2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Mixed Reality as a Bidirectional Communication Interface for Human-Robot Interaction

Eric Rosen*, David Whitney*, Michael Fishman*, Daniel Ullman and Stefanie Tellex
Department of Computer Science, Brown University
*These authors contributed equally

***Abstract*—We present a decision-theoretic model and robot system that interprets multimodal human communication to disambiguate item references by asking questions via a mixed reality (MR) interface. Existing approaches have either chosen to use physical behaviors, like pointing and eye gaze, or virtual behaviors, like mixed reality. However, there is a gap of research on how MR compares to physical actions for reducing robot uncertainty. We test the hypothesis that virtual deictic gestures are better for human-robot interaction (HRI) than physical behaviors. To test this hypothesis, we propose the Physio-Virtual Deixis Partially Observable Markov Decision Process (PVD-POMDP), which interprets multimodal observations (speech, eye gaze, and pointing gestures) from the human and decides when and how to ask questions (either via physical or virtual deictic gestures) in order to recover from failure states and cope with sensor noise. We conducted a between-subjects user study with 80 participants distributed across three conditions of robot communication: no feedback control, physical feedback, and MR feedback. We tested performance of each condition with objective measures (accuracy, time), as well as evaluated user experience with subjective measures (usability, trust, workload). We found the MR feedback condition was 10% more accurate than the physical condition and a speedup of 160%. We also found that the feedback conditions significantly outperformed the no feedback condition in all subjective metrics.**

I. Introduction

Communicating human knowledge and intent to robots is essential for successful human-robot interaction (HRI). For example, when a surgeon says "hand me the scalpel," it is crucial that the assistive robot hand over the correct utensil. In order to efficiently collaborate, humans intuitively communicate through noisy modalities such as language, gesture, and eye gaze. Failures in communication, and thus collaboration, occur when there is mismatch between two agents' mental states.

Question-asking allows a robot to acquire information that targets its uncertainty, facilitating recovery from failure states. However, all question-asking modalities have trade-offs, making choosing which to use an important and context-dependent decision. For example, for robots with "real" eyes or pan/tilt screens, looking requires fewer joints to move less distance compared to pointing, decreasing the speed of the referential action. However, eye gaze is inherently more difficult to interpret.

On the other hand, Mixed Reality Head-Mounted Displays (MR-HMD), which have been shown to reduce mental workload in HRI [16], can indicate items quickly, are very accurate given proper calibration, and are independent of the physical robot. However, visualizations may distract the user's attention more than a typical pointing or looking action. Furthermore, MR technology is still new, and users may prefer to instead interact with a robot that performs physical actions. We aim to close the gap of research on how MR compares to physical actions for reducing robot uncertainty.

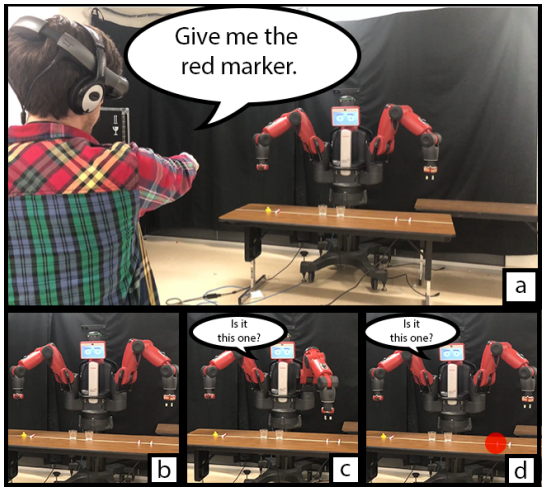


Fig. 1: An example interaction. In (a), the participant first uses speech, pointing, and eye gaze to ask for the red marker. Then the participant experiences one of three conditions: In (b), the no feedback control condition, the robot waits for more information before choosing. In (c), the physical feedback condition, the robot asks about the red marker via pointing. In (d), mixed reality feedback condition, the robot asks about the red marker via highlighting with a 3D sphere in mixed reality.

This work investigates how physical and visualization-based question-asking compare for reducing robot uncertainty under varying levels of ambiguity (Fig. 1). To do this, we first model our problem as a POMDP, termed the Physio-Virtual Deixis POMDP (PVD-POMDP), that observes a human's speech, gestures, and eye gaze, and decides when to ask questions (to increase accuracy) and when to decide to choose the item (to decrease interaction time). Then, we conduct a between-subjects user study, where 80 participants interact with a robot in an item-fetching task. Participants experience one of three different conditions of our PVD-

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

POMDP: a no feedback control condition, a physical feedback condition, or a mixed reality feedback condition. Our results show that our mixed reality model significantly outperforms the physical and no feedback models in both speed and accuracy, while also achieving the highest usability, task load, and trust scores.

II. Related Work

Previous research has investigated different communication modalities between robots and users, identifying the costs and benefits of each. A large amount of work has investigated physical robot actions used to reference objects to communicate with a human user, with two effective modes being robot eye gaze and robot pointing. Other research has opted instead to utilize a visualization-based approach, with visualizations displayed through 2D monitors, augmented reality, and mixed reality.

Eyes tend to move very quickly, and are used to both collect and communicate information. This makes eye-tracking a natural way to ground the references of other agents [1, 2, 7, 8, 12, 14, 15]. However, it is often difficult to perceive where an agent is looking, especially compared to pointing. Pointing is another natural deictic gesture that requires more effort but is easier to interpret. Admoni et al. [1] show that gaze and gesture are good at distinguishing between locationally unambiguous (far apart) items, while speech is good at distinguishing between visually unambiguous (different looking) items. However, related works [1, 2, 7, 8, 12, 14, 15] do not compare using eye gaze and pointing gestures to visualizations for reducing robot uncertainty.

Language has also been shown to be an effective means of symbol grounding, as in Chai et al. [5]. Their system enables users to use natural language to describe objects in the shared environment in order to ground them. The authors use a NAO robot with pointing and language to ask questions to clarify the human's references. Having the robot act in order to share its uncertainty to the human was shown to be important for establishing common ground. Like their work, we investigate pointing and language for disambiguation. However, we also investigate eye gaze, visualization, and question asking for mediating human-robot interaction.

Shridhar and Hsu [17] present an end-to-end system, INGRESS, to interpret unconstrained natural language commands for unconstrained object class references and perform question-asking. Their system outperforms state-of-the-art baselines, though they recognize that integration of nonverbal commands would help with requiring less complicated verbal references. Our approach, in contrast, uses a relatively simple language model, but also incorporates human gesture and eye gaze. Our model also allows the agent to ask questions via gesture, eye gaze, and visualizations for disambiguation.

Several related works have studied the usage of visual interfaces for improving communication in human-robot interactions [13, 18, 22, 23]. Sibirtseva et al. [18] perform a comparison of different visualization techniques for robot question-asking in an item-fetching domain. The authors use a semi-wizarded system to compare a 2D monitor interface, an augmented reality interface (fixed overhead projector), and a mixed reality interface for highlighting tabletop items. The authors found the mixed reality interface most engaging, but augmented reality most accurate and most preferred. They posit that technical limitations were to blame for the poor performance of MR. Our approach, in contrast, directly compares MR visualization to physical behaviors such as pointing and eye gaze. We do not compare to projector-based systems because they do not support eye-gaze tracking, whereas MR-HMDs do.

III. Background

MDPs, POMDPs, and the FETCH-POMDP are mathematical frameworks for modeling decision making. As these models form the base of our approach, we describe them further here.

A. Markov Decision Process

A Markov Decision Process (MDP) [3] is formalized as a tuple (S, A, T, R, γ). S is the set of states the agent can be in. A is the set of actions that the agent can take. T is the transition function that models the probability that performing an action a in state s lands the agent in state s': $T(s, a, s') = P(s'|s, a)$. R is the reward function that models cost of performing an action a in state: $R(a, s)$. γ is the discount factor for each subsequent action in the expected reward: $V(s) = \sum_{t=0}^{\infty} \gamma^t R_t$ where R_t is the reward obtained at time t.

B. Partially Observable Markov Decision Process

A Partially Observable Markov Decision Process (POMDP) [11] is a generalization of an MDP in which the agent does not know what state it is in, but instead maintains a belief distribution over possible states. More formally, a POMDP is defined as a tuple $(S, A, T, R, \Omega, O, \gamma)$, where all the previous definitions from MDP still apply, Ω is the set of observations, and O is the observation function that models the the probability of receiving observation o if the agent takes action a and lands in state s: $O(o, a, s) = P(o \mid a, s)$.

C. FETCH-POMDP

Whitney et al. [21] formulate a similar object-fetching task to ours as a POMDP called the FEedback To Collaborative Handoff Partially Observable Markov Decision Process (FETCH-POMDP). In the FETCH-POMDP, users are able use speech and pointing gestures to reference items on the table, and the robot is able to either point to an item to ask whether it was the desired item, wait, or pick an item that it believes is the desired item.

The authors evaluate the speed and accuracy of the FETCH-POMDP against a model with a fixed question-asking policy and a model that never asked questions, for both ambiguous and unambiguous settings. By asking questions only when it is confused, the FETCH-POMDP [21] increases interaction speed and accuracy compared to fixed

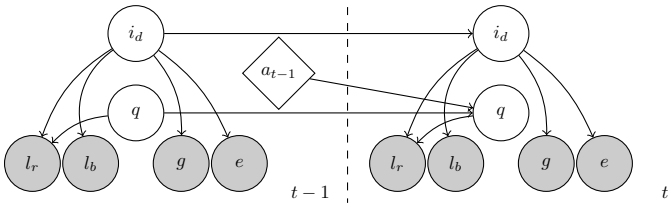


Fig. 2: A graphical model of our PVD-POMDP. Hidden variables are white, observed variables are gray (See Section IV-A for variable definitions).

question-asking policies across ambiguous and unambiguous contexts. However, the FETCH-POMDP is limited to only pointing gestures for question asking, and does not use eye gaze or MR visualizations. We hypothesize that MR visualizations are an objectively and subjectively better interface for question-asking in an item disambiguation task than physical behaviors.

IV. Technical Approach

We take a decision-theoretic approach to the item fetching problem by modeling our domain as a POMDP. This allows our robot to intelligently balance the informativeness and speed of its actions and gracefully handle its uncertainty. The Physio-Virtual Deixis Partially Observable Markov Decision Process, or PVD-POMDP, has as observations the speech, pointing gestures, and eye gaze of the user. Depending on the condition, the model enables our robot to look at, point to, and/or virtually highlight an item to ask if it is the desired item. The general intuition of our actions is that robot pointing is slower because the robot arm must move, but can be interpreted easily. Robot looking is faster because only the face and screen move, but is more difficult to interpret than pointing, especially when items are close together. MR visualizations are just as interpretable as pointing gestures because MR isolates items via highlighting, yet is faster to perform than robot looking because it requires no robot motion.

A. Model Definition

The PVD-POMDP[1] (Physio-Virtual Deixis POMDP) is given by components $\langle I, S, A, T, R, \Omega, O, \gamma \rangle$

- I is the list of all items on the table. Each item $i \in I$ has a known location (x, y, z) and set of associated words $i.\texttt{vocab}$.
- S: The state is (i_d, q). $i_d \in I$ is the human's desired item, which is hidden. q is the agent's last question, which is known. q is initialized to `null`.
- A: We divide the actions into two types: non-question-asking and question-asking. The non-question-asking actions are *wait* and *pick*(i) for $i \in I$. A *pick* action ends the interaction. The question-asking actions are *point*(i), *look*(i), and *highlight*(i) for $i \in I$. *look* is cheaper but less accurate than *point*, while *highlight* is cheaper than *look* and as accurate as *point*. However, *highlight* requires a MR-HMD for the user, while *look* and *point* do not require additional hardware than the robot itself.
- $T(s, a, s')$: i_d remains constant throughout an interaction. q is initialized to `null` and updated to a whenever a question-asking action a is taken.
- $R(s, a)$: The agent receives large positive and negative rewards for picking the right and wrong item respectively, and small negative rewards for all other actions. In decreasing magnitude of reward, the non-pick actions are *point*, *look*, *highlight*, *wait*. We calibrate these rewards roughly accordingly to how long each of the non-pick actions take: physical actions like *point* and *look* require physical robot behavior, thus take more time. *highlight* only needs to visualize on the MR-HMD, thus costs less. *wait* takes very little time.
- Ω: Each observation is composed of language, gaze, and gesture. Language is subdivided into base and response utterances. The response utterance can be positive, negative, or `null`.
- $O(o, s, a)$: The observation function can be factored into base utterance, response utterance, gaze, and gesture components. It is explained in detail in the Observation Model section below.
- γ: The discount factor is $\gamma = 0.99$.

B. Observation Model

Each observation o is a quadruple of base utterance l_b, response utterance l_r, gesture g, and eye gaze e. The components are assumed conditionally independent of each other given the state s (see Fig. 2):

$$\Pr(o \mid s) = \Pr(l_b \mid s) \Pr(l_r \mid s) \Pr(g \mid s) \Pr(e \mid s) \quad (1)$$

Following Goodman and Stuhlmüller [6], we assume each base utterance l_b has a literal interpretation probability $\Pr_{lex}(i_d \mid l_b)$ and that the speaker chooses their utterance by soft-max optimizing the probability that the listener infers the correct desired item from their base utterance. Each base utterance is interpreted as a vector l_b whose i^{th} component $l_b(i)$ is the number of words in the utterance that refer to the i^{th} object. Let U be the set of base utterance vectors and $|l_b| = \sum_{i \in I} l_b(i)$. Then we set:

$$\Pr_{lex}(i_d \mid l_b) = \begin{cases} \dfrac{(1-\alpha) l_b(i) + \alpha}{(1-\alpha)|l_b| + \alpha |I|} & |l_b| > 0 \\ \dfrac{1}{|I|} & |l_b| = 0 \end{cases} \quad (2)$$

where $\alpha = 0.02$ is a noise parameter. Let $p_l = 0.1$ be the probability a base utterance is made and $\theta = 15$ the soft-max parameter. Then:

$$\Pr(l_b \mid i_d) = \begin{cases} p_l \dfrac{e^{\theta \Pr_{lex}(i_d \mid l_b)}}{\sum_{l_b \in U} e^{\theta \Pr_{lex}(i_d \mid l_b)}} & |l_b| > 0 \\ 1 - p_l & |l_b| = 0 \end{cases} \quad (3)$$

[1]See supplemental video at https://www.youtube.com/watch?v=lXFv747b-Uc

When planning, we assume each base utterance will have at most three words to lower computation time.

The equation for $\Pr(l_r \mid s, a)$ has three components. The probability of receiving a response is $p_r = 0.6$. The probability that the human interprets the agent's question as asking about i if the agent is asking about j is $\Pr_*(i \mid j)$, which is defined in Equation 4 for *point* and *highlight*, and in Equation 6 for *look*. The probability that the human responds correctly based on their interpretation is $p_{rc} = 0.999$.

The human is assumed to always understand a *point* or *highlight* action, so the interpretation probabilities for pointing and highlighting are:

$$\Pr_p(i \mid j) = \Pr_h(i \mid j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \tag{4}$$

The interpretation probability for the *look* action uses a modified version of the model from Admoni et al. [1]. While humans have trouble identifying the exact angle of a *look*, they are very good at determining the general direction because of the robot's head motion, so we assume the human never mistakes a leftward *look* for a rightward *look* and vice versa.

Let $\text{ang}(i, j)$ be the angle between item i and item j relative to the robot's face, d_i the distance from the agent's face to item i, and $w_0 = 6$, $w_1 = 6$ noise parameters. Let $M(i, j)$ represent whether items i and j are on the same side of the robot:

$$M(i,j) = \begin{cases} 1 & i \text{ and } j \text{ are on the same side of the robot} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Then the probability $\Pr_l(i \mid j)$ that a human thinks the robot is looking at i when they are in fact looking at j is:

$$\Pr_l(i \mid j) \propto \frac{1}{d_i(1 + w_0|\text{ang}(i,j)|)^{w_1}} M(i,j) \tag{6}$$

Suppose the robot asked about item i using *point* or *highlight*. Then probability of receiving a response l_r is:

$$\Pr(l_r \mid s) = \begin{cases} p_r p_{rc} & l_r = \text{yes} \\ p_r(1 - p_{rc}) & l_r = \text{no} \\ 1 - p_r & l_r = \texttt{null} \end{cases} \tag{7}$$

Let $\Pr_l(i)$ denote $\Pr_l(i \mid i)$. If the robot asked about item i using *look*, then the probability of receiving a response l_r is:

$$\Pr(l_r \mid s) = \begin{cases} p_r(\Pr_l(i)p_{rc} + (1 - \Pr_l(i))(1 - p_{rc})) & l_r = \text{yes} \\ p_r(\Pr_l(i)(1 - p_{rc}) + (1 - \Pr_l(i))p_{rc}) & l_r = \text{no} \end{cases} \tag{8}$$

Human eye gaze e is modeled as a vector from the user's head to the point they are looking at. Gesture g is modeled as a vector from the user's the hand to the point they are pointing at. Angles are measured relative to the vector ending at the desired item. The probabilities of receiving a gaze or gesture are $p_e = 0.8$ and $p_g = 0.3$ respectively. When present, gaze and gesture are assumed to come from Gaussian distributions with mean 0 error and with and standard deviations $\sigma_e = 0.02$ and $\sigma_g = 0.06$ radians respectively:

$$\Pr(g \mid i_d) = \begin{cases} p_g \mathcal{N}(\theta_{i_d}; 0, \sigma_g^2) & g \neq \texttt{null} \\ 1 - p_g & g = \texttt{null} \end{cases} \tag{9}$$

$$\Pr(e \mid i_d) = \begin{cases} p_e \mathcal{N}(\theta_{i_d}; 0, \sigma_e^2) & e \neq \texttt{null} \\ 1 - p_e & e = \texttt{null} \end{cases} \tag{10}$$

A human's gaze is attracted to referenced items, so the robot ignores gaze observations for 1 second after asking a question.

Due to the differing noise models combined with a decision-theoretic approach, the robot considers pointing to be more costly than looking, and thus will only point at an item when the increased accuracy is worth the cost. Roughly speaking, the robot will look at an item if it is far enough away from other items that looking is unambiguous and will point at an item when it is in close proximity to other items.

C. Implementation Details

In order to observe the human's speech, we use Google's Cloud Speech to transcribe the user's speech. For gesture tracking, we use the Microsoft Kinect v2 in conjunction with OpenNI's skeleton tracker software, and calculate pointing vectors from the user's head to hand. Lastly, we use the Magic Leap One, a commercially avaliable MR-HMD, to track eye gaze.

We used Perseus, an offline POMDP planner from Spaan and Vlassis [19], as our planning algorithm. It took 6, 5191, and 724 seconds to train the control, physical, and mixed reality paradigms, respectively. Note that this training happens offline before the interaction begins, which enables the robot to act in real-time at run-time. Since human gesture and gaze are analogous, we planned using only gaze, but utilized both gaze and gesture during interaction.

For the user to understand which item the robot is asking about, the visualization presented to the user must isolate the referenced item from all the others. Our choice was to use a 3D sphere visualized over the referenced item. A sphere is the only fully rotationally invariant 3D shape, so it can be viewed equally well from all angles. We found that during our pilot studies (Section VI), users were less distracted when they moved, and generally looked directly at the item. We found 3D spheres to be the most highly regarded design in our pilot studies, and chose it as our final visualization method.

V. Evaluation

To evaluate our hypothesis, we designed an evaluation task where the robot disambiguated what item the human referred to as quickly and accurately as possible (defined in Section V-C) from an array of potential objects on a table in front of the robot. The aim of our evaluation was to investigate how communicating questions via physical robot behaviors, like looking and pointing, compare to communicating those questions via mixed reality visualizations.

We devised a user study to compare three conditions of communication modalities. In the *no feedback control condition*, the robot did not ask any questions and only decided to pick an item when it was sufficiently confident based on observations from the human (i.e: robot did not interact with the human except for when it picked an item). In the *physical feedback condition*, the robot was able to ask questions by moving, either using gesture or looking to reference items. In the *mixed reality (MR) feedback condition*, the robot was able to ask questions by visualizing a sphere over the referenced item in the user's mixed reality headset. We posited two hypotheses (H1 and H2) about the objective measures, and two hypotheses (H3 and H4) about the subjective measures:

- **H1:** The feedback conditions (physical and MR) will outperform the no feedback condition (control), as demonstrated by: (a) greater trial accuracy and (b) lower trial time.
- **H2:** The MR feedback condition will outperform the physical feedback condition, as demonstrated by: (a) greater trial accuracy and (b) lower trial time.
- **H3:** Users in the feedback conditions (physical and MR) will have a better user experience than users in the no feedback condition (control), as demonstrated by: (a) greater usability scores, (b) greater trust scores, and (c) decreased workload scores.
- **H4:** Users in the MR feedback condition will have a better user experience than users in the physical feedback condition, as demonstrated by: (a) greater usability scores, (b) greater trust scores, and (c) decreased workload scores.

A. Physical Setup

The physical setup of our experiment can be seen in Fig. 1. For the interaction, the human stood 2 meters away from a table with six items on it, and the robot stood on the other side of the table. Our item set consisted of three red expo markers, two glass cups, and one yellow rubber duck. The expo markers and glasses were identical except for their different spatial positions. The items were placed on the table in three groups of two, with the rubber duck and a marker on the far left, the two glasses in the middle, and the last two markers on the far right. The distances between the objects, from left to right, were 10cm, 40cm, 15cm, 45cm, and 10cm.

We chose the items and their locations to represent visually and spatially ambiguous scenarios. Specifically, the leftmost group is least ambiguous, as the duck is a unique item, and the marker is very far from its identical copies. The middle group is more ambiguous, as the two glasses are identical, and are somewhat close together. The rightmost group is most ambiguous, as the two markers are identical, and very close together.

The Microsoft Kinect v2 sensor was placed on top of the robot and calibrated to accurately track the pose of the human relative to the robot. The user wore the Magic Leap One HMD and headphones with a microphone in order to track the user's eye gaze and speech, respectively. The user heard the robot's question-asking through the headphones.

B. Experimental Procedure

Participants were randomly assigned to one of the three between-subjects conditions (no feedback control condition, physical feedback condition, MR feedback condition). After reading the IRB approved consent procedure, we calibrated the Magic Leap One for each user's eye gaze by using the supplied visual calibration program. We then went through the instructions for the study, and informed users there would be 18 trials with the robot. For each trial, the user was told an item number associated with an object and instructed to use speech, gesture, and eye gaze to reference the item to the robot in a clear and natural manner. If the user was in a condition with feedback, the user was told what feedback to expect from the robot (i.e., either physical or MR visualization-based question-asking). The experimenter then counted down from three to start the trial, at which point the user could reference the item; each trial ended when the robot selected an item or 30 seconds had passed. Every user was asked to reference each of the six items three times, totaling 18 trials. The order of items was randomly shuffled for each user. In each of the trials, we recorded the interaction time and whether the correct item was selected or not. After all 18 trials were completed, the user completed a series of subjective questionnaires.

C. Objective Measures

The performance of the robot in the task was evaluated using two objective measures, accuracy and time.

1) Accuracy: Accuracy was calculated as the number of correct items selected by the robot divided by the total number of trials (18 trials). We treated a trial timeout as an incorrect pick when calculating accuracy.

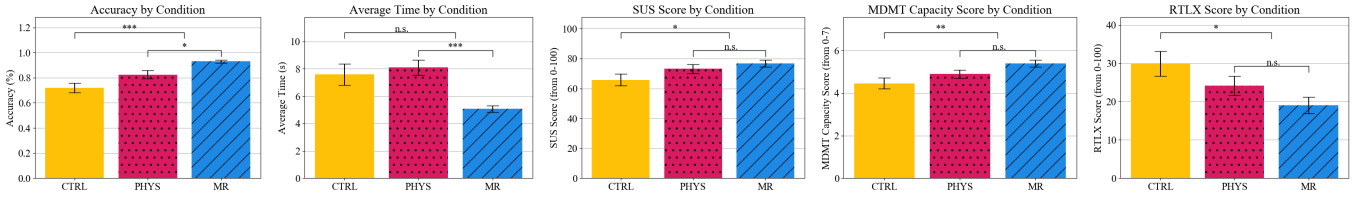
2) Time: Each trial began when the robot heard the user speak and ended when the robot picked an item; if the robot did not pick an item, the trial timed out after a 30 second period. The time measure was calculated as the average time of the interaction across all 18 trials.

D. Subjective Measures

Participants completed a series of three questionnaires to evaluate the success of the interaction on the basis of the perceived usability of the system, the trust in the robot on the task, and the task load of the interaction.

1) System Usability Scale: The System Usability Scale (SUS) is a versatile tool for assessing system usability developed by Brooke [4]. We use the SUS to evaluate user perceptions of usability of the robot system. The SUS consists of 10 Likert items, with usability scores for participants calculated by following the scoring guidelines for the SUS.

2) Multi-Dimensional-Measure of Trust: The Multi-Dimensional-Measure of Trust (MDMT) was developed by Ullman and Malle [20] to assess human trust in robots across tasks and domains. There are two superordinate dimensions of the MDMT: moral trust and capacity trust. For this study, we were interested in user evaluations of capacity trust in the robot. We used two of the four subscales from the MDMT:



(a) Accuracy (b) Time (c) SUS (d) MDMT (e) RTLX

Fig. 3: Our objective measures (a) Accuracy and (b) Time, and subjective measures (c) SUS, (d) MDMT, and (e) RTLX, shown for all three between-subjects conditions. Error bars represent stand error.

reliable (reliable, predictable, someone you can count on, consistent) and capable (capable, skilled, competent, meticulous). The MDMT consists of rating scales for each item from 0, "Not at all," to 7, "Very," with an option for "Does Not Fit." We calculated capacity scores for participants by averaging across ratings on these eight items.

3) NASA Task Load Index: The NASA Task Load Index (NASA-TLX) is an effective measure used across a variety of domains for the assessment of perceived workload [10]. We use the raw version of the NASA-TLX, often referred to as the RTLX, which is less burdensome than the original version and has been successfully utilized in numerous research studies [9]. We use the RTLX to evaluate the user workload associated with interacting with the robot system. The RTLX consists of six rating scales on different dimensions, with each scale spanning 0-100 in 5-point increments. We calculated workload scores for participants by averaging across ratings on all six scales.

VI. Results

Participants were recruited from the authors' academic institution, with participants required to be at least 18 years old and able to see without glasses (contacts were acceptable). We first conducted a pilot study with 10 participants to test the system. We then conducted the main study with a total of 83 participants. Three participants were excluded from data analysis (two for failure to follow study instructions, and one due to system technical error). Analysis was performed on the data from 80 participants: 27 in the no feedback control condition, 26 in the physical feedback condition, 27 in MR feedback condition. Please see Fig. 3 and Table I for data.

	Acc.	Time	SUS	MDMT	RTLX
CTRL	.72±.20	7.59±4.05s	65.74±20.34	4.45±1.33	29.88±16.74
PHYS	.82±.17	8.10±2.86s	73.27±15.66	4.90±0.99	24.13±12.84
MR	**.93±.07**	**5.07±1.25s**	**76.76±12.71**	**5.40±0.85**	**19.01±11.37**

TABLE I: A table of the means and standard deviations of all five of our metrics for all three conditions (CTRL = Control, P = Physical, MR = Mixed Reality). Bolded numbers are the best for that metric.

A. Objective Measures

The two objective dependent measures (accuracy, time) were correlated ($p < .001$) with each other, r = -.68. This correlation suggests that as accuracy increased, time for the task decreased. The correlation between the dependent variables also indicates that a multivariate analysis of the data is warranted to account for the relationship between the dependent variables.

A MANOVA was conducted using a pair of a priori orthogonal Helmert contrasts in order to test hypothesis H1 (that the feedback conditions would outperform the no feedback condition) and hypothesis H2 (that the MR feedback condition would outperform the physical feedback condition). An examination of the multivariate relationships of the data reveals strong support for hypothesis H1: the feedback conditions outperformed the no feedback condition. There was also strong support for hypothesis H2: the MR feedback condition outperformed the physical feedback condition.

The first Helmert contrast was significant and supports hypothesis H1, $F(2, 76) = 10.14$, $p < .001$, *multivariate* $\eta^2 = .21$. The univariate F-tests revealed that, compared to the no feedback condition, the feedback conditions were (a) higher on accuracy, $F(1, 77) = 17.69$, $p < .001$, $\eta^2 = .19$; and (b) not statistically significant different for time, $F(1, 77) = 2.21$, $p = .14$, $\eta^2 = .03$. These results indicate that the effect of increased performance in the feedback conditions is driven by higher accuracy.

The second Helmert contrast was significant and supports hypothesis H2, $F(2, 76) = 6.86$, $p < .01$, *multivariate* $\eta^2 = .15$. The univariate F-tests reveal that the MR feedback condition outperformed the physical feedback condition with (a) significantly higher accuracy, $F(1, 77) = 5.80$, $p = .02$, $\eta^2 = .07$; and (b) significantly lower time, $F(1, 77) = 13.91$, $p < .001$, $\eta^2 = .15$. These results indicate that the MR feedback condition was superior to the physical feedback condition.

B. Subjective Measures

The three subjective dependent measures (SUS, MDMT, RTLX) were all correlated (*p*s $< .001$) with each other: r = .65 for SUS and MDMT; r = -.60 for SUS and RTLX; and r = -.54 for MDMT and RTLX. These correlations suggest that usability and trust increase in tandem, and that workload decreases as both usability and trust increase. The correlations between the dependent variables also indicate that a multivariate analysis of the data is warranted to account for the relationships among the dependent variables.

A MANOVA was conducted using a pair of a priori orthogonal Helmert contrasts in order to test hypothesis H3 (that the feedback conditions would facilitate better user

experiences than the no feedback condition) and hypothesis H4 (that the MR feedback condition would facilitate better user experiences than the physical feedback condition). An examination of the multivariate relationships of the data reveals strong support for hypothesis H3: Feedback from the robot in both the physical and MR conditions facilitated better overall user experiences than no feedback. There was also a trend in the data consistent with hypothesis H4: MR feedback facilitated better user experiences than physical feedback. The means and standard deviations of all three subjective metrics for each condition are shown in Figure 3.

The first Helmert contrast was significant and supports hypothesis H3, $F(3, 75) = 3.13$, $p = .03$, *multivariate* $\eta^2 = .11$. The univariate F-tests revealed that the feedback conditions were rated as (a) significantly better on usability, $F(1, 77) = 5.66$, $p = .02$, $\eta^2 = .07$; (b) significantly higher on trust, $F(1, 77) = 7.56$, $p < .01$, $\eta^2 = .09$; and (c) significantly lower on workload, $F(1, 77) = 6.50$, $p = .01$, $\eta^2 = .08$. These results offer strong support for hypothesis H3.

The second Helmert contrast was not significant, $F(3, 75) = 1.19$, $p = .32$, *multivariate* $\eta^2 = .05$. However, the means of the measures are consistent with hypothesis H4, with ratings in the MR feedback condition greater on usability and trust than in the physical feedback condition, as well as lower on workload. None of the univariate F-tests were statistically significant, but workload and trust had noteworthy effect sizes of 2-4% explained variance: $F(1, 77) = 0.59$, $p = .45$, $\eta^2 = .01$ for usability; $F(1, 77) = 2.86$, $p = .10$, $\eta^2 = .04$ for trust; and $F(1, 77) = 1.81$, $p = .18$, $\eta^2 = .02$ for workload. Given the interesting trend but insufficient statistical confidence, future work will aim to elucidate whether there is in fact a qualitative difference between the two feedback conditions along subjective measures.

We gain some additional insight from the MANOVA by examining the semi-partial coefficients (discriminant function weights) for the three user experience measures. The semi-partial coefficients are like weights in a multiple regression and indicate which of the three measures most strongly discriminates between the conditions. When contrasting feedback to no feedback, MDMT (trust) makes the strongest contribution (.57), RTLX (workload) also makes a notable contribution (-.46), but SUS (usability) makes little unique contribution (.15) above and beyond RTLX and MDMT. Taken together, while the three measures show high correlations and share some ability to discriminate between the feedback and no feedback conditions, the MDMT is able to stand by itself as a parsimonious tool to capture user attitudes towards a robot. This is perhaps because it is a user-friendly measure, derived from natural language people use in the domain of trust [20].

VII. Discussion

The results from the objective and subjective measures in our user study paint a single, coherent story about the conditions we tested. In general, the feedback conditions (physical, MR) outperformed the no feedback condition, and the MR feedback condition (control) outperformed the physical feedback condition. The user experience of each condition roughly paralleled the performance of the system. Ultimately, we conclude that models that integrate feedback perform better and are preferred by users, and that MR is a promising modality for this communication.

In terms of objective measures, the feedback conditions (physical, MR) were more accurate than the no feedback condition (control), as was the MR condition compared to the physical condition. While the MR condition averaged less time than the physical condition, the time difference between the feedback condition and the no feedback condition was not statistically significant; this appears to stem from the reduced speed of the physical condition, which required extra time for the robot to move its end effector to offer feedback. The results thus fully support hypothesis H2 (MR feedback condition compared to physical feedback condition on objective measures), with nuanced support for hypothesis H1 (feedback condition compared to no feedback condition on objective measures). Remarkably, the MR condition was simultaneously the most accurate and the fastest, contrary to the typical speed-accuracy tradeoff. These results show particular promise for the MR feedback model, which appears to exhibit the best performance in terms of both accuracy and speed.

The subjective measures on user experience offer a similar story. Participants gave better user experience ratings across all three subjective measures (usability via SUS, trust via MDMT, workload via RTLX) in the feedback conditions (physical, MR) as compared to the no feedback condition (control). Although there was no statistically significant difference between the user experience ratings in the MR feedback condition and the ratings in the physical feedback condition, the means across all three subjective measures improve from no feedback to physical feedback, and again from physical feedback to MR feedback. As a result, we believe that the benefits of MR are worth exploring further in future work. The results thus fully support hypothesis H3 (better user experience in feedback conditions compared to no feedback condition on subjective measures), with trending support for hypothesis H4 (better user experience in MR feedback condition compared to physical feedback condition on subjective measures)

VIII. Conclusion

This work presents a robot interaction model that is able to interpret multimodal human communication and use a mixed reality interface to perform question-asking in an item disambiguation task. We approach our problem from a decision-theoretic standpoint, and ultimately offer our new model called the Physio-Virtual Deixis (PVD) POMDP. Lastly, we report the results of our user study, which compared two feedback conditions (physical, MR) to a no feedback condition, as well as compared the physical and MR feedback conditions to each other. We found statistically significant support along both objective and subjective measures in

favor of conditions that offer feedback (physical, MR) over no feedback (control), as well as statistically significant support from objective measures (and trending support from subjective measures, though not significant) in favor of a MR feedback condition over a physical feedback condition.

REFERENCES

[1] Henny Admoni, Thomas Weng, and Brian Scassellati. Modeling communicative behaviors for object references in human-robot interaction. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 3352–3359. IEEE, 2016.

[2] Adrian Bangerter. Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science*, 15(6):415–419, 2004.

[3] Richard Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, pages 679–684, 1957.

[4] John Brooke. SUS-a quick and dirty usability scale. *Usability Evaluation in Industry*, 189(194):4–7, 1996.

[5] Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. Collaborative effort towards common ground in situated human-robot dialogue. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, pages 33–40. ACM, 2014.

[6] N.D. Goodman and A. Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5, 2013.

[7] Stephanie Gross, Brigitte Krenn, and Matthias Scheutz. The reliability of non-verbal cues for situated reference resolution and their interplay with language: Implications for human robot interaction. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 189–196. ACM, 2017.

[8] Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. Eye gaze for spoken language understanding in multi-modal conversational interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 263–266. ACM, 2014.

[9] Sandra G Hart. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 904–908. Sage Publications Sage CA: Los Angeles, CA, 2006.

[10] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier, 1988.

[11] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134, 1998.

[12] Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. Exploring a model of gaze for grounding in multimodal HRI. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 247–254. ACM, 2014.

[13] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. Situated language understanding with human-like and visualization-based transparency. In *Robotics: Science and Systems*, 2016.

[14] Zahar Prasov and Joyce Y Chai. What's in a gaze?: The role of eye-gaze in reference resolution in multimodal conversational interfaces. In *Proceedings of the 13th International Conference on Intelligent User interfaces*, pages 20–29. ACM, 2008.

[15] Pernilla Qvarfordt. Gaze-informed multimodal interaction. In *The Handbook of Multimodal-Multisensor Interfaces*, pages 365–402. Association for Computing Machinery and Morgan & Claypool, 2017.

[16] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. Communicating robot arm motion intent through mixed reality head-mounted displays. *arXiv preprint arXiv:1708.03655*, 2017.

[17] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *Proceedings of Robotics: Science and Systems*, 2018.

[18] Elena Sibirtseva, Dimosthenis Kontogiorgos, Olov Nykvist, Hakan Karaoguz, Iolanda Leite, Joakim Gustafson, and Danica Kragic. A comparison of visualisation methods for disambiguating verbal requests in human-robot interaction. *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 43–50, 2018.

[19] Matthijs T. J. Spaan and Nikos A. Vlassis. Perseus: Randomized point-based value iteration for POMDPs. *CoRR*, abs/1109.2145, 2011. URL http://arxiv.org/abs/1109.2145.

[20] Daniel Ullman and Bertram F Malle. What does it mean to trust a robot?: Steps toward a multidimensional measure of trust. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 263–264. ACM, 2018.

[21] David Whitney, Eric Rosen, James MacGlashan, Lawson LS Wong, and Stefanie Tellex. Reducing errors in object-fetching interactions through social feedback. In *International Conference on Robotics and Automation, Singapore, May*, 2017.

[22] Tom Williams, Nhan Tran, Josh Rands, and Neil T Dantam. Augmented, mixed, and virtual reality enabling of robot deixis. In *International Conference on Virtual, Augmented and Mixed Reality*, pages 257–275. Springer, 2018.

[23] Tom Williams, Matthew Bussing, Sebastian Cabroll, Elizabeth Boyle, and Nhan Tran. Mixed reality deictic gesture for multi-modal robot communication. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 03 2019.