Structured Interaction Models for Robot Learning in Human Spaces

Robots are becoming increasingly prevalent in our daily lives. However, these autonomous agents work best in isolation, such as restricted areas in warehouses. Enabling these robots to coexist with humans remains an unsolved challenge, because subtle and dynamic interactions among different agents are difficult to infer, which poses significant challenges for robot operation. **To enable broader deployment of robots in human spaces, I leverage the underlying structures in interactive scenarios to improve robot decision-making.**

Past and Ongoing Work

My work tackles both **implicit** human-robot interactions through motions and **explicit** interactions through language. I design structured frameworks that unify human intention prediction, interaction reasoning, and planning.

1) Interaction-aware crowd navigation: Robot navigation alongside dynamic agents, including pedestrians and vehicles, is important for last-mile delivery and autonomous driving. Previous works used reinforcement learning (RL) to learn navigation policies [1, 2, 3]. However, these methods ignore human future intentions or the interactions among agents, resulting in shortsighted and impolite robots [4, 5]. For the first time, I formulate the crowd navigation as a heterogeneous spatio-temporal graph, which captures various interactions among agents through space and time [6, 7, 8, 9]. From the graph, I derive a novel policy network with attention mechanism, enabling the robot to attend to important humans, such as those nearly colliding with the robot. To avoid shortsighted behaviors, I propose an intention-aware RL framework that allows the robot to avoid the intended paths of humans [8, 10, 7]. My experiments in robot navigation tasks among dense pedestrian and vehicle crowds show that my planner leads to a safe, longsighted, and social-aware robot. This finding has inspired follow-up research in behavior prediction [11, 12] and spatio-temporal networks [13, 14, 15] for navigation.

2) Language-conditioned interactions: To fulfill human commands such as "bring me some water", the robot must associate human intentions with the surrounding world. However, the development of existing visual-language models is done by engineers. Thus, it is difficult for non-experts to tailor these models based on their needs [16]. To bridge this gap, I propose a visual-audio representation that is data-efficient and intuitive for non-experts to fine-tune after the robot is deployed in novel environments [17, 18]. For robots to fulfill commands, I utilize visual-audio representations to select goals for planning [19, 17, 18, 20]. My robots demonstrate good generalization to novel environments in multiple public benchmarks and high user satisfaction in user study. My work highlights the synergies between language understanding and planning for command-following robots. Furthermore, my pipeline brings end users into the development loop of visual-language models, improving the generalization and accessibility of these models in real-world applications.

Research Agenda

To live in human environments, the robot must handle various types of interactions simultaneously and improve itself continuously. To this end, my research agenda involves two directions: 1) *Developing a unified interaction model for both implicit and explicit interactions.* Such a structured interaction model can unlock more robot capabilities, such as yielding to a human and saying "please go ahead" in a narrow corridor simultaneously. 2) *Lifelong learning from non-experts.* After deployment in human spaces, robots will be able to collect a large amount of interaction data. Using these data, I plan to develop user interfaces and algorithms to allow everyone to train and customize their robots without too much expertise.

REFERENCES

[1] C. Chen, Y. Liu, S. Kreiss, and A. Alahi, "Crowd-robot interaction: Crowd-aware robot navigation with attention-based deep reinforcement learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 6015–6022.

[2] A. Cosgun, L. Ma, J. Chiu, J. Huang, M. Demir, A. M. Anon, T. Lian, H. Tafish, and S. Al-Stouhi, "Towards full automated drive in urban environments: A demonstration in gomentum station, california," in *IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1811–1818.

[3] D. Isele, R. Rahimi, A. Cosgun, K. Subramanian, and K. Fujimura, "Navigating occluded intersections with autonomous vehicles using deep reinforcement learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 2034–2039.

[4] P. Trautman and A. Krause, "Unfreezing the robot: Navigation in dense, interacting crowds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2010, pp. 797–803.

[5] Z. Huang, R. Li, K. Shin, and K. Driggs-Campbell, "Learning sparse interaction graphs of partially detected pedestrians for trajectory prediction," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1198–1205, 2022.

[6] S. Liu, P. Chang, W. Liang, N. Chakraborty, and K. Driggs-Campbell, "Decentralized structural-rnn for robot crowd navigation with deep reinforcement learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 3517–3524.

[7] S. Liu, P. Chang, H. Chen, N. Chakraborty, and K. Driggs-Campbell, "Learning to navigate intersections with unsupervised driver trait inference," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2022, pp. 3576–3582.

[8] S. Liu, P. Chang, Z. Huang, N. Chakraborty, K. Hong, W. Liang, D. Livingston McPherson, J. Geng, and K. Driggs-Campbell, "Intention aware robot crowd navigation with attention-based interaction graph," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 12 015–12 021.

[9] N. Chakraborty, A. Hasan, S. Liu, T. Ji, W. Liang, D. L. McPherson, and K. Driggs-Campbell, "Structural attention-based recurrent variational autoencoder for highway vehicle anomaly detection," in *IFAAMAS International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2023.

[10] Y.-J. Mun, M. Itkina, S. Liu, and K. Driggs-Campbell, "Occlusion-aware crowd navigation using people as sensors," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 12 031–12 037.

[11] K. Lee, J. Li, D. Isele, J. Park, K. Fujimura, and M. J. Kochendorfer, "Robust driving policy learning with guided meta reinforcement learning," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023, pp. 4114–4120.

[12] A. Hasan, N. Chakraborty, H. Chen, J.-H. Cho, C. Wu, and K. Driggs-Campbell, "PeRP: Personalized residual policies for congestion mitigation through co-operative advisory systems," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2023.

[13] H. He, H. Fu, Q. Wang, S. Zhou, W. Liu, and Y. Chen, "Spatio-temporal transformer-based reinforcement learning for robot crowd navigation," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2023.

[14] W. Wang, R. Wang, L. Mao, and B.-C.

Min, "Navistar: Socially aware robot navigation with hybrid spatio-temporal graph transformer and preference learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 11 348–11 355.

[15] J. A. Ansari, S. Tourani, G. Kumar, and B. Bhowmick, "Exploring social motion latent space and human awareness for effective robot navigation in crowded environments," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 1–8.

[16] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," Microsoft, Tech. Rep. MSR-TR-2023-8, February 2023.

[17] P. Chang, S. Liu, and K. Driggs-Campbell, "Learning visual-audio representations for voice-controlled robots," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9508–9514.

[18] P. Chang, S. Liu, T. Ji, N. Chakraborty, K. Hong, and K. R. Driggs-Campbell, "A data-efficient visual-audio representation with intuitive fine-tuning for voice-controlled robots," in *Conference on Robot Learning (CoRL)*, 2023.

[19] P. Chang, S. Liu, H. Chen, and K. 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, pp. 5580–5587.

[20] S. Liu, A. Hasan, K. Hong, R. Wang, P. Chang, Z. Mizrachi, J. Lin, D. L. McPherson, W. A. Rogers, and K. Driggs-Campbell, "Dragon: A dialogue-based robot for assistive navigation with visual language grounding," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3712–3719, 2024.