Analysis of feature point matching technology in SLAM based on binocular vision

Mingda Yu*

School of Mechanical and Energy Engineering, Tongji University, Shanghai, China

* Corresponding Author Email: 2052032@tongji.edu.cn

Abstract. With rapidly developing computing technology, intelligent robots have been used extensively in many areas of life, improving work efficiency and lowering labor costs. At the same time, with the increasing number of robot application scenarios, the demand for robot intelligence has also increased. Computer vision-based simultaneous localization and mapping (SLAM) technology is a key technology that helps robots achieve real-time positioning and navigation in order to improve their intelligence. Feature point matching is an important component of this technology. This paper mainly analyzes the current development status of feature point matching technology in visual SLAM. Firstly, a brief introduction was given to the three widely used SLAMs, and the advantages and disadvantages of different SLAM technologies were analyzed. Second, a brief presentation and explanation of the working principle of binocular SLAM technology were provided. Then, the basic principles and key points of feature point recognition and matching techniques in several different algorithms for visual SLAM were summarized. Analyzed the advantages and disadvantages of various algorithms, as well as improvements based on this algorithm. Finally, summarize and provide suggestions for the future development of visual SLAM technology.

Keywords: Computer vision; SLAM; Feature point matching algorithm.

1. Introduction

With the rapid technical development, the rapid upgrading of computing power, and the continuous popularization of robots in daily life, the uniqueness of Simultaneous Localization and Mapping (SLAM) technology has been explored and widely applied. SLAM technology is applied to mobile robots to achieve autonomous localization and environment construction of robots; Applied to drones, it can also be used for terrain exploration and target recognition; Applied to sweeping robots, it can model and track indoor environments and clean rooms [1].

The current SLAM applications can be roughly divided into two categories in terms of hardware: laser SLAM and visual SLAM. Among them, the laser SLAM (sensor is LiDAR) calculates the distance based on the time when the laser beam is emitted and returned, thereby simulating the surrounding environment; Visual SLAM (sensor is camera) detects the surrounding environment through the images taken by the camera. In the current era of rapid development of CPU and GPU computing power, visual SLAM, which has cheaper hardware devices and richer information transmission, has an advantage in processing image information and subsequent information processing. Meanwhile, due to the fact that laser SLAM works by emitting and receiving lasers through LiDAR, there will be large blind spots due to distance during scanning and composition in open outdoor areas. The information returned by the LiDAR sensor is mostly the distance and orientation of the surrounding environment entities and obstacles from the car. The SLAM system with the camera as the main sensor is called the Visual SLAM (VSLAM) system. The data source of visual SLAM is image information, which can be processed through subsequent recognition and Semantic information processing, and is more versatile. The core issue in visual SLAM is how to estimate camera motion based on images. The visual odometer was born to solve this problem, and its implementation methods in visual SLAM mainly include two methods: method of direct and method of characteristic points. The method of characteristic points is widely used for its simplicity and directness. The approach to the characteristic points method is about selecting some representative points from the image, which will unchanged even after the change of camera's

perspective. Afterwards, identify the same points in each image and discuss the issue of camera pose estimation based on these points. In the above process, the selection of feature points and matching algorithms will directly affect the image information processing and the estimation of camera pose changes. When the feature points are too dense, redundant, and have poor uniformity, it will result in low SLAM positioning accuracy and computational efficiency. When feature points are too sparse, difficulties will be encountered in matching feature points [2].

In summary, feature point detection and matching technology between video frames is an important technical support in visual SLAM. This article reviews and analyzes the current development status of feature point matching methods in the field of visual SLAM technology to address this issue. Firstly, the main classifications and corresponding characteristics of SLAM technology are introduced. Subsequently, a comprehensive overview was provided on the technical composition of visual SLAM. Then, the feature point matching algorithm in visual SLAM is introduced, and its recent citation and development are elaborated in conjunction with relevant literature. Finally, a summary of the entire article is provided and its future development direction is prospected.

2. SLAM classification

SLAM refers to synchronous positioning and map construction. The two issues mentioned above are respectively interdependent. Accurate map construction is the technological cornerstone for robots to achieve high-precision positioning. At the same time, high-precision real-time positioning also provides guarantees for the accurate construction of maps. Currently, there are three main types of SLAM classified based on hardware devices: visual SLAM, laser SLAM, and RGB D SLAM [1].

2.1. Visual SLAM

Visual SLAM mainly uses data collected by cameras for synchronous positioning and map construction. Simple structure, convenient installation, and low cost is the advantage of the Visual SLAM At the same time, compared with the laser radar, the visual SLAM has more abundant information and can process Semantic information in the later stage.

However, at the same time, visual SLAM also has the drawbacks of low accuracy, frequent accumulated map construction errors, and the need for loop detection. And visual SLAM has strict requirements for light and environment, making it difficult to work in the dark or without texture [1, 3].

2.2. Laser SLAM

Building scenes based on LiDAR is an ancient and reliable method, and LiDAR is also the most commonly used SLAM sensor for precise scenes. Lidar calculates the distance information between the robot body and surrounding obstacles by calculating the time interval between emitting and receiving lasers. Common LiDAR, such as SICK, Velodyne, Rplidar, etc., can be applied to laser SLAM. Due to the long development and application time, the relevant technology of LiDAR is relatively mature, with the advantages of high accuracy and small error. And the mapping results obtained from LiDAR scanning can be used to achieve path planning. However, due to the limitations of its working principle, LiDAR often cannot work well on open terrain. At the same time, the information obtained through LiDAR scanning lacks semantic meaning and cannot be processed more deeply [1, 4].

2.3. RGB-D SLAM

The image information collected by SLAM based on RGB D depth camera is two images of RGB and depth. Common RGB-D SLAMs include Kinect Fusion, Elastic Fusion, Kintinous, RGBD SLAM2, and RTAB SLAM. Compared to other binocular cameras, texture information is ignored and only the encoded depth projected by the infrared camera is calculated. But at the same time, due

to the fact that sunlight can wash away the infrared code, RGB-D cameras cannot be used outdoors [5].

3. SLAM Technology Based on Binocular Vision

Among many machine vision systems, binocular stereo vision is one of the important methods. Based on the existing two cameras, the geometric information of the object is obtained through the principle of parallax, and the positional deviation of the object in different cameras is derived. By using the above method, the position change of the target object can be estimated to achieve localization function [6]. The principle of the binocular vision model is shown in Fig. 1.

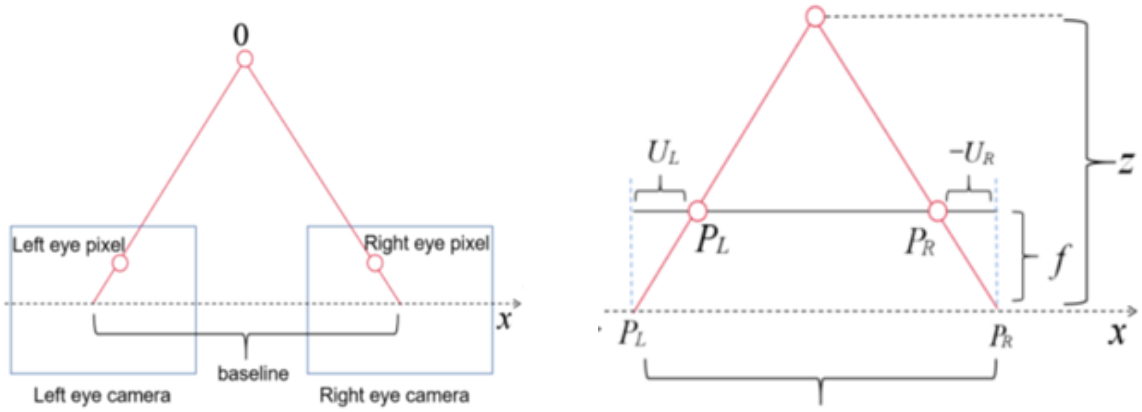

Fig. 1 Binocular Visual Model [7].

As shown in Fig. 1, after the detection and matching of target points are completed using a binocular vision system, the positional relationship between them relative to the camera can be perceived. The traditional binocular vision SLAM technology has developed relatively well. The three are approximately two type's method form in the visual SLAM: method of characteristic point and method of direct. People favored the method of characteristic due to its more prominent accuracy [8]. The operation mode and process are shown in Fig. 2. The keyframe image information captured by the camera will be sent to the visual odometer. The visual odometer will determine the position and posture of the robot by analyzing and processing relevant image sequences.

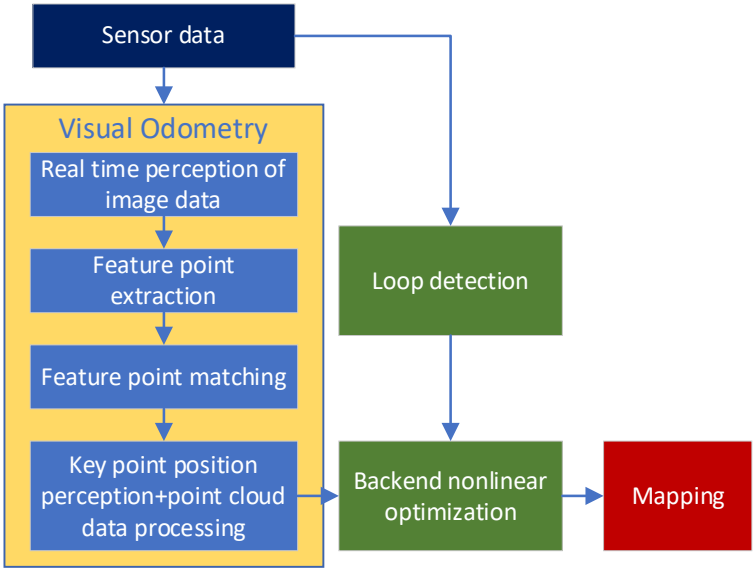

Fig 2. Traditional Binocular SLAM Process.

In visual odometry, feature point extraction is performed on real-time captured images based on different descriptors of different algorithms, and perform feature point matching on binocular captured images. The keyframe feature points obtained by a binocular camera will be matched to obtain a series of key points and their point clouds. The visual odometer will calculate and analyze the pose changes of the robot through the key point cloud obtained from each frame, and transmit the data to the backend optimization for mapping. Simultaneously perform loop back detection to eliminate errors generated during the process [9].

4. Feature point matching algorithm in visual SLAM

Due to different methods and descriptors for extracting feature points from camera captured images, subsequent feature point matching methods also differ. At present, the commonly used algorithms for feature point extraction and matching in visual SLAM include classic ORB algorithm, SIFT algorithm, SURF algorithm, and so on. The following will provide a brief introduction and analysis of existing feature point matching algorithms.

4.1. ORB-based algorithm

The ORB algorithm is a relatively mature algorithm in visual SLAM, and many improved algorithms have emerged based on the ORB algorithm framework. The traditional ORB algorithm detects feature points of target image through FAST algorithm. Identify pixel points which differ significantly from surrounding pixels in the image is the core of this algorithm. After that, label them as feature points. The FAST algorithm has several steps: first, select one pixel of the image, set its grayscale value to h, and select 16 pixels with a radius of 3 pixels and a circle centered on the pixel. Next, set the threshold t to 30% and determine that when the absolute value of the difference between the grayscale values of two points is greater than, it is considered that these two points are different. Finally, corner detection is completed by calculating and determining the 16 pixels around the pixel.

After obtaining a feature point, a specific descriptor is required to describe the relevant information of the feature point. The ORB algorithm uses BRIEF descriptors for description: the points around the feature points are paired in a descriptive manner to form descriptors. But this process will be affected by rotation and scaling. Therefore, to solve the scaling problem, ORB adopts a comparison of key frames before and after the pyramid type, put forward one method to centroid moment to determine the position of the coordinate axis in the BRIEF descriptor. Make the description of key points unaffected by image rotation.

The traditional ORB algorithm is simple and computationally fast. However, traditional ORB cannot ieally performance in handling dynamic objects and has the disadvantage of cannot control the amount of feature points. The quantity and accuracy of surrounding descriptors and feature points are the influencing factors of visual SLAM. Redundant and poorly described feature points can also contribute to diminish the accuracy of the match. The researchers suggested a number of improved methods for addressing the above questions. Jianjun Ni et al. suggested the concept of reliability based on their prior research. Reliability is defined as:

$$\mathrm{Re}\,l(P) = \frac{\sum_{i=0}^{16} \left| I_{x_i} - I_P \right|}{t} \tag{1}$$

The process of FAST angle point's selection is about comparing the pixel values of the characteristic points with the surrounding 16 points. If difference between pixel values and surrounding pixel points is greater, then the probability that that point will become a characteristic point is greater. The improved algorithm use the the definition of reliability to control the number of feature points within a predetermined range.

The algorithm then sorts all the characteristic points according to their corresponding reliability. When the number of T-points exceeds pre-determined H-threshold. The algorithm will choose the most reliable Reltop feature points to be the final Reltop feature points. Reltop is defined as:

$$\mathrm{Re}ltop = H + \beta T \tag{2}$$

Where β is the feature coefficient used to limit the number of feature points? The selection of feature points will be based on finding general FAST corners until the number reaches the preset upper limit [10].

The researchers of the team conducted several experiments and analyzed the experimental results to demonstrate the availability of FAST-SIFT algorithm. The experiment was conducted on the public dataset TUM to test the performance of visual algorithms [11]. The test results show: Improved SLAM algorithm achieve better positioning accuracy than ordinary ORB-SLAM systems and has good performance.

4.2. SIFT-based algorithm

In 1999, Canadian professor David G. Lowe proposed a new algorithm: the SIFT algorithm, which is a feature description algorithm. This algorithm has excellent registration performance and maintains stability against factors such as lighting changes, scaling, rotation, and brightness changes. It also has good robustness [12]. The general process of the SIFT algorithm is as follows: First, the original image is filtered by a function of the Gaussian nucleus. Next, perform a scale change to detect points of more stable characteristics and maintain the scale invariance. Second, approximate extremes have been found in the previous steps. Accurately locate key points through Taylor expansion, and eliminate unstable edge response points through Hessian matrix. Finally, gather the image of the Gaussian pyramid where the key points are located 3 σ the gradient and directional distribution of pixels in the neighbourhood.

However, SIFT also has drawbacks such as long computation time, large and cumbersome computation, and inability to track in real-time. At the same time, the results of SIFT detection may also contain useless feature points. Based on the SIFT algorithm, researchers have proposed many improved methods. Xixi Fu et al. Combining SIFT algorithm with FAST algorithm for recognition and localization. The SIFT algorithm will encounter difficulties in extracting feature points of edge smooth targets, and unable to detect enough feature points. The FAST algorithm can extract more feature points with less computational time without the use of scale features and directional information. This feature can effectively compensate for the shortcomings of the SIFT algorithm. The local feature detection algorithm that combines the above two algorithms has the advantages of both algorithms and compensates for their shortcomings, which can quickly and accurately extract local features of the image. The improved algorithm process is roughly as follows: first, segment the image into different blocks and remove most of the specific candidate points. Then, the ID3 algorithm is used to sequentially judge the candidate corner points and determine whether the pixels are corners. Generate a decision tree based on the pixel with the highest pixel gain to achieve correct classification of diagonal points. Subsequently, diagonal features were verified: non maximum suppression method was used to exclude non corner points, and the remaining points were extracted corner points. Use the improved algorithm for corner detection, and use it to replace and accelerate the corner detection process in the SIFT algorithm. Next, check the diagonal features: Use nonmaximum suppression methods to exclude non angular points. The remaining dots are excerpted corner dots. Use the enhanced algorithm to detect corners, and use it to replace and speed up the process of detecting corners in the SIFT algorithm. Subsequently, upon completion of the key point gradient calculation, 32 directional vectors are computed centered around the corner to describe it. And generate a gradient histogram to calculate the gradient and pixel direction within the field. Count gradient information to create a 128-dimensional characteristic vector and standardize it. Finally, the descriptor is computed based on the local characteristics of the characteristic points to match the key points [10].

This improved algorithm bypasses the computationally expensive and time-consuming steps of generating Gaussian difference pyramids and detecting extreme points in traditional SIFT algorithms through the FAST corner detection method. Improved the efficiency and accuracy of extracting image feature points. At the same time, the special point descriptor generated by SIFT method has stability and rotation invariance, making the whole feature point process process more accurate and fast [12]. This study utilized the MATLAB simulation tools to analyze the above algorithms and compare the performance of the FAST SHIFT algorithm with the FAST and SIFT algorithms. Under the same conditions of image extraction, the number of removed feature points, the time and the results of the algorithm description were compared. The results show that the improved algorithm preserves the original image features and removes some useless feature points during the process of extracting feature points. The improved algorithm extracts significantly fewer feature points than the original SIFT algorithm. After improvement of the algorithm, the front-end performance has been greatly improved. It can more accurately and quickly extract feature points. Besides, the feature points also have more scale features and directional information.

4.3. SURF-based algorithm

SURF is an enhancement of the SIFT algorithm, which improves its performance efficiency and offers the possibility of its application in real-time computer vision systems. Initially published by Herbert Bay on ECCV in 2006, it has been officially published in Computer Vision and Image Understanding in 2008 [13]. SURF is an enhancement of the SIFT algorithm, which improves its performance efficiency and offers the possibility of its application in real-time computer vision systems. In the same way as the SIFT algorithm, the basic path of the SURF algorithm can be divided into: the construction of a Hessian matrix. Generate feature points. Build a space of scale. Positioning of the characteristic spots. Allocation of primary direction of points of interest. Produce feature point descriptors. Matching characteristics [14]. The SURF algorithm is stable and keeps the invariance at rotation, scale transformation and brightness. It also exhibits a degree of stability in angle changes and noise. The downside is that real-time performance is not high, and the ability to extract attribute points for smooth edge targets is not high.

Based on the advantages of SURF algorithm, many research teams have further improved it and combined it with other algorithms. Tong Zhang et al. have addressed the issue of poor accuracy in traditional point line feature visual SLAM systems under weak indoor textures and changes in brightness, which cannot be processed in real-time. According to the point line vision, the research team put forward an inertial SLAM method, which suitable for indoor weak texture and weak lighting situations [15]. The feature extraction algorithm in this study is the SURF algorithm. The matching algorithm uses the FLNN algorithm linked to SURF parameters [16]. The experimental results of this study on public datasets are as follows: situation using the original SURF algorithm are shown in Fig. 3, the matching situation using the improved algorithm are shown in Fig.4

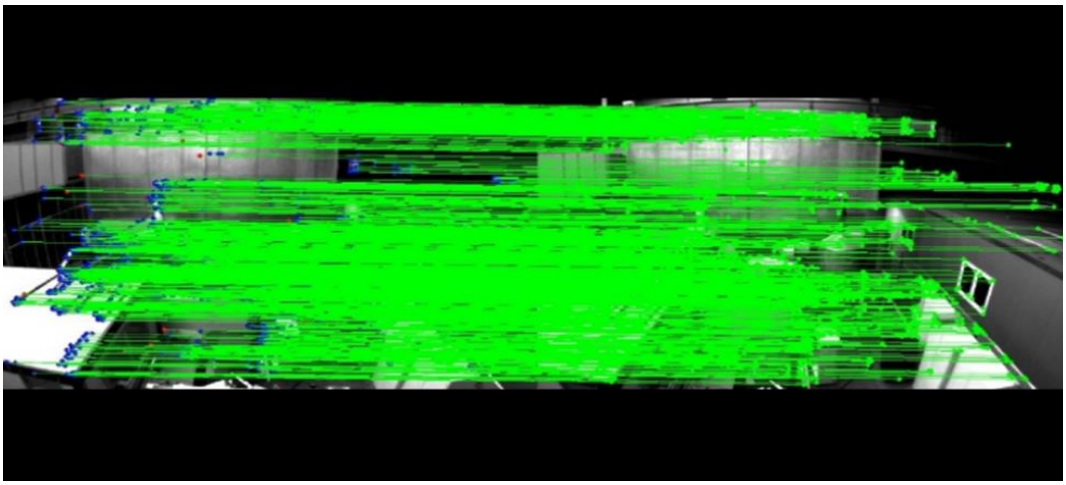

Fig 3. When using SURF algorithm for matching [16].

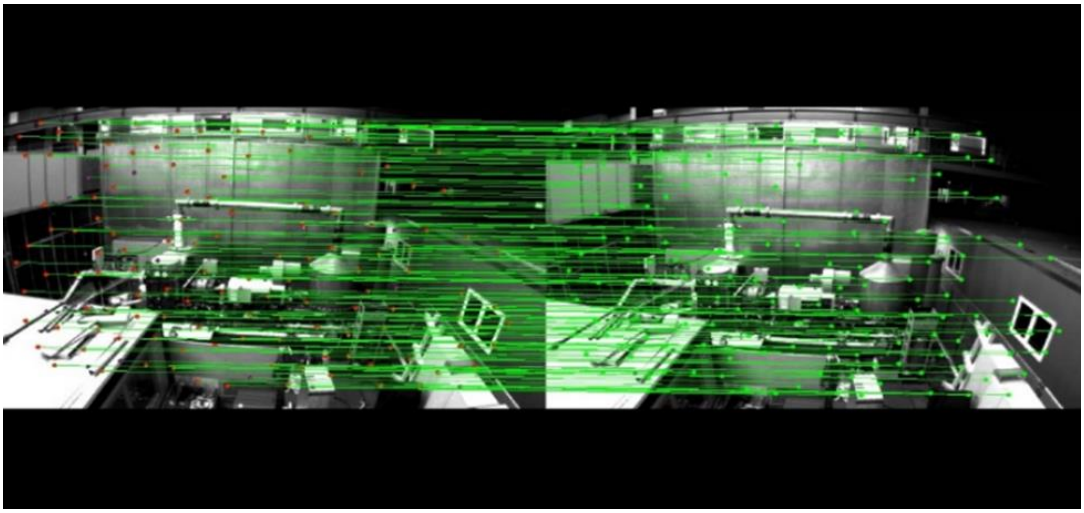

Fig 4. Effect of using improved matching algorithm [16].

The research results of the project team indicate that there is a large number of SURF based feature matching, but mismatches still exist. After adding the FLANN algorithm, system eliminated some mismatched and redundant point features and the matching results of feature points become better.

4.4. Neural network-based methods

Unlike the aforementioned algorithms based on feature points and descriptors, semantic based neural network algorithms have gradually received attention in recent years. The commonly used neural network algorithms currently include convolutional neural networks (CNN), generative adversarial networks (GAN) and recurrent neural networks (RNN).

The first algorithm to apply deep learning to object detection is the R-CNN algorithm. R-CNN follows the traditional approach of object detection, using extraction boxes to extract features from each box, image classification, and non-maximum suppression for object detection. In the feature extraction step, replace traditional features with features extracted by deep convolutional networks [17].

Mask R-CNN is a masking region convolutional neural network, which can perform pixel level segmentation of images in real-time projects. It is also a method to train and support Semantic information. The network structure is shown in Fig. 5 [18].

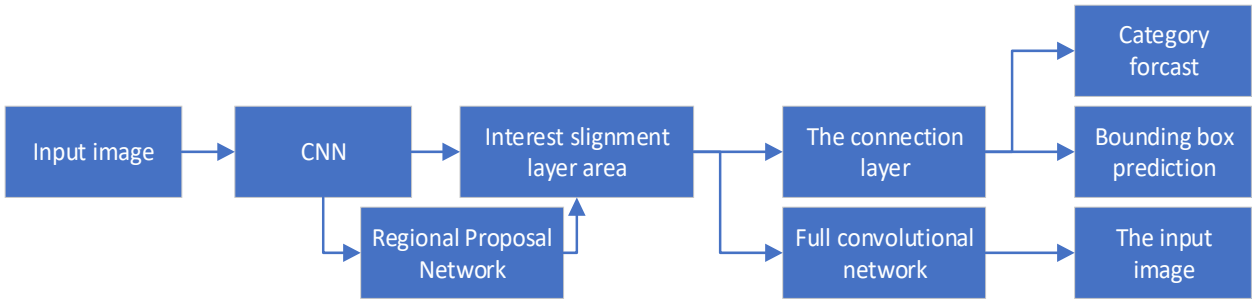

Fig 5. Mask R-CNN network structure.

In terms of advantages, Mask R-CNN introduces a Mask Head mask structure to enhance the prediction performance of the model, achieving pixel level segmentation mask prediction; Using ROI Align instead of ROI Pooling eliminates the coarse quantization of POI Pooling, resulting in good alignment between the extracted features and the input. However, at the same time, Mask R-CNN also has the disadvantage of causing some negative impacts on the model's prediction performance on the basis of low time complexity. This limitation was compensated for in the later Mask Scoring R-CNN algorithm.

Due to the power of the Mask R-CNN algorithm, some research teams have started to attempt to introduce the Mask R-CNN algorithm into the visual SLAM. The research team of Feng Zhang and al. has utilized convolutive neural networks to study visual SLAM in dynamic environments [18]. The team used the information from the R-CNN and ORB SLAM mask to determine the camera position, and filter the required feature points through the mask information generated by Mask R-CNN. The image features act as localization markers for SLAM. The system adopts the method of eliminating extremely dynamic property points to eliminate dynamic objects that cannot be recognized by convolutional neural networks. After improvement, the similarity of the descriptors of ORB characteristics was reduced. Figure 6 shows the flow diagram of the enhanced algorithm in this study.

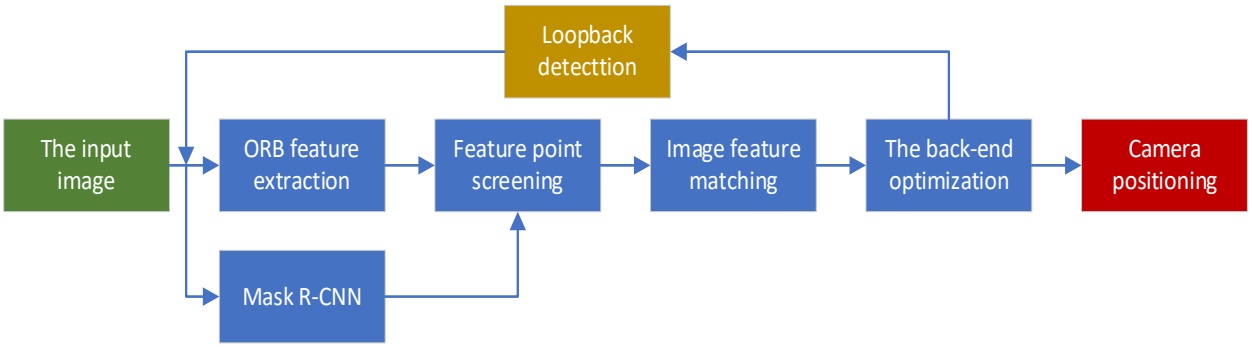

Fig 6. Flow chart of improved algorithm.

For the pre-improved algorithm, all feature points of the image are required for feature point matching, resulting in low operational efficiency [19]. Mask R-CNN divides the image into multiple different regions using bounding boxes based on semantic labeling [20]. When matching feature points, only a small range of matching calculations need to be performed within the same region. This reduces the matching error rate, improves the accuracy of pose matching, and improves the overall feature matching effect. The research team tested the improved algorithm on a publicly available dataset. After comparing with the original ORB SLAM algorithm, test data shows that in a real-time online environment, the accuracy of the algorithm combined with the neural network has been significantly improved, exceeding 96.0% [18].

5. Conclusion

From the above several cases of algorithm improvement and combination, it can be seen that the traditional ORB algorithm and SIFT algorithm framework are still relatively stable and mature technical systems. In the front-end of visual SLAM, many aspects can be improved or integrated with other algorithms to improve efficiency. For example, optimizing the algorithm for traditional ORB with excessive feature point extraction and low efficiency; for traditional ORB feature points, a more convenient and fast matching method is adopted. This can improve the operational efficiency of the entire visual SLAM.

In addition, other algorithms can be combined for algorithm fusion to improve the shortcomings of the original visual SLAM, such as integrating the YOLOv3 model into ORB-SLAM3 to detect anomalies in dynamic objects in RGB frames, and eliminate them if any. After removing outlier, the accuracy of the system SLAM algorithm is improved, and the robustness to dynamic objects is also improved [5].

As people continue to research and improve existing algorithms, the impact of visual SLAM and its derivatives will gradually expand to various fields: in the future, more and more intelligent robots equipped with visual SLAM will enter people's lives. Service robots equipped with visual SLAM can distinguish various items and corresponding semantics to help people's needs; Exploration robots can also distinguish the surrounding terrain, distance, and depth through visual SLAM, and carry out

autonomous path planning, and even collect ores; The robotic arm on the industrial production line can also accurately locate the processed products through visual SLAM and process them or identify and remove waste products. In the future, the development potential of visual SLAM remains enormous.

References

[1] XiePing Gong, LiZhong Song, and Yang Yin "Review of visual SLAM", Proc. SPIE 12329, Third International Conference on Artificial Intelligence and Electromechanical Automation (AIEA 2022),

[2] Liu, Chang, and Shuwen Dang. "The Research and Application of Improved ORB Feature Matching Algorithm." Advances in Guidance, Navigation and Control: Proceedings of 2022 International Conference on Guidance, Navigation and Control. Singapore: Springer Nature Singapore, 2023.

[3] Y. Xin and W. Li-Na, "Review of Research on Robot Visual Slam," 2022 34th Chinese Contro l and Decision Conference (CCDC), Hefei, China, 2022, pp. 6080-6085, doi: 10.1109/CCDC552 56.2022.10034316.

[4] Wei, YueQiang, et al. "A Fast 3D Laser SLAM Method for Indoor Scenarios." 2022 34th Chinese Control and Decision Conference (CCDC). IEEE, 2022.

[5] Gökcen, Berkay, and Erkan Uslu. "Object Aware RGBD SLAM in Dynamic Environments." 2022 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). IEEE, 2022.

[6] Fu, Xixi, et al. "Recognition and Location Based on Fusion of FAST Algorithm and SIFT Algorithm." 2021 IEEE 21st International Conference on Communication Technology (ICCT). IEEE, 2021.

[7] Ruan, Tianshu, V. Amrusha Aryasomyajula, and Nasser Houshangi. "Performance of monocular and stereo camera in indoor environment for Visual SLAM using ORB method." 2022 IEEE International Conference on Electro Information Technology (eIT). IEEE, 2022.

[8] Yuan, Jing, et al. "ORB-TEDM: An RGB-D SLAM approach fusing ORB triangulation estimates and depth measurements." IEEE Transactions on Instrumentation and Measurement 71 (2022): 1-15.

[9] Liu, Jia, et al. "VSLAM method based on object detection in dynamic environments." Frontiers in Neurorobotics (2022).

[10] Ni, Jianjun, et al. "An improved adaptive ORB-SLAM method for monocular vision robot under dynamic environments." International Journal of Machine Learning and Cybernetics 13.12 (2022): 3821-3836.

[11] Jiang, Rui, et al. "Road-constrained geometric pose estimation for ground vehicles." IEEE Transactions on Automation Science and Engineering 17.2 (2019): 748-760.

[12] Fu, Xixi, et al. "Recognition and Location Based on Fusion of FAST Algorithm and SIFT Algorithm." 2021 IEEE 21st International Conference on Communication Technology (ICCT). IEEE, 2021.

[13] Alcantarilla, Pablo F., and T. Solutions. "Fast explicit diffusion for accelerated features in nonlinear scale spaces." IEEE Trans. Patt. Anal. Mach. Intell 34.7 (2011): 1281-1298.

[14] Srividhya, S., S. Prakash, and K. Elangovan. "3D reconstruction of an indoor environment using SLAM with modified SURF and A-KAZE feature extraction algorithm." Intelligent Computing, Information and Control Systems: ICICCS 2019. Springer International Publishing, 2020.

[15] Zhang, Tong, et al. "A new visual inertial simultaneous localization and mapping (SLAM) algorithm based on point and line features." Drones 6.1 (2022): 23.

[16] Muja, Marius, and David G. Lowe. "Fast approximate nearest neighbors with automatic algorithm configuration." VISAPP (1) 2.331-340 (2009): 2.

[17] He,Zhiyuan, “21 projects play deep learning” Electronic Industry Press (2018): 88-90

[18] Zhang, Feng, et al. "Research of Visual SLAM in Dynamic Environment using Convolutional Neural Network." 2022 International Conference on 3D Immersion, Interaction and Multi-sensory Experiences (ICDIIME). IEEE, 2022.

[19] Yunfeng, GAO, Zhou Lun, and Lv Mingrui. "A review of indoor positioning methods for autonomous mobile robots." Transducer and Microsystem 32.12 (2013): 1-5.

[20] Steenbeek, Anne, and Francesco Nex. "CNN-based dense monocular visual SLAM for real-time UAV exploration in emergency conditions." Drones 6.3 (2022): 79.