Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090101
ISSN 2371-8838 Vol. 9 Num. 1

A Review of Common Datasets and Advanced Algorithms of Visual SLAM in Dynamic Scenes

Dazheng Wang[1,*]

[1]*Yunnan Normal University, Kunming, Yunnan, China*
[*]*Corresponding author: 503370461@qq.com*

Keywords: visual SLAM, dataset, dynamic scenes, deep learning networks, robots

Abstract: Simultaneous Localization and Mapping (SLAM) technology can help mobile intelligent robots to understand and perceive scenes in unknown environments, so it is playing an increasingly important role in the fields of intelligent robots, intelligent cars, and so on. And because camera sensors have wide applicability, visual SLAM in dynamic scenes has become a relatively popular research direction in recent years. And the SLAM algorithm needs to be tested and validated, which requires choosing appropriate datasets based on different application scenarios. Therefore, this paper comprehensively introduces the excellent open-source data sets commonly used in the research of visual SLAM. Since deep learning networks are often used to improve the performance of the SLAM system, this paper summarizes the advanced techniques in recent years for solving the common problems of visual SLAM in dynamic scenes based on deep learning networks.

1. Introduction

With the development of science and technology, the application of mobile robots in our daily life is becoming more and more extensive. They can assist people in completing many tedious tasks and they can also replace humans to carry out operations in some dangerous and extreme environments. The Simultaneous Localization and Mapping (SLAM) technology is crucial for mobile robots, which can enable robots to understand and perceive the scene in an unknown environment, and conduct autonomous positioning and map construction. With the increase of demand, the working scenarios faced by mobile robots are also more complex, and even multiple robots need to cooperate with each other to complete the work[1]. This makes the challenges faced by visual SLAM technology more and more diversified, so the SLAM algorithm needs to be continuously optimized and improved. And most working environments have dynamic objects, and these moving objects may be misidentified as part of the environment, leading to significant deviations in pose estimation results and the constructed environment map from the real situation[2], This situation affects the efficiency of robots in unknown environments and may even cause major accidents. In recent years, deep learning networks are playing an increasingly important role in the field of computer vision. Visual SLAM systems usually take pictures as input, so deep learning networks can be combined with SLAM technology to efficiently eliminate dynamic interference. And the innovation and improvement of algorithms needs to be tested and validated, which requires the use of data sets that are close to the actual scene and have rich scenes. Therefore, this article first introduces the commonly used open-

source data sets in the field of visual SLAM, and comprehensively summarizes the excellent SLAM algorithms combined with deep learning in recent years.

2. Common Datasets of visual SLAM

Some excellent open-source datasets related to visual SLAM provide rich data resources for the research in this field and help researchers verify the performance of their new visual SLAM algorithms during development process. By testing on the datasets, the accuracy, robustness, and real-time performance of the algorithms under different environmental conditions can be evaluated, which greatly improves the efficiency of research. Currently, the datasets commonly used in visual SLAM are as follows:

2.1. KITTI dataset

The KITTI dataset[3] was jointly released by the Karlsruhe Institute of Technology (KIT) and the Toyota Technological Institute at Chicago (TTI-C). It was recorded over a distance of 39.2 kilometers and includes vehicles, pedestrians, and many common outdoor objects such as trees, houses, etc,as shown in figure 1. Additionally, detailed annotations were made for some three-dimensional entities appearing in the images. In the image containing the most entities, more than ten cars can be seen, which also provides rich scene data for researchers.

Figure 1: Pictures of KITTI dataset[3]

When studying large outdoor scenes such as urban streets and highways, the KITTI dataset is one of the most commonly used datasets in the field of visual SLAM. It obtains relevant data through a data collection vehicle equipped with two gray cameras and two color cameras, which form a stereo camera sensor capable of capturing gray and color images, respectively. The pixel values of the cameras are all 1.4 million. These excellent devices ensure that the image data provided by the KITTI dataset has good lighting, clear images, and rich textures. It also includes sensors such as LiDAR and inertial measurement units, which can meet the testing requirements of multi-sensor fusion SLAM algorithms. The application scope of the KITTI dataset is very wide. Not only is it widely used in academic research, but with the growing popularity of automatic driving, many intelligent automobile enterprises also use this data set to develop and test auto drive system, to simulate the real road conditions, so as to design a more safe and reliable auto drive system.

2.2. TUM RGB-D dataset

The TUM-D dataset[4] was provided by the Technical University of Munich in 2012. The vast majority of visual SLAM research for indoor scenes selects the TUM-D dataset for simulation testing. This dataset includes RGB images with color information of the covered environment and depth images that can reflect depth information, both in the format of 640x480. Moreover, when scaling the depth map, a pixel value of 5000 in the depth map represents the distance of 1 meter between the observed object and the camera. As shown in figure 2,the TUM RGB-D dataset simulates various types of indoor scenes. Among them, the fr3_walking series simulates an office scene with walking people as interference. The scene includes common static entities in indoor environments such as desks, computers, and chairs. At the same time, two experimenters perform large-scale dynamic behaviors such as walking, sitting down, and standing up, thereby achieving the simulation of a high-dynamic scene. In contrast, the fr3_sitting series simulates a low-dynamic scene where the two experimenters perform small-scale local movements in the same scene. Furthermore, both of these two series include four different sub-datasets of camera motion poses: xyz, rpy, halfsphere, and static, which can well meet the testing and verification requirements of dynamic SLAM. In addition, for purely static scenes, relevant simulations were also conducted on the TUM dataset.

Figure 2: Pictures of TUM RGB-D dataset[4]

The data collection team of the TUM RGB-D dataset has also provided an automatic evaluation system on its official website. Researchers can upload the camera pose and trajectory data estimated by their SLAM systems. The evaluation system can calculate the absolute trajectory error (ATE), relative and relative pose error (RPE) between the provided data and the real data, and display them in both numerical and graphical forms. This provides great convenience for research in the SLAM field.

It is worth mentioning that these two measurement indicators play a crucial role in the evaluation of the SLAM system. Among them, when calculating the ATE, the estimated robot poses are paired with the real poses according to the timestamps, and then the difference between each pair of poses is calculated. It can directly reflect the global positioning accuracy of the algorithm. The RPE is used to evaluate the drift degree of the SLAM system. The calculation methods of ATE and RPE for the ith image are respectively:

$$ATE_i = Q_i^{-1} S P_i \tag{1}$$

$$RPE_i = (Q_i^{-1} Q_{i+\Delta})^{-1} (P_i^{-1} P_{i+\Delta}) \tag{2}$$

Among them, S represents the similar rotation matrix between the estimated pose P_i and the real pose Q_i, Δ represents the time interval.

2.3. Bonn RGB-D Dynamic dataset

The Bonn RGB-D Dynamic dataset[5] was provided by the University of Bonn in Germany in 2019. Originally, it was collected by Emanuele Palazzolo and others to verify their proposed 3D reconstruction algorithm ReFusion, and later the author team published the data set. This dataset mainly focuses on the simulation of high-dynamic indoor scenes. This dataset was mainly collected using the Xtion Pro Live depth camera, and accurate point cloud data was collected from the experimental site using laser scanning equipment, providing a reference for researchers to evaluate the accuracy of the SLAM algorithm.

Figure 3: Pictures of BONN dataset[5]

Compared with the TUM RGB dataset, the dynamic interference in the Bonn dataset is more complex, and the behaviors shown by the experimenters in the images are also more diverse. Human movements are not only limited to simple actions such as walking and standing up, but also includes carrying objects and beating balloons, as shown in figure 3. These actions not only make the types of movements more diverse, but also cause some previously static objects to move under the influence of humans. At the same time, due to the rapid shaking of the camera and the fast movement of people, some images in the Bonn dataset become relatively blurry at certain moments. These interference factors also make the tasks of pose estimation and map construction in SLAM more complex in this dataset. Compared with other datasets for indoor dynamic scenes, it brings new challenges to SLAM algorithms.

2.4. EuRoC dataset

The EuRoC (European Robotics Challenge) dataset[6] is an open-source dataset provided by ETH Zurich, specifically designed for research on drone flight navigation, Visual-inertial odometry (VIO), and SLAM. This dataset was collected by a small unmanned aerial vehicle（UAV） equipped with a binocular camera, IMU for visual inertial information acquisition, motion capture system, and laser scanner. EuRoC is also designed for indoor scenes. Unlike TUM and BONN datasets, the EuRoc dataset not only simulates scenes in offices, but also collects scene information in industrial factory building. Based on the complexity of the scene, this dataset collected five sequences for the factory environment. Among them, the MH_01_easy series has the least dynamic factors, while the MH_05_difficult series has the highest degree of dynamics and complexity. Researchers can choose scenarios based on their own experimental needs. Meanwhile, due to the inclusion of industrial scenarios in the dataset, both academic and industrial fields often use the EuRoC dataset for simulation and testing

2.5. Oxford RobotCar dataset

The Oxford RobotCar dataset[7]is a dataset for outdoor roads provided by the Mobile Robotics Group (MRG) of the University of Oxford in the UK. The acquisition tool of this dataset is a car,

which is equipped with a trinocular stereo camera, three monocular cameras, two 2D lidars, and a 3D lidar. In addition, the GPS system is also used to obtain the ground truth. In order to record more different road and weather conditions such as rain, night, and sunny days, the recording cycle of this dataset lasts for one and a half years. During this period, the author team drived the acquisition vehicle to travel on a road twice a week, and the total driving mileage exceeded 1000 KM. Therefore, the Oxford RobotCar dataset can simultaneously meet the research needs of large-scale scenes and long-term periods. The appearance of this dataset provides more abundant real-world scenes for researching SLAM, scene understanding, and autonomous driving algorithms.

In addition, there are also the Newer Collage dataset[8] for the campus scene, the Complex Urban dataset[9]for the urban scene, and so on. The existence of these datasets provides a solid foundation for the research and development and innovation of SLAM-related algorithms. These real and rich scene data facilitate the testing and improvement of algorithms by researchers, and play an important role in promoting the development of the SLAM field.

3. Dynamic visual SLAM Based on Deep Learning

One of the popular solutions to solve the problem of the decrease in the accuracy of the SLAM algorithm in dynamic scenes is to combine with the deep learning network. In recent years, the application of deep learning networks in the field of computer vision has become more and more widespread, and can accurately identify the entities in the image. The identification of dynamic objects combined with the prior information given by the deep learning network can greatly improve the anti-interference ability of the SLAM algorithm in dynamic scenes, and at the same time can guarantee the real-time requirement of SLAM. At present, the most widely used deep learning networks in the SLAM field mainly include: Mask R-CNN, YOLO series, etc. These networks often maintain high accuracy when performing image processing, but for SLAM, it still needs to be compensated by some other methods. This chapter will summarize the combination methods of these networks and SLAM by means of classification and induction.

3.1. Dynamic visual SLAM combined with Mask R-CNN

DynaSLAM[10] is one of the earliest algorithms that combines the traditional SLAM system with the deep learning network. This algorithm first directly eliminates the priori dynamic objects, and the movement of people may cause some priori static objects to also move, such as cups, tables and chairs, etc. These forced moving objects will also interfere with the pose estimation of the algorithm. At this time, DynaSLAM uses the multi-view geometry method for further detection and elimination.

Usually, only detecting the feature points according to the dynamic mask often leads to the omission of the points at the edge of the dynamic object. These omitted dynamic points will also lead to the decrease of the accuracy of the SLAM algorithm. To solve this problem:

Li[11]et al judge by the distance and depth value between the feature points to further detect the feature points at the edge of the dynamic object. It effectively improves accuracy of the SLAM algorithm, and also has a good advantage compared to some other advanced algorithms.

Zhang[12]et al optimize the mask at the edge through the Laplace edge detection method, and then further verify the movement of the object by through the epipolar Geometry. Compared with DynaSLAM, the ATE of this method is reduced by 36.6% on average.

MDP-SLAM was proposed by Zhang[13]et al in 2024. This algorithm uses the clustering method to divide the edge of the dynamic object, and then accurately expands the semantic mask provided by Mask R-CNN according to the division result, and finally accurately eliminates the dynamic feature points according to the semantic mask. In the high dynamic dataset f3_walking series, compared to ORB-SLAM2, the RMSE of the ATE of this algorithm is reduced from 94.25% to 98.31%.

3.2. Dynamic visual SLAM combined with YOLO series

The YOLO series has relatively high detection speed and accuracy, but most of the YOLO series only have the function of target detection, that is, only the detection box of dynamic objects can be obtained, and the dynamic objects cannot be accurately segmented. If all the feature points in the dynamic frame are eliminated, a large number of static points will also be eliminated, which may affect the accuracy of positioning and even cause tracking failure:

YKP-SLAM [14]uses K-means clustering to calculate the depth mean value, and judges the dynamic area according to the depth value in the dynamic detection boxes provided by YOLOv5,so as to achieve accurate screening of dynamic feature points. Compared with the traditional ORB-SLAM2 algorithm, YKP-SLAM has increased the accuracy of the algorithm in high dynamic scenes by 96.45%.

YG-SLAM[15]is to detect the dynamic objects in the image by combining the LK optical flow algorithm based on the detection result of YOLOv5. The prior semantic information of YOLOv5 can greatly improve the running speed of the optical flow algorithm and ensure the real-time nature of the algorithm.

In 2023, Liu[16]et al proposed the YES-SLAM algorithm. This algorithm uses the Laplace algorithm in the YOLOv7 detection boxes to obtain the edge contour information of the dynamic object. Then use the four-neighborhood algorithm to cover and fill the dynamic region according to the RGB value of the pixel point, and finally obtain the accurate dynamic region. Compared with ORB-SLAM2, the ATE of YES-SLAM in high dynamic scenes has decreased by 96.8%.

In addition, since YOLOv8 itself integrates the instance segmentation network, when YOD-SLAM [17]is processed, the segmentation mask provided by YOLOv8 is expanded through the depth information to ensure that the mask can accurately cover the dynamic object. Moreover, when the people walk too far and the segmentation network fails, the algorithm redraws the missing mask by combining the mask of adjacent frames and the central depth value. At the same time, when the priori static object is too close to the dynamic object, it is considered that this static object is not reliable. These compensation conditions make YOD-SLAM show excellent accuracy in dynamic scenes.

3.3. Dynamic visual SLAM combined with other networks

In addition to the Mask R-CNN and YOLO series mentioned above, there are also many excellent deep learning networks used to solve the problems of SLAM in dynamic scenes. For example, Chang[18]et al adopted the YOLACT instance segmentation network and combined it with the dense optical flow network to achieve the elimination of dynamic points;After obtaining the segmentation mask of YOLACT++, Li[19]et al used the Mahalanobis distance and depth value to further judge the suspected dynamic feature points, and at the same time used the method of epipolar constraint and clustering to make up for the deficiency of the Mahalanobis distance. Hu [20] et al used the DeepLabv3(+) network to obtain the mask, and then used the method of multi-view geometry to verify the motion state of the object; Wen[21]et al calculated the moving speed of the feature points between two images and fused the calculation result with the segmentation result of the SegNet network, effectively eliminating the influence of dynamic feature points.

4. Conclusions

This paper first introduces some commonly used open-source datasets in the field of visual SLAM, briefly introduces the sensor parameters used in the collection of each dataset, and discusses in detail the scene characteristics of each dataset and the environmental information contained therein. At the same time, simple picture displays of the scenes in some datasets are carried out. Subsequently, this

paper introduces advanced SLAM algorithms based on Mask R-CNN, the YOLO series and other deep learning networks, and the combination of these advanced SLAM algorithms and deep learning network is introduced in detail. When dealing with the impact of dynamic disturbances on SLAM algorithms, if only deep learning networks are used, these dynamic points cannot be accurately eliminated, so it often needs to be combined with other methods, such as geometric methods, optical flow methods and other mathematical methods. With the rapid development of different sensors, SLAM systems will adopt the fusion of multiple sensors in the future, such as the fusion of inertial measurement units and visual sensors, so as to further enhance the perception ability of mobile robots to the environment. Moreover, most open-source datasets often use many kinds of sensors to collect data, so they can meet the experimental requirements of multi-sensor fusion.

References

[1] Li Z, Xu B, Wu D, et al. A YOLO-GGCNN based grasping framework for mobile robots in unknown environments[J]. Expert Systems with Applications, 2023, 225: 119993.

[2] Wang Y, Tian Y, Chen J, et al. A survey of visual SLAM in dynamic environment: the evolution from geometric to semantic approaches[J]. IEEE Transactions on Instrumentation and Measurement, 2024, 73, 1-21.

[3] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The kitti vision benchmark suite; proceedings of the 2012 IEEE conference on computer vision and pattern recognition, IEEE, F, 2012 [C].

[4] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems; proceedings of the 2012 IEEE/RSJ international conference on intelligent robots and systems, IEEE, F, 2012 [C].

[5] PALAZZOLO E, BEHLEY J, LOTTES P, et al. ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals; proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, F, 2019 [C].

[6] Burri M, Nikolic J, Gohl P, et al. The EuRoC micro aerial vehicle datasets[J]. The International Journal of Robotics Research, 2016, 35(10): 1157-1163.

[7] Maddern W, Pascoe G, Linegar C, et al. 1 year, 1000 km: The oxford robotcar dataset[J]. The International Journal of Robotics Research, 2017, 36(1): 3-15.

[8] RAMEZANI M, WANG Y, CAMURRI M, et al. The newer college dataset: Handheld lidar, inertial and vision with IEEE, ground truth; proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), F, 2020 [C].

[9] Jeong J, Cho Y, Shin Y S, et al. Complex urban dataset with multi-level sensors from highly diverse urban environments[J]. The International Journal of Robotics Research, 2019, 38(6): 642-657.

[10] 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.

[11] LI J, LUO J. Approach to 3D SLAM for Mobile Robot Based on RGB-D Image with Semantic Feature in Dynamic Environment [J]. Journal of Intelligent & Robotic Systems, 2023, 109(1): 15.

[12] Zhang X, Wang X, Zhang R. Dynamic Semantics SLAM Based on Improved Mask R-CNN[J]. IEEE Access, 2022, 10: 126525-126535.

[13] Zhang X, Shi Z. MDP-SLAM: A Visual SLAM towards a Dynamic Indoor Scene Based on Adaptive Mask Dilation and Dynamic Probability[J]. Electronics, 2024, 13(8): 1497.

[14] Liu L, Guo J, Zhang R. YKP-SLAM: A Visual SLAM Based on Static Probability Update Strategy for Dynamic Environments[J]. Electronics, 2022, 11(18): 2872.

[15] Yu Y, Zhu K, Yu W. YG-SLAM: GPU-Accelerated RGBD-SLAM Using YOLOv5 in a Dynamic Environment[J]. Electronics, 2023, 12(20): 4377.

[16] Liu H, Luo J. YES-SLAM: YOLOv7-enhanced-semantic visual SLAM for mobile robots in dynamic scenes[J]. Measurement Science and Technology, 2023, 35(3): 035117.

[17] Li Y, Wang Y, Lu L, et al. YOD-SLAM: An Indoor Dynamic VSLAM Algorithm Based on the YOLOv8 Model and Depth Information[J]. Electronics, 2024, 13(18): 3633.

[18] Chang J, Dong N, Li D. A real-time dynamic object segmentation framework for SLAM system in dynamic scenes[J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70: 1-9.

[19] Li J, Luo J. YS-SLAM: YOLACT++ based semantic visual SLAM for autonomous adaptation to dynamic environments of mobile robots[J]. Complex & Intelligent Systems, 2024: 1-22.

[20] Hu Z, Zhao J, Luo Y, et al. Semantic SLAM based on improved DeepLabv3⁺ in dynamic scenarios[J]. IEEE Access, 2022, 10: 21160-21168.

[21] Wen S, Li X, Liu X, Li J, Tao S, Long Y, Qiu T. Dynamic slam: A visual slam in outdoor dynamic scenes. IEEE Transactions on Instrumentation and Measurement. 2023 Sep 20. DOI: 10.1109/TIM.2023.3317378