Designing for Long Horizon Social Interactions

Ifrah Idrees, Stefanie Tellex

Brown University, RI, USA

Abstract

Conversational assistive robots have the potential to guide humans in accomplishing a wide range of daily tasks, such as cooking meals, performing exercises, and operating machines, etc. However, for effective interaction, the robot must be capable of deducing human's goal from their interactions with the surrounding environment. While inferring the world and user state, the robot has limited capability to perceive human intentions, a challenge further complicated by noisy sensors that contribute to partial observability of the environment. This problem compounds as the robot accumulates more noisy observations about the user and environment over extended durations. In this paper, we will elaborate on our research to enable the robot to engage in accurate real-time inference and modeling of both the world and the user's state, especially over extended periods. Subsequently, we will delve into the prospective paths for addressing uncertainty in extended social interactions within the framework of cultivating an anticipatory dialogue about the future at the crossroads of AI and HRI. Our work equips home robots with enhanced situational awareness for long-term social interactions.

Introduction

Conversational assistive robots can aid people to accomplish various tasks such as cooking meals, performing exercises, or operating machines. Imagine a scenario where the human mistakenly performs sub-optimal actions while cooking dinner, e.g., forgetting to turn off the stove. In such a situation, an assistive robot must be able to suggest corrective next steps based on its understanding of the world and the user (Erol, Hendler, and Nau 1994; Wang and Hoey 2017). An assistive robot guide to turning off the stove will be effective in the given case. To interact effectively, the robot must infer the human's goal and current step in the activity given the observations of the state of the appliances, e.g. (the microwave is off, the stove is on) and the state of the attributes of objects (e.g., dishes are dirty). Such an assistive robot will benefit people with dementia or cognitive impairment. It can also be helpful to an operator trying to build a machine, a child with autism trying to do their homework, or a child learning to do a chore.

The problem of inferring the world and user state is challenging because the sensors are noisy, and the robot has partial observability of the environment and the human's intention. The uncertainty in state estimation compounds as the robot collects more noisy observations about the user and environment over long periods of time (Adu-Bredu et al. 2021). Existing approaches do model the user and the world state but do not handle uncertainty from both the world sensors' observation and human language for state estimation over extended periods. Our proposed work aims to introduce methods to manage these various sources of uncertainty. Our work aims to enable the robot to perform accurate and time-efficient online inference of the world and user state in a partially observable environment, especially over extended periods.

Handling Different Sources of Uncertainty

Long-term social interactions require an estimation of the world and user state given a sequence of observations of human's interaction with the environment. In this section, we will describe the different sources of uncertainty for estimating the world state, the user state for long-horizon assistive robotics. We will also outline our related research fronts for handling uncertainty in extended social interactions, contextualized within the objective of cultivating a forward-thinking dialogue about the future at the confluence of AI and HRI.

Uncertainty in the World State

Home-service robots have a great potential to assist human users by retrieving spatial-temporal [1] information about the objects in the environment from their long-term observations. Imagine a home robot monitoring the environment over long periods of time. Such a robot will have a massive amount of observations of the objects in the environment. The human user can then ask the home robot to assist them in finding objects in the environment by asking a simple query such as *"What are the favorite places in the house where I can place my keys?"* To answer such a simple query, the service robots must have a situational understanding of

[1] Spatial-temporal information of object refers to the whereabouts of the object such as where the object has been identified in the physical environment of the robot and at what times

the environment over extended periods, not just for days but weeks and months. This requires the robot to estimate the state of objects in the environment over an extended time in a partially observable environment. Service robots with such an ability will be well-suited to help the elderly, especially those with dementia.

The previous approaches for long-term object state estimation and retrieval either assume **1.** static scenes - CLIP (Radford et al. 2021; Caron et al. 2021; Huang et al. 2022; Shah et al. 2023; Gadre et al. 2022; Jatavallabhula et al. 2023), **2.** losed set of concepts (Kang, Bailis, and Zaharia 2019; Qi et al. 2020; Li and Belaroussi 2016; Sünderhauf et al. 2017), or **3.** short-time horizon for object search (Ambruş et al. 2014; Bore, Jensfelt, and Folkesson 2015). For long-term object retrieval, the above assumptions leave a partially-observable robot searching over countless detections of objects in visual sensor data from many different time slices. These detections will contain partial views [2] of different object instances. Further, storing all these partial-view detections will take up a lot of memory space and time. Existing approaches will search all of these partial views of objects even when they are not irrelevant to the query.

We propose a **D**etection-based **3**-level hierarchical **A**ssociation approach, **D3A** that allows for a compact and query-able spatial-temporal state estimation representation (Idrees et al. 2021; Idrees, Reiss, and Tellex 2020). The robot is equipped with a map of the environment, and at any given time step i can gather the following sensory information $s_i =< p_{robot,i}, d_i, f_i >$, where $p_{robot,i} \in P_{robot}$ is the associated robot's pose, and $d_i \in D$ is the corresponding depth information for image frame f_i. The robot collects large amounts of sensor data $S = \{s_1, s_2, ..., s_i\}$ and uses an object detector to extract the relevant object-centric information a_i from every frame f_i. Our algorithm then performs a three-tier online incremental learning to associate objects over time by identifying keyframes that best represent the unique objects in the environment. Our spatial-temporal representation of objects in the environment enables answering of queries for object retrieval from long-term observations.

D3A demonstrates high accuracy and time efficiency when queried. For data collection, we allow mobile robot Kuri Mayfield (Robotics 2018) to patrol our uncontrolled and cluttered robotics lab environment, including the kitchen area and a general sitting area for four days. The cleaned and annotated dataset amounted for 22 hours. For consistency, we kept the illumination same (well-lit lab). We show that our queryable semantic scene representation for the clean dataset occupies only 0.17% of the total sensory data. We also discuss the retrieval performance of our system with a parameterized synthetic embedding detector. When D3A is queried for 59 ground truth objects, the ground truth object instance is found on average in the 5th return frame, while for baseline, the ground truth object can be found in the 20th frame. For effective long-term social interactions, the robot must handle uncertainty in both the world and the user state over long-horizon tasks. This brings us to user state inference discussed in the next section.

Uncertainty in the User State

As the human interacts with its surrounding environment, even in an environment where the robot has maximal sensor information, like a smart home, the robot still needs to figure out what it is human doing. To be an effective assistant, the robot must interpret the goal the human is trying to achieve and the action the human should take to complete their plan. Our approach infers this based on their interaction based on their interaction with the environment.

Previously, human progress during hierarchical tasks has been modeled using hierarchical task networks (HTNs) (Wang and Hoey 2017; Höller et al. 2018). However, these plan/goal recognition techniques do not allow the agent to leverage its ability to use language to reduce uncertainty by asking questions. Further, interpreting language input from the human is challenging because of the vast space of observations — language utterances spoken by the human. The existing solution to this problem is heuristics (Razzaq, Khan, and Lee 2017; Fasola and Matarić 2013; Kidd and Breazeal 2008), which are prone to fail as the tasks get complex and the environment sensors become complex or noisy.

We propose to solve this problem by combining the Hierarchical Task Networks (HTNs) with Partially Observable Markov Decision Process (POMDP) in our - **D**ialogue for **G**oal **R**ecognition (**D4GR**) (prounced Dagger). This is particularly challenging because the POMDP does not assume a hierarchical human mental model or task specification structure (Goldman 2021). Further, the state and observation spaces are large for modeling users and the world. For our work, we assume the person is a planner with hierarchically-described goals and subgoals. The agent is represented by a POMDP model updating its belief in human progress by asking clarification questions about noisy sensor data. An illustration of our approach's architecture and a sample example of a clarification question is shown in Fig-1.

We evaluate the performance of D4GR over various cooking tasks and blocks domain for stacking letters to make words in a simulated environment introduced by (Wang and Hoey 2017) and perform a robot demonstration. With language feedback and the world state information in a hierarchical task model, we show that our D4GR formulation has a similar goal and plan recognition accuracy as the ORACLE baseline ALWAYS-ASK method, that always ask the correct question, while asking 68% fewer questions. D4GR framework for the highest sensor noise performs 1% better than HTN in goal accuracy. The ALWAYS-ASK oracle outperforms our policy by 3% in goal recognition and 7% in plan recognition. Our framework validates that incorporating language feedback and modeling the user as a Hierarchical Task Network improves goal and plan recognition. Our approach does so by enhancing the accuracy of robots' belief of human progress and demonstrating the effectiveness of POMDPs for tracking task progress and user engagement. Our work shows promising improvements in human intent recognition and open venues for social roboticists to improve human-robot interaction. For further enhanced interactions, the robot must handle the diversity in human be-

[2] Partial views of objects refers to the partially occluded objects seen from different viewpoints by the robot

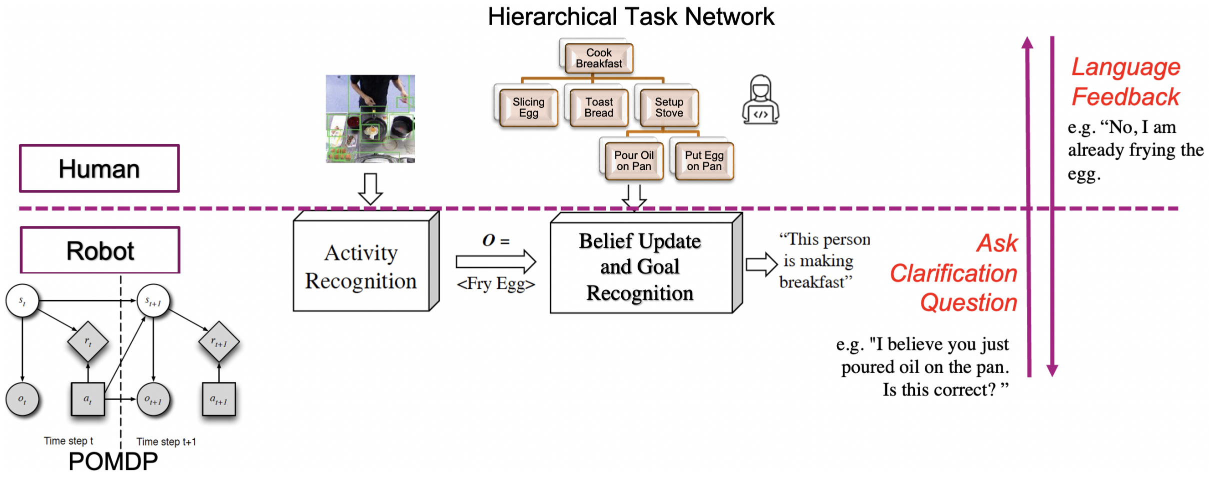
Hierarchical Task Network
Cook Breakfast
Slicing Egg
Toast Bread
Setup Stove
Pour Oil on Pan
Put Egg on Pan
Human
Robot
Activity Recognition
O =
<Fry Egg>
Belief Update and Goal Recognition
"This person is making breakfast"
Language Feedback
e.g. "No, I am already frying the egg.
Ask Clarification Question
e.g. "I believe you just poured oil on the pan. Is this correct? "
s_t
s_{t+1}
r_t
r_{t+1}
o_t
a_t
o_{t+1}
a_{t+1}
Time step t
Time step t+1
POMDP

Figure 1: Architecture of - **D**ialogue for **G**oal **R**ecognition (**D4GR**)

haviors over long periods of time. Our next section focuses on a related line of work that handles diversity in human behavior.

Variability in Human Behavior

The development of a commercial robot to coexist with people over the long term in a home environment is a challenging task. For robot behaviors that depend on spatial interaction with users, prediction of their locations, or long-term adaptation to user behavior, it can be difficult to evaluate performance effectively through on-device testing alone. Hence, simulation provides a scalable solution for testing production code.

We identify five essential prerequisites for achieving realistic simulations of human activities in the context of testing human-robot interactions. We focus on only spatial interactions between humans and robots for our proposed work. For these interactions, the identified requirements are as follows:

1. Manual control over simulated behaviors
2. Execution on a variety of different floorplans
3. Configurability for different personas or lifestyles
4. Long-term activity modeling
5. 3D human motion generation

We propose a novel framework for simulating daily user activity at scale for testing and developing commercial robots, enabling testing of robot behaviors at scale in a simulation environment. The framework enables manual daily schedule tuning for developing typical-use and corner-case tests for testing and quality assurance. Our case studies demonstrate the framework's expressive capability, and we also evaluate the approach's potential to emulate example data using both public and internally-captured datasets. In all cases, the similarity between the generated data and reference data was much higher than the random baseline, and comparable to the self-similarity within the reference set itself.

Overall, this work makes a significant contribution to the field of social robotics by providing a systematic approach for testing robot behaviors related to daily user activity. By helping to avoid training data bias, our approach has the potential to make robot behavior effective for a broad range of user households, making it a valuable tool for future research and development in AI-HRI and specifically social robotics.

AI-HRI Technical Bridge

As robotics systems engage in complex tasks across extended timeframes, effectively navigating uncertainties becomes increasingly critical. Innovations in the subfields of the AI-HRI community e.g., robot perception, planning, and decision-making algorithms, will allow robots to dynamically assess and adapt to changing environments, unforeseen events, and evolving task requirements. Within AI-HRI's long-horizon robotics community, we believe that two crucial aspects stand out: the need for generalized representations of the world and the user state, and the utilization of abstract spatial-temporal structures to enhance the modeling of these elements, ultimately leading to improved situational awareness for robots engaged in extended tasks. Based on this, our work proposes having generalized representations of the world and the user state is essential. Long-horizon robotic tasks often span diverse environments and involve interactions with various users or entities. To effectively handle these scenarios, robots must possess versatile and adaptable models of the world they operate in and accurate assessments of the states, intentions, and preferences of the users they interact with. Another proposal is to adopt spatial-temporal structures in modeling the world and the user state, which is a powerful approach. Such structures allow the robot to capture multi-level abstractions of the environment and user interactions. Spatial-Temporal representations enable the robot to discern global and local patterns, temporal dependencies, and spatial relationships within its surroundings. This abstraction approach helps the robot organize and process information more efficiently, leading to a more comprehensive understanding of its environment and the users it interacts with. By leveraging these structures, the robot can detect high-level context changes, infer complex user behaviors, and make informed decisions based on richer cues, leading to improved situational awareness.

Together, combining generalized representations and utilizing spatial-temporal structures in modeling the world and user state empowers robots to excel in long-horizon tasks. This approach equips robots with the ability to handle uncertainty, and engage in prolonged interactions while maintaining high situational awareness, making them valuable and effective partners in various real-world applications.

Conclusion

Our proposed lines of research aim to test the hypothesis that using spatial-temporal structures to model the world and user state in long-term social interactions can enable the robot to have a more accurate and time-efficient online state inference. We outline the prospective paths for handling uncertainty in extended social interactions, contextualized to cultivate a forward-thinking dialogue about the future at the confluence of AI and HRI. Our proposed direction aim to improve the robot's user experience over extended periods by performing better state estimation for long-term social interactions. In our work, we describe three sources of uncertainty - world state, user state, and variation in human behavior. First, we propose a novel algorithm that performs efficient world state estimation over long periods while handling uncertainty due to noisy sensors and partial observability. Next, we present a novel formulation for accurately inferring the latent variable of the user's goals and plans from noisy world sensors and language feedback. Finally, we outline a framework for generating scalable, configurable, and variable schedules for daily human activity, enabling testing of robot behaviors at scale in a simulation environment. Our proposed prospective paths aim to enable generalized goal inference of human progress during long-horizon task completion. Our introduced methods help manage uncertainty in the world, user, and task specifications. A home robot with enhanced state estimation capabilities can better assist humans during task completion, improving long-term social interactions. Hence, our line of work opens venues in for enhanced human-robot interaction in the AI-HRI community. Our research takes a step towards seamlessly integration of robots into users daily lives while provides valuable social support during various tasks and activities. Hence, our line of work opens venues for enhanced human-robot interaction in the AI-HRI community.

References

Adu-Bredu, A.; Devraj, N.; Lin, P.-H.; Zeng, Z.; and Jenkins, O. C. 2021. Probabilistic Inference in Planning for Partially Observable Long Horizon Problems. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3154–3161. IEEE.

Ambruş, R.; Bore, N.; Folkesson, J.; and Jensfelt, P. 2014. Meta-rooms: Building and maintaining long term spatial models in a dynamic world. In *2014 IEEE/RSJ IROS*, 1854–1861. IEEE.

Bore, N.; Jensfelt, P.; and Folkesson, J. 2015. Retrieval of arbitrary 3D objects from robot observations. In *2015 European Conference on Mobile Robots (ECMR)*, 1–8. IEEE.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Erol, K.; Hendler, J.; and Nau, D. S. 1994. HTN planning: Complexity and expressivity. In *AAAI*, volume 94, 1123–1128.

Fasola, J.; and Matarić, M. J. 2013. A socially assistive robot exercise coach for the elderly. *Journal of Human-Robot Interaction*, 2(2): 3–32.

Gadre, S. Y.; Wortsman, M.; Ilharco, G.; Schmidt, L.; and Song, S. 2022. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*.

Goldman, R. P. 2021. Solving POMDPs online through HTN Planning and Monte Carlo Tree Search. *HPlan 2021*, 57.

Höller, D.; Behnke, G.; Bercher, P.; and Biundo, S. 2018. Plan and goal recognition as HTN planning. In *2018 IEEE 30th International Conference on Tools with Artificial Intelligence (ICTAI)*, 466–473. IEEE.

Huang, C.; Mees, O.; Zeng, A.; and Burgard, W. 2022. Visual Language Maps for Robot Navigation. *arXiv preprint arXiv:2210.05714*.

Idrees, I.; Hasan, Z.; Reiss, S. P.; and Tellex, S. 2021. Where were my keys? - Aggregating Spatial-Temporal Instances of Objects for Efficient Retrieval over Long Periods of Time. *CoRR*, abs/2110.13061.

Idrees, I.; Reiss, S. P.; and Tellex, S. 2020. RoboMem: Giving Long Term Memory to Robots. *CoRR*, abs/2003.10553.

Jatavallabhula, K. M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Li, S.; Iyer, G.; Saryazdi, S.; Keetha, N.; Tewari, A.; et al. 2023. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*.

Kang, D.; Bailis, P.; and Zaharia, M. 2019. Challenges and Opportunities in DNN-Based Video Analytics: A Demonstration of the BlazeIt Video Query Engine. In *CIDR*.

Kidd, C. D.; and Breazeal, C. 2008. Robots at home: Understanding long-term human-robot interaction. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3230–3235. IEEE.

Li, X.; and Belaroussi, R. 2016. Semi-dense 3d semantic mapping from monocular slam. arXiv. *arXiv preprint arXiv:1611.04144*.

Qi, X.; Wang, W.; Yuan, M.; Wang, Y.; Li, M.; Xue, L.; and Sun, Y. 2020. Building semantic grid maps for domestic robot navigation. *International Journal of Advanced Robotic Systems*, 17(1): 1729881419900066.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Razzaq, M. A.; Khan, W. A.; and Lee, S. 2017. Intent-Context Fusioning in Healthcare Dialogue-Based Systems Using JDL Model. In *International Conference on Smart Homes and Health Telematics*, 61–72. Springer.

Robotics, M. 2018. Meet Kuri! The Adorable Home Robot.

Shah, D.; Osiński, B.; Levine, S.; et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, 492–504. PMLR.

Sünderhauf, N.; Pham, T. T.; Latif, Y.; Milford, M.; and Reid, I. 2017. Meaningful maps with object-oriented semantic mapping. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 5079–5085. IEEE.

Wang, D.; and Hoey, J. 2017. Hierarchical Task Recognition and Planning in Smart Homes with Partial Observability. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, 439–452. Springer.