Optimization method to improve visual SLAM in dynamic environment

Yufei Liu
Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan

Kazuo Ishii
Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu-ku, Kitakyushu, 808-0196, Japan
Email: liu.yufei124@mail.kyutech.jp, ishii@brain.kyutech.ac.jp

Abstract

Aiming at the problem that the robustness of the classic visual SLAM system is greatly affected by dynamic target feature points in the environment, a method is proposed to use a target detection algorithm to identify and eliminate dynamic target feature points. First, use the target detection algorithm YOLOv5 to identify the collected environmental images, and select the surrounding environment. Objects identified as dynamic targets in the environment, and then the target detection results are integrated into the feature extraction of the visual SLAM front-end, the feature points belonging to the dynamic target part of the extracted image feature points are removed, and the remaining static feature points are used to map Construction and positioning, and finally testing on the TUM data set. The results show that after using the target detection algorithm to eliminate dynamic feature points, the root mean square error of the absolute trajectory error of the visual SLAM system in highly dynamic scenes is reduced by 97.89%, effectively improving the positioning accuracy and robustness of the system.

Keywords: Visual SLAM, Deep learning, Feature detection, Position estimation, YOLOv5

1. Introduction

With the development of autonomous mobile robot platforms, robots are widely used in areas such as search and rescue operations or delivery services required at industrial sites and hotels. In these scenarios, the robot needs to understand the entire area and the precise location of the target object in the map based on the map to complete autonomous navigation. In order to achieve autonomous navigation, mobile robots need to complete two tasks: attitude estimation and map construction. Therefore, Simultaneous Localization and Mapping (SLAM) is one of the research hotspots in the field of mobile robots. It is a robot that only relies on its own sensors to obtain external information of the unknown environment to complete pose estimation and build an environment model. Technology. This technology is currently widely used in agriculture, disaster relief, auxiliary medical and other robotic fields. According to the different sensors used, SLAM is divided into visual SLAM and laser SLAM. Visual SLAM uses cameras mounted on the robot as sensors, including monocular, binocular and depth cameras. Compared with lidar, cameras are cheaper and more convenient to install and debug. Moreover, cameras can collect more semantic information in the environment. Therefore, visual SLAM has gradually become a major research hotspot in the field of autonomous navigation of mobile robots.

2. Problem Description

2.1. *Overview of visual SLAM technology*

In recent years, various visual SLAM (Simultaneous Localization and Mapping) techniques have emerged, such as ORB-SLAM, ORB-SLAM2, VINS Mono, RTABMap, PTAM, LSD-SLAM, DSO, among others, playing a crucial role in the field of autonomous robot navigation [1]. The general algorithmic process of visual SLAM begins with the acquisition of images using a camera to generate a sequence of images. Subsequently, features are extracted from adjacent frames, and the motion of the camera is estimated by minimizing pixel intensity values through feature point matching. Pixel matching methods between images include optical flow and direct methods. Optical flow involves extracting image feature points and estimating camera motion using triangulation or Epipolar Geometry. Direct methods, on the other hand, directly use pixel blocks or extract image corner points to calculate motion estimation based on grayscale values.

Following the motion estimation, noise is filtered from the data to obtain the optimal pose estimation. Maximum a posteriori probability is then used to estimate the global map. Campos et al. [2], building on the research by Artal et al., released ORB-SLAM3 in 2020, which is a feature-based SLAM system and currently represents a notable approach in the field.

2.2. *Impact of Dynamic Points on Visual SLAM*

ORB-SLAM3, as one of the classic algorithms in visual SLAM, consists of two main components in its algorithmic pipeline: the front end and the back end. The front end, also known as visual odometry, processes captured images through feature extraction and matching. It solves the Epipolar Geometry relationship between corresponding feature pixels, thereby estimating the pose parameters of camera translation and rotation. The back end includes nonlinear optimization, map estimation, and loop closure detection. Nonlinear optimization, differing from the previous use of Kalman filtering, employs Bundle Adjustment (BA) to simultaneously optimize the six degrees of freedom of camera pose parameters and the poses of landmarks in space.

The effectiveness of the Bundle Adjustment (BA) method in optimizing camera pose results depends on whether the extracted matching feature points from the captured images exclusively static object features are. However, real-world robot operating environments often include dynamic objects. In scenarios with dynamic targets, both feature-based SLAM algorithms and direct methods SLAM algorithms struggle to differentiate between feature types in regions with moving objects. The matching points of dynamic target feature points can lead to misalignments, resulting in reduced accuracy in the front-end visual odometry's estimation of camera pose. This, in turn, causes a loss of camera pose tracking, leading to significant deviations between the robot's motion trajectory and the constructed environmental map.

3. Elimination of the dynamic points

3.1. *Dynamic SLAM based on geometric methods*

Currently, there are two main methods for mitigating the impact of dynamic target feature points in detection scenes. One approach is the traditional geometric-based method. Sun et al. [3] detect moving objects by comparing differences between adjacent frames, but this method suffers from poor real-time performance. Wang et al. [4] filter matching feature points in adjacent frames using Epipolar Geometry. They also perform clustering on depth images collected by RGB-D cameras to identify independent dynamic objects in the scene. The accuracy of this algorithm is significantly influenced by the pose transformation matrix between adjacent frames. If there are many highly dynamic objects in the scene, it can lead to substantial algorithmic errors. Lin et al. [5] use a fusion of image depth information and visual ranging to detect moving objects. While this approach can determine the position of moving targets in the scene, the algorithm's accuracy is compromised due to the uncertainty of depth information and cumulative errors in calculating the transformation matrix between adjacent frames.

The underlying principle of these methods assumes that the feature points of dynamic objects deviate from the standard constraints obtained in a static scene through processes such as triangulation, fundamental matrix estimation, epipolar lines, and reprojection error. During the pose estimation process, these dynamic feature points are treated as outliers. The correctness of feature matching can be determined by examining whether the extracted feature points violate these constraints, and dynamic points are subsequently excluded. However, the accuracy of this method depends on the proportion of static feature points in the scene. In scenarios with a high density of dynamic objects, it can significantly impact the reliability of pose estimation and the accuracy of map construction.

3.2. *Dynamic SLAM based on deep learning*

With the development of deep learning, many researchers have introduced deep learning algorithms into visual SLAM to mitigate the impact of dynamic targets and improve its accuracy. The DynaSLAM (Dynamic SLAM) algorithm proposed by Berta Bescos et al. [6] optimizes ORB-SLAM2 using Mask R-CNN (Region-based Convolutional Neural Network). By combining geometry with deep learning, DynaSLAM filters out dynamic feature points in the scene. The algorithm has shown excellent results on the TUM dataset, but its use of Mask R-CNN for pixel segmentation leads to low real-time detection efficiency, limiting its practical application in real-world environments. DDL-SLAM (Dynamic Deep Learning SLAM) [7] uses DUNet (Deformable Unity Networking) and semantic masks obtained through multi-view geometry to detect dynamic objects. It then employs an image restoration strategy to reconstruct the background obscured by dynamic objects. Since the calculation of dynamic object masks occurs at the pixel level, this method also falls short of achieving real-time performance. In contrast to Mask R-CNN, another more efficient object detection model is YOLOv5 (You Only Look Once Version 5) [8]. YOLOv5 achieves detection speeds of up to 45-155 frames per second (fps), which is 9-30 times faster than the maximum 5 fps achieved by Mask R-CNN. If the results of YOLOv5's object detection can be integrated into dynamic visual SLAM algorithms, it can partially compensate for the low efficiency of Mask R-CNN.

To address the issue of reduced localization accuracy and robustness of visual SLAM in dynamic environments due to the influence of dynamic targets, this paper introduces a target detection thread based on the ORB-SLAM3 algorithm to detect dynamic targets. Using the YOLOv5 target detection network to identify objects in input images, the semantic information of dynamic target objects in the images is determined. Additionally, while visual SLAM extracts feature points in the front end, a module is added to the tracking thread of visual SLAM to eliminate dynamic feature points based on the results of target detection. Finally, only the remaining static feature points are used to estimate the pose change matrix between adjacent frames, reducing the impact of dynamic targets. This ultimately achieves high-precision estimation and localization of the environmental trajectory. The improved algorithmic flowchart is shown in Fig.1.

4. Experimental verification and result analysis

4.1. *Experimental environment construction and data set selection*

his experiment involved testing the optimized visual SLAM system proposed in this paper using video sequences from the TUM dataset. The experimental results were analyzed to evaluate the localization accuracy of the SLAM system. The experiments were conducted on the Ubuntu 20.04 operating system, with a 12th Gen Intel(R) Core(TM) i9-12900H 2.50GHz CPU, an NVIDIA GeForce RTX 3060 GPU with 12GB of VRAM, and the PyTorch deep learning framework. The algorithm's performance was tested on the fr3_walking_xyz, fr3_walking_half, and fr3_walking_static dataset sequences. The fr3_walking_xyz dataset depicts two individuals walking and conversing in a fixed scene, with both the camera and people in motion, representing a high-dynamic scene. The fr3_walking_half dataset builds upon this by having the camera move along a semi-circular trajectory in the air. The fr3_walking_static dataset, on the other hand, features relatively stationary objects, representing a low-dynamic scene.

4.2. *Visual SLAM front-end feature extraction effect after integrating YOLOv5*

Capturing a frame for comparison with the original ORB-SLAM3 algorithm's front-end feature extraction, the results are illustrated in Fig. 2. The lower image shows the feature extraction before integrating YOLOv5, while the upper image demonstrates the effect after fusion. It is evident that the visual SLAM front end, after integrating YOLOv5, accurately identifies objects such as people and computers. It successfully removes feature points on the dynamic target "person" while retaining feature points on the static object "computer" within the detected bounding box of the dynamic target. This refinement enhances the precision of the results.

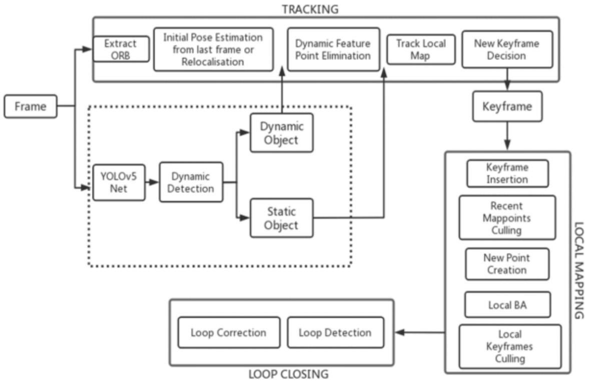
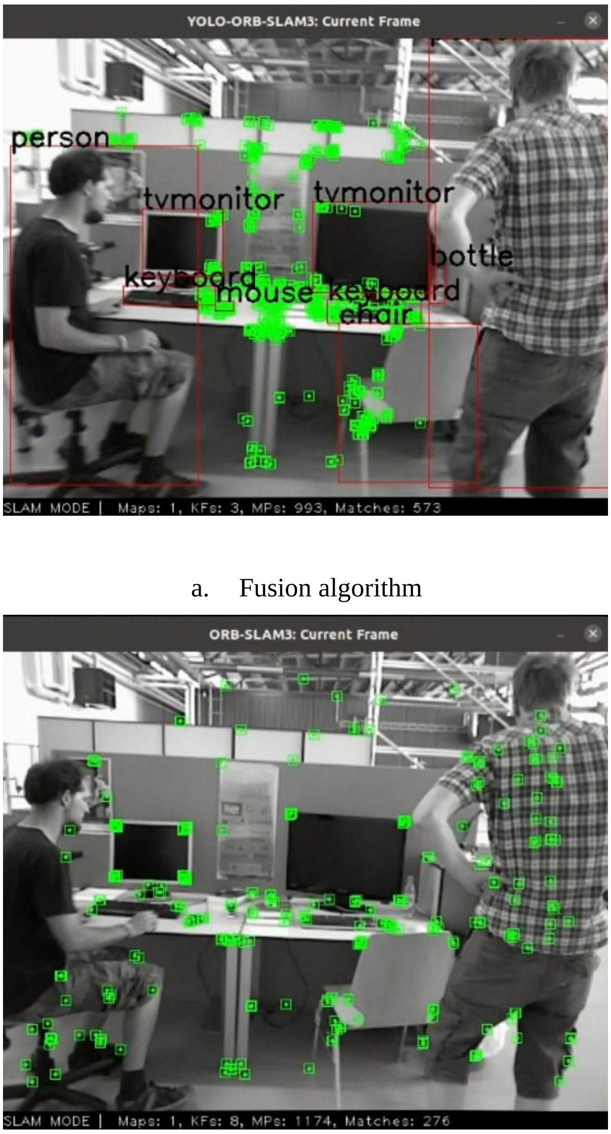


Fig.1 Algorithm framework

4.3. *Experimental data indicator analysis*



SLAM MODE | Maps: 1, KFs: 3, MPs: 993, Matches: 573

a. Fusion algorithm

SLAM MODE | Maps: 1, KFs: 8, MPs: 1174, Matches: 276

b. Original algorithm

Fig.2. Comparison of front-end feature extraction results before and after integrating YOLOv5

Absolute Trajectory Error (ATE) describes the difference between estimated poses and ground truth poses, providing an intuitive expression of the algorithm's global performance and accuracy. Relative Pose Error (RPE) calculates the differences in pose at the same timestamps, typically used for estimating odometry errors. The Root Mean Square Error (RMSE) is then employed to compute the overall value of this error. ATE and RPE are two parameters that reflect the robustness and stability of a visual SLAM system [9]. A lower RMSE value calculated from ATE and RPE indicates a better fitting performance.

Capturing a frame for comparison with the original ORB-SLAM3 algorithm's front-end feature extraction, the results are illustrated in Fig. 2. The lower image shows the feature extraction before integrating YOLOv5, while the upper image demonstrates the effect after fusion. It is evident that the visual SLAM front end, after integrating YOLOv5, accurately identifies objects such as people and computers. It successfully removes feature points on the dynamic target "person" while retaining feature points on the static object "computer" within the detected bounding box of the dynamic target. This refinement enhances the precision of the results. The improvement of the ATE and RPE performance of this algorithm compared with the original ORB-SLAM3 algorithm is shown in Table.1 and Table.2.

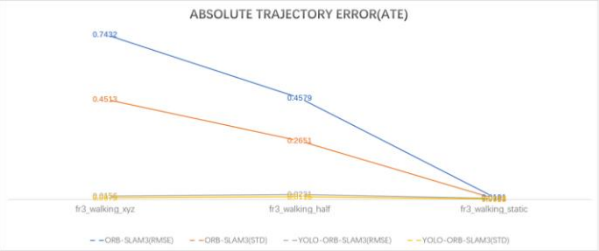


Table.1 ATE performance comparison

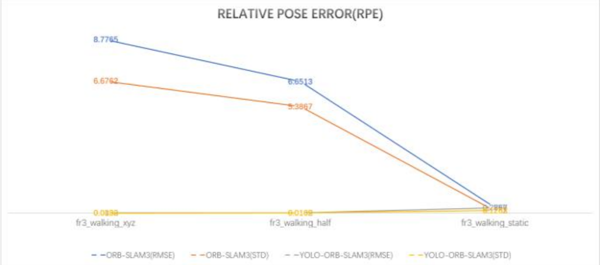


Table.2 RPE performance comparison

This indicates that the proposed algorithm not only enhances the detection capability in dynamic environments but also preserves the accuracy of map construction and localization in static scenes, as achieved by the original ORB-SLAM3 algorithm.

References

1. Zhaopeng G, Liu H, University P, et al. A survey of monocular simultaneous localization and mapping[J]. CAAI Transactions on Intelligent Systems, 2015.
2. Campos C, Elvira R, Rodríguez J J G, et al. Orb-slam3: An accurate open-source library for visual, visual‐inertial, and multimap slam [J]. IEEE Transactions on Robotics, 2021, 37(6): 1874-1890.
3. Sun Y, Liu M, Meng M Q H. Improving RGB-D SLAM in dynamic environments: A motion removal approach[J]. Robotics and Autonomous Systems, 2017, 89: 110-122.
4. Wang R, Wan W, Wang Y, et al. A new RGB-D SLAM method with moving object detection for dynamic indoor scenes[J]. Remote Sensing, 2019, 11(10):1143.
5. Lin S F, Huang S H. Moving object detection from a moving stereo camera via depth information and visual odometry[C]//2018 IEEE International Conference on Applied System Invention (ICASI). IEEE, 2018:437-440.
6. Bescos B, Fácil J M, Civera J, et al. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes[J]. IEEE Robotics and Automation Letters, 2018, 3(4) : 4076-4083.
7. Ai Y, Rui T, Lu M, et al. DDL-SLAM: A robust RGB-D SLAM in dynamic environments combined with deep learning[J]. IEEE Access, 2020, 8: 162335-162342.
8. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016:779-788.
9. Chang Z, Wu H, Sun Y, et al. RGB-D Visual SLAM Based on Yolov4-Tiny in Indoor Dynamic Environment[J]. Micromachines, 2022, 13(2):230.

Authors Introduction

Ms. Yufei Liu

She graduated from Zhejiang Normal University in China with a bachelor's degree in 2017. Now she is studying as a master's student at Kyushu Institute of Technology in Japan.

Dr. Kazuo Ishii

He received his PhD from the University of Tokyo, Japan, in 1996. In 2011, he joined Kyushu Institute of Technology and is currently a professor in the Department of Human Intelligent Systems. His research interests include information communications and marine robotics. He is a member of IEEE.