VALIDATION OF ORB-SLAM2'S APPLICABILITY ACROSS SCENARIOS

Fang Yixuan[1], Wang Guoyan[1], Fan Hongqi[1]

410073

National Key Laboratory of Automatic Target Recognition, College of Electronic Science and Technology, National University of Defense Technology, Changsha, China

Fang Yixuan – Master in Information and Communication Engineering
Wang Guoyan – PHD in Technical Sciences, Lecturer
Fan Hongqi – PHD in Technical Sciences, Research Professor

Annotation. This paper investigates the applicability of the ORB-SLAM2 across diverse scenarios based on 51 sequences from three public datasets: TUM RGB-D, EuRoC, and KITTI. Experimental results reveal that while ORB-SLAM2 demonstrates strong robustness in dynamic environments, its performance significantly deteriorates when encountering textureless regions, intense camera rotation, or extensive planar areas. In most cases within the same scenario, the stereo configuration achieves higher localization accuracy than the monocular mode. This study provides practical references for scene-specific adaptability considerations in SLAM technology applications.

Keywords. ORB-SLAM2, dynamic environments, monocular, stereo, localization accuracy

I. Introduction

Simultaneous Localization and Mapping (SLAM), proposed by Smith R.C. and Cheeseman P. in 1986 [1], is a technology that enables an agent equipped with specific sensors to construct environmental models and estimate its own motion in real time without prior environmental knowledge [2]. Based on sensor types, SLAM can be categorized into Visual SLAM (VSLAM) [3] using cameras and LiDAR-based SLAM [4] using light detection and ranging sensors. This study focuses on monocular and stereo camera-based VSLAM systems.

The evolution of VSLAM has witnessed significant algorithmic advancements. Early monocular SLAM systems primarily relied on filter-based methods [5],[6],[7],[8], which faced challenges such as high computational complexity and error accumulation. A milestone was the Parallel Tracking and Mapping (PTAM) [9] algorithm, which pioneered a keyframe-based architecture by decoupling feature tracking and map construction into parallel threads. Subsequent improvements to PTAM included the integration of edge features and enhanced relocalization techniques [10]. Among feature-based SLAM systems [5], ORB-SLAM2 [11] emerged as a representative solution due to its rapid ORB [12] feature extraction and rotation invariance, achieving high operational efficiency and stability.

However, existing research lacks systematic validation of ORB-SLAM2's applicability in complex scenarios, such as dynamic environments, weakly textured regions, and large-scale planar surfaces. This study aims to address these gaps through multi-scenario experiments, specifically:

1) Robustness analysis in dynamic environments;

2) Comparative evaluation of localization accuracy between monocular and stereo modes across diverse environments;

3) Identification of limitations in textureless areas, rapid rotational motion, and expansive planar scenes.

Ⅱ. System Overview

The ORB-SLAM2 system comprises three parallel threads: Tracking, Local Mapping, and Loop Closing, as illustrated in Figure 1.

The Tracking thread is responsible for searching feature correspondences between each frame and the local map to compute the corresponding camera pose. Based on this computation, it determines whether to

appropriately insert a new keyframe into the keyframe buffer queue of the Local Mapping thread. In monocular mode, the system initializes the map through parallel computation of both the homography matrix suitable for planar scenes and the fundamental matrix applicable to non-planar scenes [13], selecting the optimal solution via RANSAC [14].Within the tracking thread, preliminary feature matching is first performed between the received current frame and its preceding frame. Subsequently, a motion-only Bundle Adjustment (BA) [15] algorithm is employed to optimize and refine the pose estimation of the current frame.

The Local Mapping thread manages the construction process of the local map and executes all BA optimizations related to the local map. This thread processes newly inserted keyframes from the Tracking Thread. Its core task is to perform local BA optimization to achieve optimal reconstruction of the surrounding environment under the current camera pose constraints.

The Loop Closing Thread detects large-scale loops and corrects accumulated drift through pose graph optimization. For each newly inserted keyframe from the Local Mapping Thread, this thread performs loop detection to verify loop formation. It constructs a place recognition database based on the DBoW2 [16] vocabulary model, while enhancing loop detection accuracy through covisibility graph-optimized candidate keyframe selection strategy. When a loop closure is detected, the system computes the relative geometric transformation (similarity transformation [17]) between the current keyframe and the identified loop-closing keyframe.

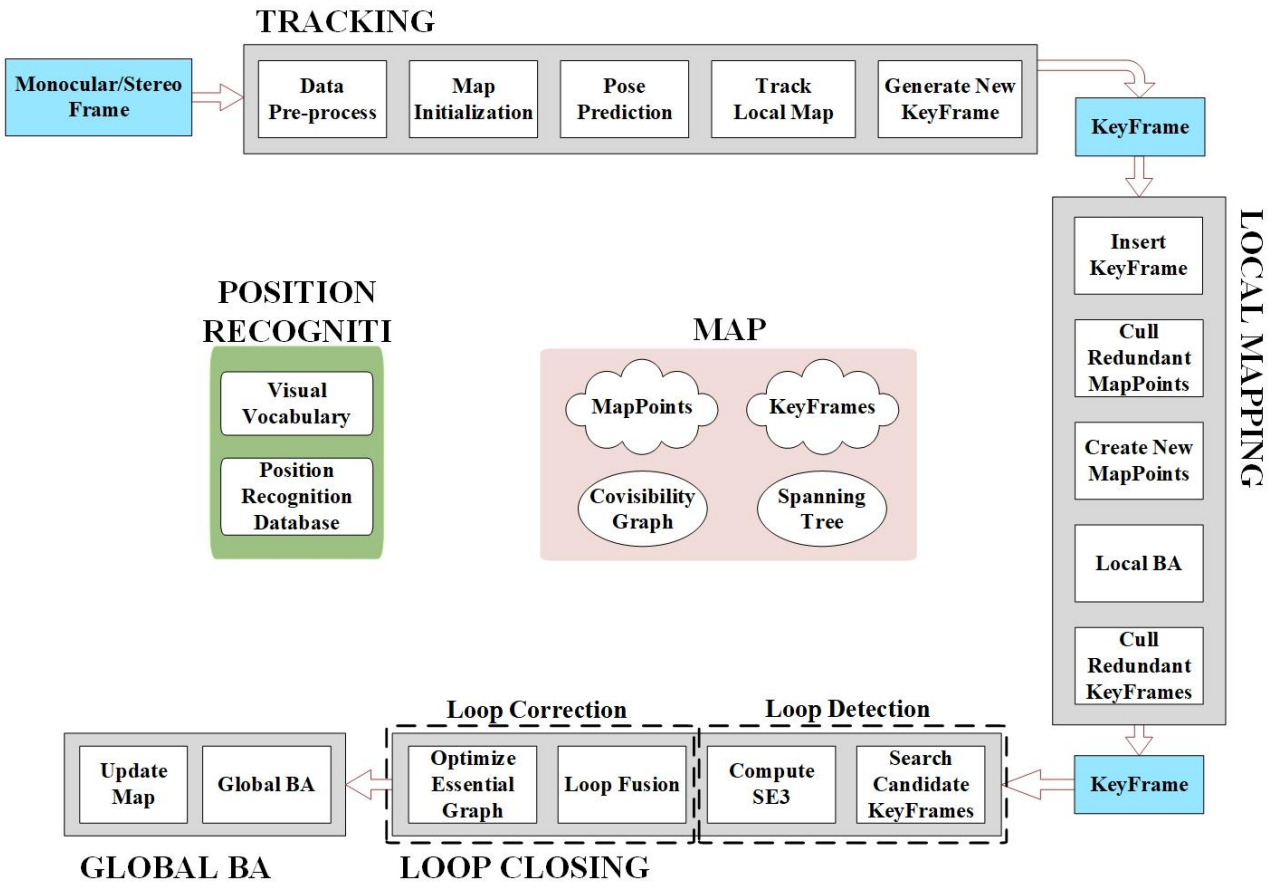

Figure 1 – ORB-SLAM2 system overview

III. Experimental Design and Analysis

The experiments were conducted on an Intel(R) Core(TM) i9-14900HX processor (2.20 GHz, x64-based) within an Ubuntu 18.04 virtual environment hosted by VMware Workstation Pro 17 on Windows 11.

Three public datasets – TUM RGB-D [18], EuRoC [19], and KITTI **Ошибка! Источник ссылки не**

найден. – were utilized to evaluate system performance. Key evaluation metrics include:

Absolute Trajectory Error (ATE), as shown in Equation (1).

$$ATE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|trans(F_i)\|^2} \quad (1)$$

where N is the total number of trajectory points, F_i is the absolute error at i-th trajectory point, expressed as $F_i = \mathrm{Q}_i^{-1} S \mathrm{P}_i$ (with Q_i being the ground-truth pose at the i-th point, P_i being the estimated pose at the i-th point, and S being the rigid transformation aligning estimated and ground-truth trajectories), $trans(F_i)$ denotes the translational component of F_i.

Relative Pose Error (RPE), as shown in Equation (2).

$$RPE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\|trans(E_i)\|^2} \quad (2)$$

where $m = N - \Delta$ is the number of available relative pose error samples, E_i is the relative pose error at the i-th point, expressed as $E_i = (\mathrm{Q}_i^{-1} Q_{i+\Delta})^{-1} (\mathrm{P}_i^{-1} P_{i+\Delta})$ (with Q_i and $Q_{i+\Delta}$ being the ground-truth poses at points i and $i+\Delta$, and P_i and $P_{i+\Delta}$ being the estimated poses at points i and $i+\Delta$), $trans(E_i)$ denotes the translational component of E_i.

To ensure the reliability of the results, each sequence in all datasets was run five times, and the median of the five results was taken as the error value.

A. TUM RGB-D Dataset

This paper uses 29 sequences from the TUM RGB-D dataset, including Handheld SLAM, Dynamic Objects, Structure vs. Texture, and Robot SLAM sequences, to conduct a detailed analysis of the monocular mode of ORB-SLAM2. The evaluation metric is the ATE.

The validation results for the Handheld SLAM sequences are presented in Table 1.

Table 1 – Handheld SLAM Sequences of the TUM RGB-D Dataset

Sequence	Description			RMSE(cm)
	Avg. translational velocity (m/s)	Avg. angular velocity (deg/s)	with loop	
fr1_floor	0.258	15.071	No	1.737
fr1_desk	0.413	23.327	No	1.360
fr1_room	0.334	29.882	Yes	6.085
fr2_360_kidnap	0.304	13.425	No	4.154
fr2_desk	0.193	6.338	Yes	0.860
fr3_long_office	0.249	10.188	Yes	1.098
fr1_360	0.210	41.600	No	5.791
fr2_360_hemisphere	0.163	20.569	No	9.335

As shown in Table 1, the monocular mode of ORB-SLAM2 achieved an absolute trajectory error of approximately 1 cm in most Handheld SLAM sequences, except in scenarios with strong camera rotation (fr1_room, fr2_360_kidnap, fr1_360, and fr2_360_hemisphere).

The results for the Dynamic Objects sequence (an office scenario) are shown in Table 2.

Table 2 – Dynamic Objects Sequences of the TUM RGB-D Dataset

Sequence	Description	RMSE(cm)
fr2_desk_person	Interaction	0.743
fr3_sit_static	Two people sitting at a desk interacting, Asus Xtion fixed	\
fr3_sit_xyz	Two people sitting at a desk interacting, Asus Xtion moving along xyz	0.932
fr3_sit_halfsph	Two people sitting at a desk interacting, Asus Xtion moving along a half-sphere trajectory	1.693
fr3_sit_rpy	Two people sitting at a desk interacting, Asus Xtion moving along rpy with strong rotation	\
fr3_walk_static	Two people walking, Asus Xtion fixed	\
fr3_walk_xyz	Two people walking, Asus Xtion moving along xyz	\
fr3_walk_halfsph	Two people walking, Asus Xtion moving along a half-sphere trajectory	1.675
fr3_walk_rpy	Two people walking, Asus Xtion moving along rpy with strong rotation	7.300

From the results in Table 2, the monocular mode of ORB-SLAM2 achieved an absolute trajectory error of less than 2 cm in most Dynamic Object sequences, except in scenarios with strong camera rotation (fr3_sit_rpy and fr3_walk_rpy) and camera stationary (fr3_sit_static and fr3_walk_static). This indicates that the system is robust to dynamic objects in monocular mode, except when the camera undergoes strong rotation and stationary.

The validation results for the Structure vs. Texture sequences are presented in Table 3.

Table 3 – Structure vs. Texture Sequences of the TUM RGB-D Dataset

Sequence	Description	RMSE(cm)
fr3_nstr_tex_far	planar, texture	9.249
fr3_nstr_tex_near	planar, texture, with loop	1.363
fr3_str_tex_far	non-planar, texture	0.922
fr3_str_tex_near	non-planar, texture	1.358
fr3_nstr_ntex_far	planar, textureless	\
fr3_nstr_ntex_near	planar, textureless, with loop	\
fr3_str_ntex_far	non-planar, textureless	\
fr3_str_ntex_near	non-planar, textureless, with loop	\

From the results in Table 3, the monocular mode of ORB-SLAM2 failed to complete initialization in textureless scenes (fr3_nstr_ntex_far, fr3_nstr_ntex_near, fr3_str_ntex_far and fr3_str_ntex_near).

However, in the Robot SLAM sequences, although the camera was not in a state of strong rotation, most sequences failed to initialize. The validation results are shown in Table 4.

Table 4 – Robot SLAM Sequences of the TUM RGB-D Dataset

Sequence	Description	RMSE(cm)
fr2_pioneer_360	warehouse, large-scale planar	\
fr2_pioneer_slam	warehouse, large-scale planar, with loop	5.038
fr2_pioneer_slam2	warehouse, large-scale planar	\
fr2_pioneer_slam3	warehouse, large-scale planar	\

By analyzing the commonalities of the Robot SLAM sequences, it is evident that such sequences often involve large-scale planar environments. Additionally, in sequences with loops (fr1_room, fr2_desk, fr3_long_office, and fr3_nstr_tex_near), the system achieved a trajectory error of approximately 1 cm, indicating strong loop closure handling capabilities.

ORB-SLAM2 demonstrated a trajectory error of less than 10 cm across all 29 sequences of the TUM RGB-D dataset. The validation results from the TUM RGB-D dataset confirm that ORB-SLAM2 is robust in dynamic scenes and effective in handling loop closures. However, it is not suitable for textureless scenes, scenarios with significant camera rotation, or scenes containing extensive planar structures.

B. EuRoC Dataset

This paper analyzed 11 sequences of the EuRoC dataset, categorized into easy, medium, and difficult

levels. The ATE of the trajectory for each sequence is presented in Table 5.

Table 5 – Results of the EuRoC Dataset

Sequence	RMSE(cm)	
	Monocular	Stereo
MH_01_easy	4.510	3.759
MH_02_easy	3.423	3.754
MH_03_medium	3.958	3.733
MH_04_difficult	7.228	12.496
MH_05_difficult	6.986	5.730
V1_01_easy	9.551	8.638
V1_02_medium	5.094	6.040
V1_03_difficult	9.787	9.753
V2_01_easy	6.080	7.097
V2_02_medium	6.048	5.920
V2_03_difficult	22.309	19.299

As shown in Table 5, in the indoor EuRoC dataset, ORB-SLAM2 showed comparable performance in both monocular and stereo modes, with most sequences achieving a trajectory error of less than 10 cm, except for the V2_03_difficult sequence. This level of positioning accuracy is sufficient for small drones used in environmental exploration.

C. KITTI Dataset

This paper analyzed 11 sequences (00 to 10) of the KITTI dataset. In addition to calculating the ATE (t_{abs}) and RPE (t_{rel}), the relative rotational error (r_{rel}) was also computed. Table 6 presents the specific results for the 11 sequences of the KITTI dataset.

Table 6 – Results of the KITTI Dataset

Sequence	$m \times m$	Monocular			Stereo		
		$t_{rel}(\%)$	$r_{rel}(deg/100m)$	$t_{abs}(m)$	$t_{rel}(\%)$	$r_{rel}(deg/100m)$	$t_{abs}(m)$
00	564×496	5.230	0.946	**7.542**	1.094	0.689	**1.288**
01	1157×1827	153.859	0.929	**533.562**	1.698	0.359	**9.623**
02	599×946	14.228	0.647	**33.502**	1.116	0.495	**5.841**
03	471×199	1.277	0.302	**0.967**	0.950	0.365	**0.755**
04	0.5×394	0.588	0.308	**0.989**	0.445	0.310	**0.187**
05	479×426	5.165	0.506	**5.367**	0.632	0.352	**0.720**
06	23×457	9.799	0.418	**13.413**	0.699	0.300	**0.784**
07	191×209	3.809	0.603	**2.127**	0.567	0.392	**0.526**
08	808×391	32.224	0.674	**52.889**	1.301	0.659	**3.721**
09	465×568	4.949	0.719	**4.858**	0.916	0.484	**3.291**
10	671×177	7.446	0.509	**8.444**	0.881	0.474	**0.100**

As shown in Table 6, in monocular mode, the trajectory error of ORB-SLAM2 is typically around 1% of the map size (sequences 00, 05, 07, 09, and 10), sometimes lower—such as 0.21% for sequence 03 and 0.25% for sequence 04—or higher, like 3.54% for sequence 02, 2.94% for sequence 06, and 6.55% for sequence 08. In stereo mode, the trajectory error is consistently less than 1% of the map size. This indicates that, in most cases, stereo mode provides higher localization accuracy than monocular mode in the same scenario.

IV. Conclusion

This paper provides an extensive experimental evaluation of the ORB-SLAM2 algorithm to determine its applicability across different environments. The results show that the ORB-SLAM2 system operates effectively in both indoor and outdoor settings, demonstrates robustness in dynamic scenes, and handles loop closures well. However, it performs poorly in textureless environments, scenarios with strong camera rotation, and scenes featuring large-scale planar surfaces. In most cases, the stereo mode achieves higher localization accuracy than the monocular mode in the same scenario. This study offers practical insights into the environmental adaptability

of SLAM technology.

References

[1] On the representation and estimation of spatial uncertainty / R. C. Smith [et al.] // The International Journal of Robotics Research, 1986. – P.56-68.

[2] A State of the Art in Simultaneous Localization and Mapping (SLAM) for Unmanned Ariel Vehicle (UAV): A Review / A. Rauf [et al.] // Electrical, Control and Communication Engineering, 2022. – P.50-56.

[3] Visual SLAM: What are the Current Trends and What to Expect? / A. Tourani [et al.] // Sensors, 2022. – P.9297.

[4] A Comparison of Modern General-Purpose Visual SLAM Approaches / A. Merzlyakov [et al.] // Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2021. – P.9190-9197.

[5] MonoSLAM: real-time single camera SLAM / A. J. Davison [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007. – P.1052-1067.

[6] Inverse Depth Parametrization for Monocular SLAM / J. Civera [et al.] // IEEE Transactions on Robotics, 2008. – P.932-945.

[7] Structure from motion causally integrated over time / A. Chiuso [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002. – P.523-535.

[8] Scalable Monocular SLAM / E. Eade [et al.] // Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006. – P.469-476.

[9] Parallel Tracking and Mapping for Small AR Workspaces / G. Klein [et al.] // Proceedings of the 2007 IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007. – P.225-234.

[10] Improving the Agility of Keyframe-Based SLAM / G. Klein [et al.] // Proceedings of the 2008 European Conference on Computer Vision, 2008. – P.802-815.

[11] ORB-SLAM2: an Open-Source SLAM System for Monocular,Stereo and RGB-D Cameras / R. Mur-Artal[et al.] // IEEE Transactions on Robotics, 2017. – P. 1255-1262.

[12] ORB: an efficient alternative to SIFT or SURF / E. Rublee [et al.] // IEEE International Conference on Computer Vision, 2011. – P.2564–2571.

[13] ORB-SLAM: A Versatile and Accurate Monocular SLAM System / R. Mur-Artal [et al.] // IEEE Transactions on Robotics, 2015. – P. 1147-1163.

[14] Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography / M. A. Fischler [et al.] // Communications of the Association for Computing Machinery, 1987. – P.726-740.

[15] Bundle adjustment a modern synthesis / B. Triggs [et al.] // Vision algorithms: theory and practice, 2000. – P.298–372.

[16] Bags of Binary Words for Fast Place Recognition inImage Sequences / D. Galvez-Lpez [et al.] // IEEE Transactions on Robotics, 2012, . – P.1188-1197.

[17] Scale drift-aware large scale monocular SLAM / H. Strasdat [et al.] // Robotics: Science and Systems, 2010. – P.73-80.

[18] A benchmark for the evaluation of RGB-D SLAM systems / J. Sturm [et al.] // Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012. – P.573-580.

[19] The EuRoC micro aerial vehicle datasets / M. Burri [et al.] // International Journal of Robotics Research, 2016. – P.1157-1163.

Vision meets robotics: The KITTI dataset / A. Geiger [et al.] // International Journal of Robotics Researc, 2013. – P.1231-1237.