

Embodied Collaborative Referring Expression Generation in Situated Human-Robot Interaction

Rui Fang, Malcolm Doering, Joyce Y. Chai
Department of Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
{fangrui, doeringm, jchai}@cse.msu.edu

ABSTRACT

To facilitate referential communication between humans and robots and mediate their differences in representing the shared environment, we are exploring embodied collaborative models for referring expression generation (REG). Instead of a single minimum description to describe a target object, episodes of expressions are generated based on human feedback during human-robot interaction. We particularly investigate the role of *embodiment* such as robot gesture behaviors (i.e., pointing to an object) and human's gaze feedback (i.e., looking at a particular object) in the collaborative process. This paper examines different strategies of incorporating embodiment and collaboration in REG and discusses their possibilities and challenges in enabling human-robot referential communication.

Categories and Subject Descriptors: H.5.2 [Information Interfaces and Presentation]: User Interfaces

General Terms: Algorithms, Experimentation

Keywords: Referring Expression Generation; Embodiment; Collaborative Referential Communication

1. INTRODUCTION

Referring Expression Generation (REG) has traditionally been formulated as a problem of generating a single noun phrase (possibly with multiple modifiers and prepositional phrases) that can uniquely describe a target object among multiple objects [5, 18] so that addressees (i.e., humans) can correctly identify the intended object given this expression. Except for a few recent works that apply computer vision to building representations of objects in scenes [8, 16, 22], most existing REG approaches were developed and evaluated under the assumption that humans and agents have access to the same kind of domain information and have the same representation of the shared world [18]. Clearly, this assumption does not hold in human-robot interaction.

In situated human-robot interaction, humans and robots have different representations of the shared environment due

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

HRI'15, March 2–5, 2015, Portland, OR, USA

Copyright 2015 ACM ACM 978-1-4503-2883-8/15/03 ...\$15.00
<http://dx.doi.org/10.1145/2696454.2696467>.

to their mismatched perceptual capabilities. The robot's representation of the shared environment is often incomplete and error-prone. When a shared perceptual basis is missing, referential communication between partners becomes difficult [3]. Specifically for the task of REG, our previous work [8] has shown that when the human and the agent have a mismatched perceptual basis traditional approaches to REG tend to break down.

To mediate mismatched representations, we have developed two computational collaborative models (i.e., an episodic model and an installment model) for REG [7] motivated by collaborative referential behaviors in human-human communication [4]. Instead of generating a single referential description, our collaborative models generate multiple small noun phrases that gradually lead to the target object with a goal of minimizing the collaborative effort. We have further evaluated these collaborative models in a web-based study using 2D images to simulate a shared environment [7].

While our previous results on collaborative models are encouraging, they have not been applied to real-time human-robot interaction in a physical environment, which is much more complex than 2D simulations. More importantly, a key characteristic of physical world interaction is *embodiment*: agents (i.e., robots) and humans both have physical bodies and they can use non-verbal modalities (e.g., gesture and eye gaze) to refer to the shared world and to provide immediate feedback. It is not clear to what degree the collaborative models can take embodiment into account and apply to physical interaction. To address these issues, we are exploring embodied collaborative models for REG. We particularly investigate the role of *embodiment* such as robot gesture behaviors (i.e., pointing to an object) and the human's gaze feedback (i.e., looking at a particular object) in the collaborative process. This paper examines different strategies of incorporating embodiment and collaboration in REG and discusses their possibilities and challenges in enabling human-robot referential communication.

2. RELATED WORK

Previous psycholinguistic studies have indicated that referential communication is a collaborative process [2, 4]. To minimize the collaborative effort, partners tend to go beyond issuing an elementary referring expression (i.e., a single noun phrase), by using other different types of expressions such as *episodic*, *installment*, *self-expansion*, etc. For example, an episodic description is produced in two or more easily distinguished episodes or intonation units as shown in the

following example from human-human communication [20].

A: below the orange, next to the apple, it's the red bulb.

An installment behavior is similar to the episodic behavior in the sense that it also breaks down referring expression generation into smaller episodes. The difference is that explicit feedback from the addressee is solicited before the speaker moves to the next episode. Here is an example of installment descriptions from [20].

A: under the pepper we just talked about.

B: yes.

A: there is a group of three objects.

B: OK.

A: there is a yellow object on the right within the group.

The generation of episodic or installment descriptions is not to minimize the speaker's own effort, but rather to minimize the collaborative effort so that the addressee can quickly identify the referent. These collaborative behaviors from human-human referential communication have motivated previous computational models for collaborative reference [6, 7, 15]. However, these models have not been applied to human-robot interaction.

As deictic gesture plays an important role in referential communication [9, 14], recent work has incorporated deictic gestures in generating referring expressions [23], referring acts [12, 13], and multimodal references [25]. More recently, different strategies of incorporating deictic gestures have been investigated in human-robot interaction [24].

Psycholinguistic findings have shown that eye gaze is tightly linked to human language comprehension and production [27]. Specifically for REG, Koller et al. [17] developed an interactive system that tracks the hearer's eye gaze to monitor hearer's interpretation of the referring expressions the system generates. However, their REG system was built on a virtual environment where the agent and the human are assumed to have the same representation of the shared world.

In contrast to previous work, our work here incorporates embodiment (i.e., deictic gestures from the robot and gaze feedback from the human partner) into the collaborative model to address the unique characteristics of human-robot interaction.

3. EMBODIED COLLABORATIVE REG

3.1 Collaborative Installment Model

In our previous work we have developed a collaborative installment model for REG [7]. Instead of generating a single referring expression to describe a target object, our installment model generates one expression (i.e., *installment*) at a time based on the human's feedback and gradually leads to the target object. We formulated the learning of the installment model as a sequential decision making problem and trained the model using reinforcement learning. The goal was to learn an optimal policy that generates installments under different situations to allow listeners to successfully identify the target object.

More specifically, we used a variety of features to represent the utility function $Q(s, a)$ for a given state s (which captures the state of interaction such as what landmark objects can be used) and a given action a (which represents a referring expression generation strategy/action - e.g., which

attributes to choose to describe the intended object) as in:

$$Q(s, a) = \theta^T \phi(s, a)$$

Where ϕ is the feature vector of each state-action pair, and θ is the weight vector associated with the features. In our previous work [7], we have applied the SARSA [26] algorithm and conducted a crowd-sourcing study with 2D images to learn the weights for the features. Table 1 summarizes some important features and the learned weights (Feature 1-22).

Once the weights for features are learned, they are applied to calculate $Q(s, a)$ for any state-action pair. During real-time generation, given a particular state s , the agent will pick the generation action a^* to maximize $Q(s, a)$.

$$a^* = \arg \max_{a \in A} Q(s, a)$$

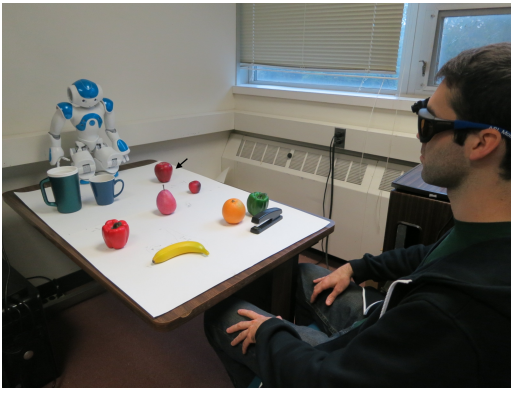
#	Feature Value Description	Learned Weights
1	normalized sum of vision confidence of all descriptors in re	0.92
2	is spatial location descriptor in re ?	0.54
3	vision confidence of spatial relation descriptor	0.52
4	vision confidence of type descriptor	0.51
5	vision confidence of spatial location descriptor	0.48
6	is type descriptor in re ?	0.21
7	number of descriptors in re	0.19
8	vision confidence of size descriptor	0.13
9	is size descriptor in re ?	0.13
10	is color descriptor in re ?	0.10
11	vision confidence of color descriptor	0.09
12	can re together with sp to lm uniquely identify an object?	0.88
13	is there a sp between o and tar ?	0.51
14	number of spatial links from lm	0.23
15	number of spatial links to o (in degree)	0.01
16	number of spatial links from o (out degree)	0.01
17	is o and lm in the same group?	1
18	is o a group?	0.97
19	is lm is a group and o in lm ?	0.96
20	is o a group and tar in o ?	0.90
21	is o and tar in the same group?	0.51
22	is o a group and lm in o ?	0.11
23	the cost of producing a pointing gesture	0.50

Table 1: Features and the learned weights from the original installment model [7].

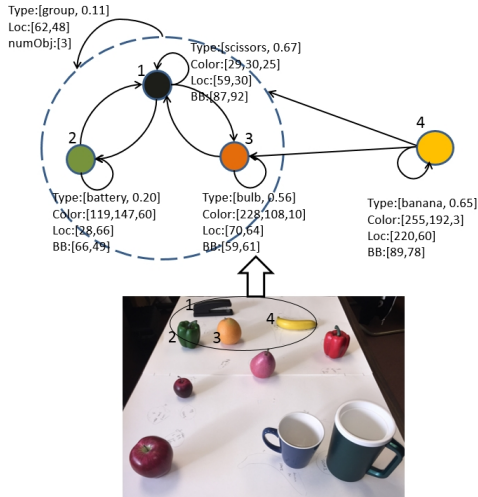
3.2 Incorporation of Embodiment in HRI

To extend the collaborative model to human-robot interaction and incorporate embodiment into the model, we explored two directions: (1) incorporating the robot's gesture (i.e., pointing) into the collaborative model to generate multimodal referring expressions; and (2) incorporating the human's real-time feedback (either verbal feedback, gaze feedback, or both) into the collaborative model.

Figure 1(a) shows an example setup where the embodied collaborative model is examined. In the setup a NAO robot sits face-to-face with a human partner who wears a mobile eye tracker as shown in Figure 1(a). The robot's internal representation of the shared environment is impoverished compared to its human partner. Figure 1(b) shows an example of the internal representation for a part of the environment. Note that Object 1, 2, and 3 are recognized as a group (by applying a perceptual grouping algorithm [11]).



(a) An example of situated setup.



(b) An example of the robot's internal representation of the shared environment.

R1:	"Do you see a group of two on the left?" (Robot points to the group)
H1:	"Yes" (The group becomes the landmark)
R2:	"That group I was just talking about, do you see an object in the back that is on the right within that group?" (Describe the intended object in relation to the landmark)
H2:	"No" (Human rejects the description)
R3:	"That group I was just talking about, do you see an object on the right that is to the right of that group?" (Robot chooses another description strategy)
H3:	"Yes"

(c) An example of collaborative referential process.

Figure 1: An example of situated setup for referential communication.

Since we are addressing referential communication to mediate perceptual differences between humans and robots, we intentionally applied only a simple computer vision algorithm [19]. Here, except for Object 4 which is correctly recognized as a banana with a confidence of 0.65, the rest of objects are mis-recognized. The numerical values related to color, location, and the size of bounding boxes are also captured by the internal representation. Figure 1(c) shows

an example of the embodied collaborative referential process for the robot to gradually lead the human to the target object (i.e., the red apple as shown by the arrow in Figure 1(a)). Note the robot takes the human's perspective when describing the objects in the scene.

Incorporation of robot's gesture in referring acts.

Pointing gestures from a robot are combined with verbal descriptions to generate referring expressions. Here, we treat the cost of gesture generation as a feature and incorporate it with other features into the collaborative models [7]. Since it is expensive to conduct a large scale in-lab study to learn feature weights from real-time human-robot interaction, we directly adopt the features and their learned weights from our web-based study [7] as shown in Table 1 (Feature 1-22). We then explicitly add the cost of the gesture generation as an additional feature and set its weight to 0.5 (Feature 23).

The cost of a pointing gesture depends on several factors: the distance from the robot to the target object, the size of the target object, adjacency of other objects to the target object, etc. Inspired by previous work on the costs of pointing, [10] and [21], we define the cost of a pointing gesture to an object as follows:

$$cost = \begin{cases} \log_2\left(\frac{1.5 \times Distance}{Size} + 1\right) & \text{If the object to be described is in a group} \\ \log_2\left(\frac{Distance}{Size} + 1\right) & \text{Otherwise} \end{cases}$$

Given a situation s , the robot will apply the features and their associated weights in Table 1 to calculate $Q(s, a)$ and then choose the generation action a^* that maximizes $Q(s, a)$. Note that incorporating gesture does not necessitate that the generated referring expressions will always include gestures. Whether gesture is used or not depends on how it weighs against other features extracted from the environment. For example, in Figure 1(c), $R1$ includes the pointing gesture as part of the referring expression, while $R2$ does not.

Incorporation of eye gaze as feedback.

Previous psycholinguistic studies have shown that human eye gaze directly links with language comprehension [27]. Immediately after hearing a referring expression, the hearer's eyes move to the objects being referred to. Motivated by this finding we incorporate the user's real-time gaze and verbal information as intermediate feedback in the installment model.

To incorporate human gaze feedback, we must find how the gaze is distributed among each of the objects in the scene over the time immediately following an RE such as in Figure 1(c). Specifically, starting from the onset of an RE uttered by the robot, we capture at least 200 gaze readings (taking on average 7.7 sec) from the participant and calculate a distribution showing which objects the gaze is drawn to most often (i.e., which objects draw the most gaze readings). Based on this distribution, we applied a simple criterion to identify the *focused object*: the object with the highest number of readings where its reading number is at least twice as high as the reading number of the second highest object.

The gaze feedback can be incorporated into the installment model with or without verbal feedback.

- **Gaze Only Feedback** If a focused object can be identified based on the gaze, then the installment is considered successful and the focused object becomes the

landmark for generating the next installment. Otherwise, the installment is considered unsuccessful and a different RE will be generated based on the policy applicable to the current situation.

- **Gaze and Verbal Feedback Combined** To incorporate both the gaze and verbal information from the human as intermediate feedback, we must handle all cases where the gaze feedback and verbal feedback conflict, such as when the human responds “yes” but the gaze indicates that no focused object is identified. We manage these cases in the following ways: when the user responds “No” then the gaze information is disregarded and the robot follows the verbal feedback. When the user says “Yes” then there are two cases: (1) if no focused object is identified, then the robot’s intended object becomes the next landmark. (2) if a focused object is identified, then the focused object becomes the next landmark.

Figure 1 shows an example of how (verbal) feedback from the user determines which object/group becomes the landmark in the next RE. The robot starts by asking, “Do you see a group of two on the left?”. After the user responds in the affirmative, that group becomes the landmark in the next utterance. The robot describes the next object in relation to this landmark, “... do you see an object in the back that is within that group?”

4. EMPIRICAL STUDIES

To evaluate the embodied collaborative REG model described above, we conducted an empirical study in human-robot interaction.

4.1 Design Factors

We designed our experiments to take into consideration the following two aspects of embodiment:

Incorporation of robot gestures. To evaluate whether the incorporation of the robot’s gestures into the collaborative model facilitates referential communication, we manipulate the following two conditions:

- **Verbal Only** refers to the condition where only verbal referring expressions are generated based on the original installment model without incorporating the gesture feature.
- **Verbal and Gesture** refers to the condition where pointing gestures are incorporated into REG as described in Section 3.2.

Incorporation of human feedback. We examine different strategies of incorporating human feedback (verbal and/or eye gaze) into the installment model for REG.

- **None** refers to the condition where no explicit feedback from the human is incorporated. This is similar to the episodic model developed in our previous work [7]. The robot generates a sequence of episodes (i.e., noun phrases) that leads to the target object with the assumption that each episode is accepted by the human.
- **Verbal Only** refers to the condition where only verbal information is used as the feedback from the human

partner. In this condition the human is allowed to respond “yes”, “no”, or “repeat” (if they did not hear clearly). If the response is “yes” the robot assumes that the human understands the generated RE and the current intended object becomes the landmark for the next step. If the response is “no” the robot assumes the human does not understand the intended object and will try another RE. Lastly, if the response is “repeat” the robot repeats the last generated RE.

- **Gaze Only** refers to the condition where only gaze is used as the feedback from human. In this condition the gaze of the human is incorporated as described in Section 3.2. The gaze is tracked and used as feedback to the installment model.
- **Gaze and Verbal** refers to the condition where both gaze and verbal information are used as the feedback from the human. The gaze and verbal feedback are combined as described in Section 3.2.

These design factors lead to $4 \times 2 = 8$ experimental conditions. In all conditions, whether or not the gaze is incorporated in REG, it is always tracked for later analysis.

4.2 Experimental Tasks

We designed an object identification game to evaluate embodied collaborative REG. Figure 1(a) shows an example of the situated setup for the experiment. We positioned a human subject and a NAO robot face-to-face at a table which held ten everyday objects. The subject wore an ASL mobile eye tracker. The NAO robot would generate referring expressions in an installment fashion to describe these objects using different strategies (depending on the experimental condition). The human subject’s goal was to identify which object the NAO robot was talking about. At the end of each game, we asked the human subject what target object was intended by the robot. We compared the object identified by the human with the true target object to compute the object identification accuracy for a particular generation strategy.

4.3 Procedures

We designed four scenes with ten objects of various types in different configurations. As described in Section 3.2, these four scenes were processed by a simple computer vision algorithm [19] and a perceptual grouping algorithm [11], and the results were given to the robot as its internal representations of the scenes (as shown in Figure 1(b)). Computer vision algorithms have advanced significantly in recent years. However, they are far from being perfect. It is likely, in a new situation or a changing environment, the robot which is equipped with modern CV algorithms can only correctly recognize a few objects (not all objects) in the shared environment with the human. This is the situation our work intends to address. Therefore, we intentionally applied a very simple CV algorithm so that the mis-recognition rate would be high. Overall, about 73% of the objects are mis-recognized or mis-segmented in our experiments.

We recruited a total of 24 human subjects to participate in our study. These subjects were not aware of any part of the system or the study design. During the experiments, each participant played 24 object identification games - 6 games on each of four separate scenes. Each scene was paired with

one of the four feedback conditions such that the participant would interact with the robot under each condition. Moreover, in each of the 6 games on a scene, one target object was randomly chosen. Half of these 6 games were played under the **Verbal and Gesture** condition and the other half under the **Verbal Only** condition. For each participant the scenes were presented in one of the 24 possible permutations so that the order would not have an effect on the experiment. Furthermore, the orders of the targets were also randomized. Based on this design each participant played three games in each of the eight experimental conditions, resulting in a total of 72 games played in each condition.

4.4 Empirical Results

To help illustrate the results, we reiterate some key terms that are used throughout this section: a target object is the object the robot intends for the human to identify at the end of interaction. An intended object at each installment is an object the robot refers to on the path to the target (it may not be the target object). A landmark object is the object on the path (i.e. an intended object) that has been confirmed or accepted by the human. We use *session* to refer to the process that leads to the target object in one game. Each session often contains several installments.

4.4.1 The role of robot gesture

Table 2 shows the overall accuracy of object identification given expressions where gesture is not incorporated (i.e., **Verbal Only**) and gesture is incorporated (**Verbal and Gesture**). As shown in the table, the number of successes in identifying the target object at the end of the session is much higher in the **Verbal and Gesture** condition, which brings to 0.76 overall accuracy. This is significantly higher than the accuracy of 0.50 when gesture is not incorporated ($\chi^2 = 43.11, p < 0.0001$).

Robot Gesture	Verbal Only	Verbal and Gesture
Sessions	288	288
Successes	144	220
Accuracy	0.50	0.76

Table 2: Overall accuracy in object identification when gesture is incorporated (Verbal and Gesture) and is not incorporated (Verbal Only).

Table 3 further breaks down the accuracy of identifying the target object under eight experimental conditions. There is a significant difference among these conditions ($\chi^2 = 56.97, p < 0.0001$). Our results have shown that when robot gesture and human verbal feedback are incorporated in REG (i.e., the **Verbal and Gesture** condition for robot gesture and the **Verbal Only** condition for human feedback), it achieves 88% accuracy. This performance is significantly higher than most of the other conditions as indicated by a Tukey post-hoc test (at $p < 0.05$). Furthermore, each model that incorporates robot gesture into its generation strategy performs significantly better than its corresponding model without gesture.

Particularly, when only human gaze feedback is incorporated in REG (i.e., the **Gaze Only** condition), robot gesture has shown to be very useful. During the experiments, we observed that

several participants appeared to have difficulty understanding which object was the landmark when hearing the expres-

Human Feedback	Robot Gesture	
	Verbal Only	Verbal and Gesture
Verbal Only	0.53	0.88
Gaze Only	0.40	0.72
Gaze and Verbal	0.47	0.67
None	0.60	0.79

Table 3: Comparison of object identification accuracy under eight experimental conditions.

sion “that object you were just looking at...”. This is because gaze feedback during comprehension is often a subconscious response to the interpreted object and sometimes the participants were not consciously aware which object they just looked at. Therefore in this condition, gesture is useful for maintaining the common ground between the dialog participants by indicating the landmark object to the human.

4.4.2 The role of human gaze

Table 4 shows the overall accuracy of target identification under the four different feedback conditions. There is only marginal difference among these conditions ($\chi^2 = 10.81, p = 0.013$). Although we were hoping that the incorporation of eye gaze as intermediate feedback might improve the performance of the collaborative installment model, the empirical results have shown otherwise. When the gaze information is incorporated (in both the **Gaze Only** condition and the **Gaze and Verbal** condition), the accuracy of identifying target objects is lower.

	None	Verbal	Gaze	Gaze and Verbal
Sessions	144	144	144	144
Successes	100	101	81	82
Accuracy	0.69	0.70	0.56	0.57

Table 4: Overall accuracy in object identification under four different human feedback conditions.

To help further understand the role of gaze feedback, we separated out the cases where the feedback indicated by gaze coincides with the intended object and examined whether this would be a strong signal of understanding and whether it improves the accuracy of identifying the target object in the end. Specifically, for the **Gaze Only** condition and the **Gaze and Verbal** condition, we calculate how often gaze feedback actually provides a focused object that is used as a landmark object and how often this focused object in fact is the intended object. Table 5 shows the results.

	Gaze Only		Gaze and Verbal	
	Success	Fail	Success	Fail
Installments	185	235	204	269
LM by gaze	164	200	140	140
LM Rate	0.886	0.851	0.686	0.520
Match intended object	39	21	27	11
Matching Rate	0.238	0.105	0.193	0.079

Table 5: Statistics of landmarks established by gaze in the Gaze Only condition and the Gaze and Verbal condition.

This table shows the number of installments where the landmark used for generation was established by the gaze feedback for both successful sessions (i.e., the target object was correctly identified at the end of the session) and failed

sessions (i.e., the target object was not correctly identified in the end). For the **Gaze Only** condition, it shows that for both successful and failed sessions the landmark was established by the gaze feedback with about the same rate (88.6% vs. 85.1%, $\chi^2 = 1.124$, $p = 0.289$).

More interestingly, the rate of landmarks established by gaze which matched the intended object (i.e., the participant gazed at the object intended by the robot) is twice as high in successful sessions as it is in failed sessions (23.8% vs. 10.5%, $\chi^2 = 11.545$, $p = 0.0007$). This appears to show that when the human is able to correctly identify the current intended object and the gaze feedback works properly, it is more likely to lead to successful target identification.

For the **Gaze and Verbal** condition, we see similar statistics as with the **Gaze Only** condition. One difference is that here the rate of landmarks established by gaze feedback (only happens when the human verbal response is “yes”) is much higher for successful games than failed games (68.6% vs. 52.0%, $\chi^2 = 13.201$, $p = 0.0003$).

Furthermore, as with the **Gaze Only** condition the rate of landmarks established by gaze which matched the intended object is higher in successful sessions than it is in failed sessions (19.3% vs. 7.9%, $\chi^2 = 7.795$, $p = 0.005$). This again shows that when the human is able to correctly identify the current intended object and the gaze feedback works properly, it is more likely to result in successful target identification.

4.4.3 Effects on number of installments

There is a significant difference among the feedback conditions for the average number of installments per session ($F(3, 572) = 7.583$, $p < 0.0001$) as shown in Figure 2. Sessions in the **Verbal Only** condition lasted slightly longer, with an average of 2.36 installments per session, than sessions in the no feedback (i.e., **None**) condition, with 2.07 installments per session. Sessions in both the **Gaze Only** condition and the **Gaze and Verbal** condition lasted significantly longer.

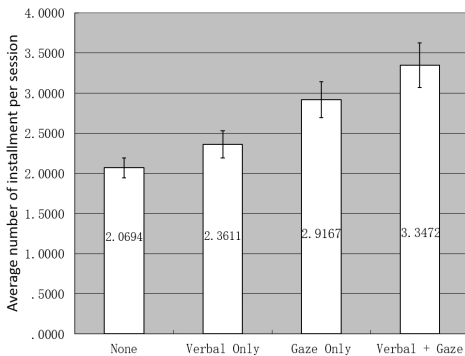


Figure 2: The average number of installments per session in each of the feedback conditions

On the other hand, statistical tests have shown that the incorporation of gesture does not have a significant effect on the number of installments per session.

4.4.4 Gesture on gaze-indicated focus

We have also examined the role of the robot’s gesture on the human’s focus on objects as indicated by eye gaze. More specifically, we compared the proportion of different eye gaze

patterns in the conditions where gesture is incorporated and not incorporated. The gaze responses after each RE were classified into three non-mutually exclusive classes¹ as follows:

- **IntendedObject** The object with the highest gaze fixations is indeed the intended object or contained in the intended group.
- **LandmarkObject** The object with the highest gaze fixations is indeed the landmark object or contained in the landmark group.
- **OtherObject** The object with the highest gaze fixations is neither (contained in) the intended object/group nor the landmark, i.e., the focused object is other than the intended or landmark object.

	Verbal Only	Verbal and Gesture
Total Num	656	739
IntendedObject	206	327
Rate (Intended)	0.31	0.44
LandmarkObject	212	252
Rate (Landmark)	0.32	0.34
OtherObject	298	248
Rate (Other)	0.45	0.34

Table 6: Rate of different gaze patterns, under the Verbal Only condition and the Verbal and Gesture Condition

Table 6 shows the rate of different gaze patterns, when robot pointing gesture is incorporated and is not incorporated in REG.

We can see that the use of gesture towards the intended object results in a higher percentage of **IntendedObject** class, with a rate of 0.44 under the **Verbal and Gesture** condition and 0.31 under the **Verbal Only** condition ($\chi^2 = 24.29$, $p < 0.0001$). Moreover, the use of gesture towards the intended object also results in a lower percentage of **OtherObject** class ($\chi^2 = 20.55$, $p < 0.0001$). These results show that the incorporation of robot gesture helps the human visually locate the intended object.

4.4.5 Gesture and gaze fixation time

Gesture also has an effect on how long it takes for the human to find the robot’s intended object. Figure 3 shows the effects of incorporating gesture on the length of time it takes for the participant’s gaze to focus, or fixate, on the intended object. We find the length of time after the first gaze reading is taken at which the intended object obtains the highest number of readings (as described in Section 3.2) and after which no other object obtains a higher number of readings than the intended object. We call this length the fixation time.

As shown in Figure 3, gesture can direct the human’s gaze quickly to the intended object. There is a significant difference between the two conditions ($t = 2.43$, $p < 0.015$). When gesture is incorporated in REG, the participant’s gaze fixates on the intended object in less time (3.73 seconds) than when gesture is not incorporated (4.17 seconds). This

¹It is possible for a gaze response to be classified into both the **IntendedObject** class and the **LandmarkObject** class if the target is contained in a landmark group.

is likely because the robot points in the direction of the intended object, immediately reducing the search space to a subset of objects in the scene and allowing the human to identify the object being referred to more quickly.

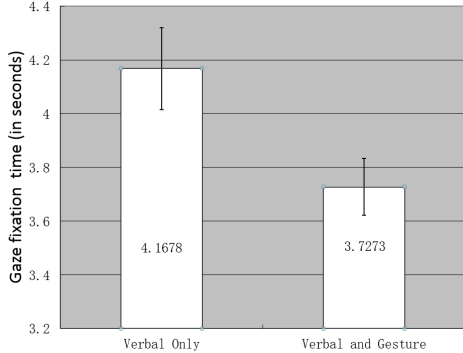


Figure 3: The average fixation time (in seconds) for the conditions with and without gesture.

5. DISCUSSION

One of the important characteristics of situated interaction is embodiment. Humans and robots have physical bodies in the shared environment. Humans can refer to objects by generating a referential expression, we can also simply point to the object, or do a combination of both. The decision of how to describe objects and when to incorporate gesture into the referring act is complex. As an initial step in our investigation, we incorporated robot pointing gesture into our collaborative installment model which was developed to mediate perceptual differences between humans and robots. We empirically set the weight for the gesture feature to take embodiment into account and directly adopted the weights learned for other features [7]. Certainly the 2D images used in our previous crowd-sourcing study are simplifications of the 3D physical environment in human-robot interaction. Ultimately, how the gesture should interact with other visual features to generate referring expressions needs to be learned in physical-world settings. We leave this to our future work.

Our empirical results have demonstrated that incorporation of gesture into the collaborative model significantly outperforms models without gesture. Our observations have shown that gestures are more useful when spatial descriptions such as “on the left/right” or “in the front/back” are used. Gestures remind the user which frame of reference is being used (human’s perspective vs robot’s perspective). Furthermore, in a scene where everything is technically in the “front” of the user, gesture disambiguates unclear spatial relations. Gesture is also very useful in the **Gaze Only** condition to remind the human which landmark object they have previously looked at.

Our experimental results have shown that eye gaze data obtained from the ASL Mobile Eye tracking system might not be suitable for making fine-grained decisions regarding the object of the human’s gaze. The human’s head movements may cause the coordinates of gaze readings to shift to objects nearby and thus affect the accuracy of identifying the focused objects. In addition, our current method of identifying focused objects and incorporating gaze feedback

is rather ad-hoc. It is not clear whether this approach can reliably reflect the focus of attention. As a result, in the conditions with eye gaze as feedback, the rate of successful sessions (where the target is identified), is lower than the conditions without eye gaze. This result seems conflicting with previous studies by Koller et al. [17]. However, the setting in [17] is human-computer interaction mediated through a computer screen, which is very different from our situated interaction. The former represents a simplified setting where a monitor-mounted gaze tracker tends to be more reliable. Better use of the Mobile Eye tracker or other devices in the physical environment (e.g., Kinect) needs to be explored as well as the granularity for representing gaze feedback (e.g., focused object versus focused region).

We currently use some heuristics to process human gaze feedback. Often, however, the verbal response from the human does not match the gaze response. For example, there were situations where the human says “Yes” but the gaze feedback does not indicate any focused object. We have recorded a large amount of gaze data from participants simultaneously with yes/no verbal feedback. In the future we plan to use this data to explore how to better classify human gaze patterns.

We have also observed some adaptation of gaze behaviors during our experiments. Some participants, realizing that the robot pays attention to their gaze feedback, adapted their gaze behavior by fixating on objects longer and saccading between objects less often. However, only about half of the participants exhibited adaptation behavior while there was no adaptation in the other half. This observation demonstrates that, to better utilize human gaze feedback, one potential solution is to design effective robot behaviors to solicit human adaptation for reliable gaze feedback.

6. CONCLUSIONS

In human-robot interaction, the robot’s representation of the shared environment can be significantly different from the human’s representation. The robot needs to be able to describe its internal representation of the shared environment so that the human understands what it is talking about. This is particularly important for establishing common ground and supporting successful interaction [1, 28]. Toward this goal, this paper explores embodied collaborative referring expression generation which incorporates non-verbal modalities in the collaborative process of referential communication. We have conducted experiments to evaluate different embodied collaborative models through situated, physical, human-robot interaction. Our empirical results have shown that the incorporation of non-verbal modalities such as robot gesture into the embodied collaborative model allows humans to better identify target objects. However, the incorporation of human eye gaze as intermediate feedback does not perform well. Our results may indicate human gaze is by nature noisy. Especially in situated interaction, it is difficult to reliably capture gaze fixations. More in-depth investigations on human gaze feedback are needed in the future.

7. ACKNOWLEDGMENTS

This work was supported by IIS-1208390 from the National Science Foundation and N00014-11-1-0410 from the Office of Naval Research.

8. REFERENCES

- [1] 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 *Proceedings of 9th ACM/IEEE International Conference on Human-Robot Interaction*, Bielefeld, Germany, 2014.
- [2] H. Clark and A. Bangerter. *Changing ideas about reference*, pages 25–49. Experimental pragmatics. Palgrave Macmillan, 2004.
- [3] H. Clark and S. Brennan. Grounding in communication. *Perspectives on socially shared cognition*, 13:127–149, 1991.
- [4] H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22:1–39, 1986.
- [5] R. Dale. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19:233–263, 1995.
- [6] D. DeVault, N. Kariaeva, A. Kothari, I. Oved, and M. Stone. An information-state approach to collaborative reference. In *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions*, 2005.
- [7] R. Fang, M. Doering, and J. Y. Chai. Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 1544–1550, 2014.
- [8] R. Fang, C. Liu, L. She, and J. Y. Chai. Towards situated dialogue: Revisiting referring expression generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 392–402, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.
- [9] C. J. Fillmore. Towards a descriptive framework for spatial deixis. In R. J. Jarvella and W. Klein, editors, *Speech, Place, and Action*, pages 31–59. Wiley, Chichester, 1982.
- [10] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 74:381–391, 1954.
- [11] A. Gatt. Structuring knowledge for reference generation: A clustering algorithm. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, pages 321–328, 2006.
- [12] A. Gatt and P. Paggio. What and where: An empirical investigation of pointing gestures and descriptions in multimodal referring actions. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 82–91, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.
- [13] A. Gatt and P. Paggio. Learning when to point: A data-driven approach. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING '14)*, 2014.
- [14] S. Goldin-Meadow. The role of gesture in communication and thinking. *Trends Cogn. Sci.*, 1999.
- [15] P. A. Heeman and G. Hirst. Collaborating on referring expressions. *Computational Linguistics*, 21:351–382, 1995.
- [16] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [17] A. Koller, M. Staudte, K. Garoufi, and M. Crocker. Enhancing referential success by tracking hearer gaze. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '12*, pages 30–39, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [18] E. Kraehmer and K. V. Deemter. Computational generation of referring expressions: A survey. *computational linguistics*, 38(1):173–218, 2012.
- [19] C. Liu, R. Fang, and J. Y. Chai. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL '12*, pages 140–149, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
- [20] C. Liu, R. Fang, L. She, and J. Chai. Modeling collaborative referring for situated referential grounding. In *Proceedings of the SIGDIAL 2013 Conference*, pages 78–86, Metz, France, August 2013. Association for Computational Linguistics.
- [21] I. S. MacKenzie, A. Sellen, and W. A. S. Buxton. A comparison of input devices in element pointing and dragging tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '91*, pages 161–166, New York, NY, USA, 1991. ACM.
- [22] M. Mitchell, K. van Deemter, and E. Reiter. Generating expressions that refer to visible objects. In *Proceedings of NAAC-HLT 2013*, pages 1174–1184, 2013.
- [23] P. Piwek. Saliency in the generation of multimodal referring acts. In *Proceedings of the 2009 International Conference on Multimodal Interfaces, ICMI-MLMI '09*, pages 207–210, New York, NY, USA, 2009. ACM.
- [24] A. Sauppé and B. Mutlu. Robot deictics: How gesture and context shape referential communication. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction, HRI '14*, pages 342–349, New York, NY, USA, 2014. ACM.
- [25] I. F. V. D. Sluis. *Multimodal Reference, Studies in Automatic Generation of Multimodal Referring Expressions*. PhD thesis, Tulburg University, 2005.
- [26] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.
- [27] M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information during spoken language comprehension. *Science*, 268:1632–1634, 1995.
- [28] S. Tellex, R. Knepper, A. Li, D. Rus, and N. Roy. Asking for help using inverse semantics. In *Proceedings of Robotics: Science and Systems*, Berkeley, USA, July 2014.