Advances in Intelligent Systems Research, volume 146
2nd International Conference on Artificial Intelligence: Technologies and Applications (ICAITA 2018)

Progress in the Research of Visual SLAM

Binbin Xu, Pengyuan Liu and Junning Zhang
The Army Engineering University of PLA, Shijiazhuang, China

***Abstract*—Visual SLAM is the process that a camera sensor is used to locate the body of the motion and build a map of the surrounding environment at the same time. First the types of sensors and the framework of visual SLAM are introduced in this article. And then it introduces the word bag model of visual mileage based on feature point method and direct method, nonlinear optimization of back end and loop detection. Finally the future development expectation is demonstrated.**

Keywords—visual odometry; backend optimization; deep learning; loop detection

I. INTRODUCTION

After thirty years development of machine vision, people achieve exciting results in observing the world using the robot and understanding things especially in the military field. In the auxiliary maintenance of military equipment, using visual positioning maintenance personnel position helps us repair equipment accurately. In the military virtual reality and augmented reality simulation training, we can locate the virtual object in the real world and block the relationship through the camera location, so that the trainers can immerse themselves in it. In military reconnaissance, effective visual positioning helps UAVs and underwater robots to gauge their position more accurately and make up for GPS limitations. The realization of all the military applications above require visual SLAM technology.

SLAM (Simultaneous Localization and Mapping) is usually translated as "simultaneous positioning and map construction." We estimate the pose of the camera by the movement of the camera, use the motion information to restore the three-dimensional scene structure, and estimate the scale of the object at the same time. According to the sensor type Vision SLAM can be divided into monocular camera, binocular camera and depth camera. Three kinds of sensors have their own limitations. Monocular camera can not determine the true scale of the object because of the single image, that is, the scale uncertainty. Binocular camera configuration and calibration are complex, and need large amount of computation, GPU and EPGA to accelerate in real-time calculation of image distance information. Depth camera can get the depth information of the scene, reduce the calculation of depth, better robustness,so it becomes the mainstream of the current visual sensor[1].

Visual SLAM can be specifically divided into five parts, as shown in the figure1. First, the sensor reads the image information; visual odometer estimates the movement of the camera through keyframe images[2]. The back end optimizes the pose and loopback feedback of the visual odometer to obtain a globally consistent trajectory and map. Loopback detects whether the robot has returned to the position to make a judgment, if it comes back to the previous position, the information is also back to the end. Finally, based on the estimated trajectory, the measurement map and topology map are established.

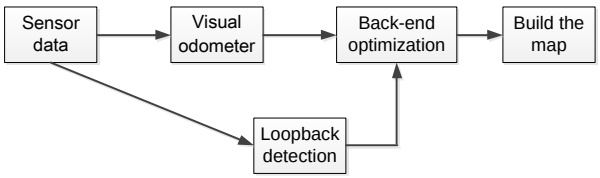

FIGURE I. VISUAL ODOMETER

II. VISUAL ODOMETER

The visual odometer is one of the most important components of visual SLAM and estimates camera motion by calculating the keyframe pose. According to whether feature points are extracted or not, it can be divided into two categories, based on the feature point extraction method and the direct method based on optical flow method[6]. According to the density of map points, it can be divided into sparse, semi-dense and dense[7]. Visual odometer is mainly to solve the problem of computer vision, the traditional pixel-based computational method is computationally intensive, does not meet the real-time. Therefore, significant feature points can be extracted from the image, these points remain stable when the camera is moving, and the pose and orientation of the camera are calculated on the basis of feature points. For the selection of feature points to meet the recurrence, easy to distinguish, locally related features, feature extraction methods are SIFT, SURF and ORB.

Feature points are usually composed of the key points and descriptive posture.The key point is the location of the feature points in space, including the direction, size and distance. Descriptors are based on a manual calculation, including a variety of key points of information. The classical SIFT feature point method, for example, we first extract the SIFT key points, and then calculate the SIFT descriptor, calculate the distance between the two image descriptors to meet the distance constraints to determine the same point[8].

The quality of the feature point method depends on accuracy and real-time. SIFT feature point method of high accuracy, consider the image transformation process of light, size, rotation factor. It has excellent robustness, but the computational complexity increases, and can not meet the real-time visual SLAM. The advantage of the FAST key point is that the calculation speed is fast, and the feature point is selected by judging whether the local pixel gray transform is obvious. However, in the complex environment, the robustness is poor and the feature points get together. At present, the improvement of the key points of FAST mainly focuses on the directionality and rotation so as to improve the accuracy on the basis of real-time performance.

Sparse feature point extraction based on feature points is the current mainstream method[10] . The idea is to replace the

entire image with some salient features and use them as road signs in visual SLAM[11]. Since the description of feature points can be consistent after a certain degree of movement of the camera, tracking and matching them often results in a more robust effect. In recent years, some easy-to-calculate feature extraction algorithms such as ORB and BRISK describe the popularity of algorithms. It gradually replaces the Harris corner or the Computed SIFT, SURF, which previously underperformed, and became the primary option for visual odometry.

On the other hand, the direct method of estimating camera motion on the assumption of pixel gray invariance[12](Direct Method) has also been rapidly developed in recent years. The direct method evolves from optical flow and can estimate the camera motion by minimizing the photometric error (minimizing the reprojection error of the feature points in the feature point method) without mentioning the feature (or without calculating the feature description) Pixel's spatial location. Due to the skipping of the steps of feature description and matching, the direct method, can often run at extremely fast speeds. At the same time, the direct method also has the ability to calculate semi-dense and even dense maps, which is not possible by the characteristic point method. However, the direct method is easily affected by light and has poor stability.

III. Back-end Optimization

The role of back-end is to make the entire trajectory in a long time to maintain optimal conditions. Because of the noise, what we do at the backend is to determine the distribution of x and y for the x and y in the equation of motion and observation equation, knowing the motion data and the observed data, to estimate the optimal value. Estimate the optimal value of the popular is that as time changes, the cumulative error is increasing, the estimated position of the deviation from the correct location of more and more, then we need to optimize the location of the point to make it back To the right place.

In the early back-end optimization to Kalman filter and Extended Kalman Filter [13](EKF) . Based on the Markov property, the current state is only related to the state at the previous moment, and the Gaussian distribution is used to get the optimal state estimation. However, the traditional Kalman filter has defects. First of all, the trajectory of the robot does not necessarily satisfy the Markov property. In particular, when the loopback occurs, the current status is affected by the status of the starting position. Secondly, the Kalman filter is only optimized once at the position, and the function distribution does not necessarily satisfy the Gaussian distribution.

The great development of visual SLAM in the 21st century is to transform the mainstream back-end processing method from the filter method represented by EKF into an optimization method based on nonlinear optimization. There are several obvious benefits to nonlinear optimization. First, you can handle more information. The filter approach must assume Markovian, marginalizing the past state. And the optimization method can take into account all the movement and observation data, using more information, so sometimes called full SLAM or batch SLAM. Second, the optimization method allows iteratively estimating the state, re-linearizing each iteration point. Therefore, optimization can often find a more accurate solution. Third, SLAM optimization problems can be naturally described using graph models or probability maps, efficiently handling loopback detection. The structure of the graph model can directly correspond to the sparse structure of the normal equation coefficient matrix in the BA problem and accelerate the solving process. Overall, the introduction of optimization methods has quickly become the mainstream method in visual SLAM.

If the optimization variables are abstracted into nodes, and the errors introduced by the equations of motion and the observation equations are abstracted into edges, a graph model corresponding to SLAM can be obtained, which is called graph optimization[15]. If we look at the problem from the perspective of a probabilistic graph model, we can think of SLAM as a Bayesian network or factor graph with nodes constrained by random variables, conditional probabilities defined by motion and observational data. Solving the maximum posteriori estimates for these graph models or networks is equivalent to solving the SLAM problem. Researchers have also developed dedicated tools for optimization of graph optimization factor maps, which are widely used in current SLAM systems.

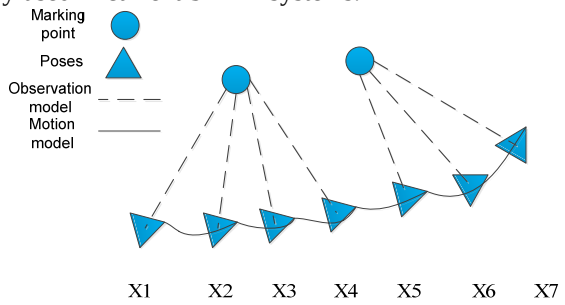

FIGURE II. GRAPH OPTIMIZATION MODEL

IV. Loopback Detection

Front-end estimation of pose and back-end optimization, can not completely eliminate the error. With the accumulation of time, the current location and the expected location will have a greater deviation, you need loop detection. Loopback detection eliminates errors and builds globally consistent trajectories and maps. The loop detection is equivalent to the error correction system. When the pose error of the robot appears, it feeds back the information to the front end and the back end, re-estimates the pose and continues to optimize. The following is a schematic diagram of cumulative error generation and elimination.

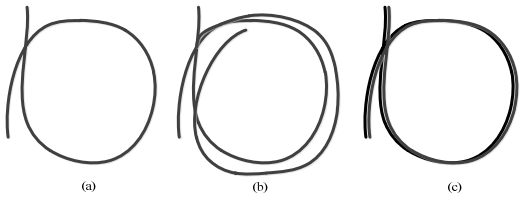

FIGURE III. ERROR ELIMINATION PROCESSTION

The mainstream method based on image appearance detection is the word bag model. Bag-of-Words In general, the characteristics of an image are divided into small pieces, and the "words" representing the features are composed into dictionaries. Dictionary generation is similar to the clustering

problem, using the simplest K-means algorithm to solve it. Since the efficiency of the algorithm itself is not high, then proposed K-means ++[12] and extended k-mean[13] . The extended k-mean algorithm is simple and practical, first select the first level of clustering, and then the first level of clustering down to the second level of clustering, and so on to form K-tree, clustering classification Finer and more efficient.

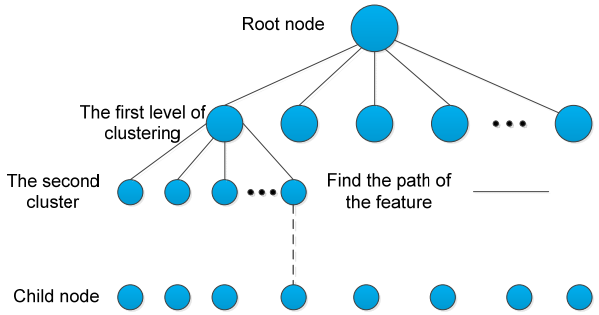

FIGURE IV. K CLASS DENDROGRAM

With the development of deep learning theory, video target tracking system based on deep learning gradually gets more research and application. The most typical application in the field of target recognition tracking is convolutional neural network (CNN). As shown in the figure, the convolution model reduces the number of parameters in the neural network greatly and increases the calculation speed by the characteristics of weight sharing and local receptivity.

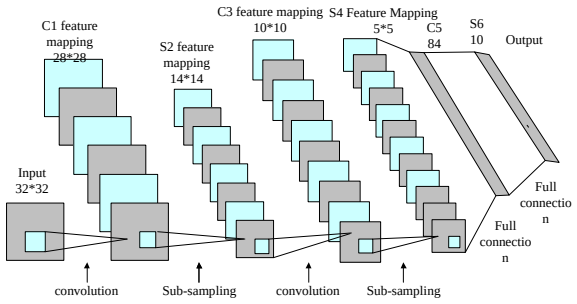

FIGURE V. CONVOLUTION NEURAL NETWORK

V. SUMMARY

On one hand, the development of miniaturization of visual SLAM allows it to run on small devices such as mobile phones and to realize underlying applications such as sports, navigation, teaching and entertainment. On the other hand, the use of high-performance computer hardware, makes it possible to achieve accurate three-dimensional reconstruction, scene understanding and other upper applications. As a good system, just one sensor is not enough, often a combination of a variety of sensors is required. For example, the vision-inertial navigation SLAM scheme, inertial sensor pose estimation is very good in a short time, and the camera can solve the inertial sensor drift problem.

Visual SLAM is currently based primarily on feature points and pixel levels, complex and imprecise. To solve this problem, researchers combined SLAM with deep learning. Some researchers combine object recognition with SLAM to construct a map of artificial markers, introduce the marker information into the back-end optimization, and combine the feature points and label information to become the semantic SLAM. With the development of deep learning, we use neural networks to identify, detect and segment images and introduce them into SLAN for object recognition and segmentation[14]. SLAM pose estimation and loopback detection have broad prospects.

REFERENCES

[1] Huang Hao. “Multi-view Imaging System Based on Kinect” Tianjing: TianjinUniversity.(2013).

[2] Wang Yalong, Zhang Zhiqi, Zhou Lili. “3D mapping for indoor environment with RGB-D camer” Application Research of Computers,08(2015)2533-2537

[3] Quan Meixiang,Piao Songhao,Li Guo. “An overview of visual SLAM” Transactions on Intelligent Systems ,11(2016)769-776.

[4] LIN Huican,LU Qiang, “The Sparse and Dense VSLAM:ASurvey”Robot,38(2016)622-631.

[5] Wang Song. “SIFT based image matching algorithm research” Xian: Xidian University.(2013).

[6] Zhou Qinuan. “ The realization of monocular vision SLAM in indoorenvironment ” Xiamen: Xiamen University.(2013).

[7] Wang Song. “SIFT based image matching algorithm research” Xian: Xidian University.(2013).

[8] Engel J,Koltun V,Cremers D, Direct sparse odometry. ArXiv preprint arXiv: 1607. 02565. 2016.

[9] Li Yagui. “Research on Robotic SLAM Algorithm Based on Stereo Vision” Haerbin: Harbin Institute of Technology.2015.

[10] Koller D, Friedman N. Probabilistic graphical models: principles and techniques[M].MIT press,(2009).

[11] Wei Shaoqing. “Research on the consistency of indoor mobile robot based on EKF-SLAM algorithm”Shijiazhuang: Hebei University of Science and Technology.(2013).

[12] D.arthur, S.Vassilvitskil. “K-mean++: The advantages of careful seeding, ”in Proceedings of the eighteenth annual ACM-SLAM symposium on Discrete algorithms[J].Society for Industrial and Applied Mathematics, 2007,pp,1027-1035.

[13] D.Galvez-Lopez, J.D Tardos. “Bag of binary words for fast place recognition in image sequences” IEEE Transactions On Robotics, vol .2012, 28,no,5,pp.1188-1197.

[14] J. Deng, W. Dong, R, SCcher,L-J ,Li,and L. FeiFei .“Imagenet: A large-scale hierarchical image database” CVPR09,(2009).