Location and Map Reconstruction based on Monocular SLAM with CNN Learned Information

Di Cui [a], Jianjun Fang [b], Dawei Luo [c]

School of BUU, Beijing Union University, Beijing 100101, China

[a]945449009@qq.com, [b]jianjun@edu.buu.cn, [c] 1361722864@qq.com

Abstract. a method based on Convolutional Neural Network for depth prediction and monocular SLAM (simultaneous localization and mapping) is proposed for the problem of time-consuming and scale uncertainty. Firstly, the RGB image is extracted and matched, and the 3D information of SLAM feature points is obtained by the depth prediction of the convolutional neural network. Then the camera position is solved by the linear optimization. Finally, the motion trajectory and the three–dimensional dense point cloud are potted by loop closure and optimized global pose. Experimental results based on standard test set show that the method of information fusion based on convolutional neural network depth prediction and monocular SLAM can improve the accuracy of SLAM system mapping.

Keywords: monocular sensor; simultaneous localization and mapping; Convolutional Neural Network; Deep learning; feature matching; linear optimization.

1. Introduction

With the research and development of intelligent robots, Simultaneous Localization and Mapping (SLAM) has attracted more and more attention. It is an implementation version of the structure from motion (SFM), and the sensors used for visual simultaneous positioning and mapping (V-SLAM) are mainly cameras. There are many papers about SLAM, and the feature-based visual SLAM method refers to The feature points are detected and extracted from the input image, and the pose estimation of the camera is calculated based on 2-D or 3-D feature matching and the environment is constructed. If the entire image is processed, the computational complexity is too high, and the feature is widely used because it effectively reduces the amount of calculation while saving important information of the image. Classified from front-end visual odometers, we are divided into monocular cameras [1,2,3] and depth cameras [4,5]. For monocular SLAM, the ORB SLAM [3] method uses the sparse ORB feature points from the input image to perform pose estimation and scene reconstruction through feature point matching between frames. BA(Bundle Adjustment) and The method of pose optimization. While the above mentioned most popular visual SLAM algorithm, ORB-SLAM [3] is the key to extract key points through "direct orientation", and has not added the constraints of depth information. Therefore, this paper proposes a SLAM algorithm based on convolutional neural network and applies it to front-end pose estimation and back-end optimization for the problem of algorithm initialization time-consuming and scale uncertainty due to the inability of monocular camera to obtain depth information.

2. Related Work

In this section, we describe the principles and algorithms of SLAM based on convolutional neural networks. Firstly, for this algorithm, we propose a framework as shown in Figure 1. The system consists of two parts: the image front end, including feature extraction, feature point depth information, 3D feature point filtering, feature matching, motion estimation and closed loop. Detection part; back-end optimization includes pose global optimization, map construction and motion trajectory.

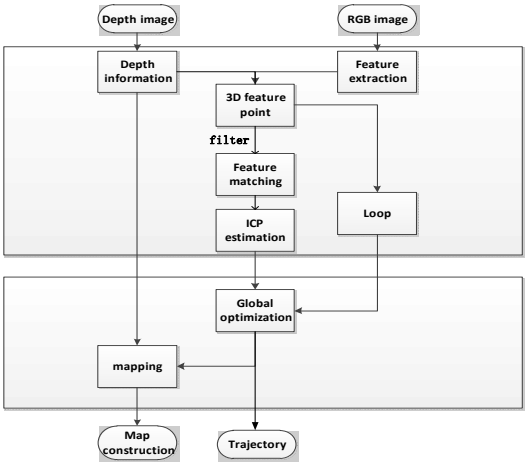

Figure 1. CNN and monocular SLAM fusion framework

3. ORB-SLAM Design Theory

Representative part of the general image, with weight Refolding, distinguishability, and efficiency, we call it image features. Taking Figure 2 as an example, we can use the corners, edges and blocks in the image as representative parts of the image.

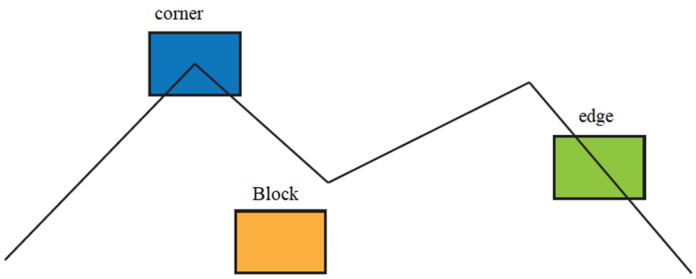

Figure 2. Image Features

FAST original feature points having no directional information and scale problems exist, in order to increase scale invariance, ORB-SLAM [3] to construct an image pyramid, and detecting corner points at each level of the pyramid. In order to obtain rotation invariant properties, Rublee use of the concept of the luminance centroid (Intensity Centroid) to calculate the direction of the feature point. In a small image block B, the image block is defined moments:

$$m_{pq} = \sum_{x,y \in B} x^p y^q I(x,y), \ p,q = \{0,1\}$$

I (x, y) represents the pixel value at any point above the image, p and q are two parameter values, and their values are 0 or 1. The centroid of the image block can be found by the moment:

$$C = (\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}})$$

Connecting the center O of the image block geometric centroid C, giving a direction vector, then the direction of the feature point can be defined as:

$$\theta = \arctan(m_{01} / m_{10})$$

By the above method, FAST having the corner points will be described with a rotating scale, greatly increasing its robustness expression between different images.

4. Improvement of ORB-SLAM

The SLAM algorithm based on convolutional neural network can directly obtain 3-d information. This paper designs an algorithm for this feature, adding more constraints for feature point extraction and matching. The idea is as shown in Algorithm 1. When the feature point depth information threshold satisfies the algorithm 1, it is determined to be a feature point, and the feature matching of the following figure description is performed. Instead, define it as a non-feature point and cull it. And Many of the most recent depth predictions have depth prediction architectures. This paper uses the most advanced method proposed in the literature [6], based on the Residual Extension Network (ResNet) architecture [7] to the complete convolutional network. We use small convolution instead of large convolution in FCRN. On the one hand, we can reduce the chessboard effect. On the other hand, we keep the edge information as much as possible. And using the method of [6], it can improve its speed, reduce parameters, and improve real-time performance, which are all needed by SLAM.

algorithm 1: Filter feature points based on depth information
1: Input: RGB images I_{R1} and I_{R2}, depth images I_{D1} and I_{D2}, camera parameters.
2: Extracting feature points using Oriented FAST;
3: Obtaining feature point depth information from I_{D1} and I_{D2}; 4: Assume feature point of the image I_{R1} is P={p1,...,pn};
5: Taking the feature point p1 as an example, taking 15 pixels on the circle with radius 3 as the center, and setting the average depth of 15 pixels to be Dp1';
6: Dp1 with Dp1 'into the formula: $$D_{pi}^{'} > 0 \&\& D_{pi} > 0 \&\& 125 > \frac{D_{pi}^{'} * 100}{D_{pi}} > 75$$
7: The feature points in Equation 6 will be satisfied, and the Fast Approximate Nearest Neighbor (FLANN) algorithm performs feature matching on I_{R1} and I_{R2} to obtain matches; 8: After obtaining the three-dimensional point is calculated rotation matrix R and translation matrix T; 9: If this happens: $$\frac{D_{pi}^{'} * 100}{D_{pi}} < 75 \,||\, \frac{D_{pi}^{'} * 100}{D_{pi}} > 125$$
10: It is considered that the depth similarity gap is too large, and it is judged that pi does not have the feature point property, discarding the point and performing feature matching on the points in accordance with Equation 5 to obtain matches; 11: After obtaining the three-dimensional point pair, calculating the rotation matrix R and the Translation matrix t ;

The monocular SLAM with CNN depth prediction feature matching effect is shown in the figure 3(a), 3(b). And the feature extraction algorithm efficiency comparison is shown in the table 1.

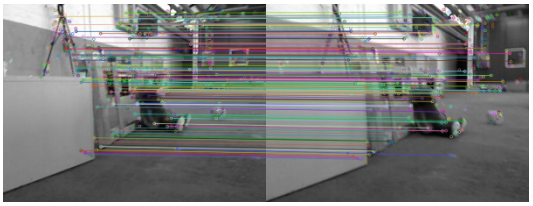

Figure 3(a) ORB feature matching

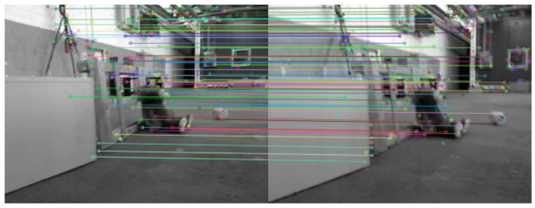

Figure 3(b) Feature matching after screening

Table 1. Feature extraction algorithm efficiency comparison

algorithm	Extraction time (ms)	Number of feature points	effective points	Matching accuracy
SIFT	364	500 470	150 78	52.0%
SURF	130	499 365	166 98	59.0%
ORB	15	500 500	172 55	31.9%
OURS	18	480 469	196 102	52.1%

The upper limit of the number of extracted features experimental threshold 500. As can be seen in the largest number of the same scene, the ORB algorithm extracted by the feature point Table 1, followed by the SIFT method, followed of OURS minimum SURF algorithm. However, the matching rate is the highest on the OURS and the ORB algorithm is the lowest. In terms of extraction time, the extraction time of SIFT algorithm and SURF algorithm exceeds 34ms, while the frequency of monocular camera is 30Hz. Therefore, in the absence of GPU processing acceleration, real-time performance cannot be satisfied. In summary, our algorithm satisfies the high matching accuracy rate and satisfies the real-time requirements of the SLAM system.

5. Conclusion

In order to improve the feature extraction and matching accuracy in the monocular SLAM algorithm, this paper proposes a method based on convolutional neural network for depth prediction and monocular vision SLAM for information fusion and mapping. By convolving the neural network to overcome the absolute scale problem of the monocular camera, a three-dimensional dense map is constructed, which improves the accuracy of map construction. The method of this paper can accurately locate the position of the robot. The root mean square error of the pose estimation trajectory is centimeter level. In an ideal environment, the error can be reduced to the millimeter level. In terms of depth estimation, it is also achieved with higher accuracy.

References

[1]. Liu H M, Zhang G F, Bao H J. A survey of monocular simultaneous localization and mapping [J]. Journal of Computer-Aided Design and Computer Graphics, 2016, 28 (6): 855-868.

[2]. Bonin-Font F, Ortiz A, Oliver G. Visual navigation for mobile robots: a survey [J]. Journal of Intelligent and Robotic Systems, 2008, 53 (3): 263 -296.

[3]. R. Mur-Artal, J. M. M. Montiel, and J. D. Tards. Orb-slam: A versatile and accurate monocular slam system[C]// IEEE Transactions on Robotics.IEEE:2015.

[4]. R. A. Newcombe, A. J. Davison, S. Izadi, et al. KinectFusion: Real-time dense surface mapping and tracking[C]// IEEE International Symposium on Mixed and Augmented Reality.IEEE:2011.

[5]. M. Keller, D. Lefloch, M. Lambers, et al. Real-Time 3D Reconstruction in Dynamic Scenes Using Point-Based Fusion[C]// International Conference on 3D Vision.3DV:2013.

[6]. I. Laina, C. Rupprecht, V. Belagiannis, et al. Deeper depth prediction with fully convolutional residual networks[C]// IEEE International Conference on 3D Vision. 3DV:2016.

[7]. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition.CVPR: 2016.