Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 1, No. 1, 2022

A Survey on Visual Simultaneously Localization and Mapping

Zijie Zhang [1,2], Jing Zeng [1,2]

[1] School of Automation and Information, Sichuan University of Science and Engineering, Yibin, 644000, China.
[2] Artificial Intelligence Key Laboratory of Sichuan Province, Yibin, 644000, China.

[*] **Corresponding author**: Zijie Zhang (Email: 459741358@qq.com)

Abstract: Visual simultaneous localization and mapping (VSLAM) is an important branch of intelligent robot technology, which refers to the use of cameras as the only external sensors to achieve self-localization in unfamiliar environments while creating environmental maps. The map constructed by slam is the basis for subsequent robots to achieve autonomous positioning, path planning and obstacle avoidance tasks. This paper introduces the development of visual Slam at home and abroad, the basic methods of visual slam, and the key problems in visual slam, and discusses the main development trends and research hotspots of visual slam.

Keywords: VSLAM, Intelligent robot, Development trend.

1. Introduction

A basic task of a mobile robot is to determine its location under the condition of a given environment map. However, the environment map is not available from the beginning. When a mobile robot enters an unknown environment, it needs to build a 3-D environment map through its own sensors and determine its location in the map at the same time. This is the Slam (simultaneously localization and mapping) problem, it is a real-time version of motion recovery architecture (SFM) [1]. The slam of camera as the only external sensor is called visual Slam (VSLAM). According to the number and category of cameras, visual slam can be divided into monocular slam, stereoscopic slam, RGB-D slam [2]

Although the SLAM problem has received considerable in-depth research, and many scholars have also made a detailed overview of this research, there have been new developments in the research on this problem in recent years, such as the movement from 3DOF to 6DOF, the map from 2D to 3D, the environment has the trend of gradually changing from indoor to outdoor, the scale of the map continues to expand, and the research on lifelong mapping has sprung up [3].

2. Development of VSLAM

In the past 10 years, with the improvement of computer computing power, visual slam has achieved rapid development. The landmark achievement of visual slam is mono SLAM [4] proposed by Andrew Davison, which is the first monocular SLAM Based on EKF [5] method. It can achieve real-time but cannot determine how much drift, and can create sparse maps online under the probability framework. DTAM [6] is a monocular SLAM algorithm based on direct method proposed in 2011. This method obtains the six degrees of freedom pose of the camera relative to the dense map through the whole image alignment of the frame rate, which can achieve the real-time effect on the GPU. PTAM [7] is the first algorithm proposed by Georg Klein to process slam with multithreading, which divides tracking and mapping into two separate tasks and processes them in two parallel threads. Kinect fusion [8] is the first Kinect based algorithm that can build dense 3D maps on GPU in real time. This method only uses the depth information obtained by Kinect camera to calculate the position and posture of sensors and build an accurate environment 3-D map model [5]. LSD-SLAM [9] proposed in 2014 is a direct monocular slam method, which directly processes the pixels of the image. Compared with the previous monocular visual odometer based on the direct method, it can not only calculate its own posture, but also build a global semi dense and accurate environmental map. The tracking method is directly operated on sim3, so that the scale drift can be accurately detected, and can be run in real time on CPU, orb-SLAM [10] is a relatively complete monocular SLAM algorithm based on key frames, which divides the whole system into three threads: tracking, map creation, and closed-loop control. Feature extraction and matching, sparse map creation, and location recognition are all based on orb features. Its location accuracy is very high, and it can run in real time.

3. Classification of VSLAM

Visual SLAM algorithm can be divided into feature-based slam method and direct method according to the different use of image information.

3.1. Feature Based VSLAM Method

Feature based visual slam method refers to the detection and extraction of feature points of the input image, and the calculation of camera pose and the mapping of the environment based on 2-D/3-D feature matching. If the whole image is processed, the computational complexity is too high. Because the feature can save the important information of the image and effectively reduce the amount of calculation, it is widely used.

The early implementation of monocular vision slam is achieved by means of filters. The extended Kalman filter (EKF) is used to realize simultaneous positioning and map creation. Its main idea is to use the state vector to store the three-dimensional coordinates of the camera pose and map points, use the probability density function to express the uncertainty, and finally obtain the mean and variance of the

updated state vector from the observation model and recursive calculation [11]

Then monocular visual SLAM Based on key frames gradually developed, of which the most representative is parallel tracking and mapping (PTAM) [12], which proposes a simple and effective method to extract key frames, and divides positioning and map creation into two independent tasks, which are carried out on two threads.

3.2. Direct VSLAM method

The direct slam method refers to the direct operation of the intensity of pixels, which avoids the extraction of feature points. This method can use all the information of the image. In addition, more environmental geometric information is provided, which is helpful for the subsequent use of the map. And it has higher accuracy and robustness to the environment with fewer features.

In recent years, monocular vision mileage calculation method based on direct method has been proposed. ENGELJ [12] constructs a semi dense inverse depth map for the current image, and uses dense image alignment method to calculate camera pose. Building a semi dense map is to estimate the depth values of all pixels with large gradient in the image. The depth values are expressed as Gaussian distribution, and when a new image arrives, the depth values are updated. KELC proposes an RGB-DSLAM [13] method, which combines the intensity error and depth error of pixels as the error function, and solves the optimal camera pose by minimizing the cost function. This process is realized by G20, and a key frame extraction and closed-loop detection method based on is proposed, which greatly reduces the path error.

4. Core Issues in VSLAM

4.1. Feature Detection and Matching

At present, point features are used most frequently, and the most commonly used point features are SIFT (scale invariant feature transform) feature [14], Surt (speed up robot features) [15] feature and orb (oriented fast and rotated brain) feature. SIFT features have been developed for more than 10 years and have achieved great success. SIFT features are distinguishable. Because its descriptor is represented by high-dimensional vector (128 dimensions), and it has rotation invariance, scale invariance, radiation transformation invariance, it is also robust to noise and illumination changes. SIFT feature is used in visual slam, but the vector dimension of SIFT feature is too high, resulting in high time complexity. Surf feature has scale invariance and rotation invariance, and the algorithm speed is 3 to 7 times higher than that of SIFT feature. Orb feature is the combination of fast feature detection operator and brief descriptor, and some improvements have been made on its basis. The biggest advantage of OB feature is its fast computing speed, which is 1O0 times that of SIFT feature and 10 times that of surf feature. The reason is that fast feature detection speed is very fast. In addition, the brief descriptor is a binary string, which greatly reduces the matching speed and has rotation invariance, but does not have scale invariance [16].

In the environment with a large number of lines and curves, when using point features, a lot of information in the environment will be abandoned. In order to make up for this defect, visual slam methods based on edge features and visual slam methods based on region features are also proposed.

4.2. Key Frame Selection

The frame to frame alignment method will cause large cumulative floating, because there will always be errors in the process of pose estimation. In order to reduce the error caused by the frame to frame alignment method, a slam method based on key frames is proposed. At present, there are several methods to select keyframes. Method 1: when all the following conditions are met, this frame is inserted into the map as a key frame. After N frames from the previous key frame, at least map points can be seen in the current frame, and the accuracy of pose estimation is high. The second method is to create a new key frame when the number of common feature points seen in the two images is lower than a certain threshold. Method 3 is a method of selecting key frames based on entropy similarity. Because the simple threshold is not applicable to different scenes, calculate an entropy similarity ratio for each frame. If the value is less than a predefined threshold, the previous frame is selected as a new key frame and inserted into the map. This method greatly reduces pose floating.

4.3. Closed Loop Detection

Closed loop detection and location identification to determine whether the current location is an environmental area that has been visited before. Error accumulation will inevitably occur in the process of 3D reconstruction, and the realization of closed loop is a means of elimination. Vision is the main sensor in position recognition algorithm. Among the image to image matching methods, bag of words method has been widely used because of its effectiveness. Word bag refers to the technology of converting the content of an image into digital vectors using visual vocabulary tree. The feature of the training image set is extracted, and its feature descriptor space is discretized into clusters by K-median clustering method. Therefore, the first node layer of the dictionary tree is created. The lower layer is obtained by repeating this operation for each cluster until a total of layers are obtained. Finally, w leaf nodes, namely visual vocabulary, are obtained.

4.4. Map Optimization

For a robot working in a complex and dynamic environment, the rapid generation of 3-D map is very important, and the created environment map plays a key role in the subsequent positioning, path planning and barrier performance, so accurate map creation is also very important. After the closed-loop detection is successful, add closed-loop constraints to the map and perform closed-loop correction.

The closed-loop problem can be described as a large-scale bundle adjustment problem, that is, to optimize the camera pose and the 3-D coordinates of all map points, but the optimization calculation complexity is too high, so it is difficult to achieve real-time.

An executable method is to optimize the closed loop through the pose graph optimization method. The graph whose vertex is the camera pose and the edge represents the relative transformation between the poses is called the pose graph. The pose graph optimization is to distribute the closed loop error along the graph, that is, evenly distribute it to all the poses on the graph. Graph optimization is usually implemented by LM algorithm [18] in graph optimization framework g2o (general graph optimization) [17].

5. Development Trend and Research Hotspot of VSLAM

5.1. Multi-sensor Fusion

The camera can capture rich details of the scene, while the inertial measurement unit (IMU) has a high frame rate and is relatively small, so it can obtain accurate short-term estimation. The two sensors can complement each other, so they can get better results when used together.

The initial pose estimation based on the combination of vision and IMU is solved by filtering. The measured value of IMU is used as the predicted value, and the measured value of vision is used for updating. Then, a real-time fusion method of IMU and monocular vision based on EKF is proposed. It is a measurement model that can represent the geometric constraints when a static feature is observed by multiple cameras. The measurement model is optimal and does not need to include the 3-D coordinates of the feature in the state vector of EKF. If the fusion problem is divided into two threads for processing, the inertial measurement and feature tracking between continuous images are locally processed in the first thread to provide high-frequency position estimation, and the second thread contains an iterative estimation of intermittent beam adjustment, which can reduce the impact of linear error. Many results have proved that the visual SLAM Based on optimization is better than the SLAM Based on filtering in accuracy. If the error of IM is integrated into the re projection error of the road sign in the form of full probability, a joint nonlinear error function will be optimized, in which the state before the marginalization is achieved through the key frame to maintain a fixed size optimization window, so as to ensure real-time operation [19].

Even in the system of vision and M-center fusion, when the robot moves violently, due to its increased uncertainty, it will still lead to positioning failure, so the robustness of the system needs to be further improved. And in order to achieve real-time positioning in real life, its computational complexity also needs to be improved.

5.2. Integration of Slam and Deep Learning

With the great success of deep learning in the field of computer vision, there is great interest in the application of deep learning in the field of robotics. Slam is a large system, in which there are many sub modules, such as closed-loop detection, stereo matching, etc., which can obtain better results through the use of deep learning.

The end-to-end deep neural network architecture can quickly extract the inter frame motion information of image sequences. Compared with the traditional inter frame estimation algorithm, the learning-based method replaces the cumbersome formula calculation, without manual feature extraction and matching, which is simple and intuitive, and the online operation speed is faster. At the same time, compared with the traditional closed-loop detection (position recognition) algorithm, the method based on deep learning uses deep neural network to extract image features, express image information more fully, and has stronger robustness to environmental changes such as light and season.

6. Conclusion

Over the past decade, visual slam has made amazing development, but using only the camera as the only external sensor for simultaneous positioning and 3D map reconstruction is still a very challenging research direction. There is still a long scientific research way to go if you want to carry out self-positioning in real time and build an environmental map similar to that seen by human eyes. In order to make up for the lack of visual information, visual sensors can be fused with inertial sensors (i-center), lasers and other sensors, and better results can be obtained through the complementarity between sensors. In addition, in order to be applied in the actual environment, the robustness of slam needs to be very high, so that it can be processed accurately in various complex environments, and the computational complexity of slam cannot be too high, so as to achieve real-time results.

References

[1] Quan Xiangmei, park Songhao Overview of visual slam [J] Journal of intelligent systems, 2016, 11 (6): 768-776

[2] Liu Haomin, Zhang Guofeng. Overview of simultaneous localization and map construction methods based on monocular vision [J] Journal of computer aided design and graphics, 2016, 28 (6): 35-42

[3] Zhao Yang, Liu Guoliang, Tian Guoliang, et al Overview of visual SLAM Based on deep learning [J] Robot, 2017, 39 (6): 77-85

[4] DAVISON A J, REID I D, MOLTON N D, et al. Mono-SLAM: real-time single Camera SLAM[J]. IEEE transactions on pattern analysis and machine intelligence, 2007, 29(6):1052-1067

[5] NEWCOMBE R A, LOVEGROVE S J. DAVISON A J. DTAM: Dense tracking and mapping in real-time[C]// International Conference on Computer Version. Bacelona, Spain, 2011: 2320-2327.

[6] Li M, Mourikis A I. High-precision, consistent EKF-based visual-inertial odometry[J]. The International Journal of Robotics Research, 2013, 32(6): 690-711.

[7] Newcombe R A, Lovegrove S J, Davison A J. DTAM: Dense tracking and mapping in real-time[C]//2011 international conference on computer vision. IEEE, 2011: 2320-2327.

[8] Izadi S, Kim D, Hilliges O, et al. KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera[C]//Proceedings of the 24th annual ACM symposium on User interface software and technology. 2011: 559-568.

[9] Engel J, Schöps T, Cremers D. LSD-SLAM: Large-scale direct monocular SLAM[C]//European conference on computer vision. Springer, Cham, 2014: 834-849.

[10] KLEIN G, MURRAY D. Parallel Tracking and Mapping foe Small AR Workspaces[C]// IEEE and ACM International Symposium on Mixed and Augmented Reality. Nara, Japan, 2007: 225-234.

[11] Pire T, Fischer T, Castro G, et al. S-PTAM: Stereo parallel tracking and mapping[J]. Robotics and Autonomous Systems, 2017, 93: 27-42.

[12] Avignone III F T, Elliott S R, Engel J. Double beta decay, Majorana neutrinos, and neutrino mass[J]. Reviews of Modern Physics, 2008, 80(2): 481.

[13] Scherer S A, Zell A. Efficient onbard RGBD-SLAM for autonomous MAVs[C]//2013 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2013: 1062-1068.

[14] Ng P C, Henikoff S. SIFT: Predicting amino acid changes that affect protein function[J]. Nucleic acids research, 2003, 31(13): 3812-3814.

[15] Baughman R P, Lower E E, Tami T. Upper airway. 4: Sarcoidosis of the upper respiratory tract (SURT)[J]. Thorax, 2010, 65(2): 181-186.

[16] Chen Weidong, Zhang Fei Research Progress on synchronous self localization and map creation of mobile and robot [J] Control theory and application, 2005 (3): 455-460

[17] Grisetti G, Kümmerle R, Strasdat H, et al. g2o: A general framework for (hyper) graph optimization[C]//Proceedings of the IEEE international conference on robotics and automation (ICRA), Shanghai, China. 2011: 9-13.

[18] Moré J J. The Levenberg-Marquardt algorithm: implementation and theory[M]//Numerical analysis. Springer, Berlin, Heidelberg, 1978: 105-116.

[19] Jixiucai, Zheng Zhiqiang, Zhang Hui Analysis and control of robot positioning error in SLAM problem [J] Journal of automation, 2008 (3): 323-33