Artificial Intelligence for Human-Robot Interaction
Papers from the AAAI 2015 Fall Symposium

Robotic Social Feedback for Object Specification

**Emily Wu, Yuxin Han, David Whitney, John Oberlin,
James MacGlashan, Stefanie Tellex**

Humans to Robots Laboratory

Brown University

Providence, RI 02912

Abstract

Issuing and following instructions is a common task in many forms of both human-human and human-robot collaboration. With two human participants, the accuracy of instruction following increases if the collaborators can monitor the state of their partners and respond to them through conversation (Clark and Krych 2004), a process we call *social feedback*. Despite this benefit in human-human interaction, current human-robot collaboration systems process instructions in non-incremental batches, which can achieve good accuracy but does not allow for reactive feedback (Tellex et al. 2011; Matuszek et al. 2012; Tellex et al. 2012; Misra et al. 2014). In this paper, we show that giving a robot the ability to ask the user questions results in responsive conversations and allows the robot to quickly determine the object that the user desires. This social feedback loop between person and robot allows a person to create an internal model for the robot's mental state and adapt their own behavior to better inform the robot. To close the human-robot feedback loop, we employ a Partially Observable Markov Decision Process (POMDP) to produce a policy which will lead to the determination of the object in the shortest amount of time. To test our approach, we perform user studies to measure our robot's ability to deliver common household items requested by the participant. We compare delivery speed and accuracy both with and without social feedback.

Introduction

When humans collaborate on a task—for example, repairing a car, or cooking a meal—both participants continually signal back and forth, communicating their current understanding of the task and the actions needed to achieve the goal. Clark describes communication as a *joint activity*, similar to playing a duet or performing a waltz (Clark 1996). In our work, we call this back and forth signaling *social feedback*, and the goal is to use social feedback to improve the speed and accuracy of human-robot interactions.

Robotic research into establishing common ground is just beginning, but has already shown promise. In (Chai et al. 2014), they developed a system to establish new names for objects visible to the robot. (Williams 2015) describes an approach to understanding underlying semantics in human dialog. There are natural areas for improvement since both works rely on hand-coded rules or logical predicates, limiting easy expansion to new domains.

Figure 1: The human's view when using our system to interact with the robot. The user and the robot participate in a conversation so that the robot can determine which of the six objects the user desires. In counter-clockwise order the objects are: the brown mug, the wooden bowl, the blue bowl, the green spoon, the white spoon, and the white brush.

We propose an approach that will estimate the human's state and choose actions in real time. Our system takes multimodal observations as input, namely speech and gesture, and responds with speech according to a policy generated by solving a POMDP representation of the world. We use an approximate solving technique called Belief Sparse Sampling (Kearns, Mansour, and Ng 2002). Performing inference is expensive, so we must make optimizations to use the policy in real time, both in the model structure and by caching the policy with k-nearest neighbors (KNN).

Our research focuses on an important part of physical collaboration: object delivery. Object delivery is an essential capability for robots (Huang, Cakmak, and Mutlu 2015). Our setup for object delivery is as follows: a human requests a series of objects from the robot one at a time. The human makes their request using speech and gesture (pointing). In

previous work, the robot had only two actions, deliver the object and wait. The robot collected information from the human until its estimation of the desired object crossed a threshold, then delivered the estimated object. This approach was successful if the observations were unambiguous to the robot, but if the robot was unsure, it was unable to communicate that fact to the human and could only wait passively for more information. Our research provides a system that not only interprets the speech and gesture of the human to determine which object they desire, but also provides a flexible, non-rule based approach to dynamically generate social feedback.

Our model allows for incremental interpretation of speech and gesture, as implemented in our prior work. However, to circumvent practical issues in synchronizing human-robot dialogue, the robot waits until the human has finished each utterance before taking an action. Future work will focus heavily on adjusting our approach to appropriately handle barge-in by both human and robot.

To evaluate our proposed approach, we conducted a user study where participants asked for objects from the robot. We measured speed and accuracy and compared the difference between trials that use social feedback and trials that do not.

Related Work

Work demonstrating the importance of social feedback in human-human communication has been done in the field of psycholinguistics. In (Clark and Krych 2004), one human (labeled the builder) builds a Lego model according to instructions given by another human (labeled the director). In the feedback-free trials, the director's instructions were prerecorded, and the resulting models were very inaccurate (in fact no model was completely correct). In the feedback trials, errors were reduced by a factor of eight. Our goal is to enable a robot to collaborate with a human in this way.

Other work with collaborative robots exists, for example, (Foster et al. 2012) have done research with a bar-tending robot. This robot follows a rule-based state estimator, and delivers drinks from fixed positions behind the bar to multiple users based on their speech and torso position. We expand the scope of the problem: we do not use a rule-based state planner, our items are not in fixed positions, and our gesture model uses pointing instead of torso position.

In (Bohus and Horvitz 2014), a robotic building guide directs guests to find specific rooms. Our project addresses a similar domain, requiring the interpretation of users' requests, but differs in the task and the type of communication necessary to accomplish that task.

Other work involving robotic object delivery also exists. Some approaches have no social feedback and will either deliver the wrong item or do nothing if given a request it does not understand (Tellex et al. 2011; Matuszek et al. 2012; Tellex et al. 2012; Misra et al. 2014). Language only feedback models also exist (Chai et al. 2014; MacMahon, Stankiewicz, and Kuipers 2006; Tellex et al. 2011; Matuszek et al. 2012; Guadarrama et al. 2014; Hewlett, Walsh, and Cohen 2011; Misra et al. 2014), and several gesture only models (Waldherr, Romero, and Thrun 2000; Marge et al. 2011).

(Matuszek et al. 2014) shows promising work in fusing language and complex gesture to understand references to multiple objects at once. We build off this work by including social feedback.

In the field of computational linguistics, previous work exists in resolving referring expressions incrementally, such as (Schlangen, Baumann, and Atterer 2009; Kruijff et al. 2007; Gieselmann 2004). Other work in that community also incorporates gesture, and/or eye gaze (Kennington, Kousidis, and Schlangen 2013; Kennington, Dia, and Schlangen 2015), but the given work does not incrementally update gesture along with speech. (Chai, Prasov, and Qu 2011) provides work towards resolving referring expressions in a different domain, but does not address the task of acting on the results of these referring expressions. In (Kruijff, Brenner, and Hawes 2008), they propose a system for planning to ask for clarifications, which covers a wide scope of knowledge failures. In this work, we are interested only in a small subset of these clarifications, and address the problem of how and when these clarifications should be used in a concrete human-robot collaboration task.

POMDP approaches to dialog (Young et al. 2013) are quite common, but treat dialog as a discrete, turn-taking interaction. The Dialog State Tracking Challenge (Williams et al. 2013) a notable driving force for computer dialog understanding, treats dialog in this turn-based way. Although the behavior of our system resembles turn-taking, our model treats dialogue as an incremental process and future implementations will make use of this.

Alternative approaches to POMDPs include cognitive architecutres such as SOAR (Laird 2012) or DIARC (Schermerhorn et al. 2006). By taking a probabilistic approach, we can seemlessly fuse information from multiple sources and explicitly reason about the robot's uncertainty when choosing actions.

Technical Approach

The goal of our work is to enable the robot to correctly determine which object the human desires from their speech and gesture.

The robot might misinterpret a person's speech and gesture; to recover from these failures, the robot chooses speech actions of its own, which change the human's belief about what the robot knows, which in turn shapes the user's subsequent speech and gesture. In actuality, we do not know how the robot's actions affect the human's belief or how the human's belief affects their subsequent actions. However, if we make certain assumptions about how the robot's actions affect what we observe from the human, we can formulate the model as a POMDP.

POMDP Overview

To solve a POMDP, an agent must perform state-estimation and policy generation. The state estimator calculates a belief state, which is a probability density function (pdf) over all possible states, and the policy generator chooses an ac-

Variable	Explanation
$s = \langle \mathcal{O}, \omega \rangle$	A single state, which is made up of the given tuple
$\mathcal{O} = \{x_1, ..., x_D\}$	Set of all objects
$\omega \in \mathcal{O}$	Object desired by user
$S = \{s_1 \dots s_N\}$	Set of all states
$x = \{name, vocab, position\}$	An object, defined by a name, vocabulary, and position
a	A possible robot action. Speech and picking
$A = \{a_1 \dots a_k\}$	Set of all robot actions
$T(s, a, s^{t+1}) = p(s^{t+1}\|s^t, a^t)$	Transition function, probability of entering new state given current state and action
$o = \langle l, g, \mathcal{O} \rangle$	A single observation, made up of observed language, gesture, and objects
$\Omega = \{o_1 \dots o_M\}$	Set of all possible observations
$O(o^{t+1}) = p(o^{t+1}\|s^{t+1}, a^t)$	Observation function, probability of an observation given the state and previous action
$R(s, a) \in \mathbb{R}$	Reward function
$\gamma \in [0, 1]$	Discount factor, discounts future rewards

Table 1: POMDP Variables.

tion that maximizes the agents expected cumulative reward according to a given reward function.

State Estimator The state estimator assumes an initial belief, and uses Bayesian mathematics to update its belief over time. In order to perform this update, the state estimator has a model of how the true state emits observations and how actions affect the state. These two models are called the observation function and the transition function.

Policy Generator The policy generator chooses a set of actions that maximizes the expected value of its reward over time. The reward for a state-action pair is given by a reward function.

POMDP Definition

We define our POMDP by the tuple $\{S, A, T, R, \Omega, O\}$.

- S is the set of states. In this problem, a state is a tuple of two items $\langle \mathcal{O}, \omega \rangle$. $\mathcal{O}$ is the is the set of objects available for the robot to deliver. Each element $x \in \mathcal{O}$ is an object with a name, unigram vocabulary, and position. An example value $\mathcal{O}$ could be the set of objects {redBowl, greenSpoon}. An example $x \in \mathcal{O}$ for a red bowl would be (redBowl, {red, bowl, plastic}, (1.0, 2.0, 0.0)). The object the human desires is denoted $\omega \in \mathcal{O}$. While $\mathcal{O}$ is considered a known variable, ω is hidden, making our POMDP a Mixed Observability MDP (Ong et al. 2010)
- A is the set of actions. The robot can deliver an object, do nothing, or ask a question about a property of the desired object.
- $T = p(s^{t+1}|s^t, a^t)$ is the transition function. It calculates the probability of transitioning from the current state to the next state given the current state and current action. We make the assumption that the human participant does not change the object they desire unless their object is successfully delivered.
- R is the reward function. The reward function takes as input a state and action, and gives a real-valued reward. The reward for delivering the correct object is 10 whereas delivering the incorrect object yields a penalty of -80 because it can have negative side effects and is time consuming. Doing nothing yields -1 as a penalty for time passing. Talking yields -4 to penalize bothering the user. These values were chosen empirically and have a natural interpretation when considered relative to the penalty for time passing.
- Ω is the set of possible observations, $\langle l, g, \mathcal{O} \rangle \in \Omega$. l is the human's speech, g is the human's gesture, $\mathcal{O}$ is as defined above.
- $O = p(o^{t+1}|s^{t+1}, a^t)$ is the observation function which describes how states emit language and gesture from the human.

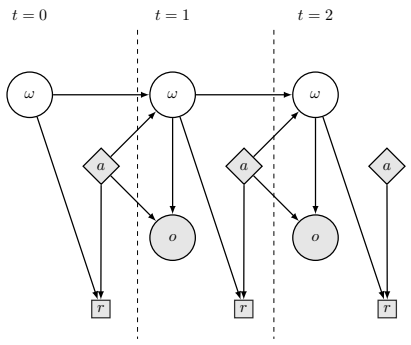

Figure 2: Graphical model of proposed POMDP over three timesteps. Gray nodes are observable.

A concise list of the POMDP variables is shown in Table 1.

Observation Function The observation function calculates the probability of an observation given the state and previous action.

$$O(o^{t+1}, s^{t+1}, a^t) = p(o^{t+1}|s^{t+1}, a^t) \tag{1}$$

We can expand Equation 1.

$$O(o^{t+1}) = p(\mathcal{O}^{t+1}, l^{t+1}, g^{t+1}|\mathcal{O}^{t+1}, \omega^{t+1}, a^t) \tag{2}$$

The set of objects on the table is only dependent on itself, so we can now factor Equation 1 to separate $\mathcal{O}$.

$$O(o^{t+1}) = p(\mathcal{O}^{t+1}|\mathcal{O}^{t+1})p(l^{t+1}, g^{t+1}|\mathcal{O}^{t+1}, \omega^{t+1}, a^t) \tag{3}$$

In our problem we assume no error in observing $\mathcal{O}$. Therefore the first term is equal to one. We can simplify and remove it from the equation.

$$O(o^{t+1}) = p(l^{t+1}, g^{t+1}|\mathcal{O}^{t+1}, \omega^{t+1}, a^t) \tag{4}$$

If we assume conditional independence of speech and gesture, we can factor Equation 4 one step further. Conditional independence in this case means that, conditioned on the object that the user desires, the speech and gesture observations are independent of each other.

$$O(o^{t+1}) = p(l^{t+1}|\mathcal{O}^{t+1}, \omega^{t+1}, a^t)p(g^{t+1}|\mathcal{O}^{t+1}, \omega^{t+1}, a^t) \tag{5}$$

It may seem inaccurate to assume speech and gesture are conditionally independent, but empirically, we observe that when the true state is known, language and gesture are largely but not completely independent. This assumption simplifies our model and allows us to separate the observation function into a language model and a gesture model.

Language model In the language model, we observe two types of speech: General speech is interpreted according to a unigram model, while yes/no responses are handled separately.

For most speech input, we use a unigram model. For each word in the observed speech, we calculate the probability that, given the state, that word would have been used to describe the state. We assume here that $\mathcal{O}$, the set of objects available, does not affect which words the participant would speak, though in practice, humans do tailor their speech in response to different objects.

$$\begin{aligned} p(l^{t+1}|\mathcal{O}^{t+1},&\omega^{t+1}, a^t) \\ &= p(l^{t+1}|\omega^{t+1}, a^t) \\ &= p(c|\omega^{t+1}, a^t) \prod_{w \in l^{t+1}} p(w|\omega^{t+1}, a^t) \end{aligned}$$

Where $p(c|\omega^{t+1}, a^t)$ describes the probability that the human chooses to communicate given the state and action. We assume that this is independent of ω, and depends only on the action; if the robot asks a question, the human is more likely to respond:

$$p(c|a^t) = \begin{cases} 0.8 & \text{if } a^t \text{ is a question} \\ 0.2 & \text{otherwise} \end{cases} \tag{6}$$

The probability of a particular word being used to describe a particular object is determined by consulting the object's vocabulary. Specifically,

$$p(w|\omega^{t+1}, a^t) = \frac{\text{Number of times } w \text{ is used to describe } \omega}{\text{Total counts for words describing } \omega}$$

If w is not part of the object's vocabulary, we assign it a small probability ϵ.

While a unigram model is rudimentary, it serves as a good starting point for our work and produces adequately accurate results. In the future, more sophisticated language models will be considered.

The user may also say "yes" or "no" in response to a question that the robot has asked, in which case we handle the utterance differently. In the case where the robot has not asked a question, the probability of the user answering "yes" or "no" is assigned probability ϵ. If the robot has asked a question,

$$\begin{aligned} p(l^{t+1} &= \text{“}yes\text{”}|\mathcal{O}^{t+1}, \omega^{t+1}, a^t) \\ &= p(c|a^t)p(l^{t+1} = \text{“}yes\text{”}|\omega^{t+1}, a^t) \end{aligned}$$

$$p(l^{t+1} = \text{“}yes\text{”}|\omega^{t+1}, a^t) = \begin{cases} 1 & \text{if } a^t.\text{text} \in \omega.\text{vocab} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

$$\begin{aligned} p(l^{t+1} &= \text{“}no\text{”}|\mathcal{O}^{t+1}, \omega^{t+1}, a^t) \\ &= p(c|a^t)p(l^{t+1} = \text{“}no\text{”}|\omega^{t+1}, a^t) \end{aligned}$$

$$p(l^{t+1} = \text{“}no\text{”}|\omega^{t+1}, a^t) = \begin{cases} 0 & \text{if } a^t.\text{text} \in \omega.\text{vocab} \\ 1 & \text{otherwise} \end{cases} \tag{8}$$

We assume that the probability the user says "yes" or "no" is independent of $\mathcal{O}$.

In addition, to account for the user not answering quickly enough, the state includes the the last question asked by the robot.

Gesture Model The gesture model operates on the vector defined by the participant's shoulder and hand, called v_t. The intersection of v_t with the plane of the table is considered to be the target of the pointing gesture, p_t. We assume the participant samples this point from a 2 dimensional Gaussian distribution centered at the location of the object the participant is indicating, with a hand-chosen variance σ. Therefore, the probability of a gesture given a state is as follows:

$$p(g^{t+1}|\mathcal{O}^{t+1},\omega^{t+1},a^t) \tag{9}$$
$$= p(g^{t+1}|\omega^{t+1}) \tag{10}$$
$$\propto \mathcal{N}(\mu = (\omega.x, \omega.y), \sigma)[(p_t.x, p_t.y)] \tag{11}$$

Again we assume that the gesture the participant chooses is independent of the other objects on the table and their placement, though this is not the case in practice.

We also track an additional vector called $u_{t,x}$ defined by the angle between the participant's shoulder and an object $x \in \mathcal{O}$. For each $x \in \mathcal{O}$, we calculate the angle between v_t and $u_{t,x}$. If the smallest angle is over a certain threshold, then no gesture was performed and the term is not factored into the overall observation model.

Transition Function The transition function calculates the probability of moving to a particular new state given the current state and action.

$$T(s^{t+1}, s^t, a^t) = p(s^{t+1}|s^t, a^t) \tag{12}$$

We substitute the corresponding variables into (12).

$$T(s^{t+1}) = p(\mathcal{O}^{t+1}, \omega^{t+1}|\mathcal{O}^t, \omega^t, a^t) \tag{13}$$

The set of desired objects on the table and the desired object only depend on themselves and the last action taken, so we can factor (13).

$$T(s^{t+1}) = p(\mathcal{O}^{t+1}|\mathcal{O}^t, a^t)p(\omega^{t+1}|\omega^t, a^t) \tag{14}$$

The two terms from (14) have simple stepwise functions if a^t is not a pick action.

$$p(\mathcal{O}^{t+1}|\mathcal{O}^t, a^t) = \begin{cases} 1 & \text{if } \mathcal{O}^{t+1} = \mathcal{O}^t \\ 0 & otherwise \end{cases} \tag{15}$$

$$p(\omega^{t+1}|\omega^t, a^t) = \begin{cases} 1 & \text{if } \omega^{t+1} = \omega^t \\ 0 & otherwise \end{cases} \tag{16}$$

If a^t is a pick action, the transition model for $\mathcal{O}$ is defined as follows:

$$p(\mathcal{O}^{t+1}|\mathcal{O}^t, a^t) = \begin{cases} 1 & \text{if } \mathcal{O}^{t+1} = \mathcal{O}^t \setminus \{a^t.\text{object}\} \\ 0 & otherwise \end{cases} \tag{17}$$

If a^t is a pick action and the user's desired object ω is picked, ω transitions as follows.

$$p(\omega^{t+1}|\omega^t, a^t) = \begin{cases} 0 & \text{if } \omega^{t+1} = \omega^t \\ 1/|\mathcal{O}^{t+1}| & otherwise \end{cases} \tag{18}$$

Otherwise, it is the same as if some other action had been taken.

Actions The robot must choose among the actions available to it. With social feedback enabled, these actions are:

- Pick up an object and deliver to the human
- Ask a question of the form "Would you like a ⟨*object property*⟩ object?"
- Wait (do nothing)

With social feedback disabled as in the baseline, these actions are:

- Pick up an object and deliver it to the human
- Wait (do nothing)

Asking a question serves both as an information gathering action (as it prompts the human to respond with additional information) and also a means for the robot to express its uncertainty about which object is desired. These questions directly mirror the unigram model of speech that we use to interpret human language, i.e., each question asks about a one-word property of the object, such as its color (e.g., blue), material (e.g., wooden) , or noun description (e.g., cup). Asking about a particular property allows the robot to reference all objects with that property at the same time. An additional benefit of using this simple question format is that if we assume the human interprets speech with the same model as the robot, its effect on its belief about the robot's knowledge is easy to predict. We will explore this in future work.

We also give the robot the action to do nothing for a given timestep. This is important in preventing the robot from inundating the human with speech utterances.

Policy Generation

The optimal policy can be exactly calculated, but is intractable due to the size of the state space. Fortunately, approximate solutions exist. Currently, a representation of this POMDP has been specified with the Brown-UMBC Reinforcement Learning and Planning (BURLAP) library(MacGlashan), a learning and planning library that can solve POMDPs. For a solver, we use Belief Sparse Sampling, a finite horizon planning algorithm used on the POMDP's belief MDP that allows us to specify how deep to search and the number of observations to sample at each timestep.

However, even using Belief Sparse Sampling, it was still not fast enough to allow the robot to respond in real time. As the number of objects grows, the number of states and the number of actions needed to address those states grow. For example, adding 1 additional object increases the number of states by 1 but increases the number of actions from anywhere between 1 (1 additional pick action) to several dozen (1 for each new unigram description of the object in its vocabulary). Each of these contributes to the branching factor, causing the the computation to quickly grow intractable even with a limited search depth.

In big-O notation, the runtime of calculating the next decision can be expressed as $O((s * a * o)^d)$, where s is the number of states, a is the number of actions, o is the number of sampled observations, and d is the search depth. We can rewrite a as $a = s + h * s + 1$, because we have 1 pick action per state, some constant number h property questions

for each object, and one wait action. o and d are both chosen as constants, which gives a big-O runtime of $O((s^2)^d)$.

We combat this exponential growth by limiting the number of actions per object to 2 (which should be enough for the robot to express which object it is referring to) as well as condensing the pick actions into a single macro action. This macro pick action allows the robot to pick the object with the greatest belief, and saves it from considering pick actions that are likely to be incorrect. Our new run time is $O((s * a * o)^d)$ where $a = 1 + 2s + 1$ (1 macro pick action, at most 2 question actions per state, and 1 wait action). While the big-O runtime remains the same, the improvement to constant factors is significant when adding new objects to the domain.

Even with these measures to reduce the branching factor, the system did not run fast enough for real time at the search depth needed for meaningful decisions. In order to obtain real time responses, we run a simulated environment of our domain, with simulated object locations and simulated user input. We cached the results of several thousand decisions made by Belief Sparse Sampling, storing the belief state and the action chosen. Then, at interaction time, we use these results to make a decision about which action to take. Specifically, we perform a KNN classification where the feature vector is composed of the belief values of the POMDP's belief state and the class label is the chosen action. At runtime, we are given an input vector of the current belief state, which we classify and then perform the action given by the class label. This allows for nearly instantaneous responses, enabling us to make real time decisions.

Making these optimizations is not ideal. For example, adding or changing the set of objects on the table requires rebuilding the KNN cache. In the future we will likely move to alternative POMDP solvers which should give better performance.

Evaluation

We evaluated the POMDP with social feedback against a baseline system which is unable to speak and only able to listen. Each system was evaluated by six users. We report four metrics: the number of picks that delivered the correct object (picks correct), the number of total attempted picks, the average time from the end of the first utterance referring to an object until the end of the delivery of that object (EUED time), and the fractional amount of time spent delivering an incorrect object (IOD time).

User Study Protocol

Participants for this user study were acquired through convenience sampling by inviting volunteers from the Brown University CIT building and the campus area surrounding it. We gathered 6 participants. Each participant contributed two trials, a baseline trial and a social feedback trial. During the baseline trials, all feedback from the robot was excluded. During the social feedback trials, the robot communicated with the participants through voice and animated eye gaze which indicated the directions of the objects. The order of these two trials was randomly assigned to each participant.

The participants interacted with a modified Baxter robot to perform a pick and place task with six everyday objects: a brown mug, a wooden bowl, a blue bowl, a green spoon, a white spoon, and a white brush. The six objects were placed on a large table, behind which the robot was stationary. The six objects and their starting locations stayed unchanged for all participants.

At the beginning of each study, the user was given a headset microphone, calibrated the Kinect, and faced Baxter (Figure 1). The headset microphone left their arms free to gesture. They were instructed to choose an object on one side of the table and then to indicate the desired object through a combination of pointing and natural language instructions. They were told to continue indicating until the desired object had been delivered. If the robot delivered an incorrect object, the user continued interacting until the correct object was delivered. Next they were instructed to do the same for an object on the other side of the table. Once the second object was delivered, the user was finished with the first half of the study.

Each participant first completed the above study with social feedback either disabled (baseline) or enabled. The participants were then given a follow up survey to record their subjective reactions towards the robot. The survey measured the participants' impressions on the robot concerning its level of friendliness, reliability, capability, communication skills and their perception of safety. After the survey, the participant completed a second study with social feedback either enabled or disabled, so that after the second study the participant had completed the trial once with feedback on and once with feedback off. Half the participants tried with feedback first. After the second trial, the user was dismissed. There were no trials where the first object delivery left no objects on the opposite side.

We expect to continue the survey as part of ongoing research to support our exploration of human-robot interaction. The detailed results and analysis of this survey will be included in our future studies.

Results

The results of our user studies are summarized in Table 2. When using social feedback we observed a four second average decrease in the EUED time but we incurred one additional incorrect pick. Since we performed a statistically small number of trials, we view these results more as a proof of concept than as a rigorous evaluation.

Discussion

Social feedback enables logical conversations between human and robot which allow the robot to quickly determine the user's object of choice (Table 3). In the first example, the user indicates that they want a spoon. This is an ambiguous statement because there are two spoons, one of them white and the other green. The model can capture the fact that asking about the color of the spoon will determine the object. The robot asks if the user wants a white object, and when the user indicates that they do not, the robot knows that the object in question is the green spoon. Similarly, wanting a bowl is ambiguous because there is a blue bowl and a

Metric	No Feedback	Feedback
Picks Correct	12/13	12/14
EUED Time	17.33s	13.33s
IOD Time	2.59%	3.18%

Table 2: Experimental Results. EUED time is time from the end of the first utterance in the request until the end of delivery of the desired object and IOD time is time spent delivering incorrect objects.

Human	Robot
I want a spoon.	
	Would you like a white object?
No.	
	(Picks the green spoon.)
I want a bowl.	
	Would you like a blue object?
Yes.	
	(Picks the blue bowl.)
(Points ambiguously at the wooden bowl and brown mug.)	
	Do you want the cup?
No.	
	(Picks the wooden bowl.)

Table 3: Example dialogue from interactions with our system. The robot chooses to ask more informative questions because the search depth of the planner is three rather than two.

wooden bowl, so the robot asks about the color of the object, determining that the user wants the blue bowl. This type of inference is not limited to speech: ambiguous gestures can be resolved by asking about one of the candidate objects, as in the case of the wooden bowl and brown mug.

By asking questions, the agent divides and conquers the belief space. Constructing a POMDP for an arbitrary set of objects is easier than specifying search rules in multimodal interaction spaces with complicated relationships between objects. Solving a POMDP allows those rules to be discovered rather than specified, and the burden is transferred from the modeler to the solver.

Without social feedback, users must resort to repeating instructions with no idea whether their input is received by the robot or whether it is interpreted correctly. This might lead to frustration and a lower long term tolerance of the system. When the robot can reason about and issue social feedback and tell the user what it believes, the user and the robot can engage in a natural dialogue which quickly determines the item of choice. We predict that because the interaction is more like a human-human interaction, users will be able to engage with the system for longer periods of time with less frustration.

Conclusion

Social feedback is key to human-human interaction. This research hopes to give robots the ability to participate in social feedback, making human-robot interaction easier, faster, and more accurate. Our approach uses a POMDP that takes input from a human in the form of speech and gesture, estimates the human's mental state, and chooses social feedback actions to facilitate the determination of the desired object. Our results suggest that by incorporating social feedback we can expect a decrease in response time, but more experimentation is required.

Future Work

We plan to expand our current approach by implementing robotic gesture as an additional feedback action the robot can take. This will add to the the number of actions and make planning more expensive, but by implementing robotic speech and gesture as collapsed actions similarly to our pick action, we should maintain computational tractability. Moving beyond a unigram speech model would allow more sophisticated references involving multiple objects, such as "the bowl behind the spoon." Finally, a placement action would allow us to support more complex collaborative tasks.

In addition, we would like to explore the potential for richer communication by having the robot track the human's belief about its current state. Adding this additional state variable would greatly increase the computational complexity of this task, but also allow it to make decisions that consider how its actions are perceived by the human.

Acknowledgment

This paper is a collaboration between Brown University's Humans to Robots Laboratory and the Rhode Island School of Design's Industrial Design department. We would like to thank Dr. Claudia B. Rébola, associate professor at the Rhode Island School of Design, for organizing this collaboration and providing valuable guidance in designing the user experience and conducting user studies.

This work was in part supported by the National Science Foundation under "NRI: Collaborative: Jointly Learning Language and Affordances," grant # ISS-1426452. It was also supported by the Army Robotics CTA under "Robotics CTA: Perception, Human-Robot Interaction," grant # 40228388.

References

Bohus, D., and Horvitz, E. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, 2–9. New York, NY, USA: ACM.

Chai, J. Y.; She, L.; Fang, R.; Ottarson, S.; Littley, C.; Liu, C.; and Hanson, K. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, 33–40. ACM.

Chai, J. Y.; Prasov, Z.; and Qu, S. 2011. Cognitive principles in robust multimodal interpretation. *CoRR* abs/1109.6361.

Clark, H. H., and Krych, M. A. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50(1):62–81.

Clark, H. H. 1996. *Using Language*. Cambridge University Press.

Foster, M. E.; Gaschler, A.; Giuliani, M.; Isard, A.; Pateraki, M.; and Petrick, R. P. A. 2012. Two people walk into a bar: dynamic multi-party social interaction with a robot agent. In *International Conference on Multimodal Interaction, ICMI '12, Santa Monica, CA, USA, October 22-26, 2012*, 3–10.

Gieselmann, P. 2004. Reference resolution mechanisms in dialogue management.

Guadarrama, S.; Rodner, E.; Saenko, K.; Zhang, N.; Farrell, R.; Donahue, J.; and Darrell, T. 2014. Open-vocabulary object retrieval. In *Robotics: Science and Systems*.

Hewlett, D.; Walsh, T. J.; and Cohen, P. 2011. Teaching and executing verb phrases. In *Development and Learning (ICDL), 2011 IEEE International Conference on*, volume 2, 1–6. IEEE.

Huang, C.-M.; Cakmak, M.; and Mutlu, B. 2015. Adaptive coordination strategies for human-robot handovers. In *Proceedings of Robotics: Science and Systems*.

Kearns, M.; Mansour, Y.; and Ng, A. Y. 2002. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning* 49(2-3):193–208.

Kennington, C.; Dia, L.; and Schlangen, D. 2015. A discriminative model for perceptually-grounded incremental reference resolution. In *Proceedings of the 11th International Conference on Computational Semantics*, 195–205. London, UK: Association for Computational Linguistics.

Kennington, C.; Kousidis, S.; and Schlangen, D. 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. In *Proceedings of the SIGDIAL 2013 Conference*, 173–182. Metz, France: Association for Computational Linguistics.

Kruijff, G.-J.; Lison, P.; Benjamin, T.; Jacobsson, H.; and Hawes, N. 2007. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Language and Robots: Proceedings from the Symposium (LangRo'2007)*. University of Aveiro.

Kruijff, G.; Brenner, M.; and Hawes, N. 2008. Continual planning for cross-modal situated clarification in human-robot interaction. In *Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on*, 592–597.

Laird, J. 2012. *The Soar cognitive architecture*. MIT Press.

MacGlashan, J. Brown UMBC Reinforcement Learning and Planning Library. http://burlap.cs.brown.edu/.

MacMahon, M.; Stankiewicz, B.; and Kuipers, B. 2006. Walk the talk: connecting language, knowledge, and action in route instructions. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 1475–1482. AAAI Press.

Marge, M.; Powers, A.; Brookshire, J.; Jay, T.; Jenkins, O. C.; and Geyer, C. 2011. Comparing heads-up, hands-free operation of ground robots to teleoperation. *Robotics: Science and Systems VII*.

Matuszek, C.; Herbst, E.; Zettlemoyer, L.; and Fox, D. 2012. Learning to parse natural language commands to a robot control system. In *Proc. of the 13th Intl Symposium on Experimental Robotics (ISER)*.

Matuszek, C.; Bo, L.; Zettlemoyer, L.; and Fox, D. 2014. Learning from unscripted deictic gesture and language for human-robot interactions.

Misra, D.; Sung, J.; Lee, K.; Saxena, A.; Sung, J.; Selman, B.; Saxena, A.; Sung, J.; Ponce, C.; Selman, B.; et al. 2014. Tell me dave: Context-sensitive grounding of natural language to mobile manipulation instructions. In *Robotics: Science and Systems, RSS*.

Ong, S. C. W.; Png, S. W.; Hsu, D.; and Lee, W. S. 2010. Planning under uncertainty for robotic tasks with mixed observability. *Int. J. Rob. Res.* 29(8):1053–1068.

Schermerhorn, P. W.; Kramer, J. F.; Middendorff, C.; and Scheutz, M. 2006. Diarc: A testbed for natural human-robot interaction. In *AAAI*, 1972–1973.

Schlangen, D.; Baumann, T.; and Atterer, M. 2009. Incremental reference resolution: The task, metrics for evaluation, and a Bayesian filtering model that is sensitive to disfluencies. In *Proceedings of the SIGDIAL 2009 Conference*, 30–37. London, UK: Association for Computational Linguistics.

Tellex, S.; Kollar, T.; Dickerson, S.; Walter, M. R.; Banerjee, A.; Teller, S.; and Roy, N. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proceedings of the AAAI*.

Tellex, S.; Thaker, P.; Deits, R.; Kollar, T.; and Roy, N. 2012. Toward information theoretic human-robot dialog.

Waldherr, S.; Romero, R.; and Thrun, S. 2000. A gesture based interface for human-robot interaction. *Autonomous Robots* 9(2):151–173.

Williams, J.; Raux, A.; Ramachandran, D.; and Black, A. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, 404–413.

Williams, T. 2015. Toward more natural human-robot dialogue. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*, HRI'15 Extended Abstracts, 201–202. New York, NY, USA: ACM.

Young, S.; Gašić, M.; Thomson, B.; and Williams, J. D. 2013. POMDP-based statistical spoken dialog systems: A review.