A Multimodal Human-Robot Interaction Manager for Assistive Robots

Bahareh Abbasi[1*], Natawut Monaikul[2*], Zhanibek Rysbek[1], Barbara Di Eugenio[2], and Miloš Žefran[1]

***Abstract*— With rapid advances in social robotics, humanoids and autonomy, robot assistants appear to be within reach. However, robots are still unable to effectively interact with humans in activities of daily living. One of the challenges is in the frequent use of multiple communication modalities when humans engage in collaborative activities. In this paper, we propose a Multimodal Interaction Manager, a framework for an assistive robot that maintains an active multimodal interaction with a human partner while performing physical collaborative tasks. The heart of our framework is a Hierarchical Bipartite Action-Transition Network (HBATN), which allows the robot to infer the state of the task and the dialogue given spoken utterances and observed pointing gestures from a human partner, and to plan its next actions. Finally, we implemented this framework on a robot to provide preliminary evidence that the robot can successfully participate in a task-oriented multimodal interaction.**

I. Introduction

The success of human-robot interactions in service and healthcare robotics depends on the understanding between the human and the robot during activities of daily living (ADLs). The main challenge in these applications is that the interaction is inherently multimodal [1], and employs a back-and-forth emblematic of human conversation, which includes additional modalities such as force exchanges or pointing gestures besides speech. While the role of each modality such as speech and gestures in the interaction is well established, a planning and execution framework to manage multiple modalities remains a challenge.

For a robot to interact with a human through language and physical actions, these modalities need to be processed in a unified manner, despite their different levels of abstraction. Similarly, the robot must be able to generate responses in the form of an utterance, a physical action, or a combination of both. Furthermore, the robot should have some knowledge of the task at hand in order to help the user navigate through the task successfully.

*First two authors contributed equally to this work.

[1]B. Abbasi, Z. Rysbek, and Miloš Žefran are with the Robotics Lab, Electrical and Computer Engineering Department, University of Illinois at Chicago, Chicago, IL 60607 USA.

[2]N. Monaikul and B. Di Eugenio are with the Natural Language Processing Lab, Computer Science Department, University of Illinois at Chicago, Chicago, IL 60607 USA.

This work has been supported by the National Science Foundation grants IIS-1705058 and CMMI-1762924.

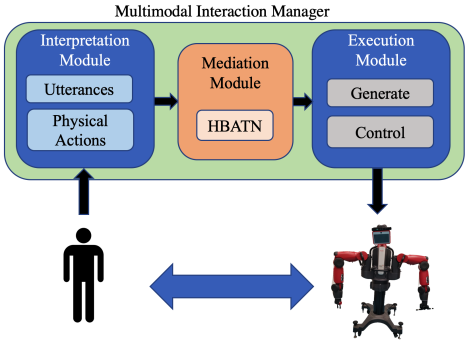

Fig. 1: The architecture for multimodal human-robot interaction that governs the robot's language/physical responses.

In this work, we propose a framework – a Multimodal Interaction Manager – for an assistive robot that bridges the gap between the symbolic processing of language and the low-level control of physical actions. As shown in Fig. 1, our framework is composed of three components. The Interpretation Module takes the observations perceived by the sensors, including spoken utterances and physical actions, and generates their symbolic interpretation, which consists of a dialogue act [2] (roughly, the intention of the speaker), non-verbal communication type, and the objects they refer to. Then, the Mediation Module, which contains an integrated dialogue and task model, establishes the state of the interaction given the interpreted multimodal user inputs. Once these states have been identified, the model is used to find the desired robot response, which is again a combination of language and physical actions. Finally, the desired robot response is communicated to the Execution Module to be realized.

Conceptually, our main contributions are (1) a framework for multimodal human-robot interaction that interprets and performs actions in a collaborative task; and (2) a Hierarchical Bipartite Action-Transition Network (HBATN) that models both agents simultaneously to maintain the state of a task-driven multimodal interaction and plan subsequent moves.

Our particular focus is an interaction scenario in which a human and a robot work together to find an object in the environment, what we call the *Find* task. Another important contribution of this work is an implementation of the proposed Multimodal Interaction Manager in the Robot Operating System (ROS) and a preliminary user study that shows that such a robot can engage with a

human in a multimodal exchange and help complete the *Find* task with a high success rate. Although there have been similar implementations on social robots [3], [4] and in virtual agents [5], [6], the innovation of our work is in implementing the full cycle in Fig. 1 in a robot to not only interact with a user via speech and pointing gestures, but to also manipulate its environment to assist the user in completing a given task.

The rest of the paper is organized as follows: in Sec. II, we review the literature on modeling multimodal human-robot interactions; in Sec. III, we discuss the corpus that provides a basis for our work and our annotation process; the Multimodal Interaction Manager and its components are described in Sec. IV; we present results of a preliminary evaluation of our framework in Sec. V; finally, concluding remarks and future work are provided in Sec. VI.

II. Related Work

Our defined *Find* task involves an interaction in which the collaborators must understand what the other is saying and where they are pointing, as well as manipulate surrounding objects, to search for a target object and respond appropriately. The challenge in implementing this in a robot is to develop a unified framework integrating both speech and physical actions, while also maintaining a natural interaction that moves towards completing the task.

The work that is most related to our research are autonomous agents (both physical and virtual) that support language and physical actions. Researchers in multimodal human-robot interaction have explored techniques for interpreting multimodal input in dialogic interactions [1] and generating non-verbal behaviors (such as eye gaze, pointing, and iconic gestures) that accompany speech to mimic human-like communication [7]. Studies have shown that implementing these co-verbal behaviors in robots indeed affect users' perceptions, promoting trust, likeability, and engagement [3], [7]–[10]. Assistive robots that are capable of both interpretation and generation have also been proposed, such as in social robotics [4], [11], [12], or in personal mobile robot agents [13]. These robots, however, do not manipulate their environments directly with physical actions.

Assistive robots generally perform or aid in the completion of some kind of task, and various models have been explored to represent and learn these tasks. A common approach is to decompose a high-level task as a set of low-level subtasks that must be completed in some fashion to achieve the high-level goal. This hierarchical representation has been used for learning task-oriented dialogue policies [14], as well as how to perform physical actions from demonstrations using Hierarchical Hidden Markov Model (HHMM) [15]. Another common formalism for representing tasks is a Hierarchical Task Network (HTN) [16], [17]. For example, [18], [19] have used HTNs to enable a robot to learn the subtasks required to complete the higher-level tasks through demonstration and interactive input. HTNs have also been used to encode the higher-order structure of dialogic interactions in the Disco dialogue manager [20], [21].

Move	Actor	Utterance	DA	Pointing	H-O
1	ELD	Can I get a tray?	Instruct		
2	HEL	Where are the trays?	Query-w		
3	ELD	Try the bottom shelf.	Instruct	cab6	
4	HEL	--	--		open(cab6)
5	HEL	Is this the tray?	Query-yn		hold(tray1)
6	ELD	Yes, that's it.	Reply-y		

Fig. 2: A sample annotated interaction from the *Find* subcorpus as a sequence of moves.

Importantly, HTNs have been used to integrate dialogue and physical actions into a single framework for multimodal tasks by treating both utterances and physical actions as units that can make up subtasks of a higher-level task. Many implementations of this framework in robots primarily focus on actions such as gaze and pointing that do not manipulate the environment directly [3], [4]. Other implementations can engage in dialogue [22] or analyze natural language input [23], [24] to learn to perform such physical actions; however, the focus is on learning the mapping from natural language commands to physical actions, and not necessarily guiding the user through the task. There are also implementations in virtual agents that can directly manipulate their virtual environments to assist in multimodal tasks [5], [6], [25]. The novelty of our work is in the integration of dialogue and physical actions that interact with the user and the environment with the ultimate goal of completing a physical collaborative task.

III. Multimodal Interaction Corpus

An emerging application of robotics is elderly care, especially in the context of aging-in-place [26], in which a typical day consists of many ADLs that may be facilitated by robotic assistance. In this section, we give an overview of a relevant human-human multimodal interaction corpus that our framework is based on and how we further annotate it to build our interaction manager.

A. Corpus Description

The ELDERLY-AT-HOME corpus [1] is a publicly available corpus of multimodal collaborative human-human interactions between gerontological nursing students and elderly persons residing in an assisted living facility. The interactions involved performing assisted ADLs, such as putting on shoes and preparing dinner, which are generally comprised of multiple subtasks.

The corpus has been annotated for dialogue acts (DAs) such as *instruct* or *acknowledge* for each utterance (see [1] for the full list of 13 DA tags), which capture the intention of the speaker [27]; pointing gestures (when and to what someone is pointing); and haptic-ostensive (H-O)

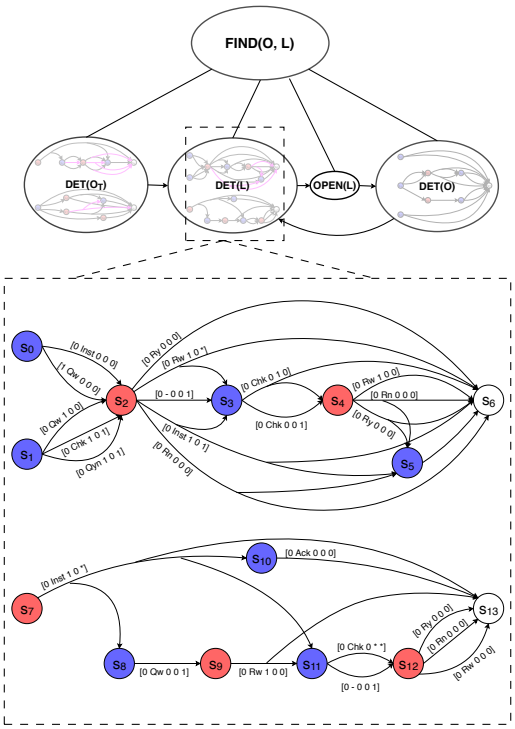

Fig. 3: The hierarchical *Find* task model decomposed into multiple subtasks. Each subtask is an AcTNet modeling both agents. (*) represents 0 or 1.

actions such as *open* and *hold*, which manipulate objects and can also often perform a referring function [28].

Furthermore, a subcorpus has been created that specifically extracts those interactions which are instances of the collaborative *Find* task, such as looking for a pot or plates (as subtasks of preparing dinner) [1], [29]. This *Find* subcorpus is composed of 137 instances (totaling about 1 hour and 24 minutes). In most of these instances, the elderly participant (ELD) would ask for an object, and the helper (HEL) would try to find it by asking follow-up questions. In many cases, the participants would perform multiple rounds of questions and answers to achieve the goal.

B. Corpus Annotation

In this work, we developed a task model of the *Find* task from the *Find* subcorpus. While the corpus is annotated on multiple levels (for utterances, pointing gestures, and H-O actions) that frequently overlap, [29] defines a *move* as any combination of the three levels that forms a unit. Because we would like to integrate language and physical actions in a single framework, we used this notion of a move to represent a multimodal action taken by a participant and transformed the multilevel-annotation corpus into a corpus of sequences of moves that represent these multimodal interactions (using the algorithm provided in [29]). An example of such an interaction is given in Fig. 2.

We carefully examined the resulting corpus to develop a task model for the *Find* task that decomposes the task into smaller but still high-level subtasks. The *Find* task can be described as retrieving an object O potentially from location L. We found that the interactions typically involved four main interconnected subtasks (see Fig. 3):

1) *Det(O_T)*: determining the object type that needs to be found.
2) *Det(L)*: determining a potential location for O.
3) *Open(L)*: opening L to look inside.
4) *Det(O)*: looking for O and confirming it with the other person if it is found.

We then annotated the corpus to label instances of each of these subtasks. We treated the annotation procedure as a data segmentation problem, where the start of any subtask can be at any move. As an example, in the sample interaction in Fig. 2, *Det(O_T)* starts at move 1, *Det(L)* starts at move 2, *Open(L)* starts at move 4, and *Det(O)* starts at move 5. Note that there are instances in which *Det(O_T)* and *Det(L)* can be accomplished in a single move (e.g., "Can you get a spoon from the drawer?"), and that a subtask can be repeated in a single trial since L may not be the location in which the object can be found. In the corpus, the participants looped through steps 2-4 twice on average since the object could not be found in the first attempt.

Two annotators manually segmented 38 trials into subtasks. To measure inter-annotator agreement on deciding where each subtask begins, 9 trials were independently segmented by each annotator prior to the 38. Computing Cohen's kappa on the 115 possible points for segmentation showed extremely high agreement ($\kappa = 0.9443$).

IV. Multimodal Interaction Manager

Our goal is to develop a multimodal interaction manager, as in Fig. 1, that allows a robot to collaborate with a human partner on a task using spoken utterances and physical actions. To this end, we designed data-driven models of each subtask, which we describe in detail in this section. using the annotated data described above

We first establish a representation for multimodal actions performed by the human that are observed and interpreted by the robot in the Interpretation Module. We define a vector containing the following features for each observed move:

1) whether or not an object type is uttered
2) the DA tag
3) whether or not a location is uttered
4) whether or not an H-O action is performed
5) whether or not a pointing gesture is performed

Each of the features is binary, except the DA tag, which is categorical. For observed moves that are only physical actions (no utterances), the DA tag feature can be empty. Then given an observed human move as a vector, the Mediation Module should be able to produce an appropriate responding move (also as a vector).

The challenge in modeling each subtask is that because there are two agents (ELD and HEL), the fact that

each agent's actions affect the other's actions must be considered. Here, we propose to model this problem using an Action-Transition Network (AcTNet), in which both agents' states and actions are incorporated in a single graph, and transitions are triggered by action vectors. Fig. 3 shows an AcTNet for *Det(L)*, where blue nodes represent states in which HEL should perform the next move, red nodes for states in which ELD should perform the next move, and white nodes for end states.

Using the segmented data in the corpus, we derived AcTNets for each subtask and counted frequencies of possible paths through the AcTNets, i.e., possible sequences of moves in each subtask. The frequencies determined the probability of taking each path, which was used to model the stochastic behavior of interactions. The Mediation Module then uses these models to navigate each subtask, deciding which action to perform next given an interpreted action. We note that the purpose of AcTNets in our framework then is to help estimate the current state of the human participant given their actions, as well as plan the next robot move from the resulting robot state.

To better illustrate this, we will walk through an example using Fig. 3. Suppose the robot (as HEL) is in state s_0. It can select one of the possible actions using the calculated probabilities from the training data. Suppose the action [1, Query-w, 0, 0, 0] is selected (the robot asks a question about where the object is located). The human is then estimated to be in state s_2, and the robot's next state depends on the vector of observations from the human. If the human answers the robot's question with an utterance containing a location name – [0, Reply-w, 1, 0, 0] – the robot chooses s_3 or s_6, again based on the probabilities from the training data.

When more than one human state is possible after a robot action, a state is chosen by comparing the observed human action with all of the possible human actions from the set of estimated states. The state with the closest matching action (defined by matching feature values, with preference given to matching DA tags) is selected as the most likely human state. If no action is reasonably similar to the observed action (here, defined as no matching DA tags), then the robot asks the user to repeat their action, after which another failure would result in stochastically choosing the human state.

When the end of a subtask is reached, the next state depends on the next subtask, which depends on whether or not the subtask was successfully completed, e.g., whether or not the *Det(L)* interaction ended with a location determined. If the subtask is successfully completed, then the next subtask is given in the hierarchical task decomposition; otherwise, the subtask must be repeated. Note that *Det(O)* is completed when an object is determined to be the target object or when an object is determined to not be the target object. In the former case, the *Find* task is complete, whereas in the latter case, a new object or location must be chosen. For each subtask, either ELD or HEL can initiate the interaction. Thus, for each subtask, there are two AcTNet components.

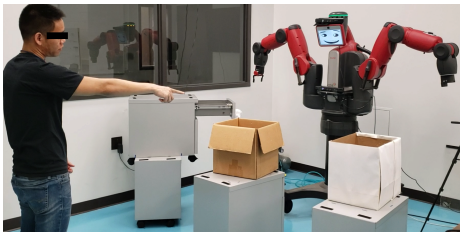

Fig. 4: A snapshot of a trial in which a human and Baxter interact to find a cup.

Because the *Find* task is decomposed hierarchically into subtasks, and each subtask is modeled by an AcTNet, we consider our Mediation Module a Hierarchical Bipartite Action-Transition Network (HBATN). As desired, this unified framework integrates physical actions (both pointing and H-O) and language while maintaining the state of the interaction in order to move towards completing the goal and provide appropriate responses. Another advantage of this framework is in its flexibility: once the model is generated, the robot can take on any role and interact with the human accordingly.

We also note that the proposed architecture is general enough that it can be applied to other task-oriented multimodal human-robot interactions, such as collaborative manipulation [30] and object handover [31], [32], by decomposing the tasks analogously and defining appropriate action vectors. While the topology of each AcTNet was derived from the data by human inspection for expediency in this work, automating this process using well-established techniques similar to those used for learning Markov models [33] will be the subject of future work.

V. Experimental Evaluation

A. Experimental Setup

To be able to test our Multimodal Interaction Manager and evaluate the efficacy of the HBATN, we implemented the framework on a robot so that it could participate in the *Find* task, in which the robot acted as HEL and human participants acted as ELD. For our experiment, we used the robot Baxter from Rethink Robotics.

In our experimental setup, Baxter was equipped with a Kinect sensor mounted on its chest. In front of Baxter were three separate containers (two boxes and one drawer) in which objects could be placed out of view. We used four objects (two balls and two cups) of different colors as targets of the *Find* task. Each location contained only one object at a time. Participants stood facing Baxter and spoke into a hands-free microphone. A snapshot of this setup is given in Fig. 4. A video sample of a trial with Baxter is also included with this paper.

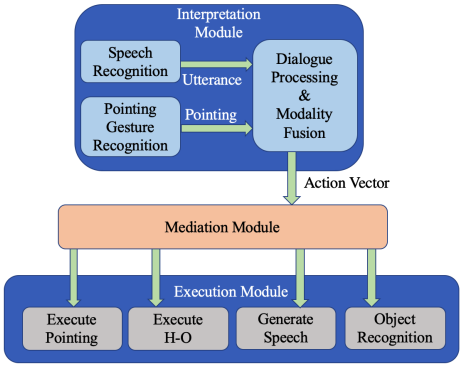

Fig. 5: The subcomponents of the implemented modules of the Multimodal Interaction Manager.

B. Implementation

To realize the full cycle of a multimodal human-robot interaction as in Fig. 1, we needed to implement modules for interpretation and execution. The Interpretation Module is responsible for processing human input perceived via multiple sensory sources and providing the observed action vectors to the Mediation Module to determine the state of the interaction and decide on the response. The Execution Module is responsible for converting an action vector from the Mediation Module that represents the next robot move into physical control of the robot and generated speech.

Fig. 5 shows the different subcomponents of our implementation of the modules. Because the participants acted as ELD, and H-O actions by ELD were very rare in the *Find* corpus, the physical actions we focus on recognizing are pointing gestures; however, the HBATN of the Mediation Module still allows for H-O actions, which we plan to explore in our future work. Thus, the Interpretation Module is composed of three primary parts: a speech-to-text component, a pointing gesture recognition component, and a component for processing the utterance and gesture together (which we call *Dialogue Processing & Modality Fusion*). We utilized the Google Cloud Speech-to-Text API for speech recognition. For pointing gesture recognition, we developed a model based on the user's skeleton information from the Kinect sensor: first, a vector connecting the user's hand and shoulder coordinates is computed; then, predefined thresholds on the vector's distance, roll angle, and pitch angle are used to determine if and to which of the three locations the user is pointing.

This sensory information is then sent to the *Dialogue Processing & Modality Fusion* component, which includes the multimodal DA classifier described in [1]. This classifier uses linguistic features extracted from the utterance, features representing the dialogue history, and features for the presence/absence of pointing or H-O actions to assign a DA tag to each utterance. The component also extracts object and location names mentioned in the utterance relevant to the task. This information, along with the pointing gesture (if it exists), is then combined into an action vector (as defined above), which can be passed to the Mediation Module.

The Execution Module is also composed of multiple subcomponents: pointing, H-O action, and speech generation, and an object recognition component to allow the robot to determine what object is in a given location, as HEL would frequently need to do in the *Find* task. The output of the Mediation Module is an action vector defining the robot's next move; each execution component uses this vector to perform its respective action.

The robot performs a pointing gesture by moving its arm to predefined configurations near the location of interest. As far as haptic actions, we equipped the robot with the ability to open a drawer by programming a demonstrated trajectory. We also used Baxter's display screen to show an image acquired from Baxter's built-in hand camera to the user, which simulated the H-O action of conspicuously holding an object up to the user. When a location is checked for an object, the image captured by the hand camera is first sent to a deep learning-based image classifier (using features from ResNet-50 [34]) to determine if the object type matches that requested by the user. This object recognition is necessary for updating the state of the task according to whether or not the object is of interest.

For speech generation, we use the Festival speech synthesis system [35]. Since the action vector encodes utterances as DA tags, we created template sentences that mirrored those in the *Find* task corpus for each possible action. For some actions, different sentences were used depending on whether or not a pointing gesture accompanied the utterance to provide a more natural response. It is these template sentences that Festival converts into speech output.

C. User Study

In a preliminary user study, seven healthy adult participants who were naive to our system were recruited (two females and five males). Each subject performed four trials in which they interacted with Baxter to find one of the four objects, giving a total of 28 trials. At the beginning of each trial, we showed the target object to the participant, and then secretly placed it in one of the three containers; the other two containers each contained one other object (leaving the fourth object out of the trial). The participants were asked to have Baxter help locate the object by talking and pointing. No script was provided to the participants.

D. Evaluation Procedure

We measured the performance of our implemented framework and the quality of the interaction using a variety of metrics. We report the average length of the interactions as the mean duration and the mean number of moves. Because the main goal of the interactions was to find the target object, we report the percentage of

Avg. Duration	Avg. # Turns	Successful Trials	Non-Understandings	WER	SSREs	Wrong Pointing	SSRE & Wrong Pointing	Wrong DAs	Avg. User Rating
1m 45s	15.6	85.7%	11.7%	16.3%	23.4%	28.9%	1.2%	11.1%	4

TABLE I: Performance results of the Multimodal Interaction Manager on the *Find* task with a human partner.

successful trials, i.e., trials in which the target object was successfully located. Since our model encodes the state of the task, including when the task is completed, trials in which Baxter continued the interaction when the object was already found counted as failed trials.

To evaluate the Mediation Module specifically, we recorded instances in which Baxter did not adequately understand the participant, which we defined as the number of times Baxter needed to ask the participant to repeat their action (i.e., the action could not be matched to an action in the HBATN) and the number of times a subtask needed to be repeated due to unsuccessful completion (e.g., a location was not determined at the end state of *Det(L)*). We call these *non-understandings*, consistent with the definition in [36], since these cases reflect Baxter not being able to interpret the user's action. We report the percentage of Baxter's turns across all trials that constituted a non-understanding.

We also investigated the performance of our Interpretation Module, as errors in this module can propagate to the Mediation Module. For the speech recognition component, we report the word error rate (WER) – the minimum edit distance (at the word level) between all of the recognized utterances and the true utterances [37]. We also calculated the percentage of human turns for which a speech recognition error altered the utterance in such a way that a subtask could not be successfully completed, which we call *serious speech recognition errors* (SSREs). For example, if the utterance "can I get a cup" is recognized as "can I get a cop," *Det(O_T)* cannot be successfully completed because neither of our object types are mentioned; however, if the utterance is recognized as "can again a cup," *Det(O_T)* can still be successfully completed, despite the recognition error.

For the pointing recognition component, we computed the percentage of pointing gestures performed that were either not recognized or not tagged with the correct intended location. Additionally, to see if errors frequently stacked, we also recorded moves with both an SSRE and a pointing recognition error. The performance of the *Dialogue Tools & Modality Fusion* component was assessed by comparing the DA tags produced by the classifier with manually-labeled DA tags of all utterances.

Finally, we measured the overall quality of the interaction by asking the participants to rate their experience on a 5-point Likert scale according to their expectations – a score of 1 meant the experience was "significantly worse than expected," and a score of 5 meant the experience was "significantly better than expected."

E. Results

The results of our preliminary user study are given in Table I. We see that despite the fact that over 20% of the participants' turns had a speech or pointing recognition error, only 11.7% of Baxter's turns indicated a non-understanding. This is likely due to a correctly recognized pointing gesture compensating for an SSRE or vice versa, as suggested by the 1.2% co-occurrence rate of both errors. This highlights an important advantage of our multimodal framework: when one sensory input is imperfect, it can be supported by evidence from other communication modalities.

Since many of the features of the DA classifier rely on the words of the utterance itself, the WER and SSRE rate can influence DA classification accuracy. However, the classifier still maintains a reasonably low error rate with its dialogue history and gesture features. The classifier's performance could have also contributed to the lower percentage of non-understandings.

Overall, these results, as well as the high percentage of successful trials and the average user rating of 4 (the experience was "better than expected"), suggests that our framework shows promise in successfully navigating a multimodal task-oriented interaction with a human.

VI. Conclusion

In this work, we investigated an interaction scenario in which a human and an assistive robot work together to locate a target object – the *Find* task. We first proposed an architecture for multimodal human-robot interaction with three main components: the Interpretation Module, the Mediation Module, and the Execution Module. Crucially, this framework integrates language and physical actions, both in interpretation and execution, while participating in a physical collaborative task with a human user. We then described our Hierarchical Bipartite Action-Transition Network for the Mediation Module, which maintains the state of the task-driven multimodal interaction, models both agents simultaneously, and plans subsequent moves.

Our findings from a preliminary study, in which we implemented our framework with Baxter to perform the *Find* task with participants, provide evidence that this framework can successfully interact with a human via speech, pointing, and H-O actions to complete a collaborative task. As future work, we plan to develop a method for automatically segmenting datasets of multimodal interactions like the *Find* corpus into pre-defined subtasks to derive the transitions and probabilities in the HBATNs for a given task. We also plan to extend our framework to more complex collaborative tasks that

involve physical interactions with the user as well as language, such as object handover and collaborative manipulation.

References

[1] L. Chen, M. Javaid, B. Di Eugenio, and M. Žefran, "The roles and recognition of haptic-ostensive actions in collaborative multimodal human–human dialogues," *Computer Speech & Language*, vol. 34, no. 1, pp. 201–231, 2015.

[2] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[3] C. L. Sidner, C. Lee, and C. Kidd, "Engagement during dialogues with robots," in *AAAI Spring Symposia*, 2005.

[4] J. Hemminghaus and S. Kopp, "Towards adaptive social behavior generation for assistive robots using reinforcement learning," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '17. ACM, 2017, pp. 332–340.

[5] P. Hanson and C. Rich, "A non-modal approach to integrating dialogue and action." in *AIIDE*, 2010.

[6] C. Rich, C. L. Sidner, and N. Lesh, "COLLAGEN: Applying collaborative discourse theory to human-computer interaction," *AI Magazine*, vol. 22, no. 4, pp. 15–25, 2001.

[7] C.-M. Huang and B. Mutlu, "Learning-based modeling of multimodal behaviors for humanlike robots," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '14. ACM, 2014.

[8] C. Stanton and C. J. Stevens, "Robot pressure: the impact of robot eye gaze and lifelike bodily movements upon decision-making and trust," in *International Conference on Social Robotics*. Springer, 2014, pp. 330–339.

[9] J. K. Lee and C. Breazeal, "Human social response toward humanoid robot's head and facial features," in *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2010, pp. 4237–4242.

[10] M. Salem, F. Eyssel, K. Rohlfing, S. Kopp, and F. Joublin, "To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability," *International Journal of Social Robotics*, vol. 5, no. 3, pp. 313–323, 2013.

[11] H. Admoni and B. Scassellati, "Nonverbal behavior modeling for socially assistive robots," in *2014 AAAI Fall Symposium Series*, 2014.

[12] H. Admoni, T. Weng, B. Hayes, and B. Scassellati, "Robot nonverbal behavior improves task performance in difficult collaborations," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE Press, 2016, pp. 51–58.

[13] S. Rosenthal, J. Biswas, and M. Veloso, "An effective personal mobile robot agent through symbiotic human-robot interaction," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, vol. 1, 2010, pp. 915–922.

[14] B. Peng, X. Li, L. Li, J. Gao, A. Celikyilmaz, S. Lee, and K.-F. Wong, "Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

[15] M. Patel, C. H. Ek, N. Kyriazis, A. Argyros, J. V. Miro, and D. Kragic, "Language for learning complex human-object interactions," in *2013 IEEE International Conference on Robotics and Automation*, 2013, pp. 4997–5002.

[16] K. Erol, J. Hendler, and D. S. Nau, "HTN planning: Complexity and expressivity," in *AAAI*, vol. 94, 1994, pp. 1123–1128.

[17] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[18] A. Mohseni-Kabir, C. Rich, S. Chernova, C. L. Sidner, and D. Miller, "Interactive hierarchical task learning from a single demonstration," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015, pp. 205–212.

[19] A. S. Clair, C. Saldanha, A. Boteanu, and S. Chernova, "Interactive hierarchical task learning via crowdsourcing for robot adaptability," in *Refereed workshop Planning for Human-Robot Interaction: Shared Autonomy and Collaborative Robotics at Robotics: Science and Systems, Ann Arbor, Michigan. RSS*, 2016.

[20] B. Nooraei, C. Rich, and C. L. Sidner, "A real-time architecture for embodied conversational agents: beyond turn-taking," *ACHI*, vol. 14, pp. 381–388, 2014.

[21] C. Rich and C. L. Sidner, "Using collaborative discourse theory to partially automate dialogue tree authoring," in *International Conference on Intelligent Virtual Agents*. Springer, 2012, pp. 327–340.

[22] J. Y. Chai, Q. Gao, L. She, S. Yang, S. Saba-Sadiya, and G. Xu, "Language to action: Towards interactive task learning with physical agents," in *IJCAI*, 2018, pp. 2–9.

[23] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[24] J. Thomason, S. Zhang, R. J. Mooney, and P. Stone, "Learning to interpret natural language commands through human-robot dialog," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[25] M. Wollowski, C. Berry, R. Winck, A. Jern, D. Voltmer, A. Chiu, and Y. Shibberu, "A data-driven approach towards human-robot collaborative problem solving in a shared space," in *2017 AAAI Fall Symposium Series*, 2017.

[26] K. D. Marek, L. Popejoy, G. Petroski, D. Mehr, M. Rantz, and W.-C. Lin, "Clinical outcomes of aging in place," *Nursing Research*, vol. 54, no. 3, pp. 202–211, 2005.

[27] J. Austin, *How to do things with words*. Harvard University Press, 1962.

[28] M. Foster, E. Bard, M. Guhe, R. Hill, J. Oberlander, and A. Knoll, "The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue," in *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*. ACM, 2008, pp. 295–302.

[29] L. Chen, "Towards modeling collaborative task oriented multimodal human-human dialogues," Ph.D. dissertation, University of Illinois at Chicago, 2014.

[30] E. Noohi, M. Žefran, and J. L. Patton, "A model for human–human collaborative object manipulation and its application to human–robot interaction," *IEEE transactions on robotics*, vol. 32, no. 4, pp. 880–896, 2016.

[31] S. Parastegari, E. Noohi, B. Abbasi, and M. Žefran, "A fail-safe object handover controller," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 2003–2008.

[32] ——, "Failure recovery in robot–human object handover," *IEEE Transactions on Robotics*, vol. 34, no. 3, pp. 660–673, 2018.

[33] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] P. A. Taylor, A. Black, and R. Caley, "The architecture of the festival speech synthesis system," in *The Third ESCA Workshop in Speech Synthesis*, 1998, pp. 147–151.

[36] G. Hirst, S. McRoy, P. Heeman, P. Edmonds, and D. Horton, "Repairing conversational misunderstandings and non-understandings," *Speech Communication*, vol. 15, no. 3-4, pp. 213–229, 1994.

[37] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall, 2008, vol. 2.