A Framework for Inferring Belief States in Partially-Observable Human-Robot Teams

Jack Kolb and Karen M. Feigh

Abstract— **We propose a framework for robots to estimate the belief state of a human teammate in real-time in 3D partially-observable domains. Recent research has shown potential for robots to use a predicted belief state – or team mental model – to inform downstream planning and coordination tasks in human-robot teams. However, no works have applied mental models to realistic domains where the robot operates in 3D space and does not have perfect observability over the environment. Our framework leverages recent advancements in scene graph construction towards addressing the open problem of estimating a human user's situation awareness. We aim to evaluate the framework on a 3D simulation platform and a real-world robot.**

I. Introduction

The cognitive engineering community holds that people internally represent their environments via a structure termed a mental model [11]. The mental model contains all task-relevant information the person is aware of, such as the locations of environment objects, their higher-level semantic relationships, and the current task goals [2]. In human-human teams, we estimate the mental model of our teammates to inform our own actions, planning, and communication. Practically, this team mental model facilitates important aspects of team dynamics including theory of mind, implied objectives, non-verbal coordination, and selective communication [8], [9], [14].

There is significant research interest in enabling robots to apply mental models to human-robot teams. Central to the mental model is a belief state, the core knowledge representation of the world state, which is analogous to level 1 situation awareness. Researchers have taken two approaches to modeling a user teammate's belief states – latent representations through end-to-end deep learning architectures that embed the belief state in a task-oriented process, and explicit representations that aim to precisely model the user's awareness of environment objects [1], [6], [13].

The two representation types present a tradeoff. Implicit representations are tailored to specific inference goals and can model abstract user preferences. However, the end-to-end training makes it difficult to use the embedded belief state for applications outside of the model's intended application. Alternatively, explicit representations can be used to inform a range of downstream inference and reasoning tasks, however are challenging to define, model, and reliably construct. While the literature has primarily found success in using implicit representations [3], recent research has worked towards defining and verifying frameworks for explicit representations [1], [4], [12].

We propose a system for robots to leverage inferred dynamic scene graphs as a knowledge representation for a team mental model. Fig. 1 shows the system architecture and target evaluation domains.

II. Methods

We represent a belief state β as a *dynamic scene graph* $\mathcal{G}$ containing a set of nodes $\mathcal{N}$ and edges $\mathcal{E}$. $\mathcal{G}$ is a directed acyclic graph with three node layers, corresponding to *buildings*, *rooms*, and *objects*. Edges link objects to rooms and rooms to buildings. Object nodes have properties associated with their recognition by the robot's perception system, including point cloud information, object class, and last-known pose. The graph is continuously updated as the robot navigates a scene, by resolving conflicts in object permanence and maintaining an up-to-date model of the environment. The robot's sensors produce visual, depth, and pose observations $\mathcal{O}^{robot}$ that is used to maintain the robot's scene graph β^{robot}.

The team mental model uses two dynamic scene graphs – the robot's graph representing the true state of the environment from the robot's perspective (β^{robot}), and the teammate's graph that represents what the robot thinks the user teammate is aware of (β^{pred}). The robot constructs the teammate's graph by projecting the portions of the robot's graph that the robot believes the user has observed.

Predicted observations $\mathcal{O}^{pred}$ use two mechanisms:

1) When directly observing the teammate (e.g., the robot is facing the teammate), the robot uses the teammate's pose and field of view to identify the subset of the robot's scene graph used to update the teammate's graph.
2) When observing the teammate between two points (e.g., the robot last saw the teammate in another room), the robot estimates the path taken by the teammate to the current pose, and identifies the subset of the robot's scene graph that the teammate would have observed along that path. There are several opportunities for the robot to estimate the path, from classical A* to generative models (e.g., a GAN) trained on prior user behavior. We use A* in this initial implementation.

The resulting subgraph β^{pred} represents the robot's prediction of the user teammate's knowledge of environment objects.

This work is supported by Amazon Consumer Robotics.

Jack Kolb and Karen M. Feigh are with the Institute of Robotics and Intelligent Machines, Georgia Institute of Technology, North Avenue, Atlanta 30332, USA {kolb, karen.feigh}@gatech.edu

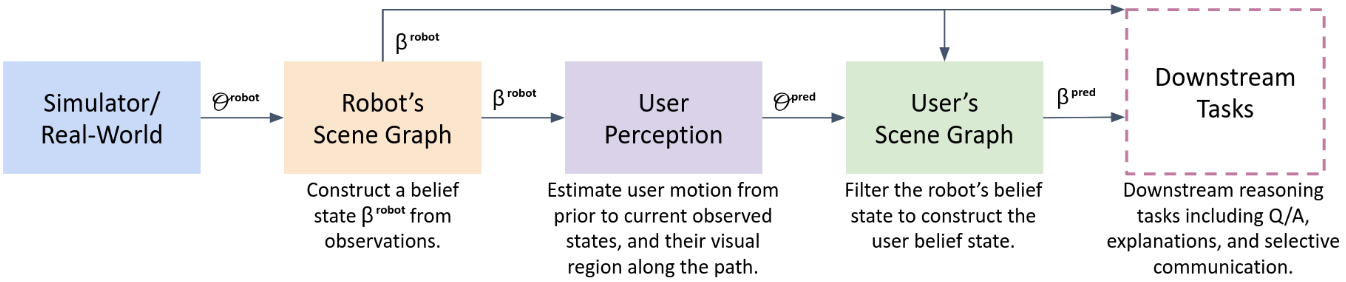
Simulator/ Real-World
𝒪 robot
Robot's Scene Graph
β robot
β robot
User Perception
𝒪 pred
User's Scene Graph
β pred
Downstream Tasks
Construct a belief state β robot from observations.
Estimate user motion from prior to current observed states, and their visual region along the path.
Filter the robot's belief state to construct the user belief state.
Downstream reasoning tasks including Q/A, explanations, and selective communication.

Fig. 1. Architecture of the proposed framework. A simulator or real-world robot obtains robot-centric camera, depth, and pose feeds, $\mathcal{O}^{robot}$. A scene graph constructor creates a belief state (dynamic scene graph) from the observations, β^{robot}. The belief state is then filtered by predicted observations of the human teammate, $\mathcal{O}^{pred}$, which informs the teammate's predicted belief state β^{pred}.

III. Implementation

We are implementing the framework in two domains, a 3D simulation platform and a real-world robot. While a real-world domain more accurately represents conditions that robots with team mental capabilities will be deployed in, the simulation environment enables reliable testing and extensive data collection. We are using VirtualHome [10] as our simulation platform, which was chosen for its ease-of-development, built-in navigation stack, and extensive library of common household items. To keep the domain general-purpose across a range of robot platforms, the framework only requires two inputs: an RGBD camera stream, and the robot's pose.

To evaluate our framework we implemented a "clean up" task where objects in the household are randomly scattered from their known initial positions. The human and robot agents must independently rearrange the objects back. While many household tasks only require a single vantage point to capture the majority of the scene (e.g., collaborative cooking mostly occurs within a small area), *clean-up* is constrained by partial observability as the teammates frequently move in-and-out of sight. This characteristic encourages a strong scene graph construction and presents opportunities for false beliefs about the environment.

To construct the dynamic scene graph we are comparing two scene graph generators: an off-the-shelf scene graph generator (Hydra [7], with HRNet [15] as a semantic segmentation network), and a custom generator that uses Detectron2 [16] for image segmentation. The resulting scene graph is the robot's belief state, which is then processed using estimated observations of the other agent to obtain the predicted teammate's belief state. The system outputs the two scene graphs for downstream use.

IV. Conclusion

If successful, this framework will have broad utility for several areas of human-robot interaction and human-AI teaming – enabling robots to estimate the situation awareness of user teammates, and directly inform theory of mind inference tasks.

The human factors community defines situation awareness as a task-oriented subsection of a mental model [5]. In this model, situation awareness has three levels – an awareness of the raw classes and locations of environment elements, an awareness of the present contextual meaning of those elements, and an awareness of how the environment is likely to change in the near-future. The constructed dynamic scene graph is a prediction of the user's current belief state, which is analogous to a prediction of the user's level 1 situation awareness. Downstream work can leverage this scene graph to expand the scope of the robot's predicted teammate mental model, by predicting contextual relationships between elements or using the graph's history to inform future plans [13].

Additionally, the dynamic scene graph dyad – the robot's belief and the predicted teammate's belief – presents an interesting opportunity for inference tasks reliant on a theory of mind. A classic example of theory of mind is the false belief task, where a robot observer predicts whether a user is aware of an important environment feature that recently changed. In this framework, downstream work can use the overlaps and disparities between the two belief states to identify potential false beliefs of the user teammate, note areas of the environment where the user may know more than the robot, and infer additional information about the user such as their goals, intentions, and rationales.

References

[1] Matthew L Bolton, Elliot Biltekoff, and Kevin Byrne. Fuzzy mental model finite state machines: A mental modeling formalism for assessing mode confusion and human-machine "trust". In *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*, pages 1–4. IEEE, 2022.

[2] Janis A Cannon-Bowers, Eduardo Salas, and Sharolyn Converse. Shared mental models in expert team decision making. *Current issues in individual and group decision making. Lawrence Erlbaum, Hillsdale, NJ*, pages 221–246, 1993.

[3] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

[4] Gwendolyn Edgar, Matthew McWilliams, and Matthias Scheutz. Improving human-robot team performance with proactivity and shared mental models. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2322–2324, 2023.

[5] Mica R Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society annual meeting*, volume 32, pages 97–101. Sage Publications Sage CA: Los Angeles, CA, 1988.

[6] Felix Gervits, Terry W Fong, and Matthias Scheutz. Shared mental models to support distributed human-robot teaming in space. In *2018 aiaa space and astronautics forum and exposition*, page 5340, 2018.

[7] Nathan Hughes, Yun Chang, and Luca Carlone. Hydra: A real-time spatial perception system for 3d scene graph construction and optimization. *arXiv preprint arXiv:2201.13360*, 2022.

[8] Janice Langan-Fox, Sharon Code, and Kim Langfield-Smith. Team mental models: Techniques, methods, and analytic approaches. *Human factors*, 42(2):242–271, 2000.

[9] Beng-Chong Lim and Katherine J Klein. Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 27(4):403–418, 2006.

[10] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8494–8502, 2018.

[11] William B Rouse, Janis A Cannon-Bowers, and Eduardo Salas. The role of mental models in team performance in complex systems. *IEEE transactions on systems, man, and cybernetics*, 22(6):1296–1308, 1992.

[12] Matthias Scheutz, Scott A DeLoach, and Julie A Adams. A framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making*, 11(3):203–224, 2017.

[13] David Schuster, Scott Ososky, Florian Jentsch, Elizabeth Phillips, Christian Lebiere, and William A Evans. A research approach to shared mental models and situation assessment in future robot teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 55, pages 456–460. SAGE Publications Sage CA: Los Angeles, CA, 2011.

[14] Aaquib Tabrez, Matthew B Luebbers, and Bradley Hayes. A survey of mental modeling techniques in human–robot teaming. *Current Robotics Reports*, 1:259–267, 2020.

[15] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020.

[16] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.