Research on Loop Closing for SLAM Based on RGB-D Images

Hongwei Mo, Kai Wang, Haoran Wang and Weihao Ding

Abstract This paper mainly studies a loop closing detection method based on visual SLAM. We used RGB-D image as data source. The main idea is to construct a word bag based on DBoW3. Using rBRIEF makes it possible to perform feature extraction after the image is rotated. And added the elimination of mis-match links to improve the accuracy of detection. In order to ensure the reliability of the loop closing test results, the matching image is also verified. RGB-D image is rich in information and can synchronously extract the depth and color information of the main objects in the scene. The depth information directly reflects the distance information of each object in the scene.

Keywords SLAM · Loop closing · RGB-D · DBoW3

1 Introduction

SLAM is the abbreviation of Simultaneous Localization and Mapping. It refers to the subject carrying a specific sensor. In the absence of environmental prior information, it establishes an environmental model during the exercise process and estimates its own movement. Because of different sensors, there are laser-based SLAM and vision-based SLAM. This paper is focus on visual SLAM. The advantage of visual SLAM is that the camera generates rich texture information, which has great advantages in repositioning and scene classification. Especially in recent years, rapid development of science and technology has brought a new vision SLAM.

Visual SLAM system gave up the expensive laser and inertial measurement units, replaced it with a cheaper camera. The visual SLAM algorithm can build a 3D map of the world in real time and track the position and orientation of the camera. So, SLAM involved positioning technology, tracking technology and path planning technology. Loop closing detection is an important part of the entire SLAM system. The close loop

H. Mo (✉) · K. Wang · H. Wang · W. Ding
School of Automation, Harbin Engineering University, Harbin 150000, China
e-mail: mhonwei@163.com

Y. Jia et al. (eds.), *Proceedings of 2018 Chinese Intelligent Systems Conference*,
Lecture Notes in Electrical Engineering 528,
https://doi.org/10.1007/978-981-13-2288-4_70

detection module can provide some long-distance constraints apart from adjacent frames. The key of loop closing detection is how to effectively detect whether the camera has passed the same place. If it can be successfully detected, it can make the back end get more effective data, and finally get a globally consistent estimate. Therefore, in order to eliminate the error, loop closing detection is necessary in visual SLAM. It is also the key to ensure the quality of the construction and repositioning after losing position information, and the loop closing detection is more accurate constraint method than BA (Bundle Adjustment).

2 Related Work

There has been no major breakthrough in the loop closing detection method in recent years. Sivic and Zisserman [1] and others proposed in 2003 that the bag of words in text information retrieval converts the continuously changing features into discretized "words", and then use the statistical histogram of words to describe the scene. Newman and Ho [2] proposed to extract the feature descriptors from the pictures in 2005 and store them in a database. When matching, the images in the database are used for comparison and a value is returned. This value is used to detect whether a closed-loop event has occurred. This method relies on accurate position estimation and accurate positioning. Niser and Stewenius [3] proposed in 2006 that the local feature descriptors were hierarchically quantified into a vocabulary tree. Using the tree structure to increase the clustering speed on the basis of BoW, it was better at that time. Angeli and Filliat [4] proposed an online method for visual recognition of the effect of closed-loop detection on the size of the dictionary, using local shape and color information to monitor previous scenes, extending the BoW method, and constructing A quantitative dictionary that uses the classified image and Bayesian formula to estimate the likelihood of a loop closing. Cadena and Galvez-Lopez [5] used the complementary features of binocular cameras in 2010 to test loop closing detection in conjunction with the pouch method, ignoring geometrically discontinuous pictures, and experimentally verifying that CR-matching performed on the appearance of the scene. Liu [6] proposed a method for describing children using a compact feature in 2012. This method uses the low-dimensional descriptors available in a single image to perform image matching and PCA (Principal Component Analysis) for dimensionality reduction. At the same time, the computing efficiency of the computer can also be improved, and the correlation between different pictures can be found by using a particle filter method. Cumminsk and Newman [7, 8] proposed in 2007 the use of a probabilistic framework model to calculate two similar obstacles and whether they are the starting position. This method assumes that the map is known to achieve positioning, and in 2011 A SLAM construction method in a large scene is proposed [8], which is called a sparse approximation FAB-MAP model. The minimum spanning tree is used to describe the relationship between words, and the context information is used to further reduce the perceived ambiguity, achieving an ideal effect and becoming a test. The benchmark of the visual closed-loop detection

method. The improved closed-loop detection method is compared with the FAB-MAP method to verify the real-time, high-efficiency and robustness of the method. Tully and Kantor [9] proposed a hybrid map representation method in 2012. They use topological maps globally to optimize maps, locally use maps that are more suitable for human observation (such as grid maps and point clouds), and then use recursive Bayesian leaves to solve the closed-loop detection problem.

3 System Description

The hardware platform uses TurtleBot2 mobile robot, its mainly includes Kobuki mobile base, Kinect vision sensor, 2200 mAh battery. It need install ROS (robot operating system) on Ubuntu 14.04 as a development platform. And take a notebook to run the SLAM algorithm. The Kinect is used to capture RGB-D images. TurtleBot2 is an open source hardware platform and mobile base station. When using ROS software, TurtleBot2 can handle vision, positioning, communication and mobility, It can autonomously move to designated place.

4 RGB-D Feature

The RGB-D camera can observe the three-dimensional position of the landmark points each time. It can easily track of image feature pairs with more information. As show follows.

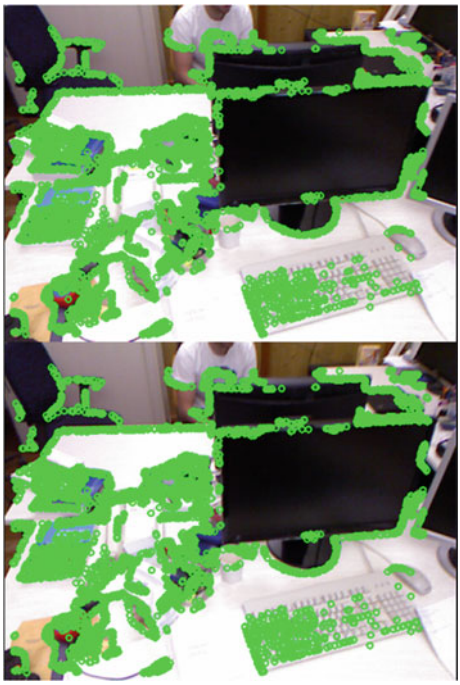
We need to get each pixel depth value in depth images, and the pixel should be marked by the index. If a pixel is marked, we should find the corresponding coordinate in color images.

5 DBoW3 Algorithm

BoW (bag of words) was originally used in text categorization and later used in image feature extraction and target detection. DBoW3 can sort image features and convert images into visual words bag. It uses the similar image features of hierarchical tree structures to gather together on physical storage to create a visual dictionary. It will generate an image database with sequential and reverse indexing. As shown in Fig. 1, the word in the dictionary is the leaf node in the tree. The weight of the word is stored in the inverted index. The contents of the direct index are mainly the characteristics of the image and the nodes in the dictionary tree related to the features.

In this paper, the rBRIEF used to make the binary descriptor space discretized, and generates a more concise dictionary, ranks the word bag by rank, and the whole dictionary is a above tree structure. Here rBRIEF is used instead of BRIEF in order to let the descriptor have rotation invariance. And find the direction of FAST key point as the direction of BRIEF. So, we will get a coordinate matrix S:

Fig. 1 RGB-D feature



$$S = \begin{pmatrix} x_1, x_2, \ldots, x_n \\ y_1, y_2, \ldots, y_n \end{pmatrix} \tag{1}$$

In this way can easily get the coordinates of the picture after the rotation transformation.

$$S_\theta = R_\theta S \tag{2}$$

In order to generate a dictionary tree, a large number of features are extracted from the training images. Their corresponding descriptors are discrete into K_w binary clusters according to the K-means++ algorithm. These clusters are the first level nodes of the dictionary tree, and the K-means++ algorithm for each node is carried out again. In order to generate second level nodes, L_w (L indicates the number of pairs of feature points) steps are followed according to this step, and eventually a W word dictionary tree is generated.

The K-means++ algorithm steps are as follows (Table 1).

As show in Fig. 2, a bag of word dictionary is established (Fig. 3).

The establishment process is as follows (Fig. 4).

According to the relevance of the word in the training center, each word is assigned a weight, those which appear frequently and do not have much effect on different images, assign a small weight, and have a significant allocation right for the words that distinguish the significant image.

Table 1 K-means++ algorithm

K-means++ algorithm step
Step 1: We select a sample point randomly in the data set as the first initialized cluster center Step 2: We calculate the distance between each sample point in the sample and the cluster center that has been initialized, and select the shortest distance between them Step 3: We select the sample with the largest distance as the new cluster center and repeat the above steps until the Kth cluster centers have been find Step 4: For the k initial cluster centers, we calculate the final cluster center using the K-means algorithm

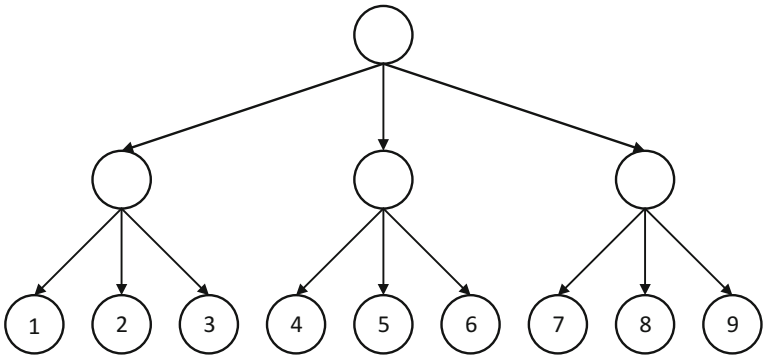


Fig. 2 Vocabulary tree

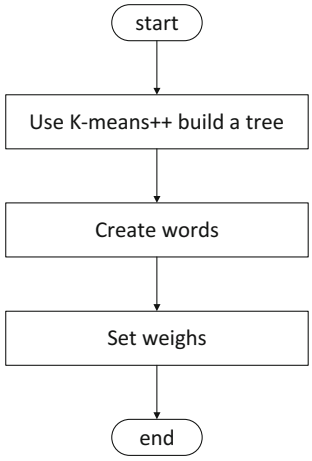


Fig. 3 The algorithm flow to create a dictionary

Given any eigenvalue f_i, as long as search in the dictionary layer-by-layer, and finally can find the corresponding word w_j, so, f_i and w_j come from same object.

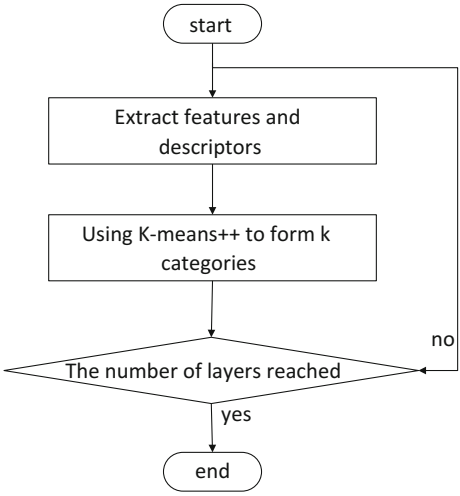


Fig. 4 The algorithm flow to create a tree

In order to distinguish the similarity of the two images, TF-IDF (Term Frequency-Inverse Document Frequency) is used here. Assuming that the number of all features is n, the number of features of the leaf node w_i is n_i, the IDF is:

$$IDF_i = \log \frac{n}{n_i} \tag{3}$$

TF refers to the frequency of occurrence of a feature in a single image. So, TF is:

$$TF_i = \frac{n_i}{n} \tag{4}$$

And the weight is:

$$weight_i = TF_i \times IDF_i \tag{5}$$

In this way, it can describe an image with a single vector. The similarity between two images can be calculated by the L1 norm of two vectors:

$$s(v_1 - v_2) = 2 \sum_{i=1}^{N} |v_{1i}| + |v_{2i}| - |v_{1i} - v_{2i}| \tag{6}$$

6 RANSAC Algorithm

In this paper, the most important step is use RANSAC (Random Sample Consensus) algorithm to eliminate mis-match after extract features. It is based on a set of sample

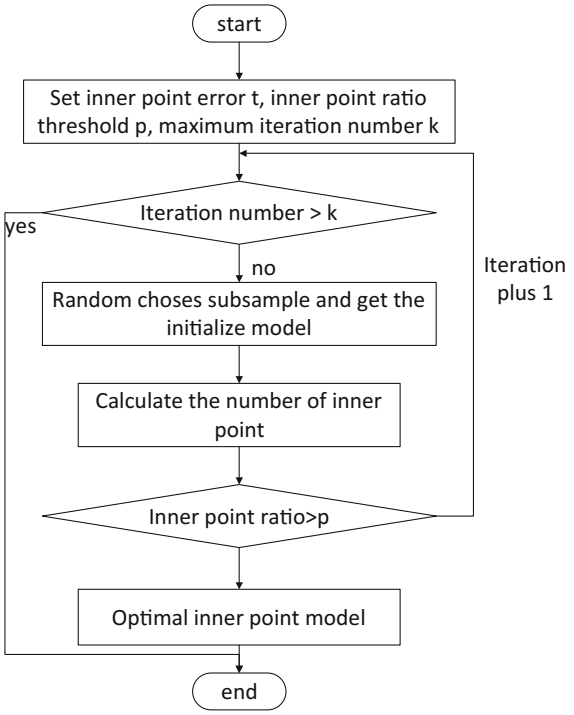


Fig. 5 The RANSAC algorithm flow

data sets containing abnormal data, calculating the mathematical model parameters of the data, and obtaining effective sample data algorithm. It first proposed in 1981 by Newman and Ho [2]. RANSAC algorithm is often used in computer vision. For example, in the field of stereo vision, the matching points of a pair of cameras and the calculation of the fundamental matrix are simultaneously solved.

The main step as follows (Fig. 5).

The input data of RANSAC algorithm is a set of observation data, a parameterized model that can be interpreted or adapted to observation data and some credible parameters. The model corresponds to the rotation and translation of one point cloud data in another space to another point cloud data. The first step is to get the point pair in a point cloud, and to use its invariant feature (distance of two points, normal vector angle of two point) as the index value of the hash table to search for a pair of corresponding points in the other point cloud, and it will calculate the parameter values of the rotation and translation. Then the transformation is applied to find other local points. And the algorithm need to recalculate the rotation and translation to the next state after finding the inner point. Then it iterates the above process to find the final location.

Apply RANSAC algorithm can remove points outside of consistency. The advantage of RANSAC is its robust estimation of model parameters. For example, it can estimate high-precision parameters from data sets containing a large number of out-

Fig. 6 Feature matching without RANSAC

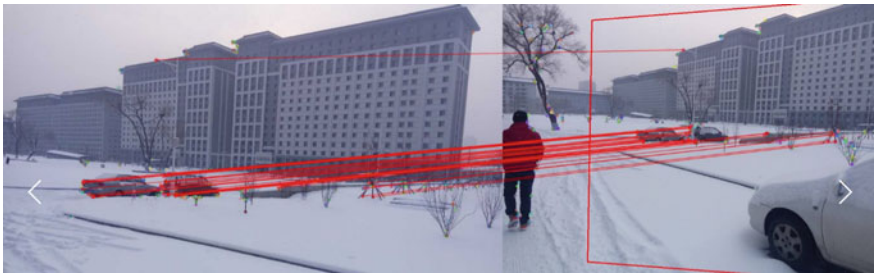


Fig. 7 RANSAC eliminate mis-match

side points. The disadvantage of RANSAC is that there is no upper limit on the number of iterations that it calculates, and if the upper limit of the number of iterations is set, the result may not be the best result or even the wrong result. RANSAC has only a certain probability to get a credible model, and the probability is proportional to the number of iterations. Another drawback of RANSAC is that it requires setting the threshold associated with the problem.

In the feature matching, the following situation occurs (Fig. 6).

As is show above, although most of the matches are correct, there are some matching errors, which constitute "contaminated observation data" and also the application conditions of RANSAC (Fig. 7).

Table 2 shows the similarity matrix obtained after entering the above algorithm with ten pictures. It can be seen that after setting a certain threshold, the algorithm cam be more accurately judged whether it is a loop closing.

7 Conclusions

In visual SLAM system, loop closing detection is an extension of data association. It is a process from point to point to face to face. In this paper, the image database structure used in the closed loop detection algorithm has a direct index in addition to the rank word bag and the inverted index. This structure makes the efficiency of

Table 2 Similar matrix

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}
I_1	1	0.251869	0.264077	0.258784	0.250996	0.268458	0.269992	0.257847	0.28638	0.241537
I_2		1	0.236969	0.264267	0.252454	0.245609	0.251969	0.2404556	0.254931	0.23764
I_3			1	0.227906	0.23721	0.238696	0.261195	0.21486	0.23448	0.235304
I_4				1	0.281962	0.242712	0.266108	0.225938	0.241647	0.238004
I_5					1	0.255106	0.248995	0.236341	0.213317	0.240724
I_6						1	0.239303	0.231276	0.2399	0.228895
I_7							1	0.246679	0.25424	0.2588312
I_8								1	0.217322	0.216933
I_9									1	0.228173
I_{10}										1

the geometric verification better. Using of binary descriptor rBRIEF makes it faster to extract image features and to calculate the distance between descriptors. When reference the scale, illumination and camera rotate, the BRIEF descriptor is unstable. So, using rBRIEF can improve the SLAM system performance.

Geometric verification is also needed for candidate loop closing images. And using RANSAC algorithm to find correspondence of local features between two images. The matching images are validated for determining the similarity between the two images when use the geometric verification, and the correct data association is also needed after the verification.

In an unknown environment, the mobile robot's awareness of the surrounding environment is a very basic function. Therefore, SLAM is an important direction to solve this problem. Loop closing detection can improve the accuracy of building a map and make the robot's autonomous positioning more accurate. Using RGB-D information can improve the efficiency of loop closing detection.

References

1. J. Sivic, A. Zisserman, Video google: a text retrieval approach to object matching in videos, in *2003 IEEE International Conference on Computer Vision*, vol. 2 (Nice, France, 2003), p. 1470
2. P. Newman, K. Ho, SLAM-loop closing with visually salient features, in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation* (Washington, America, 2005), pp. 635–642
3. D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (New York, America, 2006), pp. 2161–2168
4. A. Angeli, D. Filliat, Fast and incremental method for loop-closure detection using bags of visual words. IEEE Trans. Rob. **24**(5), 1027–1037 (2008)
5. C. Cadena, D. Galvez-Lopez, Robust place recognition with stereo cameras, in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 25, no. 1 (Taipei, Taiwan 2010), pp. 5182–5189
6. Y. Liu, H. Zhang, Visual loop closure detection with a compact image descriptor, in *2012 IEEE/RSJ International Conference on Intelligent Robots & Systems*, vol. 57, no. 1 (Vilamoura, Algarve, 2012), pp. 1051–1056
7. M. Cummins, P. Newman, Probabilistic appearance based navigation and loop closing, in *IEEE International Conference on Robots & Automation* (Roma, Italy, 2007), pp. 2042–2048
8. M. Cummins, P. Newman, Appearance-only SLAM at large scale with FAB-MAP 2.0. Int. J. Robot. Res. **30**(9), 1100–1123 (2011)
9. S. Tully, G. Kantor, A unified bayesian framework for global localization and SLAM in hybrid metric topological maps. Int. J. Robot. Res. **31**(31), 271–288 (2012)
10. F.M. Campos, L. Correia, Mobile robot global localization with non-quantized SIFT features, in *IEEE 15th International Conference on Advanced Robotics: New Boundaries for Robotics* (Tallin, 2011), pp. 582–587
11. R. Mur-Artal, J.D. Tardos, ORB-SLAM2: an open-source system for monocular, stereo, and RGB-D cameras. IEEE Trans. Rob. **PP**(99), 1–8 (2016)
12. C. Kerl, J. Sturm, Dense visual SLAM for RGB-D cameras, in *2014 IEEE/RSJ International Conference on Intelligent Robots & Systems*, vol. 8215, no. 2 (Chicago, America, 2014), pp. 2100–2106
13. Y. Latif, C. Cadena, Robust loop closing over time for pose graph SLAM. Int. J. Rob. Res. **32**(14), 1611–1626 (2013)