Open World GPS Goal Visual Navigation Approach (Draft)

Zishuo Wang, Joel Loo, Yuwei Zeng, Nielsen Cugito, David Hsu

Architectural Principles. We developed a modular system addressing the key challenges underlying the task: **(i)** visual navigation with a monocular camera in **(ii)** open-world human environments, with **(iii)** low-frequency, high-latency sensing and control. Unreliable sensor streams coupled with noisy proprioception made accurate depth and scale estimation in the monocular setting challenging. To tackle **(i)**, the choice was made to forgo 3D metric geometry estimation and focus instead on traversability estimation in 2D image space, relying on semantic image cues. To generalise over the diverse scenes and appearance variations of **(ii)**, visual features pretrained on large-scale datasets are used, and augmented with fine-tuning on select portions of the FrodoBots-2K data. Owing to hardware limitations and the unpredictability of latency, **(iii)** was harder to directly address. The system instead focuses on handling navigation failures induced by suboptimal path-finding and poor trajectory tracking, which arise from the poor communications. This is achieved by augmenting the navigation pipeline with robust failure detection and recovery.

At a high level, the system (Figure 1) consists of *perception*, *control* and *failure detection and recovery* modules. The perception module estimates traversability from RGB input, and also issues an egocentric direction vector to the next waypoint. The control module selects kinodynamically feasible trajectories aligned with the waypoint vector and generates control commands. The failure detection and recovery module is a supervisory monitor taking in raw RGB and predicted traversability from perception to detect failures, overriding the control module to execute heuristic recovery behaviours when necessary.

Perception. Given the need to operate in open-world human environments without reliable depth sensing due to the monocular setting, visual traversability prediction based on scene semantics was used. The perception module takes an RGB image as input, and outputs a traversability mask based on the input image, with traversability scores in [0, 1]. Internally, a fast traversability estimator generates an initial mask, which is then further postprocessed with clustering heuristics to identify and strongly penalise likely obstacles. The estimator uses pretrained DINO-ViT visual features which enable strong generalisation over diverse environments, and allow for sample-efficient training and finetuning to adapt to new scenes.

To train an estimator for the wheeled FrodoBot configuration while capturing preferences on different terrains, a pipeline for automatically labelling data from FrodoBots-2K was developed. Based on the fixed egocentric camera view, the region where the robot is currently teleoperated onto as the traversable region is segmented with Segment Anything Model [1] prompted with the bottom central pixels. The Side Adapter Network [2] filters out low-quality images with motion blur and overexposure, by checking and discarding images with no detected traversable areas.

Trajectory generation and control. Kinodynamically feasible trajectories are chosen and tracked with a modified Dynamic Window Approach (DWA). DWA simplifies system design by unifying local planning and trajectory tracking, since it generates trajectories parameterised with velocities to directly command the robot with. Its inputs are an egocentric heading toward the next subgoal and the 2D traversability mask, and it outputs (v, ω). Firstly, reactive obstacle avoidance is improved by modifying DWA's search space to use more complex trajectory primitives. Trajectory primitives are extended from simple arcs to multi-segmented arcs. Similar to MPCs, each trajectory is rolled out for t_{sim} but only followed for $t_{track} < t_{sim}$. Secondly, trajectories are projected onto the traversability mask using camera intrinsics, to evaluate kinematic feasibility in the absence of bird's-eye view geometry information. A traversability score is summed from pixel values in the mask that lie within the trajectory inflated by the robot's projected footprint.

Failure detection and recovery. The inevitability of failures in the open world is a key principle of the system's design, necessitating a module to recover from navigation failures. This monitors RGB input and traversability masks for failures, then activates heuristic recovery behaviours which override the navigation layer to reset the robot. It maintains a severity level based on failure frequency which balances between caution and aggressiveness of corrective action. The module's strategy is to take successively bolder *local* actions to perturb the robot out of the failure state.

Two common failure modes are: **(i)** suddenly encountering untraversable areas (*e.g.* when blocked by a dynamic obstacle); **(ii)** getting stuck in local minima (*e.g.* taking a wrong turn into a dead-end). Detection of these modes is approximated by detecting overall low traversability across the mask, and detecting that the robot is immobile despite being commanded to move. Upon failure detection, the module alternates among *backtracking* and *perturbation* behaviours. Backtracking executes cached actions open-loop, while perturbations are local traversability-aware actions generated by DWA with reduced goal weighting. The magnitude of these actions increases with severity level. Competition results empirically (Figure 1) show recovery to be crucial for escaping local minima in cluttered urban spaces (*e.g.* benches, bushes

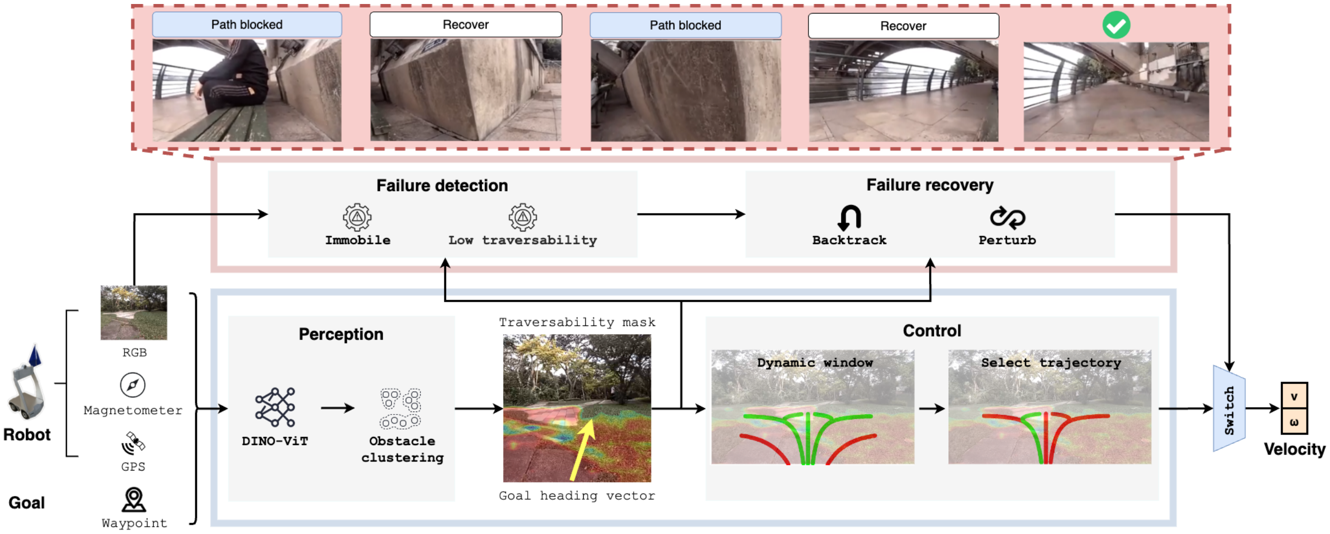

Fig. 1: The system deals with purely monocular navigation across diverse locations via traversability estimation with pretrained models coupled with selection of kinodynamically feasible trajectories in image space, without explicit 3D geometry reconstruction. Open-worldness and latency lead to inevitable failures, addressed by a high-level failure recovery system for monitoring and execution of heuristic recovery behaviours when necessary.

etc.) and handling challenging areas with mixed terrains.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[2] M. Xu, Z. Zhang, F. Wei, H. Hu, and X. Bai, “Side adapter network for open-vocabulary semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2945–2954.