Monocular 3D Mapping using DepthAnythingV2 and OctoMap: A Hybrid Approach for Occupancy Grid Reconstruction

Werikson Frederiko de Oliveira Alves
Department of Informatics
Postgraduate Program in Computer Science
Federal University of Viçosa
Viçosa-MG, Brazil
Email: werikson.alves@ufv.br

***Abstract*—The abstract goes here.**

I. Introduction

The reconstruction of 3D environments from monocular visual data remains a challenging problem in computer vision and robotics. While Simultaneous Localization and Mapping (SLAM) methods have matured significantly over the past decades, many rely on stereo vision systems or depth sensors such as LiDAR, which may increase the overall cost, complexity, and power consumption of robotic systems [1]. An alternative approach involves estimating depth maps directly from RGB images using deep learning models, enabling 3D perception with only monocular input.

In this work, we propose a novel pipeline that combines monocular depth estimation with 3D occupancy mapping. Specifically, we employ the DepthAnythingV2 model, a state-of-the-art depth estimator based on Vision Transformers [2], to infer depth from a monocular camera (e.g., Intel RealSense D435). The resulting depth maps are transformed into point clouds, which are incrementally integrated into an OctoMap [3], a probabilistic framework for volumetric mapping based on octrees. This hybrid approach aims to generate semantically coherent 3D occupancy maps in a low-cost and scalable way.

Practical applications of this system include autonomous robot navigation in unstructured environments, remote inspection via aerial vehicles, and scene understanding for augmented reality systems. Its ability to operate solely with monocular input makes it suitable for embedded and resource-constrained platforms, such as drones and micro-robots.

The main objective of this project is to develop an efficient and lightweight mapping solution capable of transforming monocular RGB video into consistent 3D occupancy maps. Evaluation will be based on both qualitative and quantitative metrics, including:

- Comparison between SLAM trajectory and ground-truth odometry.
- Visual comparison of the generated occupancy grid (using OpenCV) with real-world images.
- Processing time required to generate the map.
- Qualitative visual analysis of map coherence with the real scene.

The diagram illustrating the pipeline - from image acquisition, depth estimation, to occupancy map generation - is shown in Figure 1.

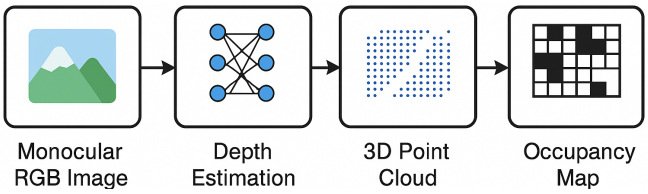

Fig. 1. Proposed Monocular 3D Mapping Pipeline

II. Related Work

Traditional SLAM techniques such as ORB-SLAM [4] and RTAB-Map [5] have achieved robust performance in various scenarios but typically depend on stereo vision or RGB-D inputs. In contrast, monocular SLAM approaches often suffer from scale ambiguity and drift. Recent advances in deep learning-based monocular depth estimation, particularly models like MiDaS [6] and DepthAnythingV2 [7], offer promising alternatives by predicting dense depth maps from single RGB frames with high generalization capacity.

The use of learned depth maps for mapping has been explored in recent literature. For example, [8] combined monocular SLAM with learned depth priors to enhance mapping accuracy. However, many of these methods still rely on bundle adjustment or additional motion priors. Our work leverages direct depth prediction and integrates the resulting 3D points into the OctoMap framework [3], which provides a memory-efficient, probabilistic volumetric representation suited for online robotic applications.

This work differentiates itself by integrating DepthAnythingV2 — a recently proposed model trained on a diverse collection of datasets — into a mapping pipeline without

requiring geometric consistency enforcement. Additionally, OctoMap offers incremental updates and hierarchical representation, enabling real-time operation and compact storage.

REFERENCES

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1309–1332, Dec 2016. [Online]. Available: https://doi.org/10.1109/TRO.2016.2624754

[2] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 10 371–10 381. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.00987

[3] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: an efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, no. 3, pp. 189–206, April 2013. [Online]. Available: https://doi.org/10.1007/s10514-012-9321-0

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós, "Orb-slam: A versatile and accurate monocular slam system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, Oct 2015. [Online]. Available: https://doi.org/10.1109/TRO.2015.2463671

[5] M. Labbé and F. Michaud, "Rtab-map as an open-source lidar and visual simultaneous localization and mapping library for large-scale and long-term online operation," *Journal of Field Robotics*, vol. 36, no. 2, pp. 416–446, 2019. [Online]. Available: https://doi.org/10.1002/rob.21831

[6] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1623–1637, March 2022. [Online]. Available: https://doi.org/10.1109/TPAMI.2020.3019967

[7] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 21 875–21 911. [Online]. Available: https://depth-anything-v2.github.io/

[8] K. Tateno, F. Tombari, I. Laina, and N. Navab, "Cnn-slam: Real-time dense monocular slam with learned depth prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6565–6574. [Online]. Available: https://doi.org/10.1109/CVPR.2017.695