Computer Life (CPL) ISSN: 1819-4818
DELIVERING QUALITY SCIENCE TO THE WORLD

Kinect Based 3D Visual Simultaneous Localization and Mapping

Hanhai Shi [a], Shuwen Dang [b], Fajiang He [c], Qingqu Wang [d]

Shanghai University of Engineering Science, China.

[a]18606448961@163.com, [b]345485638@qq.com, [c] mikehfj@sues.edu.cn, [d] wqqhao@163.com

---

Abstract: This paper study the visual slam algorithm, which is developed rapidly with depth camera as sensor in recent years. Compared to the traditional visual SLAM scheme used in the K-d or K-means tree algorithm, we propose a closed-loop detection scheme based on K-means + + algorithm, which corrects the cumulative error of correction system and improves the system stability and accuracy. In the front-end of the system, we choose the point features-method based on ORB feature. The back-end consist of pose graph optimization and closed loop detection, the pose graph optimization is realized by g2o general solver and closed-loop detection adopts the bag of words model based on two-dimensional image feature. And the experimental results succeeded in registering the three-dimensional environment point cloud map and get an accurate motion path.
Keywords: SLAM; ORB; K-means++; graph optimization; loop closure.

---

1. INTRODUCTION

Simultaneous Localization and Mapping(SLAM) is the process of estimating itself posture and building an environmental map simultaneously according to sensor information. It is considered to be the key to realize autonomous navigation of mobile robot. Since it was first proposed in 1986, SLAM has been a hot research topic in the field of robotics. According to the different types of sensors and installation methods, SLAM can be divided into laser and visual categories. The research of laser SLAM is early, and both theory and engineering are relatively mature. Since it was first proposed in 1986, SLAM has been a hot research topic in the field of robotics. According to the different types of sensors and installation methods, SLAM can be divided into laser and visual categories. The research of laser SLAM is early, and both theory and engineering are relatively mature. However, visual SLAM is still at the level of technology accumulation and has not yet been applied in large-scale commercial applications. Early SLAM studies mainly used the filter method to minimize the noise of movement posture and signpost. After the 21st century, scholars began using SFM (Structure from Motion), based on the optimization theory to solve the SLAM problem, and get good results, this method has become a mainstream method in the study of visual SLAM. Visual sensors are classified as monocular, binocular and RGBD (rgb-depth). RGBD camera would be the biggest characteristic is the depth of the direct access to environmental information, compared with monocular and binocular cameras can save a lot of calculation, it is very attractive for high requirement of real-time SLAM[1-3].

Kinect 2.0 was selected as the sensor in this paper, and the system was constructed according to the basic framework of the front-end and back-end visual SLAM. System implementation process is shown in figure 1, including the sensor data acquisition and preprocessing, RGB image feature extraction and matching, the construction and fusion splicing of point cloud image, image motion transformation between estimation, based on the position of Pittsburgh figure of optimization and the close-loop testing of the model based on word bag.

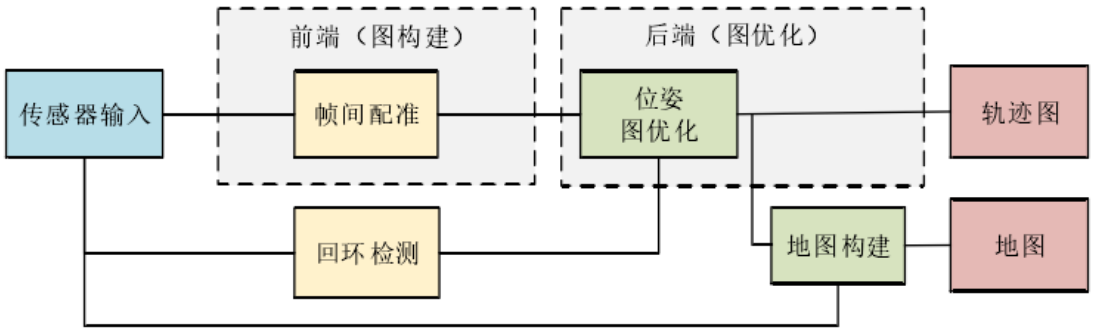


Fig.1 system implementation framework

2. DATA ACQUISITION

2.1 Kinect 2.0 data acquisition

Kinect 2.0 is a low-cost, highly popular depth camera that Microsoft launched in 2010 as a human-computer interaction device for somatosensory interaction. In 2013, Microsoft released the second generation of Kinect, which has greatly improved its performance compared with the first generation. Currently, there are two Development platforms for Kinect 2.0, Microsoft's official Software Development Kit and a third-party hacker driver libfreenect2 [4]. Since the Kinect SDK can only be used in Windows environment, most of SLAM's research is conducted in Linux environment, and the Kinect SDK itself is not open source, which is a great constraint for developers. Therefore, libfreenect2 was chosen as the driver platform for obtaining Kinect data. Libfreenect2 is provided by the OpenKinect community and is used primarily for Kinect 2.0 development. Libfreenect2 supports cross-platform enablement. It can help us to transfer the original data of Kinect. To meet later development needs, this article also USES an interface program iai_kinect2 in a ROS (Robot Operating System) environment. Iai_kinect2 is used to connect ROS and Kinect 2.0 Numbers. It also provides Kinect 2.0 calibration tools and data display tools.

2.2 Kinect 2.0 camera calibration

With the advantages of high precision and simple operation, the camera calibration technology based on 2D plane target represented by zhang zhengyou checkerboard calibration method has been widely used. Calibration principle is through the phase plane and plane between the matching points for the construction of single board should be matrix, and the homographic matrix to decompose again after receive the corresponding rotation matrix and translation vector t R. This method assumes that the checkerboard plane is Z = 0 in the world coordinate system, and then the relationship between the two planes is as follows:

$$S\begin{bmatrix} u \\ v \\ l \end{bmatrix} = A[r_1 \quad r_2 \quad r_3 \quad t]\begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix} = A[r_1 \quad r_2 \quad t]\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \tag{1}$$

$$\mathrm{H} = \mathrm{A}[r_1 \quad r_2 \quad t] \tag{2}$$

Equation (2) is the single-response matrix to be solved. In equation (1), A is the internal reference matrix of the camera, r is the rotation matrix, and t is the translation matrix. The singlet matrix here has 8 degrees of freedom, and two valid constraints can be constructed through a set of matching points. In theory, only four pairs of matching points are needed to solve all the parameters in the single response matrix [5-7].

3. FEATURE EXTRACTION AND MATCHING AND ATTITUDE ESTIMATION

After obtaining data from the sensor, the RGB image sequence should be first extracted for feature extraction and feature matching between adjacent images, so as to obtain the matching point pairs needed to calculate camera pose.

3.1 Feature point extraction

Feature points are composed of key points and description sub-parts. The key point is the position of the feature point in the image. Some feature points also have orientation, size and other information. The descriptor is usually a vector. In some way, the information about the pixels around the key points is described. As long as the descriptors of two feature points are close enough in vector space, they can be considered the same feature points.

Currently, the commonly used feature extraction algorithms are SIFT [8], SURF and ORB. The SIFT algorithm was originally developed by David g. Lowe in 1999 and improved in 2004. SIFT algorithm has the characteristics of rotation invariance, scale invariance, illumination invariance and good anti-noise. SURF (Speeded up Robust Features) algorithm was proposed by Bay in 2006 [9]. It is an improvement of the SIFT algorithm, which is significantly faster than the SIFT algorithm. SURF also has scale and rotation invariance, but its computing capacity is still large, and it has certain requirements on hardware, which is not dominant in real time. Rublee ORB algorithm was presented in 2011 [10], called the Oriented FAST and Rotated BRIEF, is a very efficient feature extraction method. ORB algorithm adopts FAST (Features from Accelerated Segment Test) feature detection operator and BRIEF (Binary Robust Independent Elementary Features) feature description operator. While the ORB maintains rotation and scale invariance, the speed increases significantly.

The effect of SIFT, SURF and ORB feature extraction algorithms was actually experienced, and open source visual library named OpenCV was used for experimental simulation in the project. RGB artwork used in the experiment is shown in figure 2 came from the mobile car loading device of a view taken by the camera in the laboratory, the RGB artwork for SIFT, SURF and ORB, output characteristic points after feature extraction effect comparison in table 1.

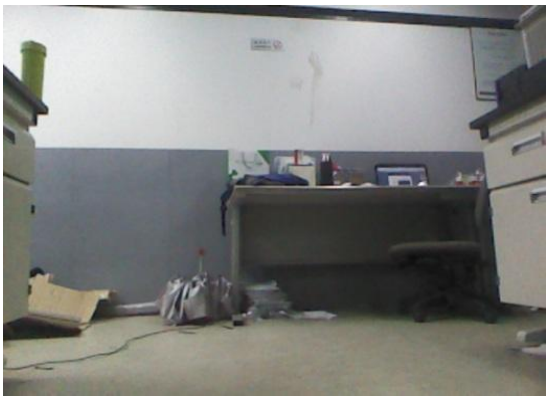


Fig.2 RGB original image

Table 1. Comparison of test results of feature points

Detection algorithm	Average number of feature points	The elapsed time(s)
SIFT	983	0.6441
SURF	1455	0.2251
ORB	562	0.0268

It can be seen that ORB algorithm has obvious advantages in speed. In this paper, ORB algorithm is selected as feature extraction method.

Fig.3 Violent matching results

Fig.4 Results after threshold screening

3.2 Feature point matching

Feature matching is a key step in a visual SLAM. More broadly, feature matching solves the data association problem in SLAM.

Consider the image at two moments, if the feature point x_t^m, (m=1, 2 ,…,M) is extracted from the first frame image. In the second frame image, the feature point x_{t+1}^m (n=1, 2 ,…,N) is extracted from. For the corresponding relation of each element, this paper USES the method of violent matching. That is to measure the distance between each feature point x_t^m and all x_{t+1}^m descriptors, and then sort them and take the closest distance. A pair of them as a match point. Description sub distance represents the degree of similarity between two features. For binary descriptor (such as BRIEF descriptor), Hamming distance [10] is used for measurement. The hamming distance between two equally long strings is the number of characters in the corresponding position of two strings.

The resulting matching pair error is very large, we first set a threshold to screen all matching pairs. In general, the threshold is set to four times the shortest distance of the matching pair. Any matching pair greater than the threshold is eliminated.

4. CLOSED LOOP DETECTION

If only considering the motion transformation between adjacent frames, the error produced by a frame is passed to the next frame, so constantly, will inevitably bring the whole system seriously accumulated error, finally calculate the trajectory of severe drift. This highlights the importance of closed loop detection.

Closed-loop detection refers to the robot's ability to identify the scene it has reached. If the detection is successful, the accumulated error can be significantly reduced. Closed-loop detection module can provide in addition to the adjacent frames after older constraints, provide more effective data for the backend posture figure, thus improving the precision of robot pose estimation and map building. This is of great importance to the long-term and large-scale SLAM system, as well as to the exploration and navigation in complex and dynamic environments. Figure 5 shows the influence of the closed loop detection module on the positioning accuracy. Red is the actual track, in the left figure, no closed-loop module is added, and in the right figure, the result after the closed loop is added, it can be seen that the closed-loop module has a significant correction effect on the trajectory of the robot.

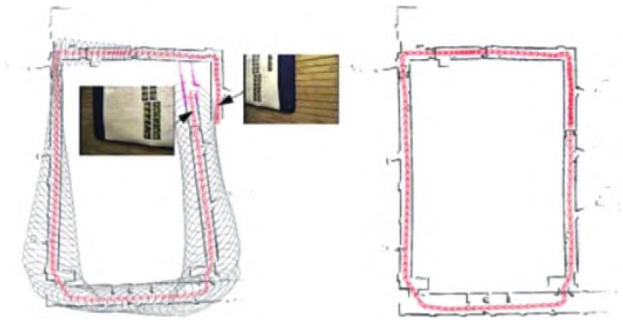


Fig.5 Effect diagram of closed loop detection

The word bag model is essentially a problem to detect the similarity of observation data. In the word bag model, we extract the features in each image, cluster their feature vectors, and establish the category database. The word bag model can be understood as a dictionary with feature description as element. If it's an ORB feature, it's an ORB dictionary. If it's a SIFT feature, it's a SIFT dictionary.

In SLAM system based on graph optimization, the dictionary is often constructed by Kd tree or k-means algorithm. For K - means algorithm itself, such as the need to specify beforehand clustering

number, each has a different clustering result, efficiency is not high and the initialization is too sensitive, in this paper, the retrieval efficiency higher K - means + + algorithm constructs a dictionary. K-means + + [11-12]is a very simple and effective method. It was developed on the basis of k-means algorithm. K-means algorithm is the most common and simplest clustering algorithm. The algorithm principle is shown in FIG. 6. (b) the samples are distributed to the nearest central vector, and these samples are used to construct non-intersecting clusters; (c) use the central vector of each cluster as the new center; (d) repeat (a) and (b) until it converges.

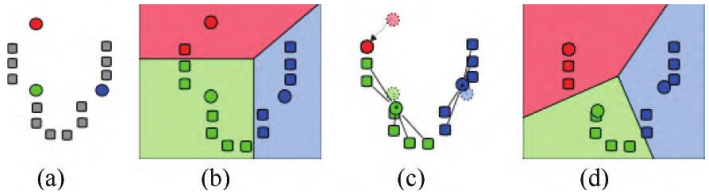


Fig.6 K-means clustering process diagram

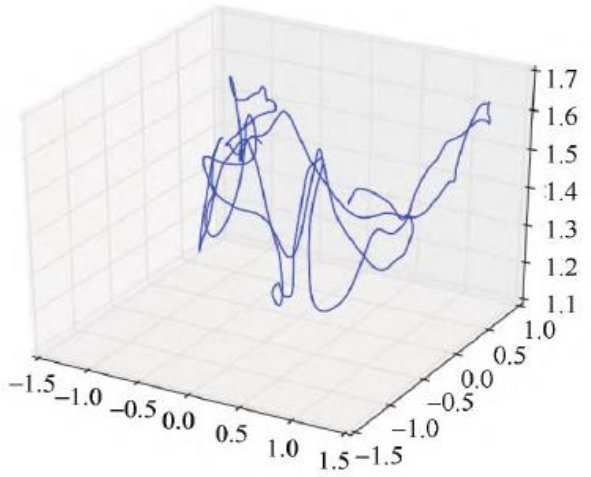
5. EXPERIMENT AND RESULT ANALYSIS

The TUM data set provided by TUM was used for the experiment. The TUM data set provides an accurate track of RGBD image data and data acquisition process, which is suitable for experimental research.

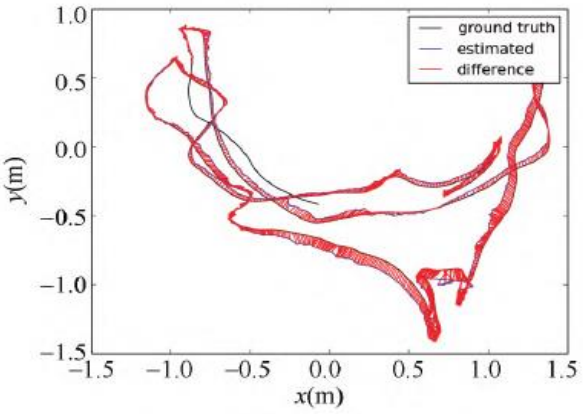
The data packet selected in the experiment is fr1 /room, which contains 1362 RGB image sequences and 1360 depth image sequences. It is attached with the acquisition time of image acquisition and the real trajectory of the robot. Because two image sequence is not one-to-one correspondence, so still need according to the acquisition time when using the matching, data set the official website to provide another called the associate. The file can help us to finish this part of the matching work. It should also be noted that the image resolution provided here is 640 x 480.

FIG. 7 shows the experimental results of this experiment. (a) the figure shows the environment point cloud map of the data set built by the system, consisting of 1049432 points; (b) the graph is the trajectory of the robot calculated by the system; (c) the figure shows the comparison between the actual trajectory of the robot and the estimated trajectory of the system, and the error is maintained at the centimeter level. In addtion by the figure (c) can be seen in the error distribution of the whole system is evener, explain the cumulative error is corrected well, no, the phenomenon of the drift, this also verify the effectiveness of the closed-loop detection module. The resulting environment map is clear and complete enough that the estimated trajectory is exactly the same as the actual trajectory.

(a) Point cloud figure



(b)Estimate the trajectory



(c)Comparison chart

Fig.7 The experimental results

6. CONCLUSION

This paper summarizes the development of visual SLAM, and proposes a closed-loop detection scheme based on k-means + + algorithm for RGBD camera, and constructs a 3d vision SLAM scheme with front-end and back-end frames. The closed loop detection module has successfully solved the problem of accumulated errors of the system and avoided the occurrence of map dislocation and trajectory drift. The experimental results of the data set verify the feasibility of the SLAM system and the validity of the closed loop module.

ACKNOWLEDGEMENTS

Thanks for the support of the science and innovation project of Shanghai University of Engineering Science, the project fund is Research on indoor navigation system based on IMU and Lidar (E3-0903-17-01106).

REFERENCES

[1] Thrun S, Leonard JJ. Simultaneous localization and mapping[M]. Springer Handbook of Robotics, Springer Berlin Heidelberg,2008: 871-889.

[2] Cadena C, Carlone L, Carrillo H, et al. Simultaneous localizationand mapping: present, future, and the robustperceptionage[J]. IEEE Transactions on Robotics,201632(6): 1-18.

[3] Smith R C, Cheeseman P. On the representation and estimation of spatial uncertainty[J]. TheInternational

Journal of Robotics Research,1986,5(4): 56-68.

[4] Zennaro S, Munaro M, Milani S, et al. Performance evaluation of the 1st and 2nd generation Kinect for multimedia applications[C]/ /Multimedia and Expo (ICME),2015 IEEE International Conference on, IEEE,2015: 1-6.

[5] Zhang Z. A flexible new technique for camera calibration[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2000,22(11): 1330-1334.

[6] Xin Guanxi. Research on simultaneous localization andmapping based on RGB-D Camera[D]. Harbin Institute of Technology, 2016.

[7] Chang Ming, Li Liang, Chen Zhiqiang. Study on the application of parameters of the converged binocular stereo vision system in CT scanning[J]. Chinese Journal of Stereology and Image Analysis,2011(1): 89-95.

[8] Lowe D G. Distinctive image features from scale-invariant key points[J]. International Journal of Computer Vision,2004,60(2): 91-110.

[9] Bay H, Ess A, Tuytelaars T, et al. Speeded-up robust features (SURF)[J]. Computer Vision and Image Understanding,2008,110(3): 346-359.

[10] Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF[C]/ /Computer Vision (ICCV) ,2011 IEEE International Conference on, IEEE,2011: 2564-2571.

[11] Gao Xiang, Zhang Tao. Fourteen Lessons of Visual SLAM: Form Theory to Practice[M]. Beijing: Publishing House of Electronics Industry,2017.

[12] Arthur D, Vassilvitskii S. K-means + +: The advantages of careful seeding [C]/ /Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics,2007:1027-1035.

[13] Wang Zhong, Liu Guiquan, Chen Enhong. A K-means algorithmbased on optimized initial center points[J]. Pattern Recognition and Artificial Intelligence,2009,22(2): 299-304.