Towards a Multi-modal, Multi-layer Mapping Framework for Autonomous Robotics – an Outline

Peteris Racinskis[1*], Janis Arents[1] and Modris Greitans[1]
[1]*Robotics and Machine Perception Laboratory, EDI, Riga, Latvia*
*Contact: peteris.racinskis@edi.lv

***Abstract*—In this paper, we outline the structure and necessary elements of a multi-modal, multi-layered (metric, topological, semantic) mapping framework intended to be used in autonomous robotics — with a focus on applications in completely unstructured (i.e., outdoor) environments. A concise overview of the current state of the art in relevant areas is used to serve as the basis for a provisional system architecture, while current limitations are used to motivate key directions for novel research. Specifically, we intend to tackle the challenges in terrain segmentation, SLAM in unstructured outdoor environments, fusing advances in open-set segmentation of 3D point clouds with hierarchical mapping techniques.**

***Keywords*—SLAM, Autonomous Robotics, Mobile Robotics, Semantic Mapping, Open-set Segmentation,**

I. Introduction

Mapping is a crucial aspect of enabling fully autonomous operation in mobile robots. Using the same terminology as in [1] a metric map allows the robot to localize itself in the environment and avoid obstacles. This becomes topological when a searchable graph structure is used to describe free space to aid planning tasks. Semantics then provide additional information such as the locations of distinct objects, and spatial relationships between them. Fig. 1 illustrates one possible conceptualization of an autonomous robot's control system and the role a mapping framework serves within it.

In many cases, the assumption that a previously constructed map of the environment is available does not hold. In this case, it's necessary to use the methods of Simultaneous Localization and Mapping (SLAM) to incrementally build one based on sensor measurements collected during exploration, the fundamentals of which are described in much greater detail in [2]. Currently, a multitude of software packages already exists to tackle the SLAM problem. Metric map construction methods exist on a continuum from the purely visual [3] through visual-inertial [4] up to LiDAR-based ones [5], [6], which are more prevalent in the more variable, less structured outdoor environment. Others focus on the construction of hierarchical metric-topological-semantic maps given segmented image data [7], which contrasts with approaches that directly infer semantics from point clouds [8].

Specifically with regards to the former — the use of semantically segmented image data — of great interest are the great strides that have been made in computer vision over the preceding decade. Methods for mapping text and images to a shared latent space have been developed [9], and recently these have been adapted for use in image-based point cloud segmentation [10]. This bears great promise when one considers achievements in using large language models to break down high-level natural language commands into actionable plans for robot execution [11]. It must, however, be noted that in some areas even producing segmentations in discrete class sets remains a challenge, specifically in the domain of terrain segmentation [12].

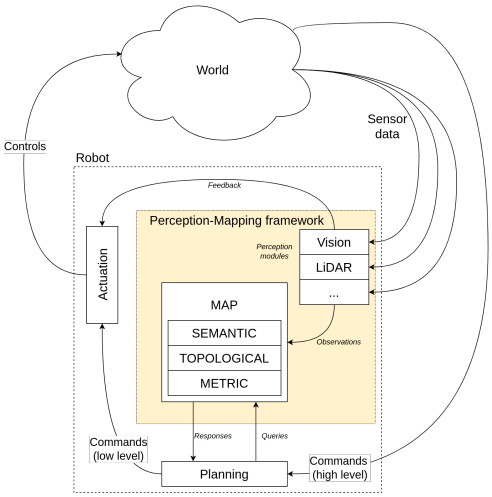

Fig. 1. A conceptual schematic of an autonomous robot control system. Highlighted in yellow are the aspects intended to be tackled as part of this research project - constituting a mapping framework.

II. Approach Overview

With this background in mind, we seek to build a mapping framework that advances the state-of-the-art of the semantic mapping field in the following ways:

- robust multi-modal perception modules with a focus on outdoor applications, specifically with regards to terrain traversability segmentation;
- integration between the latest open-set segmentation techniques and hierarchical map representations

To attain the objectives stated above, we plan to build out a framework consisting of several perception modules and a SLAM system integrating these into a unified, hierarchical representation with metric, topological, and semantic layers.

This research is funded by the Latvian Council of Science, project "Smart Materials, Photonics, Technologies and Engineering Ecosystem" project No VPP-EM-FOTONIKA-2022/1-0001.

A. Multi-modal terrain traversability segmentation

As in [12], where a vision transformer model is augmented with a highly engineered auxiliary loss computed directly on self-attention, we expect to heavily rely on visual (RGB) data for semantic segmentation of images. For practical path planning purposes, it is likely that discrete categories of terrain will have to be mapped even if the complete system also performs open-set segmentation as in [10]. We expect that integrating observations made from Unmanned Aerial Vehicles equipped with radar and/or hyperspectral sensors will help detect obstacles obscured to cameras mounted on a ground-based robot and discern otherwise invisible features (e.g., waterlogged soil). Higher-quality input data may help offset the inherent difficulties of terrain segmentation that have thus far required augmenting standard vision model architectures with highly engineered additional functionality.

B. SLAM in unstructured environments

Given the much greater variance in lighting conditions, greater distances, and prevalence of other sources of sensor ambiguity in unstructured outdoor environments, it is likely that purely image-based SLAM methods as in [3] will prove insufficient for our application. Already existing visual-inertial SLAM frameworks such as [4] provide a sensible starting point for our own SLAM software implementation. Extending these to make use of LiDAR data is likely to be a priority. Prior work in outdoor SLAM often relies on some assumed structure in the environment — such as the tree trunk features studied in [5], [6]. Taking a step back from the pure SLAM problem and considering external sources of positioning data (e.g., GPS) may be required for robust localization in environments without any reliable invariants.

C. Open-set segmentation in hierarchical maps

Assuming reliable estimates of the robot's pose are available along with depth measurements, creating 3-dimensional semantic maps from planar segmentations becomes a matter of backprojection, which is well understood [10]. However, existing approaches stop at the construction of point clouds, which cannot be traversed or queried in computationally efficient ways. A spatially grounded graph structure as in [7] provides a much better basis for use with search-based planning algorithms. Furthermore, work has already been done in extracting potentially arbitrary complex relationships between objects in a map through the use of graph neural networks [8]. However, the latter only studied inference on point clouds without any prior semantic annotation. We seek to explore the integration of open-set semantics at the point cloud level with graph-based, searchable maps. One possibility worth exploring is the propagation of embeddings through graph neural networks in constructing true semantic maps of arbitrary relationships between objects. Ultimately, the goal is to have a map able to readily interface with natural language processing conditioned high-level planning engines such as [11], queryable with respect to text-to-image embedding vectors in addition to discrete categories.

III. Discussion

Surveying the state-of-the-art in semantic SLAM, image segmentation and language-conditioned planning for robotics has revealed two clear directions where further research is needed, and these in turn motivate the structure of our proposed outdoor-oriented mapping framework. Comparing the maturity of indoor and outdoor approaches, however, indicates that operating in the less structured, more variable outdoors is significantly more challenging at the level of perception.

References

[1] C. Galindo, A. Saffiotti, S. Coradeschi, P. Buschka, J. A. Fernandez-Madrigal and J. Gonzalez, "Multi-hierarchical semantic maps for mobile robotics," 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2005, pp. 2278-2283, doi: 10.1109/IROS.2005.1545511.

[2] Thrun, Sebastian, Wolfram Burgard, and Dieter Fox. 2006. Probabilistic robotics.

[3] R. Mur-Artal, J. M. M. Montiel and J. D. Tardós, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," in IEEE Transactions on Robotics, vol. 31, no. 5, pp. 1147-1163, Oct. 2015, doi: 10.1109/TRO.2015.2463671.

[4] 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," in IEEE Transactions on Robotics, vol. 37, no. 6, pp. 1874-1890, Dec. 2021, doi: 10.1109/TRO.2021.3075644.

[5] Li, Qingqing, Paavo Nevalainen, Jorge Peña Queralta, Jukka Heikkonen, and Tomi Westerlund. 2020. "Localization in Unstructured Environments: Towards Autonomous Robots in Forests with Delaunay Triangulation" Remote Sensing 12, no. 11: 1870. https://doi.org/10.3390/rs12111870

[6] F. Nie, W. Zhang, Y. Wang, Y. Shi and Q. Huang, "A Forest 3-D Lidar SLAM System for Rubber-Tapping Robot Based on Trunk Center Atlas," in IEEE/ASME Transactions on Mechatronics, vol. 27, no. 5, pp. 2623-2633, Oct. 2022, doi: 10.1109/TMECH.2021.3120407.

[7] A. Rosinol, M. Abate, Y. Chang and L. Carlone, "Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping," 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 2020, pp. 1689-1696, doi: 10.1109/ICRA40945.2020.9196885.

[8] S. -C. Wu, J. Wald, K. Tateno, N. Navab and F. Tombari, "SceneGraphFusion: Incremental 3D Scene Graph Prediction from RGB-D Sequences," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 2021, pp. 7511-7521, doi: 10.1109/CVPR46437.2021.00743.

[9] Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision." International Conference on Machine Learning (2021).

[10] Jatavallabhula, Krishna Murthy, Ali Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Varma Keetha, Ayush Kumar Tewari, Joshua B. Tenenbaum, Celso M. de Melo, M. Krishna, Liam Paull, Florian Shkurti and Antonio Torralba. "ConceptFusion: Open-set Multimodal 3D Mapping." ArXiv abs/2302.07241 (2023): n. pag.

[11] Ahn, Michael et al. "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances." Conference on Robot Learning (2022).

[12] T. Guan, D. Kothandaraman, R. Chandra, A. J. Sathyamoorthy, K. Weerakoon and D. Manocha, "GA-Nav: Efficient Terrain Segmentation for Robot Navigation in Unstructured Outdoor Environments," in IEEE Robotics and Automation Letters, vol. 7, no. 3, pp. 8138-8145, July 2022, doi: 10.1109/LRA.2022.3187278.