УДК 004.896 + 007.52 + 004.93’11

ONLINE INTEGRATION OF VISUAL, LANGUAGE, AND GEOMETRIC DATA FOR AN OPEN-VOCABULARY SEMANTIC SLAM SYSTEM

Mohrat M., Mahmoud J., Iumanov M. A. (ITMO University)
Scientific supervisor – Professor, Doctor of Technical Sciences, Kolyubin S. A.
(ITMO University)

Introduction. Modern robotic systems operating in dynamic and unpredictable environments require not only precise mapping and localization (SLAM) but also an understanding of their surroundings. Traditional SLAM algorithms such as ORB-SLAM3 [1] are typically limited to geometric reconstruction, which often lacks ability to address tasks that involve interaction with various objects. In recent years, increasing attention has been given to semantic SLAM—an approach that combines geometric maps with information about object types and attributes [2]. However, many existing methods such as RVWO [3] which is specified to a closed-set of moving objects can’t handle an open-set of categories which is essential for general-purpose robotics.
In this work, we propose a system that tightly integrates visual data from a camera, semantic and language descriptions model, and geometric features derived from the SLAM process. This integration not only enhances mapping and localization accuracy but also enables interactive, on-the-fly object searches via natural language queries, which is a valuable feature for various robotic systems.

Main part. The proposed system consists of a classical RGB-D SLAM backbone that delivers real-time camera pose estimation and geometric mapping. When each keyframe is detected, the data is passed to a semantic processing module where segmentation masks are generated using the semantic model of MobileSAMv2 [4] and embedded representations of objects are extracted via the visual-language model of MobileCLIP [5]. This integration enables open-vocabulary segmentation, allowing the system to process unseen objects before, without being limited by a fixed set of classes. To overcome the high computational load typically, the proposed pipeline employs a multi-threaded architecture in which SLAM operations of localization, mapping, and loop closure run in parallel with semantic processing.
To handle the challenges caused by dynamic environments where objects such as people for indoor environments and vehicles for outdoor environments, the proposed pipeline integrates semantic-geometric clustering that utilize depth information alongside segmentation masks and language features to differentiate dynamic objects from the static background. This shall enhance localization accuracy and minimize mapping errors. Furthermore, by using language models, our proposed pipeline supports open-vocabulary in an interactive manner that enables the end-user to make arbitrary queries in natural language to search the mapped space semantically for relevant matches. Experimental evaluations on datasets including Replica [6] and ARIA [7] show that the proposed method achieves high trajectory and high semantic understanding accuracy both are executed in real time.

Conclusion. To sum up, the proposed solution demonstrates that the tight integration of multi-modal information such as visual, language, and geometric data in an online manner can empower metric-semantic SLAM systems for robots to navigate and interact more effectively with their environment. The multi-threaded architecture of the pipeline and the utilization of SoTA real-time models deliver robust and efficient performance for operation even in dynamic environments.

References.

1. C. Campos, R. Elvira, J. J. G. Rodriguez, J. M. M. Montiel, and J. D. Tardos, “Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam,” IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874–1890, Dec. 2021.
2. A. Rosinol, M. Abate, Y. Chang, and L. Carlone, “Kimera: An open-source library for real-time metricsemantic localization and mapping,” in 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 1689–1696.
3. J. Mahmoud, A. Penkovskiy, H. T. Long Vuong, A. Burkov and S. Kolyubin, "RVWO: A Robust Visual-Wheel SLAM System for Mobile Robots in Dynamic Environments," *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Detroit, MI, USA, 2023.
4. C. Zhang, D. Han, S. Zheng, J. Choi, T.-H. Kim, and C. S. Hong, Mobilesamv2: Faster segment anything to everything, 2023. arXiv: 2312.09579 [cs.CV]. [Online]
5. P. K. A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, and O. Tuzel, Mobileclip: Fast image-text models through multi-modal reinforced training, 2024. arXiv: 2311.17049 [cs.CV].
6. J. Straub, T. Whelan, L. Ma, et al., The replica dataset: A digital replica of indoor spaces, 2019. arXiv: 1906.05797 [cs.CV].
7. X. Pan, N. Charron, Y. Yang, et al., “Aria digital twin: A new benchmark dataset for egocentric 3d machine perception,” in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Oct. 2023, pp. 20 133–20 143.

Mohrat M. (author) signature

Kolyubin S. A. (scientific supervisor) signature