Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.071011
ISSN 2371-8838 Vol. 7 Num. 10

Overview of Visual SLAM Technology: From Traditional to Deep Learning Methods

Keting Huang[1,*]

[1]*Shenyang Ligong University, Shenyang, Liaoning, China*
[*]*Corresponding author*

Keywords: Simultaneous localization and mapping; Laser SLAM; Visual SLAM; Deep learning; Map construction

Abstract: SLAM refers to the problem of simultaneous localization and mapping (SLAM) of mobile robots in unknown environments. With the development of robot technology and artificial intelligence technology, SLAM has become an important technology and is widely used in all aspects of production and life. The importance of SLAM is self-evident for autonomous vehicles. SLAM is a very strange field for most people. In the wave of artificial intelligence, more and more enterprises and universities have invested in the research of visual SLAM. In this paper, we will introduce several classic visual SLAM algorithms and discuss their applications in the robot field, propose some problems in the current SLAM research field, and look forward to the future development direction of visual SLAM research.

1. Introduction

SLAM technology is an important part of mobile robot technology. It can obtain 3D point cloud data in unknown environment, and finally realize robot positioning and map building on 3D point cloud data. Its application prospect is very broad. The history of SLAM can be traced back to the 1980s. At the end of the 1980s, the representation and estimation of spatial uncertainty by Smith[1], Cheeseman and Durrant Whyte[2] became the beginning of SLAM research. A series of probabilistic filtering methods based on lidar sensors and subsequent graph optimization methods were initially developed, mainly including EKF (Extended Kalman Filter)[3], RBPF (Particle Filter)[4], Gmapping[5], Karto SLAM[6] and Cartographer[7]. Among them, the filtering based laser SLAM[5] proposed by Giorgio Grisetti et al. in 2007 mainly uses the Gmapping algorithm, and the graph optimized laser SLAM algorithm[7] opened by Google in 2016 mainly uses the Cartographer algorithm.

At the beginning of the 21st century, Montemerlo et al. first combined Rao Blackwelled PF (particle filter) with EKF to carry out research on robot SLAM, which was later the open-source FastSLAM algorithm[8]. After that, the laser SLAM has been continuously optimized and improved. However, as more and more studies have found that the laser SLAM has limitations such as high cost, high energy consumption, low price of lidar accuracy, filter based laser SLAM cannot perform loopback detection, large linear error accumulation, and difficulty in maintaining the waypoint based covariance matrix, which are difficult to promote.

With the development of artificial intelligence technology and computer hardware, the research in SLAM field has gradually entered a climax. In the 1990s, SLAM systems based on computer vision and machine learning technology began to rise. In 2007, Andrew et al. proposed the first pure vision SLAM system Mono SLAM[9]. Mono SLAM is the first visual SLAM system based on monocular camera, which mainly uses sparse feature points of images to generate maps. The appearance of this method makes up for many defects of laser SLAM, and has been rapidly applied in the fields of robots, unmanned driving, etc. In the following ten years, with the rapid development of computer hardware and software technology, the visual SLAM algorithm is also constantly improving and developing.

In the 21st century, the visual SLAM technology has developed rapidly. On the one hand, image recognition, semantic segmentation and other algorithms based on deep learning have become the most commonly used methods in the field of visual SLAM; On the other hand, the vision SLAM system based on multi-sensor fusion is also constantly improving and optimizing. In this context, many scientists and engineers began to try to combine deep learning with SLAM. Then a series of SLAM algorithms based on deep learning emerged and gradually became the mainstream algorithm in the SLAM field.

2. Fusion SLAM of laser radar and vision

Lidar can provide high-resolution environmental information at a relatively long distance (>100m), but lidar is a "passive" sensor relative to the camera, which cannot obtain visual information. Therefore, the combination of laser radar and vision sensor can improve the robot's positioning and navigation ability.

As early as 2008, Peyman Moghadam et al. fused the measured values of the stereo vision camera system and the 2D laser rangefinder[10] to dynamically plan and navigate mobile robots in a cluttered and complex environment, verifying the effectiveness of the proposed fusion strategy of laser radar and stereo vision in mobile robot navigation. After that, SLAM based on the fusion of laser radar and vision is gradually applied and studied.

In 2015, Ji Zhang et al. proposed a general framework that combines visual odometer and laser radar odometer[11] to improve robustness to radical movements and temporary lack of visual features. In 2018, for the tracking part of SLAM, et al. used RGB-D camera and 2D low-cost laser radar to complete robust indoor SLAM through mode switching and data fusion[12]. In 2020, Lili Mu et al. proposed a new synchronous positioning and mapping (SLAM) method based on graph optimization, combined with LiDAR, RGB-D camera, encoder and inertial measurement unit (IMU)[13]. In 2022, Jun Yin et al. proposed a new 3D laser radar assisted monocular vision synchronous positioning and mapping (LAMV-SLAM) framework[14] for mobile robots in outdoor environments. In the same year, Xiaolong Cheng et al. proposed a semantic segmentation odometer and mapping method based on the visual fusion of lidar and camera data[15], which is used for real-time motion state estimation and advanced understanding of the surrounding environment.

In 2023, Sheng 'En Li and others proposed a large scene construction scheme of vision and LiDAR fusion[16] in view of the limitations of SLAM system using a single sensor, reducing the impact of low precision and low recall of closed-loop detection point cloud on SLAM construction accuracy. In the same year, Matteo Frosi et al. proposed a D3VIL-SLAM, which extends the existing LiDAR based SLAM system ART-SLAM[17] to include inertial and visual information, which can generate highly detailed 3D maps while maintaining real-time performance. In July 2023, Zexi Liu et al. proposed a relocation method of vision and laser radar sensor fusion[18], which greatly improved the relocation performance. In 2023, Florian Sauerbeck et al. proposed a new method to integrate 3D LiDAR depth measurement into the existing ORB-SLAM3 based on the

RGB-D mode[19] to improve the system running time.

However, this method has some problems. First, using lidar and vision fusion usually requires more expensive sensor hardware. The high cost of lidar and high-resolution visual sensor, the possible limitation of sensing range and angle, and the need for greater computing power make this fusion method not widely used at that time. Even so, due to the complementarity of laser radar and visual sensor, the fusion method is suitable for various environments, and can improve the performance and stability of SLAM system in complex environments. Therefore, it is irreplaceable in many applications.

3. Camera based visual SLAM

As a typical application of SLAM, the camera based visual SLAM (Computer Vision and Mapping) algorithm is very simple in principle and process, has high real-time and robustness, and is easy to implement. Camera based visual SLAM mainly refers to the simultaneous localization and map building of mobile robots based on monocular cameras. The monocular vision SLAM algorithm uses a monocular camera as a sensor to reconstruct the three-dimensional space through the monocular camera, without using other sensors, such as gyroscope, accelerometer, GPS, etc. Compared with the traditional binocular SLAM algorithm, monocular SLAM has obvious advantages in system structure, robustness and accuracy.

In 2007, Andrew et al. proposed the first visual SLAM system based on monocular camera[9]. Since then, people began to apply monocular cameras to SLAM technology. The vision SLAM algorithm based on monocular vision includes two parts: back-end optimization and front-end mapping. The back-end optimization part mainly includes camera calibration, feature point extraction and matching, feature point matching selection and optimization after feature point matching, pose solution and so on. The front-end mapping part mainly includes pre-processing the two-dimensional map and modeling the environment.

In 2014, Christian Forster et al. proposed a semi direct monocular vision mileage calculation method, called SVO (semi direct vision odometer), and released it as open source software[20].

In 2015, Ra úl Mur Artal et al. proposed a feature-based monocular synchronous positioning and mapping (SLAM) system[21] (ORB-SLAM) for real-time operation in small and large indoor and outdoor environments. In 2017, Jakob Engel et al. proposed a visual odometer method based on novel, high-precision sparse and direct structure and motion formula - direct sparse odometer (DSO)[22]. In 2018, David Schubert et al. proposed a new direct monocular VO method combined with the roller shutter model, expanding the direct sparse odometer[23]. However, as the monocular vision based SLAM algorithm directly uses monocular cameras for data acquisition, its accuracy is relatively low. In addition, it is difficult to calibrate the monocular camera because a single camera sensor is used for data acquisition. Therefore, the vision SLAM algorithm based on monocular vision has not been widely used.

In recent years, vision based SLAM algorithm has been greatly developed. In addition to several classic algorithms introduced above, there are many algorithms for positioning and mapping based on visual information (such as IMU, GPS). In 2013, Simon Lynen et al. proposed a general framework, called Multi sensor Fusion Extended Kalman Filter (MSF-EKF)[24], which can handle delay, relative and absolute measurements from an theoretically unlimited number of different sensors and sensor types, while allowing online self-calibration of the sensor suite. In 2015, Michael Bloesch et al. proposed a monocular vision inertial mileage calculation method. This algorithm achieves accurate tracking performance by directly using the pixel intensity error of image patches[25], while showing very high robustness. In 2018, Tong Qin et al. proposed a robust and multi-functional monocular vision inertial state estimator[26], which integrates monocular

camera and IMU, and fuses visual and inertial information through tight coupling, realizing robust positioning and navigation. In the same year, Xiang Gao et al. proposed the expansion of direct sparse odometer (DSO) in monocular vision SLAM system with closed-loop detection and attitude map optimization (LDSO)[27], which combines IMU information and has good real-time performance and robustness. In May of the same year, Pan Zeng et al. proposed a VIO weighing Euler pre integration method based on monocular camera and IMU[28]. These SLAM algorithms based on visual information have good performance in many scenarios.

At present, the realization of SLAM based on visual information has become a trend and trend. At present, most intelligent driving vehicles are based on monocular vision technology to achieve map building and positioning. It is believed that in the future, SLAM algorithm based on visual information will become more and more popular.

4. Visual SLAM based on deep learning

After the rise of deep learning, we can learn how to extract key information from images by training a model. This is a big challenge, because SLAM is perceived in an unknown environment, so the map is dynamic and often incomplete. Therefore, using the deep learning method can help us better understand the environment. For example, when we mark an object as "green", we can predict the color of the object through the depth learning model. Using this information, we can make it easier for robots to reach their destinations.

Deep learning algorithm is the mainstream recognition algorithm in the current computer vision field. It relies on multi-layer neural network to learn the hierarchical feature representation of images. Compared with traditional recognition methods, it can achieve higher recognition accuracy. In the field of visual SLAM, hierarchical image feature extraction methods, represented by deep learning technology, have emerged in recent years, and have been successfully applied to SLAM inter frame estimation and closed-loop detection. Based on the classification method of the existing literature, the semantic information is divided into three categories, which are the low-level semantic information of line and surface, the pixel-level or image-level semantic information of object category, and the attribute semantic information of object 's ' moving ' and ' static ' state.

In 2013, R F. Salas Moreno et al. demonstrated a new object oriented 3D SLAM framework, which takes advantage of prior information: many scenes are composed of repeated, special category objects and structures[29]. In 2017, Mc Cormac et al. proposed Semantic Fusion, which uses the Elastic Fusion algorithm to provide long-term dense correspondence between indoor RGB-D video frames[30], combined with convolutional neural network (CNN) to allocate labels from multiple pixel points, and used Bayesian inference method and conditional random field method to calculate and fuse into an effective map. In 2017, Sen Wang et al. proposed a visual odometer using a deep recursive convolutional neural networks (RCNNs) monocular VO end-to-end framework[31], which verified that end-to-end deep learning technology can become a feasible complement to traditional VO systems.

In 2018, Berta et al. proposed Dyna SLAM, which added a target detection model[32] on the basis of ORB-SLAM2, met the detection requirements of dynamic targets, repaired the images occluded by dynamic objects, and generated dense maps in static environments. In 2019, Carnegie Mellon's Yang S et al. proposed Cube SLAM, generated high-quality three-dimensional area suggestion box from two-dimensional bounding box and vanishing point random sampling, and established object level map without prior object knowledge model[33]. In the same year, Nicholson L and others from Queensland University of Science and Technology in Australia constructed the Quadric SLAM system[34], which can directly estimate the conic curve from the rectangular box detected in 2d, and then construct the ellipsoid constraint to clearly express the position, direction, size and orientation

information of objects. In 2021, Weixiang Shen et al. proposed a semantic mapping algorithm based on improved YOLOv5[35].

It can be seen that visual SLAM based on deep learning has many advantages, which is very useful for feature matching and scene understanding in visual SLAM. With the improvement of hardware performance and the optimization of deep learning algorithms, some deep learning SLAM methods have made significant progress in real-time performance, making them more suitable for practical application scenarios. Therefore, visual SLAM based on deep learning is widely used in various fields such as automatic driving.

5. Summary

The visual SLAM algorithm has made great progress in recent years, and has been widely used in many robot fields. The main principle is to extract and match the feature points using the images taken by the camera, and then optimize the feature points using the matching results to obtain accurate pose estimation. In this process, we usually use two methods: feature point extraction algorithm and feature point matching algorithm. The former uses the camera's internal parameter information, while the latter uses the camera's external parameter information.

In general, the current visual SLAM algorithms are divided into two categories: feature point matching based algorithms and image processing based algorithms. The former is mainly for image feature extraction and matching, while the latter is for image processing.

However, with the progress of technology, SLAM based algorithms are increasingly used in the field of robots. For example, visual SLAM technology has been widely used in the field of unmanned vehicles to achieve indoor positioning and navigation; In the field of UAV, visual SLAM is also widely used in autonomous flight control; In augmented reality (AR) and virtual reality (VR), visual SLAM can realize real-time modeling of the real world, thus improving the perception of augmented reality experience or virtual reality environment; In smart home and Internet of Things (IoT), visual SLAM can be used in smart home devices and Internet of Things sensors to help these devices perceive the environment, locate objects and conduct intelligent interaction, and improve life and work efficiency.

In addition, vision SLAM in the future can also develop towards the integration of deep learning and traditional methods, multimodal data fusion, cross industry applications and other trends, as well as cope with more complex environments and challenges.

References

[1] R. Smith and P. Cheesman. On the representation of spatial uncertainty [J]. Int. J. Robot. Res., 1987, 5(4): 56–68.

[2] H.F. Durrant-Whyte. Uncertain geometry in robotics [J]. IEEE Trans. Robot. Automat., 1988, 4(1): 23–31.

[3] H. Durrant-Whyte and T. Bailey. Simultaneous localization and mapping: part I[J]. IEEE Robotics & Automation Magazine, 2006, 13(2): 99-110.

[4] J. Civera, A. J. Davison and J. M. M. Montiel. Inverse Depth Parametrization for Monocular SLAM[J].IEEE Transactions on Robotics, 2008, 24(5): 932-945.

[5] G. Grisetti, C. Stachniss and W. Burgard. Improved Techniques for Grid Mapping With Rao-Blackwellized Particle Filters [J]. IEEE Transactions on Robotics, 2007, 23(1): 34-46.

[6] Gerkey, Brian P., Morgan Quigley, and Ken Conley. Karto SLAM: An Open-Source Toolkit for SLAM[C] // Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), May 2010, Anchorage, Alaska, USA.

[7] Hess, Wolfgang, Damon Kohler, Holger Rapp, and Daniel Andor. Cartographer: Real-Time SLAM for 2D and 3D Mapping[C]//In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), May 2016, Stockholm, Sweden.

[8] Montemarlo M. Fast SLAM: A Factored Solution to the Simultaneous Localization and Mapping Problem[C]//Proc of Theaaai National Conference on Artificial Intelligence. American Association for Artificial Intelligence, 2002.

[9] DAVISON J, REID D I, MOLTON D N, et al. MonoSLAM: Real-Time Single Camera SLAM[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6):1052-1067.

[10] P. Moghadam, W. S. Wijesoma and Dong Jun Feng. Improving path planning and mapping based on stereo vision

and lidar[C]//2008 10th International Conference on Control, Automation, Robotics and Vision, Hanoi, Vietnam, 2008.
[11] J. Zhang and S. Singh. Visual-lidar odometry and mapping: low-drift, robust, and fast[C]//2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 2015.
[12] Y. Xu, Y. Ou and T. Xu. SLAM of Robot based on the Fusion of Vision and LIDAR[C]//2018 IEEE International Conference on Cyborg and Bionic Systems (CBS), Shenzhen, China, 2018.
[13] L. Mu, P. Yao, Y. Zheng, K. Chen, F. Wang and N. Qi. Research on SLAM Algorithm of Mobile Robot Based on the Fusion of 2D LiDAR and Depth Camera [J].IEEE Access, 2020, 8: 157628-157642.
[14] J. Yin, D. Luo, F. Yan and Y. Zhuang. A Novel Lidar-Assisted Monocular Visual SLAM Framework for Mobile Robots in Outdoor Environments [J].IEEE Transactions on Instrumentation and Measurement, 2022, 71:1-11.
[15] X. Cheng, K. Geng, G. Yin, Y. Sun, J. Wang and P. Ding. Semantic Mapping Optimization Based on LIDAR and Camera Data Fusion for autonomous vehicle[C]//2022 6th CAA International Conference on Vehicular Control and Intelligence (CVCI), Nanjing, China, 2022.
[16] S. Li, S. Jing, Q. Yue and Y. Zhang. Static Map Building Scheme for Vision and Lidar Fusion[C]//2023 4th International Symposium on Computer Engineering and Intelligent Communications (ISCEIC), Nanjing, China, 2023.
[17] M. Frosi and M. Matteucci. D3VIL-SLAM: 3D Visual Inertial LiDAR SLAM for Outdoor Environments[C]//2023 IEEE Intelligent Vehicles Symposium (IV), Anchorage, AK, USA, 2023.
[18] Z. Liu, Y. Liu, X. Wang and D. Zhu. Visual-LiDAR Fusion Relocation for SLAM Systems[C]//2023 International Conference on Frontiers of Robotics and Software Engineering (FRSE), Changsha, China, 2023.
[19] F. Sauerbeck, B. Obermeier, M. Rudolph and J. Betz. RGB-L: Enhancing Indirect Visual SLAM Using LiDAR-Based Dense Depth Maps[C]//2023 3rd International Conference on Computer, Control and Robotics (ICCCR), Shanghai, China, 2023.
[20] C. Forster, M. Pizzoli and D. Scaramuzza. SVO: Fast semi-direct monocular visual odometry[C]//2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 2014.
[21] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós. ORB-SLAM: A Versatile and Accurate Monocular SLAM System [J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.
[22] J. Engel, V. Koltun and D. Cremers. Direct Sparse Odometry[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 611-625.
[23] Schubert D, Demmel N, Usenko V, et al. Direct Sparse Odometry with Rolling Shutter[C]//IEEE.IEEE, 2018.
[24] Lynen S, Achtelik M W, Weiss S, et al. Robust and Modular Multi-Sensor Fusion Approach Applied to MAV Navigation[C]//IEEE.IEEE, 2013.
[25] M. Bloesch, S. Omari, M. Hutter and R. Siegwart. Robust visual inertial odometry using a direct EKF-based approach [C]//2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 2015.
[26] T. Qin, P. Li and S. Shen. VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator [J]. IEEE Transactions on Robotics, 2018, 34(4): 1004-1020.
[27] X. Gao, R. Wang, N. Demmel and D. Cremers. LDSO: Direct Sparse Odometry with Loop Closure[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 2018.
[28] P. Zeng, S. Pan, S. Wang, L. Huang and F. Ye. A Weighing Euler Pre-integration Method in the Visual-Inertial Odometry[C]//2018 Ubiquitous Positioning, Indoor Navigation and Location-Based Services (UPINLBS), Wuhan, China, 2018.
[29] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, P. H. J. Kelly and A. J. Davison. SLAM++: Simultaneous Localisation and Mapping at the Level of Objects [C] //2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013.
[30] J. McCormac, A. Handa, A. Davison and S. Leutenegger. SemanticFusion: Dense 3D semantic mapping with convolutional neural networks [C]//2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017.
[31] S. Wang, R. Clark, H. Wen and N. Trigoni. Deep VO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks[C]//2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 2017.
[32] B. Bescos, J. M. Fácil, J. Civera and J. Neira. Dyna SLAM: Tracking, Mapping, and Inpainting in Dynamic Scenes [J]. IEEE Robotics and Automation Letters, 2018, 3(4):4076-4083.
[33] S. Yang and S. Scherer. CubeSLAM: Monocular 3-D Object SLAM[J].IEEE Transactions on Robotics, 2019, 35(4): 925-938.
[34] Nicholson L, Milford M, Sünderhauf N. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam [J]. IEEE Robotics and Automation Letters, 2019, 4(1): 1-8.
[35] W. Shen, Y. Jia, M. Li and J. Zhu. A New Semantic SLAM Mapping Algorithm Based on Improved YOLOv5 [C]//2021 14th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, 2021.