Mobile Manipulation in Unexplored and Unstructured Environments

Daniel Honerkamp
University of Freiburg
Email: honerkamp@cs.uni-freiburg.de

I. Introduction

Recent advances in machine learning and robotics have led to remarkable progress in static manipulation tasks [13, 14] and in the navigation of unstructured environments [7, 12]. Despite these strides, robots have remained confined to very specific applications within each domain. To truly permeate our everyday lives, we need robots that are general-purpose and flexible enough to act autonomously in human environments. For this, we need to bring these capabilities together. Mobile manipulators are robots with mobile bases and manipulator arms. In recent years, these platforms have developed the necessary hardware capabilities. However, with their flexibility comes a large action space, and their mobility means that they are acting in large areas, spanning full apartments and a myriad of objects with many possible interactions. As a result, the input and output spaces combinatorially explode, making it an open challenge to efficiently control these platforms.

In my research, I tackle these challenges both from the low-level executability and the high-level reasoning in three interconnected lines of work: (i) I develop hybrid methods to execute arbitrary end-effector motions based on known kinematics and reinforcement learning (RL), (ii) I develop sample-efficient and multimodal approaches for the exploration of unknown indoor environments, and (iii) hierarchical methods that scale these components to autonomously complete long-horizon tasks. The resulting approaches achieve efficient, autonomous behavior in large, real-world apartments and generalize to unseen tasks and environments. Importantly, the methods I developed are fully reactive and capable of acting in unexplored environments, enabling work alongside and in collaboration with humans in unstructured environments. By demonstrating their effectiveness on a wide range of different robots, I ensure their generality and usefulness. By making all tasks, benchmarks, and models openly accessible, I hope to further progress the field.

II. Low-level Motion Execution

Given their large control space, most existing approaches are unable to produce efficient whole-body motions for mobile manipulators. Instead, they restrict themselves to sequentially move the base and the arm of the robot [19, 28]. As the problem space increases, optimal control [18] struggles with local optima while the planning times of motion planners [16] quickly increase, making them unreactive and unsuitable for dynamic environments such as working alongside humans. End-to-end learning approaches either require infeasible amounts of data or learn solutions to very specific tasks [27], requiring expensive retraining to solve even simple variations.

Fig. 1. I develop efficient, generalizable and reactive mobile-manipulation policies that can act in and reason about unexplored environments.

I proposed a hybrid method that combines RL with the kinematic models of the robot [8]. I decompose the problem into an RL agent that controls the base of the robot, and an inverse kinematics solver to complete the motions to achieve desired end-effector motions. The aim of the RL agent is to move such that, given arbitrary end-effector motions, kinematics remain feasible at all times. I then extended this approach to incorporate obstacles and control the speed of the task execution [9]. By focusing on geometric modalities such as local occupancy maps, we can train the agent in procedurally generated environments without the need for an expensive simulator or real-world data. The trained agent achieves zero-shot performance to the real world, where it can solve complex, unseen tasks such as opening doors while driving through the narrow door frame, opening cabinets and drawers, and rearranging objects - while avoiding dynamic obstacles, as shown in Figure 1.

While this is efficient, we may still want to collect large datasets for the training of foundation models [17]. But teleoperation of these platforms is either cumbersome, as joysticks do not have enough inputs to control the whole body, or expensive, with specialized equipment such as Mobile-Aloha costing more than 30,000 USD [6]. Instead, I infer end-effector motions from existing inputs, such as joysticks or hand guidance, then connect them with our RL agent. This results in a zero-cost whole-body teleoperation that simplifies data collection, enables rapid learning of mobile manipulation actions, and the first dynamic approach for mobile kinesthetic teaching [11].

Since we can now learn performant policies, I then loop back to the hardware design of modular mobile manipulators. I introduce a concurrent design approach to optimize the mounting parameters of arms on mobile platforms, utilizing Bayesian Optimization to generate designs that lead to significantly higher performance across mobile manipulation tasks [24].

III. Efficient Multi-modal Indoor Exploration

To navigate large, unexplored, human-centered environments, agents have to build compact representations and integrate short- and long-term reasoning. I proposed an approach that unifies short- and long-term reasoning in a single model by predicting long-term intentions together with short-term continuous navigation commands [22]. This method is centered around an extended semantic top-down map, which serves as central memory. We train the agent to predict the direction towards the next target, which is then communicated to an RL agent that produces continuous control commands. While existing methods relied on granular discrete actions [2, 26], our approach can directly act in the continuous low-level action space of the robot controller and achieves state-of-the-art results on multi-object-search and zero-shot transfers to a real robot.

Investigating the object search literature, we identify a heavy reliance on zero-shot deployment of ground-truth trained RL policies with pretrained semantic perception models. However, this results in a large performance drop. As remedy, I incorporate uncertainty measures into the temporal aggregation and found decisions to make the policies aware that they are acting on imperfect perception [20]. This not only significantly decreases the perception gap at deployment time, but can also be incorporated without any additional finetuning or retraining.

Sound serves as a major communication signal in our world, be it spoken word or audio alarms such as a ringing telephone. To act in our world, robots need to leverage these signals. In my work [29], we extend existing audiovisual navigation tasks [3, 4] and strongly increase their complexity by designing audio-specific distractors, noise sources, and the first dynamic audio-navigation task in which the agent has to catch moving sound-emitting targets. We then introduce a method to integrate audio signals with the geometric information inherent in top-down maps through a spatial audio encoder. This resulted in strongly increased generalization to unheard sounds and a first and second place in the CVPR Soundspaces Challenges.

IV. High-level Reasoning

To complete long-term tasks, we require an additional level of reasoning to coordinate behaviors over long spatial and temporal distances. Existing ObjectNav work focuses on freely accessible objects out in the open [1, 21, 5]. But in human environments, we cannot expect these assumptions to hold. We take the next step by introducing an interactive search task in which the robot has to open doors to free pathways or search through cabinets and drawers to find the objects of interest. To address these new challenges, we build on top of our previously introduced semantic map memory [22] and design an object-centric action space in which detected instances serve as navigation points. I developed HIMOS [23], a hierarchical RL approach that learns to trade off the costs and benefits of object interactions and explorations. Lastly, by training with imperfect manipulation subpolicies, the agent learns a re-trial behavior if subpolicies do not succeed. The resulting agent readily transfers to the real world, where its modularity enables us to replace the subpolicies with completely unseen real-world versions of them.

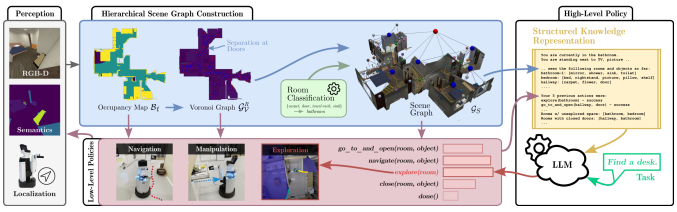

Fig. 2. Interleaving mobile manipulation with efficient scene representations and high-level reasoning to solve complex tasks over long-horizons.

While RL can learn good decision-making with enough training data, large language models were shown to absorb a lot of knowledge about human environments and be capable of high-level reasoning. However, so far, this has remained restricted to game-like environments, static table-top manipulation, or fully observable scenes [25, 14, 15]. It remains a challenge to reason over partially observed large scenes. In [10], I developed a graph-based scene understanding approach that scales to large apartments and can be tightly coupled with motion execution. We then introduce a knowledge extraction that encodes scene graphs into structured textual representations for a large language model (LLM). An overview is depicted in Figure 2. We show that this representation results in well-grounded reasoning from the language model. In contrast to previous work, the approach scales to many objects, and all components can be built up dynamically as the agent explores the scene.

V. Future Work

Feedback and Re-trial: Robots will never achieve perfect success on all tasks they attempt and neither do humans. The range of possible failure reasons is almost unlimited. Therefore, to achieve high reliability, we need to understand failure reasons, to then be able to react to them and either re-try or change plans. If the door does not open because it is locked, we need to react differently than if we failed to grasp the handle. To address this challenge, I plan to focus on multi-modal models that can summarize high-dimensional information across visual, audio, and force sensing. I then aim to develop tight feedback loops between perception and reasoning that enable event-based reactions.

Bridging Decision Making Frameworks: LLMs are good at reasoning about tasks that are well represented within their training data but may be less optimal for robot-specific aspects. On the other hand, RL can learn near-optimal policies if we can simulate them many times. Lastly, planning methods exceed in known environments if given enough budget. I plan to develop methods that draw on the strengths of these frameworks. As a first step, I plan to investigate the interweaving of RL and LLMs for embodied tasks, translating tasks into abstract plans via LLMs and then leveraging RL to translate them into actionable items.

General Open-set Tasks: While we are able to complete most combinations of rearrangements and articulated object interactions, completely arbitrary tasks require further, often unforeseeable reasoning and motion capabilities. I plan to research flexible approaches that leverage language and graph-based methods to model arbitrarily structured tasks.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *Proc. of the Int. Conf. on 3D Vision*, 2017.

[2] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Proc. of the Conf. on Neural Information Processing Systems*, 33:4247–4258, 2020.

[3] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In *Proc. of the Europ. Conf. on Computer Vision*, pages 17–36, 2020.

[4] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *International Conference on Learning Representations*, 2020.

[5] Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, et al. Robothor: An open simulation-to-real embodied ai platform. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 3164–3174, 2020.

[6] Zipeng Fu, Tony Z. Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. In *Proc. of the Conference on Robot Learning*, 2024.

[7] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pages 7606–7623, May 2022.

[8] Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Learning kinematic feasibility for mobile manipulation through deep reinforcement learning. *IEEE Robotics and Automation Letters*, 6(4), 2021.

[9] Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. N^2M^2: Learning navigation for arbitrary mobile manipulation motions in unseen and dynamic environments. *IEEE Transactions on Robotics*, 39(5), 2023.

[10] Daniel Honerkamp, Martin Büchner, Fabien Despinoy, Tim Welschehold, and Abhinav Valada. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. *IEEE Robotics and Automation Letters*, 9(10), 2024.

[11] Daniel Honerkamp, Harsh Mahesheka, Jan Ole von Hartz, Tim Welschehold, and Abhinav Valada. Whole-body teleoperation for mobile manipulation at zero added cost. *IEEE Robotics and Automation Letters*, 10(4):3198–3205, 2025.

[12] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Int. Conf. on Robotics & Automation*, London, UK, 2023.

[13] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Proc. of the Conference on Robot Learning*, 2023.

[14] Boyi Li, Philipp Wu, Pieter Abbeel, and Jitendra Malik. Interactive task planning with language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

[15] Kevin Lin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg. Text2motion: From natural language instructions to feasible plans. *Autonomous Robots*, 47(8):1345–1365, 2023.

[16] Michael Otte and Emilio Frazzoli. Rrtx: Real-time motion planning/replanning for environments with unpredictable obstacles. In *Algorithmic Foundations of Robotics XI*, pages 461–478. Springer, 2015.

[17] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models. In *Int. Conf. on Robotics & Automation*, pages 6892–6903. IEEE, 2024.

[18] Johannes Pankert and Marco Hutter. Perceptive model predictive control for continuous mobile manipulation. *IEEE Robotics and Automation Letters*, 5(4):6177–6184, 2020.

[19] F. Paus, P. Kaiser, N. Vahrenkamp, and T. Asfour. A combined approach for robot placement and coverage path planning for mobile manipulation. In *Int. Conf. on Intelligent Robots and Systems*, 2017.

[20] Sai Prasanna, Daniel Honerkamp, Kshitij Sirohi, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. Perception matters: Enhancing embodied ai with uncertainty-aware semantic segmentation. *Robotics Research*, 2024.

[21] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

[22] Fabian Schmalstieg, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Learning long-horizon robot exploration strategies for multi-object search in continuous action spaces. *Robotics Research*, 2022.

[23] Fabian Schmalstieg, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Learning hierarchical interactive multi-object search for mobile manipulation. *IEEE Robotics and Automation Letters*, 8(12), 2023.

[24] Raphael Schneider, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Task-driven co-design of mobile manipulators. *Under review*, 2024.

[25] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[26] Saim Wani, Shivansh Patel, Unnat Jain, Angel Chang, and Manolis Savva. Multion: Benchmarking semantic map memory using multi-object navigation. *Proc. of the Conf. on Neural Information Processing Systems*, 33: 9700–9712, 2020.

[27] Josiah Wong, Albert Tung, Andrey Kurenkov, Ajay Mandlekar, Li Fei-Fei, Silvio Savarese, and Roberto Martín-Martín. Error-aware imitation learning from teleoperation data for mobile manipulation. In Aleksandra Faust, David Hsu, and Gerhard Neumann, editors, *Proc. of the Conference on Robot Learning*, volume 164, pages 1367–1378, 2022.

[28] Fei Xia, Chengshu Li, Roberto Martín-Martín, Or Litany, Alexander Toshev, and Silvio Savarese. Relmogen: Leveraging motion generation in reinforcement learning for mobile manipulation. In *Int. Conf. on Robotics & Automation*, 2021.

[29] Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *IEEE Robotics and Automation Letters*, 8(2), 2023.