Situated Language Understanding with Human-like and Visualization-Based Transparency

Leah Perlmutter[1], Eric Kernfeld[2] and Maya Cakmak[1]
[1]Computer Science & Engineering Department [2]Department of Statistics
University of Washington, 98195, Seattle, Washington, USA

***Abstract*—Communication with robots is challenging, partly due to their differences from humans and the consequent discrepancy in people's mental model of what robots can see, hear, or understand. Transparency mechanisms aim to mitigate this challenge by providing users with information about the robot's internal processes. While most research in human-robot interaction aim towards natural transparency using human-like verbal and non-verbal behaviors, our work advocates for the use of visualization-based transparency. In this paper, we first present an end-to-end system that infers task commands that refer to objects or surfaces in everyday human environments, using Bayesian inference to combine scene understanding, pointing detection, and speech recognition. We characterize capabilities of this system through systematic tests with a corpus collected from people (N=5). Then we design human-like and visualization-based transparency mechanisms and evaluate them in a user study (N=20). The study demonstrates the effects of visualizations on the accuracy of people's mental models, as well as their effectiveness and efficiency in communicating task commands.**

I. INTRODUCTION

One of the key interactions between a human and a function-oriented robot is *commanding* the robot to perform a useful task. This interaction can be simplified to the press-of-a-button for single-purpose robots like the vacuum cleaner robot Roomba. But for multi-purpose complex robots, the space of possible tasks is unbounded. As a result, commanding the robot requires much richer communication, including the ability to reference the environment. For example, users of a robot that can fetch and deliver objects need to be able to refer to an arbitrary target object and delivery point. Similarly, commanding a robot that can use different cleaning tools on different surfaces requires the ability to specify a tool (*e.g.,* a feather duster) and a target surface (*e.g.,* window sills).

Enabling this type of rich communication between humans and robots is still challenging. Part of the challenge can be ascribed to errors and ambiguities in natural language processing, visual scene understanding, and recognition of non-verbal signals from humans. But more importantly, this challenge is magnified by people's inability to judge what robots can see, hear, or infer, whether or not they are correct and accurate. Our research addresses this problem of *mental model discrepancy*.

Human-human communication relies on several mechanisms to avoid, detect, and mitigate mental model discrepancies. For example, people use gaze, gestures, and facial expressions to indicate their understanding while listening to someone. They also verbally restate their understanding to confirm it. Similar mechanisms can help human-robot communication, by providing *transparency* about the robot's perception and inference processes. In fact, a large body of human-robot interaction research tackles the problem of designing human-like verbal and non-verbal communication behaviors for robots.

Fig. 1: LUCIT combines *speech recognition*, *pointing detection*, and *scene understanding* to interpret situated natural language commands. LUCIT's transparency mechanisms allow the user view to the output of these intermediate internal processes.

While human-like transparency mechanisms can capitalize on people's natural ability to interpret them, they are limited in how much they can express about the robot's complex internal processes. In this paper we explore an alternative transparency channel that is not afforded in human-human communication: *visualizations*. While visualizations may feel less natural, they are extremely powerful tools for communicating computational concepts. Many researchers and engineers rely on them while developing new robotic capabilities. The question we ask in this work is whether naive users can benefit from simple visualizations that reveal what a robot can see, hear, and infer.

In this paper, we first develop an end-to-end situated language understanding system, LUCIT, that combines scene perception, pointing detection, and speech processing. The system enables users to command a robot to perform tasks referencing objects and surfaces in realistic room-scale scenes. We characterize LUCIT's capabilities through systematic tests with data collected from 5 users. Next, we design both human-like (speech, gaze, pointing) and visualization-based transparency mechanisms for LUCIT. Through a user study with 20 participants we demonstrate the benefits and characterize the use of visualization-based transparency mechanisms.

II. RELATED WORK

Situated language understanding: Our work aims to enable users to communicate robot tasks that reference the environment. Previous work has contributed a number of approaches towards this goal. Many combine computer vision methods with alternative language understanding techniques to resolve references to the scene [10, 17, 23, 25]. Other work incorporates understanding of actions or tasks to be performed in the referenced environment [3, 9, 24, 30, 34]. Researchers have also incorporated perception of human gesture and gaze to enhance verbal language understanding [8, 20, 22, 31]. Others propose to use *dialogue* to collaboratively and incrementally move towards a common ground about a reference [1, 2, 5, 7, 14, 18]. Our system combines some of these elements in a Bayesian framework with scene and human gesture perception and different transparency mechanisms. We focus on tasks in room-scale everyday scenes, while most previous work involves manipulation on narrow tabletop scenes or navigation on a map.

Mental Models and Transparency in HRI: A large body of research in the HRI community studies verbal and non-verbal communication with robots (*e.g.,* [11]). Of particular interest to our paper are those focusing on how people's mental model of a robot influences the way they talk or gesture to a robot. Kiesler discusses the influence of robot design and social cues on people's mental models of the robot [15]. Similarly, Fischer demonstrates the role of people's preconceptions in how they talk to robots [6]. Vollmer *et al.* show that people modify their actions when demonstrating a task to a child-like robot [37]. A robot's speech can also influence how people speak to it in consecutive interactions; for example Iio *et al.* demonstrate that people align their lexicon to that of a robot [13]. Their previous work demonstrated a similar entrainment effect in pointing gestures [12]. Other work provides characterizations of how people are likely to naturally talk to robots [16, 29, 36].

Researchers have also demonstrated that mental model discrepancies can lead to interaction challenges [33]. Closely related to our work, Liu and Chai find that perceptual differences between humans and robots result in communication challenges [21]. To mediate these differences, they present a graph-based representation of the robot's perceived environment and propose to use this graph in reference resolution, with learned weights that promote matched perceptual attributes.

Transparency mechanisms have been shown to mitigate the challenges of mental model discrepancy. Pejsa *et al.* use facial expressions to provide transparency about dialog uncertainties [28]. Crick *et al.* show that people give task demonstrations that are more usable for the robot if they are given robot-like perception of the environment [4]. Our visualization-based transparency mechanisms are intended to have the same effect by revealing the robot's limitations to the user. Thomaz and Breazeal show that natural transparency mechanisms like gaze can steer the human's behavior while demonstrating a task [35]. Several others have shown positive effects of transparency mechanisms in HRI contexts [27, 32].

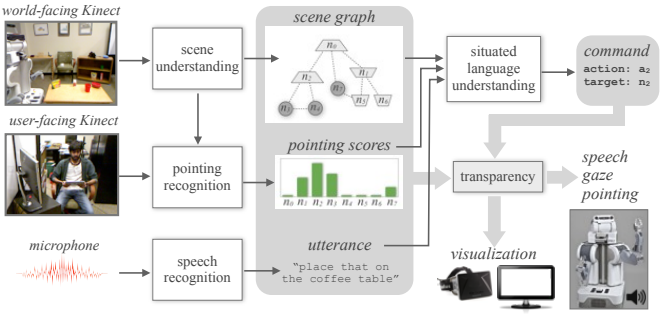


Fig. 2: Overview of the LUCIT system.

Novel interfaces for human-robot communication: While a majority of work in the HRI community is driven by the human-human interaction metaphor, some researchers have proposed clever ways in which affordances of a robot that are different from those of humans can be leveraged for effective communication. We consider our work to belong in this category. Some examples include Nguyen *et al.*'s laser pointer based interface for referencing objects in the environment [26] and the use of projections by the robot as a way to highlight parts of the environment [19].

III. SITUATED LANGUAGE UNDERSTANDING

We focus on the scenario of commanding a household robot to do tasks within a room-scale environment. The robot is a mobile manipulator capable of navigating the room, picking up and placing objects and tools from and to horizontal surfaces, and applying different cleaning tools on surfaces. Hence the robot has three parametrized actions that can be commanded:

- a_1: `pickup(`τ`)` where τ is an object,
- a_2: `place(`τ`)` where τ is a surface,
- a_3: `clean(`τ`)` where τ is a surface.

Robot actions are applicable in certain contexts. The `pickup` action requires the robot's hand to be free. The `place` action requires the robot's hand to contain an object (*i.e.,* a previously picked up object). The `clean` action requires the robot's hand to contain a particular type of object, namely a cleaning tool. The goal of our system is to infer the action and the parameter being commanded by a user.

Our system has three main perceptual components: *scene understanding*, *pointing recognition*, and *speech recognition*. Our *situated language understanding* method combines the output of these perceptual processes to infer the commanded action and its parameter. The transparency mechanisms reveal information about the three perceptual components as well as the final output of the system. We refer to our system as LUCIT which stands for "**L**anguage **U**nderstanding with **C**omplex **I**nference **T**ransparency". An overview of LUCIT is shown in Fig. 2. The next sections give more detail about each component of LUCIT.

A. Perceptual components

1) Scene understanding: LUCIT uses a fixed RGBD sensor to perceive the environment. We assume that the robot is

familiar with the scene and is given a prior distribution of surfaces of interest. Surfaces are detected more precisely using planar RanSaC. Individual objects on the surface are detected using K-means clustering. In our experiments, we assume the robot knows which objects are present in the scene and where each object should be at each step of the task sequence[1].

Once all surfaces and objects are detected the robot computes the relationships between all detected entities to create the scene graph. Similar to Liu and Chai's vision graph [21], we represent the scene with an attributed relational graph in which each node and edge can have attributes attached to it. The scene graph $G = (N, E)$ is a directed graph with p nodes $N = \{n_i | i = 1, .., p\}$ corresponding to entities detected in the environment and q edges $E = \{e_j = (n_0, n_1)_j | j = 1, .., q; n_0, n_1 \in N\}$ corresponding to relationships between those entities. Attributes associated with each node include a type $\tau \in \{\tau_s, \tau_o\}$ (τ_s:surface, τ_o:object) and a set of keywords (nouns and adjectives) that might be used to refer to the object directly. Attributes associated with edges include a set of keywords (prepositions like *on*, *above*, or *next to*) that might be used to describe the relationship between the two entities.

2) Pointing recognition: Humans often use deictic gestures to refer to entities in the environment. To harness this additional communication channel, LUCIT takes input from another RGBD sensor facing the user to track them, detect when they are pointing, and infer the target of their pointing given the layout of the scene. To enable association between pointing gestures and entities in the environment, we register the user-facing and scene-facing RGBD sensors in the same coordinate frame through a one-time calibration procedure. We use ROS openni_tracker[2] to track the person's skeleton and obtain pointing directions by drawing vectors from the user's head to each hand. Next, we trace a ray in the direction of these vectors to assign weights to entities in the environment based on their shortest distance to the pointing rays. The robot assumes that a pointing gesture is happening when a skeleton is detected and any point on a pointing ray is within a threshold distance to some object or surface in the scene.

3) Speech recognition: For speech recognition we use the Python speech recognition package[3] to capture pause-separated utterances from a microphone and the Google speech API to transcribe each utterance into text.

B. Situated Language Understanding

The language understanding problem addressed in this work is the inference of the intended command (an action and its parameter) given the state of the environment and the user's utterance and gesture. We formulate this inference problem as a Bayes Network, shown in Fig. 3. We represent the distributions over actions and targets (*i.e.,* parameters of the action) with two discrete random variables A and T. We represent the world state with the random variable W as a distribution over possible scene graphs (Sec. III-A1). We define L as the distribution over possible utterances and G over possible pointing gestures. Finally, we represent context with the variable C to capture the preconditions that affect whether an action is applicable. We model language and gesture production as generative processes with the influence relationships shown in Fig. 3.

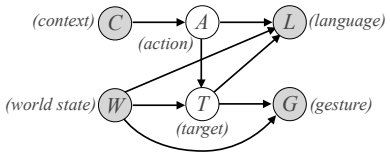


Fig. 3: Bayesian network that models the language and gesture generation process.

The problem can be stated as inferring distributions over unknowns given the observables:

$$a^*, t^* = \arg\max_{A,T} P(A, T | W, C, L, G) \quad (1)$$

Using the Bayes rule and the chain rule (based on the influence structure of the defined Bayes Network), we obtain:

$$a^*, t^* = \arg\max_{A,T} \big[P(A|C)P(T|W, A) P(G|T, W)P(L|A, W, T)\big] \quad (2)$$

This decomposition yields four components that are intuitive to model. We discuss each of them next.

1) Action-context model: The simplest part of our model is the conditional probability distribution $P(A|C)$ which is a uniform distribution over all actions whose preconditions are met in a given context as discussed in Sec. III. When the robot's hand is empty, the only action possible is `pickup`. As we do not currently differentiate cleaning tools from other objects, both `place` and `clean` are possible when the robot has something in its hand.

2) Parameterization model: Each action type (`pickup`, `place`, or `clean`) has a single parameter of particular type (`object` or `surface`). Given an action a_i and the state of the world G, we model $P(T|W, A)$ as a uniform distribution over all nodes of G whose type attribute τ is equal to the parameter type of a_i.

3) Gesture model: We represent pointing gestures as normalized vectors in 3D (specified by a pitch and a yaw) indicating the direction of pointing from the user's known location. We model the distribution over pointing gestures given the world state and the target $P(G|W, T)$ as a binomial Normal distribution whose mean is the direction of the ray from the user to the target and covariance is a diagonal matrix with empirically determined standard deviations.

4) Language model: Lastly we model the distribution over possible utterances given the state of the world and the intended action and target. For computational simplicity, rather than explicitly modeling a probability density function over all possible utterances, we approximate $P(L|A, W, T)$ with function $\phi(a, t, \ell)$ that scores an observed utterance ℓ for a

[1]This is done to avoid challenges with real-time room-scale object recognition which is outside the scope of this work.

[2]http://wiki.ros.org/openni

[3]https://pwww.webpage.orgypi.python.org/pypi/SpeechRecognition/

candidate action-target pair. We expect that an utterance will contain a verb associated with the commanded action and a noun phrase describing the target. The noun phrase can include keywords (names or adjectives) associated with the target and, possibly, prepositional phrases that specify spatial relationships of the target to other entities in the environment. Hence, we decompose the scoring function as:

$$\phi(a,t,\ell) = \phi_{verb}(a,\ell) + \phi_{noun}(t,\ell) + \sum_{t' \in R^t} \phi_{noun}(t',\ell)$$

where $\phi_{verb}(a,\ell)$ and $\phi_{noun}(t,\ell)$ are the number of occurrences of keywords associated with action a or target t in utterance ℓ; and R^t is the set of nodes in the scene graph G that have at least one edge connected to the node corresponding to target t. We do not check for the particular relationship between t and t', which proved to be unnecessary in practice.

IV. TRANSPARENCY MECHANISMS

Transparency mechanisms allow people to view additional information, besides the system's intended output, about its internal processes. LUCIT's final inference about the user's command relies on several intermediate components, highlighted in Sec. III. We implement two types of transparency mechanisms for these components, described in the following.

A. Platform

Although language understanding capabilities of LUCIT are platform-independent, transparency mechanisms can depend on the particular robot's embodiment. The robot used in this work is a PR2; a mobile manipulator with two arms and an omnidirectional base. Although not particularly anthropomorphic, PR2's 7 DOF (degrees-of-freedom) arms and 1 DOF gripper make it possible to replicate human-like pointing gestures. PR2 has a pan-tilt sensor head enables interpretable head gestures and gaze. For visualization-based transparency, we add a screen and an Oculus headset to the platform.

B. Human-like transparency

We implement three human-like transparency mechanisms.

1) Speech: The first mechanism is verbal confirmation; a common method in human-robot dialog (Sec. II). The robot generates a sentence to describe a command based on the simple template "You would like me to `verb-phrase noun-phrase`, is that correct?" The `verb-phrase` is a fixed verb phrase associated with the inferred action (a_1:"pick up", a_2: "place the object on", a_3: "clean"). The `noun-phrase` is a fixed keyword associated with the inferred target (*e.g.,* "the feather duster," "the top shelf of the bookshelf," "the chair"). Verbal confirmations provide partial transparency into the robot's speech recognition and language understanding components.

2) Pointing: As part of command confirmations the robot points to the target of the inferred command. Pointing gestures are produced by blending four arm gestures corresponding to four corners of a scene segment. The blending weights are determined by the position of the target relative to the segment

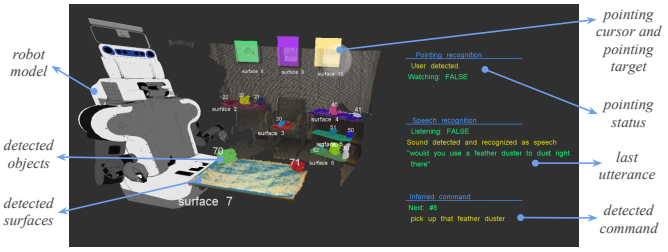


Fig. 4: LUCIT's visualization which provides additional transparency about its internal processes including speech recognition, pointing detection, scene understanding, and command inference.

corners. Pointing provides additional transparency into the robot's scene understanding component by associating a noun phrase with a spatial location.

3) Gaze: The robot's pan-tilt head can be pointed at any point in the 3D scene. We use this channel for additional transparency about the user's pointing gestures by turning its head towards the pointed target during confirmation.

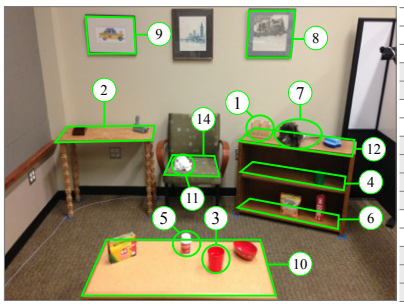
C. Visualization-based transparency

The visualization-based transparency of LUCIT has four components (highlighted in Fig. 2) that correspond to the outcomes of the three intermediate perceptual modules and the language understanding module. A snapshot of the visualization designed for this purpose is shown in Fig. 4, with four components highlighted. The 3D view of the interface displays a point cloud snapshot from the scene-facing camera. Colored planes and point cloud segments overlay the surfaces and objects detected in the scene. The unique IDs given to surfaces and objects (*e.g.,* "surface 7") appear near the targets. Pointing or gaze vectors appear as mobile cursors, and the pointing target is highlighted by lightening its color. The top right display indicates whether the human is detected and whether the robot is paying attention to pointing. The middle right display indicates whether the microphone is listening or off and it displays the exact utterance obtained from the speech recognition module. The bottom right display shows the inferred command, phrased using the `noun-phrase` and `verb-phrase` template described in Sec. IV-B1.

We explore two modalities for visual transparency: a regular 2D screen placed near the user and an immersive virtual reality (VR) headset. The elements of the interface described above are mostly the same in both interfaces, with the following differences. On the 2D screen, two large, round cursors (one for each hand) indicate pointing direction, whereas pointing with arms is disabled on the VR headset. Instead, the user can point to objects by looking at them to center them on the VR headset screen. This screen includes a smaller, cross-shaped cursor at the center. In addition, text placement is optimized for readability in each interface.

V. SYSTEM EVALUATION

We first characterize the performance and capabilities of LUCIT with data collected from people using only human-like transparency mechanisms.



#	action	target	batch
1	pick-up	*wooden box*	1
2	place	*side table*	
3	pick-up	*red cup*	2
4	place	*middle shelf*	
5	pick-up	*pill bottle*	3
6	place	*bottom shelf*	
7	pick-up	*feather duster*	4
8	clean	*house picture*	
9	clean	*taxi picture*	
10	place	*coffee table*	
11	pick-up	*white t-shirt*	5
12	place	*top shelf*	
13	pick-up	*feather duster*	6
14	clean	*chair*	
15	place	*top shelf*	

Fig. 5: The scene used in our system evaluation (Sec. V) and user evaluation (Sec. VI) in its initial configuration and the 15 commands that participants gave to the robot with targets (object or surface) annotated on the scene.

Procedure: The data collection involves 4 sessions in which the participant gives 15 pre-specified commands (Fig. 5) to the robot. The commands are described to participants visually with an image showing the scene and the robot, both before and after the command is executed. The target object or surface is highlighted in these images (possibly in the robot's gripper). The commands are grouped into 6 batches (combinations of pickup-place or pickup-clean*-place). After giving the commands in each batch, the participant is asked to demonstrate the actions corresponding to that batch, before moving on to the next batch. This is done to vary the scenes in which commands are given without having the experimenter interrupt the flow of the interaction. To further increase the variety in our data participants are given different instructions in each session. In the first two sessions they are told not to use pointing and in the last two they are asked to use it. In the first and third sessions they are told to refer to the object directly and in the other sessions they are asked to not do that.

Measurements: We characterize the performance of LUCIT with the accuracy of (a) *speech recognition* (measured by the Levenshtein distance between the ground truth utterance and the speech recognizer output); (b) *spoken target* interpretation (whether the correct target was identified based on speech); (c) *pointed target* interpretation (whether the correct target was identified based on pointing); and (d) the overall *inferred target*.

Results: Data was collected from 5 participants, for 15 commands, in 4 sessions, resulting in 300 trials (150 with pointing). Some trials involved multiple attempts before confirmation, yielding a total of about 620 utterances. Table I presents the measurements from this data.

TABLE I: Summary of system performance (mean of different accuracy metrics) in 620 utterances across four sessions (S1-4).

metric	S1	S2	S3	S4
speech	4.53	4.24	2.42	2.77
spoken target	90.12%	52.10%	78.70%	23.15%
pointed target	N/A	N/A	7.41%	14.81%
inferred target	88.89%	52.10%	64.81%	56.48%

We observe that the average accuracy of LUCIT's combined target identification has a large variance across sessions, ranging between about 50% and 90%. Note that 50% accuracy means that providing the correct command took about 2 attempts on average, and 90% accuracy means the first attempt was successful most of the time. The data reveals errors in speech recognition (Levenshtein distance around 2.5 to 4.5) which are common with off-the-shelf generic speech recognizers. Despite these errors, LUCIT can infer the referred target purely from speech. The accuracy is higher when participants refer to the object directly by naming it (sessions 1 and 3). In contrast, it took a greater number of failed attempts to verbally refer to an object without directly naming it (sessions 2 and 4), *e.g.,* by using references to nearby objects.

Pointing on its own was sufficient for identifying the target in very few attempts. Often times the pointing detector failed to track the person or it was inaccurate. Combining speech and pointing resulted in worse performance in session 3 and better performance in session 4. In some cases, pointing appeared to have distracted people from accurately specifying the target with speech (*e.g.,* "pick that up" "place it over there"), without offering an accurate alternative. In other cases where speech and pointing alone each underspecified the command, LUCIT was able to infer the correct command by combining them. For example saying "Pick up the cup" (ambiguous about which cup – red or green) combined with pointing that is inaccurate but relatively closer to the red cup, allows the system to infer that the correct command is to pick up the red cup. Participants were able to refer to targets indirectly with speech, even though it took more attempts than referring directly. For example, the target in the utterance "Put it down next to the green cup" was correctly inferred as the middle shelf of the bookshelf, based on the relationship between the "green cup" and the "middle shelf" in the scene graph.

Overall, the system evaluation demonstrated that LUCIT enables interpretation of multi-modal commands that reference the environment; however, there is room for improvement particularly in the interpretation of more complex referential commands and commands that incorporate deictic gestures.

VI. USER EVALUATION

Next, we conducted a user study to evaluate the impact of visualization-based transparency mechanisms.

A. Study Design

Our study has three conditions:

1) **Baseline:** Only human-like transparency mechanisms.
2) **Screen:** Visual transparency with a screen in addition to human-like transparency.
3) **VR:** Visual transparency with an immersive VR headset in addition to human-like transparency.

We use a within-participants design (*i.e.,* all participants interact with all conditions). All participants first interact with the baseline condition. The order of the other conditions is counter-balanced. Participants perform the same set of tasks (*i.e.,* providing 15 commands in 6 batches) in each condition (Fig. 6), in the same way as our data collection (Sec. V).

The sessions for our experimental conditions are split into two parts. In the first part the participant gives the first 4 batches of commands (*i.e.,* first 10 commands), with the transparency mechanisms specific to that condition. In the second part, the participant completes the last 2 batches (*i.e.,* last 5 commands) with the transparency mechanism taken away, defaulting back to the human-like transparency. This is done to distinguish between real-time effects of the transparency mechanisms (**P1**) and their potential after-effects (**P2**).

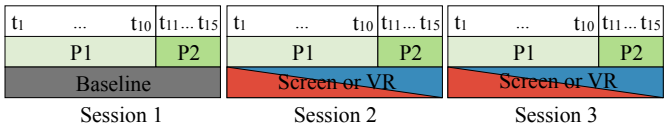


Fig. 6: Ordering of sessions, conditions, session parts, and tasks in our study design.

Note that studying after-effects requires the ability to compare performance on the same set of tasks *before* and *after* adding the visualizations. Hence we keep the baseline condition always in the first session. Counterbalancing all three conditions would have placed the baseline condition after visual transparency conditions for some users, which would have resulted in the baseline condition being subject to after-effects. Our data analysis compensates for this study design by making assumptions about learning effects across sessions.

B. Procedure

After completing a consent form, participants are taken to the study area and seated facing the scene (Fig. 5) and are introduced to the tasks of giving verbal commands and demonstrations. We ask them to practice giving the first command while the robot is off. Next they begin interacting with the robot in the first condition. After each session, participants complete a questionnaire specific to the condition. Before the two experimental conditions, participants are given additional instructions about the components of the visual interface. After all sessions, participants respond to additional questions and provide demographic information.

C. Interaction

During the study the robot is controlled by a finite-state-machine with the following behavior. At the beginning of each batch the robot says "Taking a minute to look around the room," gazes towards the scene, and updates its scene graph. Then it turns to the participant, requests a command by saying "What should I do next?", and activates the microphone for the response. If an invalid utterance (*e.g.,* silence) is detected, the robot speaks an error message (*e.g.,* "I could not hear you."), and repeats the request for a command. When a valid utterance is detected, the robot infers the command using the scene and the pointing information, and invokes a confirmation (Sec. IV-B1). If the user confirms the command, the robot moves on to the next command or demonstration. Otherwise, it requests the command again by repeating "What should I do next?" If it cannot infer the action type it says "I did not hear an action word or a verb." If multiple inferences are equally likely, it says "Please be more specific." When a demonstration point is reached the robot says "Please demonstrate the last n commands." To avoid errors in the task progression, user confirmations and ending of demonstrations are invoked by the experimenter.

D. Measurements

We measure the impact of transparency mechanisms with several metrics. Objective quantitative metrics (per command) include: *accuracy* (correctness of confirmed action and target), *task completion time* (from the robot's command request to the user's confirmation), *number of repetitions* before confirming, and *number of words used*. The user's perceived accuracy of the robot's perceptual components (scene, pointing, speech) and of their own mental model of what the robot can understand is obtained with four 7-point Likert-scale questions administered after each condition. Qualitatively, we assess the user's mental models based on an open-ended question, asking participants to describe how the robot perceives the world.

We ask additional Likert-scale questions to characterize people's perception of how different input channels (speech, pointing) contribute to the communication of commands and output channels (robot's speech, gaze, pointing, and visualizations) contribute to their awareness of the robot's understanding. In the experimental conditions, additional questions ask the participant to rate the contribution of different visualization elements in their awareness of what the robot can see, hear, or infer. Two questions ask the participant to subjectively assess whether the visualization was useful and whether it was useful even after being removed. Questions administered at the end of the study ask participants to compare the two visualization modalities and rank the three conditions.

E. Hypotheses and statistical analysis

Our study aims to test the following three hypotheses.

H1: Adding visualization-based transparency after natural transparency will positively impact communication.

H2: By improving the user's mental model, visualization-based transparency will improve communication *even after it is removed*.

H3: The medium in which visualizations are provided will not impact communication.

The dependent variables for measuring communication performance include the various task metrics mentioned in Sec. VI-D. Independent variables are the *transparency condition* (Baseline, Screen, VR), *session order* (1, 2, 3), and *part of the session* (P1, P2). To test H1 and H2, we assume that there is linear learning on a log scale across the three sessions. This means that the percent change in average task metrics due to learning between sessions 1 and 2 equals that of learning between sessions 2 and 3. We use the following *random effects model* to capture the setup of our experiment:

$$Y_{ij} = z_i + \mu_j + \beta_{learn}(s_{ij} - 1) + \epsilon_{ij} \quad (3)$$

where $\epsilon_{ij} \sim N(0, \sigma^2_{phase})$ is a random noise term describing natural variability in a single measurement Y_{ij}; $z_i \sim N(0, \sigma_i^2)$ is a random effect specific to participant i describing how their initial aptitude for the task differs from average; β_{learn} is the average effect of learning that occurs between two consecutive sessions; μ_j is a fixed, unknown parameter that corresponds to the average value of the task metric on a log scale, where $j = 1..6$ corresponds to a combination of transparency condition and session part (*i.e.,* Baseline-P1, Baseline-P2, Screen-P1, Screen-P2, VR-P1, VR-P2); s_{ij} is the session number (1, 2, 3) in which the condition j occurs for participant i.

The order counterbalancing of the Screen and VR conditions allows us to separate the learning effect (β_{learn}) from the transparency-related effects. Our model allows testing H1 with the null hypotheses $\mu_1 = \mu_3$ and $\mu_1 = \mu_5$ (for P1) and H2 with the null hypotheses $\mu_2 = \mu_4$ and $\mu_2 = \mu_6$ (for P2). Similarly, H3 is tested with the null hypotheses $\mu_3 = \mu_5$ (for P1) and $\mu_4 = \mu_6$ (for P2). Note that tests of H3 do not depend on our assumption of constant learning rate, but tests of H1 and H2 do.

VII. FINDINGS

Our study was completed by 20 participants (7F, 13M) in the age range 18-36 (M=26.55, SD= 4.93). Fig. 7 presents the objective metrics obtained from the study and Fig. 9 summarizes the responses to the questionnaire. Next we go through and highlight some observations from these results.

Real-time benefits of visual transparency. Visual transparency, particularly when provided on a *screen*, improves several objective metrics. The commands given by participants were significantly more accurate (Fig. 7a, P1) and took less time to give (Fig. 7b, P1) in the Screen condition. Both experimental conditions (Screen and VR) had a smaller number of attempts (Fig. 7c, P1) and number of words used to give a command (Fig. 7d, P1); however, the difference was statistically significant only for number of words in the Screen condition. Participants strongly agreed (with an average rating above 6 for screen on a 7-point Likert scale, Fig. 9e) that both visualizations helped improve their mental model.

After-effects of visual transparency. The effects of visual transparency did not carry through to the second part of the experimental condition sessions (P2) during which they were removed. Nonetheless, having the visualizations during P1 did not negatively impact their performance in P2. Additionally, participants had an overall positive agreement with the statement that visualizations help *even after* being removed; however, not as high as when they are available (Fig. 9e). They agreed with this statement significantly more referring to the Screen, compared to the VR.

Behavioral differences with visual transparency. When giving commands in the experimental conditions, participants started using the unique keywords assigned to objects and surfaces by the robot. Some examples include "Pick up object 72 (Screen)," "Place it on 5 (VR)", "Clean surface number 6 (Screen)." We saw that 12/20 participants in the screen condition and 11/20 participants in the VR condition used

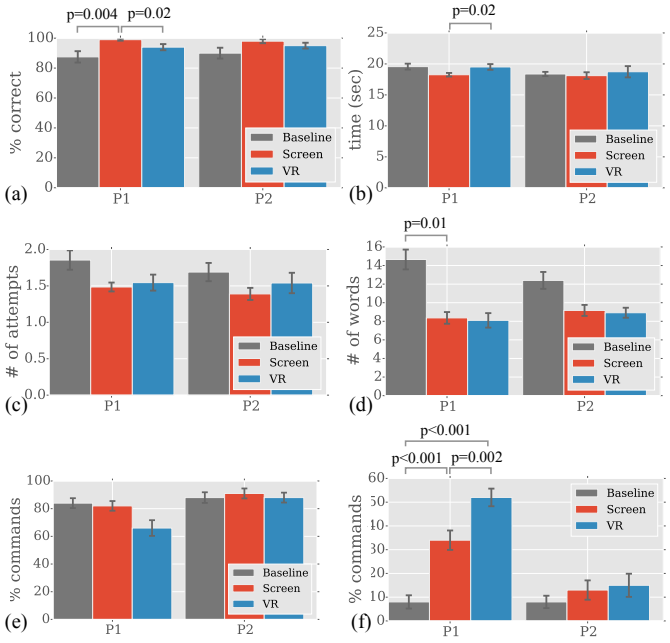


Fig. 7: Analysis of objective metrics across the three conditions and two parts of each session (P1, P2). (a) Percentage of correct commands. (b) Time taken per attempt at a command. (c) Number of attempts until a command is confirmed. (d) Number of words used in each command attempt. Percentage of commands for which (e) input language and (f) pointing contributed to the inference of the target. Significant differences indicated with p values for the null hypotheses stated in Sec. VI-E. Error bars indicate standard error.

Fig. 8: Examples of people pointing while checking the visual interface in the screen condition.

such a reference at least once. However, the participants did not exclusively use robot assigned names. On average such references were used 1.70 times (SD=2.51) for Screen and 1.95 times (SD= 2.3) for VR.

What changed most drastically between the conditions was people's use of pointing gestures. While pointing contributed to only about 10% to commands in the Baseline condition, this number increased to about 30% in the Screen condition and 50% in the VR condition (Fig. 7f, Fig. 8). This increase was statistically significant, as was the difference between the two conditions. We believe that the head movement based cursor pointing was particularly compelling since participants were already moving their heads and seeing potential targets get highlighted when they are pointed at. Although the difference was not significant, participants appeared to have used pointing more even after the visualizations were removed (P2) in the Screen and VR conditions, compared to he Baseline (Fig. 7f, P2). Consistent with these findings, participants reported that

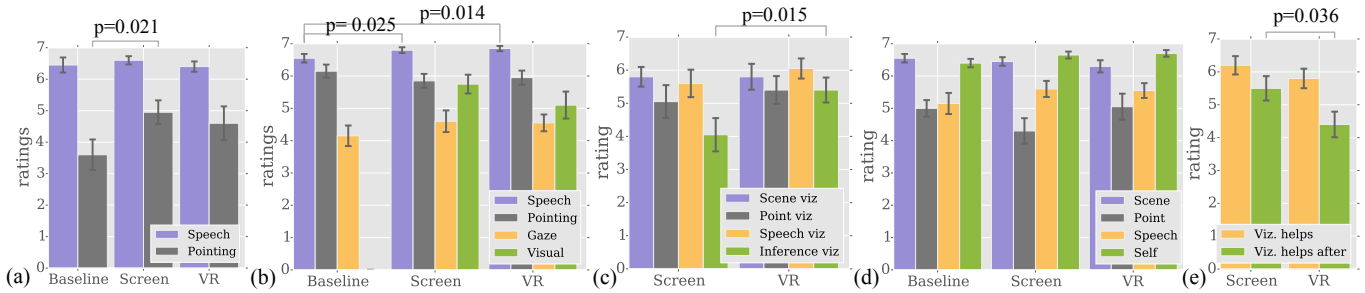


Fig. 9: Analysis of questionnaire responses about (a) how much inputs contributed to commands; (b) how much different outputs and (c) particularly, different visual interface elements contributed to the user's performance and mental model; (d) participant agreements with statements related to their mental models (see Sec. VI-D) and (e) contribution of visualizations to their mental models. Significant differences indicated with p values for paired Wilcoxon signed-rank tests. Error bars indicate standard error.

pointing contributed more to their commands (Fig. 9a). The use of pointing in VR condition may have impacted the user's input language. As seen in Fig. 7e, language contributed less to people's commands in the VR condition compared to the other conditions, but this difference was not statistically significant.

Mental models with visual transparency. We observed some differences in people's mental models between the baseline and experimental conditions. For example people became more aware that the robot could interpret their pointing gestures fairly well when introduced to the visualization. The flaws in people's mental models and the changes that occurred after switching to a visual transparency were captured well by free form questions in the questionnaire:

"*... I learned the way that the system sees my pointing. I found it was accurate on the right side of the room but it was difficult [..] on the left side of the room..*"

"*Based on my observation from this round, I see that [the robot] has a much richer perception of the environment around her, and is probably much less controlled by a human during the experiment than I had initially thought.*"

"*... understood in this round what went wrong when [it] did not understand me: for instance, mis-translating 'dust' as 'does' and thereby not recognizing the action word in the command.*"

Despite these differences, participants did not have different levels of agreement with statements regarding how well the robot could see the scene and the user's pointing and hear the user's utterances (Fig. 9d). Similarly their self assessment of the mental model accuracy was not different across conditions and was very high (Fig. 9d, Self). Furthermore, participants did not think that natural transparency mechanisms contributed less to their understanding of what the robot understood from their commands (Fig. 9b). They rated the contribution of the robot's speech particularly high. Among the different visual interface elements, participants did not favor any particular one and they did not have a different opinion across the Screen and VR conditions (Fig. 9c), except for the contribution of the inference visualization.

Subjective preferences. In the forced ranking questions, 11/20 chose Screen as their most preferred, 8/20 chose the baseline, and 1/20 chose VR. Participants who chose the baseline appeared to take less advantage of the visualizations (*e.g.,* use unique IDs for targets or point with the help of the visualized cursors). This indicates a polarization among participants. Indeed, although most participants preferred Screen over VR, 10/20 indicated the baseline as their least preferred interface.

VIII. DISCUSSION

Our study involved a particular implementation of visual transparency for LUCIT; however, several of the visualization techniques used in our work are applicable to other perceptual and inference systems on robots. That includes: renderings and annotations of geometrical sensor steams or snapshots (*e.g.,* images and pointclouds); cursors and highlighting for pointing devices; unique naming of task-related entities; text representations of raw speech input and inferred command.

LUCIT has several technical limitations, such as the lack of context-independent object recognition or use of simple keyword spotting. While upgrading the system in these aspects (*e.g.,* using a statistical language model) would improve its scalability, it would potentially introduce more errors. Nonetheless, our transparency mechanisms would still be applicable and would likely provide more benefits.

IX. CONCLUSION

We described LUCIT: a system that interprets a visual scene and a user's speech and pointing to infer their task commands. We characterized the capabilities of LUCIT with data from 5 participants. Then, through a user study with 20 participants, we investigated the effect of adding visualization-based transparency mechanisms to this system. We concluded that visualizations can help communication, especially the use of pointing gestures, and change people's mental models of the robot's capabilities. Visualizations are more effective and preferable when displayed on a screen than when they are immersive; but some people still prefer only human-like transparency mechanisms.

ACKNOWLEDGMENT

We thank Prof. Paul Sampson for his input on our statistical analysis. This work was partially funded by the National Science Foundation, grant number IIS-1525251.

References

[1] T. Brick and M. Scheutz. Incremental natural language processing for hri. In *ACM/IEEE Intl. Conf. on Human-robot Interaction (HRI)*, pages 263–270, 2007.

[2] J.Y. Chai, L. She, R. Fang, S. Ottarson, C. Littley, C. Liu, and K. Hanson. Collaborative effort towards common ground in situated human-robot dialogue. In *ACM/IEEE Intl. Conf. on Human-robot Interaction (HRI)*, 2014.

[3] D.L. Chen and R.J. Mooney. Learning to interpret natural language navigation instructions from observations. *San Francisco, CA*, pages 859–865, 2011.

[4] C. Crick, S. Osentoski, G. Jay, and O.C. Jenkins. Human and robot perception in large-scale learning from demonstration. In *ACM Intl. Conf. on Human-robot Interaction (HRI)*, 2011.

[5] R. Deits, S. Tellex, P. Thaker, D. Simeonov, T. Kollar, and N. Roy. Clarifying commands with information-theoretic human-robot dialog. *J. of Human-Robot Interaction*, 2(2), 2013.

[6] K. Fischer. The role of users preconceptions in talking to computers and robots. In *Workshop on How People Talk to Computers, Robots, and other Artificial Communication Partners*, pages 112–130, 2006.

[7] K. Fischer. How people talk with robots: Designing dialog to reduce user uncertainty. *AI Magazine*, 32(4):31–38, 2011.

[8] B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski. Using vision, acoustics, and natural language for disambiguation. In *ACM/IEEE Intl. Conf. on Human-robot Interaction (HRI)*, pages 73–80, 2007.

[9] P. Gorniak and D. Roy. Situated language understanding as filtering perceived affordances. *Cognitive science*, 31(2):197–231, 2007.

[10] Peter Gorniak and Deb Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, pages 429–470, 2004.

[11] Chien-Ming Huang and Bilge Mutlu. Robot behavior toolkit: generating effective social behaviors for robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (HRI)*, pages 25–32. ACM, 2012.

[12] T. Iio, M. Shiomi, K. Shinozawa, T. Akimoto, K. Shimohara, and N. Hagita. Investigating entrainment of peoples pointing gestures by robots gestures using a woz method. *Int. J. of Social Robotics*, 3(4):405–414, 2011.

[13] T. Iio, M. Shiomi, K. Shinozawa, K. Shimohara, M. Miki, and N. Hagita. Lexical entrainment in human robot interaction. *International Journal of Social Robotics*, 7(2):253–263, 2015.

[14] M. Johnson-Roberson, J. Bohg, G. Skantze, J. Gustafson, R. Carlson, B. Rasolzadeh, and D. Kragic. Enhanced visual scene understanding through human-robot dialog. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 3342–3348. IEEE, 2011.

[15] S. Kiesler. Fostering common ground in human-robot interaction. In *IEEE Intl. Workshop on Robot and Human Interactive Communication (ROMAN)*, pages 729–734. IEEE, 2005.

[16] E.S. Kim, D. Leyzberg, K.M. Tsui, and B. Scassellati. How people talk when teaching a robot. In *ACM/IEEE Intl. Conf. on Human-robot Interaction (HRI)*, pages 23–30. IEEE, 2009.

[17] J. Krishnamurthy and T. Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.

[18] Geert-Jan M Kruijff, Pierre Lison, Trevor Benjamin, Henrik Jacobsson, and Nick Hawes. Incremental, multi-level processing for comprehending situated dialogue in human-robot interaction. In *Symposium on Language and Robots*, 2007.

[19] D. Lazewatsky and W.D. Smart. Context-sensitive in-the-world interfaces for mobile manipulation robots. In *IEEE Intl. Symp. on Robot Human Communication (ROMAN)*, pages 989–994. IEEE, 2012.

[20] S. Lemaignan, R. Ros, E.A. Sisbot, R. Alami, and M. Beetz. Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. *International Journal of Social Robotics*, 4(2):181–199, 2012.

[21] C. Liu and J.Y. Chai. Learning to mediate perceptual differences in situated human-robot dialogue. In *AAAI Conference on Artificial Intelligence*, 2015.

[22] C. Matuszek, L. Bo, L. Zettlemoyer, and D. Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *AAAI Conference on Artificial Intelligence*, 2014.

[23] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423*, 2012.

[24] D.K. Misra, J. Sung, K. Lee, and A. Saxena. Tell me dave: Context-sensitive grounding of natural language to mobile manipulation instructions. In *Robotics: Science and Systems, RSS*, 2014.

[25] R.J. Mooney. Learning to connect language and perception. In *AAAI Conf. on Artificial Intelligence*, pages 1598–1601, 2008.

[26] H. Nguyen, A. Jain, C. Anderson, and C.C. Kemp. A clickable world: Behavior selection through pointing and context for mobile manipulation. In *IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 787–793. IEEE, 2008.

[27] N. Otero, A. Alissandrakis, K. Dautenhahn, C. Nehaniv, D.S. Syrdal, and K.L. Koay. Human to robot demonstrations of routine home tasks: exploring the role of the

robot's feedback. In *ACM/IEEE Intl. Conf. on Human-robot Interaction (HRI)*, pages 177–184, 2008.

[28] T. Pejsa, D. Bohus, M.F. Cohen, C.W. Saw, J. Mahoney, and E. Horvitz. Natural communication about uncertainties in situated interaction. In *ACM Intl. Conf. on Multimodal Interaction (ICMI)*, 2014.

[29] J. Peltason, N. Riether, B. Wrede, and I. Lütkebohle. Talking with robots about objects: a system-level evaluation in hri. In *ACM/IEEE Intl. Conf. on Human-robot Interaction (HRI)*, pages 479–486, 2012.

[30] Vasumathi Raman, Constantine Lignos, Cameron Finucane, Kenton CT Lee, Mitchell P Marcus, and Hadas Kress-Gazit. Sorry dave, i'm afraid i can't do that: Explaining unachievable robot tasks using natural language. In *Robotics: Science and Systems*, volume 2, pages 2–1. Citeseer, 2013.

[31] O. Rogalla, M. Ehrenmann, R. Zöllner, R. Becher, and R. Dillmann. Using gesture and speech control for commanding a robot assistant. In *IEEE Intl. Workshop on Robot and Human Interactive Communication (ROMAN)*, pages 454–459. IEEE, 2002.

[32] A. StClair and M. Mataric. How robot verbal feedback can improve team performance in human-robot task collaborations. In *ACM/IEEE Intl. Conf. on Human-robot Interaction (HRI)*, pages 213–220, 2015.

[33] K. Stubbs, D. Wettergreen, and P.H. Hinds. Autonomy and common ground in human-robot interaction: A field study. *Intelligent Systems, IEEE*, 22(2):42–50, 2007.

[34] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S.J. Teller, and N. Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conf. on Artificial Intelligence*, 2011.

[35] A.L. Thomaz and C. Breazeal. Transparency and socially guided machine learning. In *IEEE Intl. Conf. on Development and Learning (ICDL)*, 2006.

[36] Elin Anna Topp, Henrik I Christensen, and Kerstin Severinson Eklundh. Acquiring a shared environment representation. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 361–362, 2006.

[37] A. Vollmer, K.S. Lohan, K. Fischer, Y. Nagai, K. Pitsch, J. Fritsch, K.J. Rohlfing, and B. Wredek. People modify their tutoring behavior in robot-directed interaction for action learning. In *IEEE Intl. Conf. on Development and Learning (ICDL)*, pages 1–6. IEEE, 2009.