Structured Interaction Models for Robot Learning in Human Spaces

Shuijing Liu
Department of Electrical and Computer Engineering
University of Illinois Urbana-Champaign
Email: sliu105@illinois.edu

I. Introduction

Robots are becoming increasingly capable of performing tasks in isolation. However, how to operate them in human environments remains an open challenge. Subtle and highly dynamic interactions among different agents are prevalent in human environments. These interactions are difficult to infer, which poses significant challenges for robot planning. **To enable autonomous navigation in human environments, I propose methods that rely on structured interaction models to facilitate robot decision making.** By uncovering the structures beneath the interactive behaviors of agents, I aim to build better robots that share living spaces with humans.

Two types of interactions exist while robots navigate among humans: Implicit interactions through motions (e.g., two agents yielding to each other when crossing a corridor) and explicit interactions through communication (e.g., speech, text, interfaces). To smoothly achieve tasks with both types of human-robot interactions, robots need to predict human intentions, reason about interactions, and plan to achieve the task goal. To this end, I propose navigation systems with structured interaction frameworks, achieving synergies between interaction modeling and decision making.

Implicit interactions: Robot navigation with other dynamic agents, such as moving crowds and busy intersections, is an important yet challenging problem. In crowded spaces, dynamic agents implicitly interact and negotiate with each other. The intended goal of agents are not observable to the robot, which poses extra difficulties for crowd navigation. Previous works have used model-based methods or reinforcement learning (RL) for navigation policies [21, 8, 5, 7, 11]. However, these works suffer from two drawbacks: 1) The robot does not explicitly predict human intentions during planning, which results in shortsighted and impolite behaviors [20, 10]; 2) Different types of interactions among agents through space and time are (partially) ignored by planners, which causes performance degradation in dense crowds and highly interactive scenarios [6, 15].

Explicit interactions: Audio is a natural and intuitive way for robots to communicate with humans. There is a spur of work that enables robots to follow human commands by taking advantage of language models [1, 9, 19]. However, in these works, the design of robot systems is not tailored toward the needs of target users, which limits their applications and societal impact. In addition, when deployed in new domains, existing pipelines for command following robots are difficult or inefficient to finetune [4].

Contributions: In response to the aforementioned challenges, I propose the following research contributions: (1) To encourage the robot to be aware of human intentions during navigation, I incorporate the predictions of agent behaviors into RL planners for robots. As a result, my intention aware RL framework learns safer and socially-aware navigation policies. (2) To model heterogeneous spatial and temporal interactions among agents, I propose policy network architectures that model *interactive scenarios* as graphs. The graph decomposition ensures better scalability in complex multiagent scenarios. (3) For robots to fulfill language commands, I propose methods that incorporate visual-language representations into the robot navigation modules. My methods show promising deployment and continual improvement with non-expert users.

II. Past and Ongoing Research

A. Intention aware robot navigation with interaction graphs

1) Intention aware RL [15, 14]: I propose an intention aware RL framework that incorporates predictions of human future intentions into planning, which leads to longsighted robot trajectories. The human intentions include the future trajectories and individual traits, such as aggressiveness. Given any off-the-shelf human trajectory predictor or trait predictor, we introduce the predicted information into the robot observation space and adjust the reward function accordingly. In RL training, the predictor takes past human trajectories as input and predicts human intentions. The output predictions are used as part of robot observations, which are fed to the robot policy network. Additionally, in [15], we design a reward function that encourages the robot to keep away from both current and intended positions of observed humans, improving both safety and politeness of the robot. We conducted extensive experiments in robot crowd navigation and autonomous driving tasks. The results show that intention-aware RL leads to safer and more polite robot policies. This finding demonstrates the high correlation between prediction and planning to achieve proactive robot behaviors in human spaces. We also illustrate the importance of modifying the RL formulation for robotic reinforcement learning on real world problems.

2) Graph strucutres in policy networks [13, 15, 14]: We propose a heterogeneous spatio-temporal graph (st-graph), which captures different types of interactions among agents through both space and time. Based on the st-graph, we derive a novel neural network to learn navigation policies, which consists of two components as follows. We use attention networks to represent spatial interactions among agents at the same timestep. The attention networks enable the robot to pay more attention to important agents, which ensures good performance when the number of humans increases and the graph becomes complex. We use recurrent neural networks to represent the temporal interactions, which model the dynamic evolution of the navigation scenarios. The recurrent networks enables temporal reasoning, which is useful in modeling highly dynamic crowds or traffic. By training the network that consists of spatial and temporal components, the robot learns a safer policy compared with methods that consider partial or no interactions. In addition, we successfully transfer the policy to real world robot navigation among pedestrains. This result demonstrates the power of injecting graph structures into neural networks. By doing so, complex problems can be decomposed into smaller components which become easier to solve.

B. Visual audio grounding for command following robots

1) Visual audio representation learning [3, 4]: To communicate with humans via dialogue, the robot must be able to associate audio commands with visual observations and motor skills. We propose a novel framework that uses visual-audio representation (VAR) as RL reward for skill learning. The image and audio came from cheap raw sensor inputs without expensive hardware for state measurements. More specifically, we train VAR with contrastive learning using paired audio and image data. The data pairs are fed to audio and image processing branches, where the encoded vectors that have the same meaning are pulled closer to each other in the representation space (e.x. An audio clip of "Turn on the TV" and an image of a TV turned on). Unlike other visual-language representations that need speech recognizers to handle speech, VAR directly interprets raw sound signals, which alleviates the intermediate errors caused by speech-to-text. In addition, VAR generalizes well to different sounds including speech, environmental sound, and music. Our work highlights the importance of multi-modal representations, which serve as building blocks for smooth human-robot communication.

2) Planning with representations [2, 3, 16]: We incorporate visual language or visual audio representations into robot planners so that the robot can take actions to fulfill user commands. More specifically, after receiving a user command, the trained representations compute similarity scores between the command and image observations from the environment. Then, the similarity scores can guide the robot to approach states where the image observation matches the command with the highest scores. After deployment in everyday environments, non-experts can easily improve the planner by finetuning the representations with a few amount of visual language pairs (Sec. II-B), ensuring the reslience and robustness of our end-to-end systems in novel environments. This representation-aided planning method is compatible with both RL planners and conventional search-based planners. We demonstrate our method in various robotic tasks, including assistive navigation for blind people and embodied navigation. Our systems demonstrate good generalization performance and high user satisfaction. Our work highlights the synergies between visual-language grounding and planning for command following robots, calling for the co-design of these components. Furthermore, our results are the first to show that grounding and dialogue enhances human-robot interactions through user study with real users.

III. Future Work

For robots to serve humans in all types of scenarios, we need to iterate between training the robot and deploying it in real human environments. To facilitate this training and deployment loop, my research agenda involves three directions: 1) Developing a unified model for both implicit and explicit interactions to unlock more robot capabilities; 2) Learning social behaviors from foundational models to improve the finetuning efficiency during deployment; 3) Developing intuitive lifelong learning methods from non-expert humans.

Unified interaction models: For robots to collaborate with humans on tasks such as preparing a meal, they must interact with humans both implicitly and explicitly. For this reason, I aim to expand the spatio-temporal interaction graph to capture interactions beyond movement, such as motions, dialogue, gaze, and so on. However, as we expand the graph, we must consider computational constraints and better understand the important factors in different interaction settings. To achieve this, inspired by the successful adaptation of st-graph in complex prediction tasks [12], I plan to design st-graphs with edges that encode various types of interactions.

Interaction learning from foundational models: My previous works train predictors and planners using a separate dataset or simulator for each task. However, this "tabula rasa" paradigm can be data inefficient. The trained models also have limited generalization across different tasks [17]. On the other hand, large language models (LLMs) encompass common sense knowledge such as social norms [18]. To improve the data efficiency and generalization of my method, I plan to incorporate visual language foundation models into the intention predictor and robot planner. The proposed approach takes observed video frames and task-specific prompts as inputs and outputs human intentions or robot motion primitives. Since my previous framework makes minimal assumptions on the form of submodules, it has promising compatibility for LLMs.

Intuitive finetuning from non-experts: When a trained robot is deployed in everyday environments, its performance drops inevitably due to domain shifts. Ideally, we need data-efficient and intuitive fine-tuning algorithms that allow non-experts with little domain expertise to customize and improve the robot. Building on my previous work on planning with representations [3], I aim to propose 1) intuitive user interfaces

for non-experts to provide finetuning data from their own devices, such as phone cameras and microphones, and from direct physical teleoperation; 2) data-efficient algorithms for robot to self-improve with minimal data collected from non-experts [4, 22]. With an intuitive and data-efficient finetuning mechanism, the robot can continually improve itself in target environments and customize toward user preferences.

REFERENCES

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning (CoRL)*, 2022.

[2] Peixin Chang, Shuijing Liu, Haonan Chen, and Katherine Driggs-Campbell. Robot sound interpretation: Combining sight and sound in learning-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[3] Peixin Chang, Shuijing Liu, and Katherine Driggs-Campbell. Learning visual-audio representations for voice-controlled robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[4] Peixin Chang, Shuijing Liu, Tianchen Ji, Neeloy Chakraborty, Kaiwen Hong, and Katherine Rose Driggs-Campbell. A data-efficient visual-audio representation with intuitive fine-tuning for voice-controlled robots. In *Conference on Robot Learning (CoRL)*, 2023.

[5] Changan Chen, Yuejiang Liu, Sven Kreiss, and Alexandre Alahi. Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

[6] Yuying Chen, Congcong Liu, Bertram E Shi, and Ming Liu. Robot navigation in crowds by graph convolutional networks with attention learned from human gaze. *IEEE Robotics and Automation Letters*, 2020.

[7] Akansel Cosgun, Lichao Ma, Jimmy Chiu, Jiawei Huang, Mahmut Demir, Alexandre Miranda Anon, Thang Lian, Hasan Tafish, and Samir Al-Stouhi. Towards full automated drive in urban environments: A demonstration in gomentum station, california. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1811–1818, 2017.

[8] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 1995.

[9] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[10] Zhe Huang, Ruohua Li, Kazuki Shin, and Katherine Driggs-Campbell. Learning sparse interaction graphs of partially detected pedestrians for trajectory prediction. *IEEE Robotics and Automation Letters*, 2022.

[11] David Isele, Reza Rahimi, Akansel Cosgun, Kaushik Subramanian, and Kikuo Fujimura. Navigating occluded intersections with autonomous vehicles using deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2034–2039, 2018.

[12] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] Shuijing Liu, Peixin Chang, Weihang Liang, Neeloy Chakraborty, and Katherine Driggs-Campbell. Decentralized structural-rnn for robot crowd navigation with deep reinforcement learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[14] Shuijing Liu, Peixin Chang, Haonan Chen, Neeloy Chakraborty, and Katherine Driggs-Campbell. Learning to navigate intersections with unsupervised driver trait inference. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[15] Shuijing Liu, Peixin Chang, Zhe Huang, Neeloy Chakraborty, Kaiwen Hong, Weihang Liang, D. Livingston McPherson, Junyi Geng, and Katherine Driggs-Campbell. Intention aware robot crowd navigation with attention-based interaction graph. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[16] Shuijing Liu, Aamir Hasan, Kaiwen Hong, Runxuan Wang, Peixin Chang, Zachary Mizrachi, Justin Lin, D Livingston McPherson, Wendy A Rogers, and Katherine Driggs-Campbell. Dragon: A dialogue-based robot for assistive navigation with visual language grounding. *IEEE Robotics and Automation Letters*, 2024.

[17] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022.

[18] Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268, 2022.

[19] Dhruv Shah, Błażej Osiński, brian ichter, and Sergey Levine. LM-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning (CoRL)*, 2022.

[20] Peter Trautman and Andreas Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010.

[21] Jur Van Den Berg, Stephen J Guy, Ming Lin, and Dinesh Manocha. Reciprocal n-body collision avoidance. In *Robotics research*. 2011.

[22] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning (CoRL)*, 2022.