.e. identifying previously visited areas) is particularly important. Successful loop detection can effectively reduce accumulated errors and optimize the entire map. However, the similarity and repetition between feature points may lead to erroneous associations, increasing the difficulty of implementing closed-loop detection.

5. The Future Development of Visual SLAM

5.1. Integrating deep learning

The core of visual SLAM lies in algorithms such as feature extraction, matching, and data association. Currently, many researchers are committed to improving the efficiency and accuracy of these algorithms. Traditional feature detection methods such as SIFT and ORB have their limitations, while deep learning based feature learning methods can automatically learn and extract features in complex environments, improving the accuracy of matching. Data association is a crucial step in SLAM, especially in loop detection and dynamic environments. Future research can focus on using graph neural network methods to more accurately handle the relationships between feature points, thereby enhancing the system's adaptive capabilities.

The success of deep learning technology in the field of computer vision has brought new opportunities for visual SLAM. By building end-to-end deep learning models, it is possible to directly predict poses and maps from raw images. This method not only simplifies the traditional SLAM process, but also improves the robustness and accuracy of the system to a certain extent. We can explore the use of Convolutional Neural Networks (CNNs) for sparse and dense reconstruction, in order to obtain richer scene information at different scales. This method can effectively improve the details and accuracy of maps. By combining deep learning with traditional SLAM methods, the features extracted by deep learning can be utilized to enhance the performance of traditional algorithms, thereby achieving better performance. In order to improve the real-time performance of SLAM systems, emphasis is placed on the application of multi threading and parallel computing technologies, and overall processing speed is enhanced through GPU acceleration and cloud computing.

5.2. Multi sensor fusion

A single sensor visual SLAM may perform poorly under certain conditions, especially in low light or complex environments. Therefore, multi-sensor fusion will be an important development direction for future visual SLAM. Combining multiple sensors such as LiDAR and IMU can provide redundant information in different environments, improving the accuracy and robustness of positioning and map construction. The future SLAM system will place greater emphasis on the collaborative work of multiple sensors, optimizing the performance of SLAM by combining different types of data such as images, depth maps, and IMU data. The advancement of data fusion technology will help solve the problem of unstable feature matching and enhance the overall robustness of the system.

6. Conclusion

Early visual SLAM systems achieved localization and mapping through relatively simple feature points. In order to improve the real-time performance of the system, keyframes have been proposed; To solve the problem of feature matching, researchers have developed methods such as SIFT, SURF, ORB, etc. In order to solve the problem of accumulated errors, a closed-loop detection module was added to optimize the camera pose and update the map based on the optimization of the SLAM system backend. However, there are still several problems that are difficult to completely solve, such as visual SLAM systems running in extreme environments such as dynamic scenes, extremely dark scenes, and texture sparse scenes. With the rapid development of deep learning, artificial intelligence, GPU, etc. in the past few years, various difficulties faced by visual SLAM seem to have solutions. The shadow of deep learning has appeared in traditional functional modules, and both system quality and efficiency have been greatly improved. The improvement of GPU computing power has also greatly promoted visual SLAM systems, allowing them to process information from multiple sensors simultaneously while ensuring the real-time performance of the algorithm, without being limited to the information obtained by a single sensor. However, the research on visual SLAM has not yet reached a perfect stage, and further efforts are needed to improve the accuracy and robustness of the system to adapt to more and more complex application scenarios.

Acknowledgments

Tianjin Municipal Cooperation Project(Research and development of machining detection system for large-scale complex component intelligent mobile robot);Tianjin Jinnan District Jiebang leader project(2022JB03).

References

[1] Khairuddin, A.R.; Talib, M.S.; Haron, H. Review on simultaneous localization and mapping (SLAM). In Proceedings of the 2015.IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Penang, Malaysia, 27–29 November 2015; pp. 85–90.

[2] Smith R C, Cheeseman P. On the Representation and Estimation of Spatial Uncertainty[J]. International Journal of Robotics Research, 1986, 5(4): 56–68.

[3] Cadena C , Carlone L , Carrillo H ,et al. Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age[J].IEEE Transactions on Robotics, 2016, 32(6):1309-1332.DOI:10.1109/TRO.2016.2624754.

[4] Davison A J .Real-time simultaneous localisation and mapping with a single camera[J].Proc.ieee Int.conf.on Computer Vision, 2003.

[5] Mourikis A I , Roumeliotis S I .A Multi-State Constraint Kalman Filter for Vision-Aided Inertial Navigation[C]//Robotics and Automation, 2007 IEEE International Conference on. IEEE, 2007.DOI:10.1109/ROBOT.2007.364024.

[6] Klein, G.; Murray, D. Parallel tracking and mapping for small AR workspaces. In Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, Nara, Japan, 13–16 November 2007; pp. 225–234.

[7] Jia, Y.; Yan, X.; Xu, Y. A Survey of simultaneous localization and mapping for robot. In Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Chengdu, China, 20–22 December 2019; Volume 1, pp. 857–861.

[8] Henry P , Krainin M , Herbst E ,et al.RGB-D mapping: Using Kinect-style depth camerasfor dense 3D modeling of indoor environments[J].International Journal of Robotics Research,2013, 31(5):647-663.DOI:10.1177/0278364911434148.

[9] Henry P,Krainin M ,Herbst E ,et al.RGB-D Mapping: Using Depth Cameras for Dense 3DModeling of Indoor Environments[J].2014.DOI:10.10 07/978-3-6 42-285 72-1_33.

[10] Krainin M .Manipulator and Object Tracking for In Hand Model Acquisition[C]//ICRA 2010 Mobile Manipulation and Best Practice in Robotics Worksh ops. 2010.

[11] Engelhard N , Endres F ,Jürgen Hess,et al.Real-time 3-D visual SLAM with A hand-held camera[J]. 2011.

[12] Arshad, S.; Kim, G.W. Role of deep learning in loop closure detection for visual and lidar slam: A survey. Sensors 2021, 21, 1243.

[13] Singandhupe, A.; La, H.M. A review of slam techniques and security in autonomous driving. In Proceedings of the 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 25–27 February 2019; pp. 602–607.

[14] Saputra, M.R.U.; Markham, A.; Trigoni, N. Visual SLAM and Structure from Motion in Dynamic Environments: A Survey. ACMComput. Surv. 2018, 51, 1–36.

[15] Davison, A.J.; Reid, I.D.; Molton, N.D.; Stasse, O. MonoSLAM: Real-time single camera SLAM. IEEE Trans. Pattern Anal. Mach.Intell. 2007, 29, 1052–1067.

[16] Endres F, Hess J, Sturm J, et al. 3-D map** with an RGB-D camera[J]. IEEE transactionson robotics, 2013, 30(1): 177-187..

[17] R. Mur-Artal and J. D. Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system[J]. IEEE transactions on robotics, 2015, 31(5): 1147-1163.

[18] Engel J,Koltun V,Cremers D.Direct sparse odometry[J]. IEEE transactions on pattern analysis and machine intelligence,2017,40(3):611-625.

[19] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras," 2016.

[20] Mourikis A I , Roumeliotis S I .A Multi-State Constraint Kalman Filter for Vision-Aided Inertial Navigation[C]//Robotics and Automation, 2007 IEEE International Conference on.IEEE, 2007. DOI:10.1109/ROBOT.2007.364024.

[21] Pollefeys M .Visual 3D Modeling from Images[J].DBLP, 2004..

[22] Zhang J , Singh S .Visual-lidar odometry and mapping: low-drift, robust, and fast [J]. IEEE, 2015.DOI:10.1109/ICRA.2015.7139486.

[23] Tong Q , Peiliang L , Shaojie S .VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator[J].IEEE Transactions on Robotics, 2017, PP(99):1-17.D OI:10.1109/TRO.2018.2853729.

[24] Qin T, Shen S. Robust Initialization of Monocular Visual-Inertial Estimation on Aerial Robots[C]. IEEE RSJ International Conference on Intelligent Robots & Systems. IEEE, 2017: 4225-4232.

[25] Qin T, Li P, Shen S. Relocalization, Global Optimization and Map Merging for Monocular Visual-Inertial SLAM[J]. IEEE, 2018.

[26] Shan T , Englot B , Ratti C ,et al.LVI-SAM: Tightly-coupled Lidar-Visual-In -ertial Odometry via Smoothing and Mapping[J]. 2021.DOI:10.48550/arXiv.2104.10831.