ICRA2026 Late Breaking Results Poster presented at
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

E2O-SLAM: A Hierarchical Visual SLAM Framework Using Edge-based and Object-level Representations

Eunseon Choi*[1] and Soohee Han**[1]

***Abstract*— In this paper, we present a hierarchical simultaneous localization and mapping (SLAM) system that leverages point-level features, mid-level geometric organized edge representations [1], and high-level object semantics within a unified framework. While object-level SLAM provides semantic information and improves long-term data association, it often suffers from coarse geometric constraints and unreliable detections. In contrast, organized edge representations capture rich structural and textural information, offering stable geometric cues in low-texture or challenging environments.**

By hierarchically integrating these complementary representations, the proposed system achieves robust camera tracking, reliable data association, and consistent mapping.

I. Introduction

Simultaneous localization and mapping (SLAM) using camera sensors has been an active field of research for more than two decades, driven by the increasing demand for autonomous systems capable of perceiving and interacting with complex environments. By jointly estimating a sensor's trajectory while incrementally constructing a representation of the surrounding scene, visual SLAM (VSLAM) enables robots and intelligent agents to operate without prior knowledge of the environment. Accurate localization and mapping are fundamental requirements for a wide range of applications, including autonomous navigation, mobile robotics, augmented and virtual reality, and robot manipulation, where reliable spatial understanding directly impacts safety and task performance.

Over the years, significant progress has been made in improving the robustness, accuracy, and efficiency of visual odometry (VO) and VSLAM systems under diverse environmental conditions. Nevertheless, challenges such as illumination changes, textureless regions, motion blur, and long-term drift continue to limit their deployment in real-world scenarios. To address these challenges, a variety of methodological paradigms have been explored, leading to distinct classes of VO/VSLAM systems that differ in their feature representations, data association strategies, and optimization formulations.

Based on these design choices, VO and VSLAM approaches can be broadly categorized into three main groups: feature-based methods, direct or semi-direct methods, and learning-based methods. Feature-based approaches rely on sparse geometric primitives, such as points [2], [3], lines [4], or edges [1], [5], [6], to establish correspondences across frames and estimate camera motion through geometric constraints. Classical feature-based VO/VSLAM systems achieve accurate pose estimation and consistent mapping by extracting and tracking stable visual features under favorable imaging conditions. However, in low-light or low-texture environments, the availability and reliability of such visual features deteriorate significantly, making feature detection and tracking unreliable and often leading to tracking failure or complete system breakdown.

In contrast, direct and semi-direct methods operate directly on pixel intensities, avoiding explicit feature extraction and enabling dense or semi-dense map representations. These methods rely on the photometric consistency assumption, which presumes that the appearance of a 3D point remains constant across multiple frames. Consequently, accurate photometric calibration—accounting for camera response functions, lens vignetting, and image signal processing—is critical for stable performance. Despite careful calibration, real-world environments frequently violate this assumption due to illumination variations caused by changing lighting conditions, shadows, or exposure adjustments. Such effects cannot be fully compensated through calibration alone and often lead to substantial performance degradation in direct visual odometry. As a result, the VO/VSLAM community has increasingly revisited feature-based approaches as a more robust alternative under challenging illumination conditions.

- We present a hierarchical visual SLAM framework that unifies point-level keypoints, mid-level organized edges, and high-level object semantics, leveraging their complementary strengths.

II. E2O-SLAM

A. System Overview

E2O-SLAM integrates three modules: Hierarchical Visual Odometry and Hierarchical Mapping. It processes synchronized RGB-D frames, taking instance results either from precomputed inputs or generated online via YOLOv26 [7]. The visual ddometry module extracts point-level keypoints, mid-level organized edge features [1], and object-level ellipses derived from instance segmentation. Keypoints are tracked for motion estimation, which guides organized edge tracking, while object tracking employs a Wasserstein-distance-based association metric.

*The first two authors contributed equally to this work.

**Corresponding author.

[1]Authors are with the Department of Convergence IT Engineering, Pohang University of Science and Technology, Choengam-ro 77, Namgu, Pohang-si, Gyeongsangbuk-do, Republic of Korea. {eunseon103, sooheehan}@postech.ac.kr

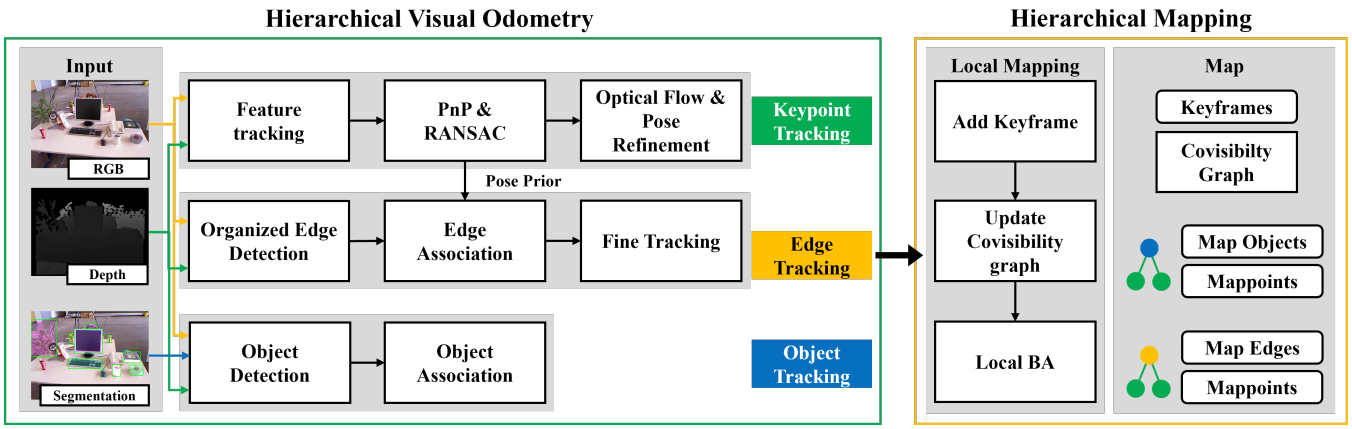

Fig. 1. Overview of the proposed E2O framework, which hierarchically integrates point-level keypoints, mid-level geometric edge representations, and high-level object semantics for robust visual SLAM.

B. Organized Edge Feature Tracking

Organized edges are extracted using the Canny edge detector [8] and clustered through a breadth-first search. We estimate relative camera motion by aligning 3D edge structures with 2D observations, formulating a 3D–2D edge registration problem where the pose is optimized by minimizing geometric residuals between reprojected 3D edge points and observed edges. coarse-to-fine tracking strategy is employed for efficient pose estimation, utilizing optical flow for initial alignment. The camera pose is then refined via nonlinear least squares optimization based solely on photometric residuals.

C. Object Tracking

Objects are modeled using constrained 3D dual quadrics [9]. Image points are projected onto a normalized plane for consistency checks using a dual conic representation. The residuals are computed using the 2nd order Wasserstein distance. For object-level data association, dual-quadric-based systems commonly rely on IoU-based matching and geometric consistency checks based on Wasserstein distance. Upon successful tracking, dual quadric parameters are optimized by minimizing accumulated Wasserstein residuals across multiple frames, refining object-level landmarks through multi-view observations.

III. Experimental Results

Despite the absence of global optimization, our method achieves competitive performance in terms of RTE, demonstrating the effectiveness of the proposed representation in local trajectory estimation. Although the ATE results remain less competitive, the RTE improvements indicate that the proposed approach provides reliable relative motion estimation.

TABLE I
ATE and RTE RMSE on TUM RGB-D Sequences

Method	fr1_rpy	fr2_desk	fr3_cabinet
ATE (m)			
ORB-SLAM3-TR	0.0377	0.0388	Fail
ORB-SLAM3-LM [3]	**0.0199**	**0.0216**	Fail
E2O-SLAM	0.0330	0.1066	**0.0861**
RTE (m)			
ORB-SLAM3-TR	0.0246	0.0312	Fail
ORB-SLAM3-LM [3]	0.0393	0.0036	Fail
E2O-SLAM	**0.0064**	**0.00317**	**0.0120**

ACKNOWLEDGMENT

The authors used ChatGPT during the preparation of this manuscript for language editing and grammar enhancement throughout the manuscript. Following the use of this tool, the authors reviewed and edited the content as needed and take full responsibility for the integrity and accuracy of the publication.

References

[1] M. Liu, X. Zuo, R. Huang, M. Zhao, J. Chen, and L. Li, "ROEVO: Robust organized edge feature-based visual odometry using rgb-d cameras," *IEEE Trans. Robot.*, vol. 41, pp. 4860–4880, 2025.
[2] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1147–1163, 2015.
[3] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual–inertial, and multimap SLAM," *IEEE Trans. Robot.*, vol. 37, no. 6, pp. 1874–1890, 2021.
[4] R. Gomez-Ojeda, F.-A. Moreno, D. Zuñiga-Noël, D. Scaramuzza, and J. Gonzalez-Jimenez, "PL-SLAM: A stereo SLAM system through the combination of points and line segments," *IEEE Trans. Robot.*, vol. 35, no. 3, pp. 734–746, 2019.
[5] Y. Zhou, H. Li, and L. Kneip, "Canny-VO: Visual odometry with RGB-D cameras based on geometric 3-D–2-D edge alignment," *IEEE Trans. Robot.*, vol. 35, no. 1, pp. 184–199, 2019.
[6] H. Zhao, F. Gu, J. Shang, X. Long, J. Dou, C. Chen, H. Pu, and J. Luo, "Towards accurate, efficient and robust rgb-d simultaneous localization and mapping in challenging environments," *IEEE Transactions on Robotics*, 2025.
[7] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO," Jan. 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[8] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, 1986.
[9] Y. Wang, C. Jiang, and X. Chen, "VOOM: Robust visual object odometry and mapping using hierarchical landmarks," in *Proc. IEEE Int. Conf. Robot. Automat. (ICRA)*, 2024, pp. 10 298–10 304.