Cooperative gazing behaviors in human multi-robot interaction

Tian Xu[1,2], Hui Zhang[3] & Chen Yu[1,2,4]
[1]Computer Science Department, Indiana University
[2]Cognitive Science Department, Indiana University
[3]Pervasive Technology Institute, Indiana University
[4]Psychological and Brain Sciences Department, Indiana University

When humans are addressing multiple robots with informative speech acts (Clark & Carlson 1982), their cognitive resources are shared between all the participating robot agents. For each moment, the user's behavior is not only determined by the actions of the robot that they are directly gazing at, but also shaped by the behaviors from all the other robots in the shared environment. We define cooperative behavior as the action performed by the robots that are not capturing the user's direct attention. In this paper, we are interested in how the human participants adjust and coordinate their own behavioral cues when the robot agents are performing different cooperative gaze behaviors. A novel gaze-contingent platform was designed and implemented. The robots' behaviors were triggered by the participant's attentional shifts in real time. Results showed that the human participants were highly sensitive when the robot agents were performing different cooperative gazing behaviors.

Keywords: human-robot interaction; multi-robot interaction; multiparty interaction; eye gaze cue; embodied conversational agent

1. Introduction

With the advance of robotics technology and prevalence of robots all around the world (Tapus et al. 2007; Goodrich & Schultz 2007; Rich & Sidner 2009), investigating and further improving collaborative behaviors among multiple robot agents has drawn keen interest among researchers. Such collaborations not only take places between teams of robots (Parker 1998; Dudek et al. 2002; Balch 2002; Farinelli et al. 2004), but more often between humans and robots (Duffy 2003; Dautenhahn 2007; Mutlu et al. 2009; Cakmak et al. 2011). For example, a human supervisor may guide a group of robot workers in an assembly line; or

Interaction Studies 14:3 (2013), **390–418**. DOI 10.1075/is.14.3.05xu
ISSN 1572–0373 / E-ISSN 1572–0381 © John Benjamins Publishing Company

a housekeeper may ask a group of robot assistants to clean a house, in which each robot is designated to finish a particular task (e.g. cleaning a floor, cleaning a desk); or a human team leader may give directions to a group of robot teammates working together on a rescuing operation (Casper & Murphy 2003). There are two major advantages to deploying multiple robots in those applications. First, a complicated task can be decomposed into a set of simple ones, each of which can be accomplished by individual robots in parallel. In the rescuing example, multiple robots can execute an exhaustive search for victims in a broad area, which may largely reduce the amount of search time (Kitano et al. 1999; Modi et al. 2005). Second, individual robots equipped with multiple advanced functions can be expensive to build while single-functioning robots can be low-cost and easy to maintain and operate (McLurkin et al. 2010; Rubenstein et al. 2012). Even though each individual robot can accomplish only a simple task, multiple robots are able to work together to perform more complex tasks (Cao et al. 1997; Parker 2008).

The above scenarios, however, pose a particular challenge in human-robot interaction: to establish smooth and effective interaction between human users and multiple robots. Every individual robot within a group will need to be aware of the user's internal state moment by moment, read it correctly, and coordinate with other robots to respond promptly and appropriately. To build robots that can infer the human's momentary cognitive state, a fundamental understanding of how humans generate real-time behaviors when reacting and responding to multiple robots in such reciprocal interaction is critical. Inferring the joint intention and cooperative activity of the whole group goes beyond just summing up each individual agent's intention (Demiris 2007). Instead, it involves an integration of mutual responsiveness, commitment to the joint activity and commitment to mutual support within a group (Bratman 1992). Thus, human multi-robot interaction poses additional questions and challenges over typical dyadic human-robot interaction, mainly because a user has to decide which robot he should attend to and interact with moment by moment, what behavior he should generate, and whether such behavior will be toward one robot, several of them or the whole group. Such decisions depend on responsive behaviors he observed from each individual robot and his inference of their "cognitive states". In such a context, the user's responsive behavior toward one robot is essentially influenced by not only the robot's behaviors that he or she directly interacts with but also the behaviors of the other robots in the entire group indirectly. To achieve the goal of building responsive social robots, we need to gain a deeper understanding of the effectiveness of the robot's behaviors on human participants during group interaction.

Eye gaze plays an important role in both human-human and human-robot group interactions (Sacks et al. 1974; Matsusaka et al. 2001; Bennewitz et al. 2005; Mutlu 2009). Vertegaal argued that eye gaze serves as a good indicator for predicting the group's overall intention from moment to moment and coordinating between speakers and listeners in a multi-party interaction (Vertegaal et al. 2001). During face-to-face interaction, people use eye gaze to monitor every member's attention within the group, show whom they are addressing, and suggest the next speaker during turn-takings (Kendon 1967; Isaacs & Tang 1994; Vertegaal et al. 2000). Mutlu and colleagues (Mutlu et al. 2009) have shown that when a single robot agent is interacting with multiple human participants, gaze cues from the robot regulate the whole communication and successfully signaled three types of listener roles proposed by Goffman: addressee, bystander and over-hearer (Goffman 1981). During the experiment, the human participants successfully interpreted and reacted to the gaze cues formed by the robot agent in over 97% of the cases.

So far, few studies have looked into the effect of gaze cues when multiple robots and one human user engage in a task. When a human participant talks to multiple robots, the robots should pay attention to the user's behaviors and carry out actions accordingly. Following the definition proposed by Clark and Carlson in 1982, the speech acts performed by the speaker when he or she is providing information and instructions to a group of addressees are categorized as informatives: "the fundamental kind of participant-directed illocutionary act is one by which the speaker jointly informs all the participants". Evidence has shown that essentially every traditional illocutionary act in group interaction is carried out in the form of informatives (Searle 1976). Similarly, during human multi-robot interaction, the human speaker has to ensure that both the directly addressed and other robot listeners recognize his or her intention. Thus, how should individual robots as addressees behave in such scenarios so that they can best facilitate group communication? For instance, should we program the robots to generate the same gaze behaviors toward the human user at the same time, all looking at the human's face or all looking away? When the human user pays attention to a particular robot, should the other robots look away from the human user to not interfere with mutual gaze between the target robot and the human, or alternatively, should they look at the human's face to compete for the human's attention? As shown in Figure 1, the focus of the present paper is to understand such **cooperative gazing behavior** in a multiparty interaction.

There are different possible cooperative gazing behaviors in the context of a human multi-robot interaction: robots can follow the human user's gaze; robots can actively direct the human user to look at itself or another location; or robots can encourage the user to pay more attention toward other robots. When the user is trying to convey information to the entire group, the user is likely to be engaged in mutual gaze with different robots from time to time, e.g. scanning through individual robots one by one while speaking. If different robot agents adopt different cooperative gazing strategies in the same interaction, would the users even be aware of different behaviors generated by individual robots within the robot group? And if so, how would the participants adjust their behaviors dynamically in real-time? How would robot agents with different cooperative gazing strategies influence each other in the same interaction? In this paper, we are taking the first step to explore these questions.

In the following sections, we will first introduce a human multi-robot interactive platform that we developed and used in the present study. Within this system, the robots' reactions are contingent on the participant's behaviors in real-time while each robot's reactive gazing behaviors are controlled independently. Built upon this interactive platform, different types of interactions are created by controlling the robots to adopt different types of cooperative gazing behaviors.

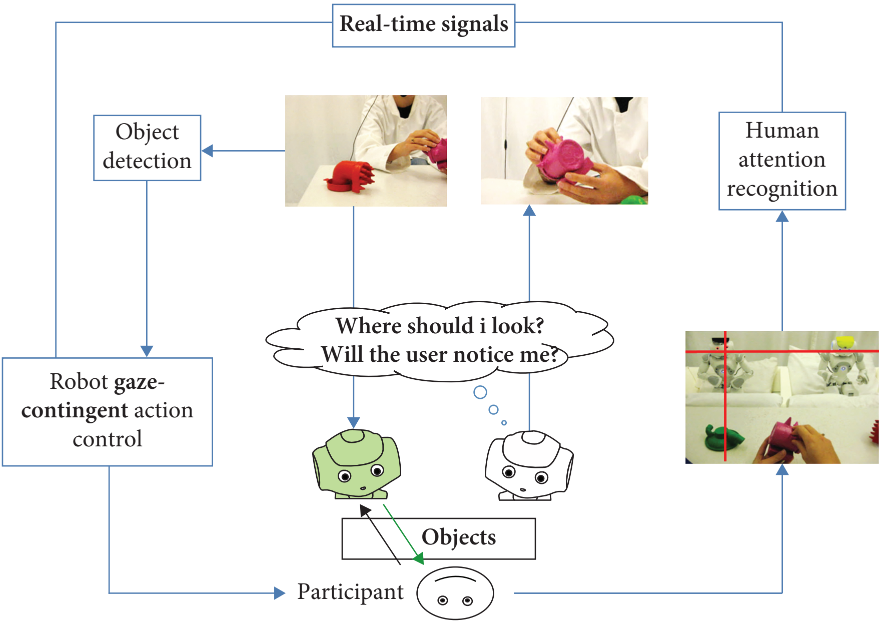

Figure 1. The system overview of our interactive platform. The human participants' behaviors were recorded, tracked and processed. Target objects attended to by the human were detected at each moment and the signals were fed into the robot action control module so that the robot agents were reacting to the participant's behaviors in real-time. The action control unit of each robot was independent from each other and different robots were programmed with different gaze-contingent behavioral strategies. The triadic interaction and the thought bubble in the center generated by the robot icon on the right side highlight the key scenario and question we were trying to explore in the designed experiment

2. A human multi-robot multimodal interactive paradigm

For this study, we implemented a multimodal gaze-contingent platform with two Nao humanoid robots manufactured by Aldebaran Robotics (Gouaillier et al. 2009). Using this platform, we asked human participants to interact with two robots at the same time to accomplish a joint task involving multiple objects on a table. Each robot has its own behavioral control unit and acted independently during the experiment. This platform allows us to tightly control the robots' responsive actions in a systematic way. As a result, we can analyze the participants' sensitivity and responsive behaviors to cooperative gaze behaviors performed by the two robots.

2.1 Gaze-Contingent platform

Figure 1 shows the overall structure of our multimodal human-robot interaction platform. In this system, the robot's gaze behavior is triggered by the participant's eye gaze in real-time. The human-robot interface detects the participant's visual attention through real-time eye tracking, and then generates gaze-contingent behaviors through the robot control component. Multiple sensing devices were used to record first-person view videos from both the participant's and the robots' perspectives, human speech, and most importantly, participant's gaze data. The two robots' head movements and gaze directions are also tracked. This rich dataset allows us to extract and analyze multimodal behavioral streams to discover fine-grained micro-level patterns in face-to-face human-robot interactions.

The Nao robot has 25 degrees of freedom. Its eye unit is made of a CMOS camera with an image resolution of 640 * 480 pixels at a sample rate of 30 frames per second with a view angle of 61° horizontally and 47° vertically. To track the participant's eye movements, a head-mounted eye tracker (Eye-Trac 6000, ASL LLC) was positioned on the participant's forehead. The eye tracker contains a scene camera to capture the first-person view video from the human's perspective at a rate of 30 frames per second. In addition, the eye tracking system outputs x and y coordinates in the first person video, indicating where the human participant is looking. Both gaze data and image frames were transferred to and processed by the human attention detection component in the platform. To facilitate the object detection process, the interaction environment was constituted by white curtains and objects with single solid colors. A method based on color blob detection was developed to automatically segment objects in the first-person view camera in real time (Yu et al. 2009). In addition, a unique visual marker was attached to each robot's head to facilitate the automatic detection of the robots' head positions in the human's view. The eye gaze location was integrated with the detected objects and robot head blobs in the first-person view to compute the target of attention – a Region-Of-Interest (ROI) at each moment indicating which target the participant was looking at. The ROI could be either one of the four objects or one of the two robots (see Figure 2 for details). The participant's current attention target was then sent to a robot action control unit to guide the robot's actions in real time. The whole procedure of image processing took about 50ms and the robot's execution of a head turn toward a target object took 250ms on average. In total, the potential lag from attention detection to gaze following was around 300ms.

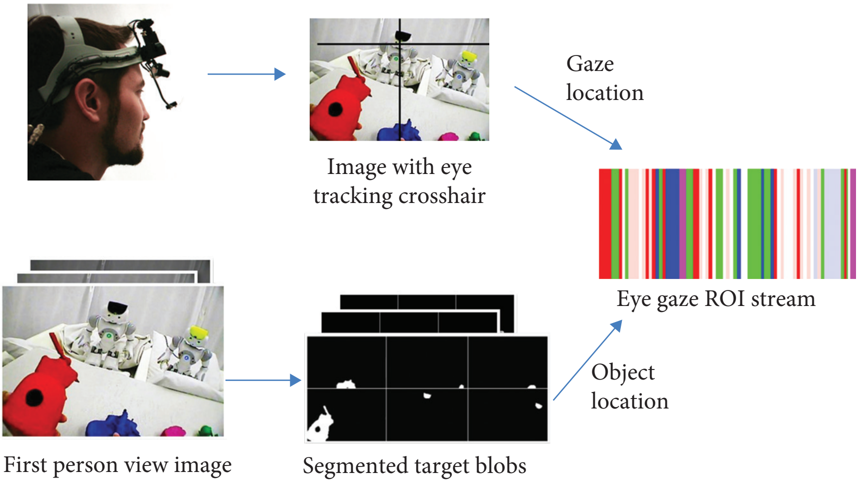

Figure 2. An overview of attention detection based on first-person view video and gaze data. Several visual targets (e.g. objects and robot heads) were detected and extracted based on color blobs in image frames. Next, (x, y) gaze location was superimposed on the first-person image to find the object of attention at each moment. For more detailed information in video processing, please see Yu et al. (2009) and Yu et al. (2010b)

A straightforward way to build a gaze-contingent robot control system is to follow exactly where the human looks moment by moment and react to the user's eye gaze shift as fast as possible. However, we noticed that in practice, this solution would not necessarily lead to real-time gaze following. At time, people briefly look at one location and quickly shift to another location. If the robot follows all of these rapid attention shifts, then the human's attention may already switch to the next ROI even before the robot completes the execution of a head turn to join the human with the detected ROI. Furthermore, even though the eye tracking system works quite well overall, it is inevitable that it may occasionally fail to track gaze direction. Therefore, the robot's attention control system needs to decide what to do at this moment without gaze data fed from eye tracking. To build a control system that can reliably respond to human gaze, we decided to keep track of not only the current gaze data point but also a set of 30 data points in the past which constitute a buffer of 1 second. Only after more than 50% of data points in the buffer indicated the same ROI would the control system send a motor command to switch the robot's attention. The 50% criterion was based on the analysis of empirical data collected from pilot experiments as well as the results reported on mean visual fixations during various viewing tasks (Le Meur et al. 2010). The buffered mechanism added an additional 350ms lag in the robot's response but gave us much more robust following behaviors from the robot.

By taking all the lags in this real-time control system into account, a thorough test revealed that the robot's response time was 657 ms on average. Human response times for attentional shifts range from 250ms to 350ms under different conditions in visual cuing experimental paradigms (Posner 1980). The gaze cueing effect emerged with varying stimulus-onset asynchronies (SOAs) from 300ms to 1005ms with modified Posner's cueing paradigms (Frischen et al. 2007). But these data were reported with on-screen visual stimulus in laboratories; the reaction time is probably longer in naturalistic social environment when interacting with other agents. A reaction time of 657ms can be considered as comparable to a human's response time during face-to-face social interaction. During our post-experiment interviews, most participants reported that the whole interaction was smooth and the robotic agents were engaging and actively following their attention.

2.2 Experiment design

We employed a word learning task where the human participants were asked to teach two robots the names of a set of objects. This task was selected for several reasons: (1) it has an explicit goal that allows participants to naturally interact with the robots while being constrained enough to make real-time processing of the robots' actions feasible; (2) it has been used successfully in a variety of behavioral studies investigating multimodal human-human interactions (Yu et al. 2009; Smith et al. 2010; Yu & Smith 2012) and human-robot interaction (Yu et al. 2012); (3) it allows us to investigate the fine-grained temporal patterns and relationships between human eye gaze and human speech as part of the larger joint attention processes; (4) beyond understanding the principles of human multi-robot interactions, the task itself has its own applied utilities. Instead of teaching each robot individually, it is more efficient for a human user to teach a group of robots so that all of them can acquire new knowledge simultaneously through social interaction.

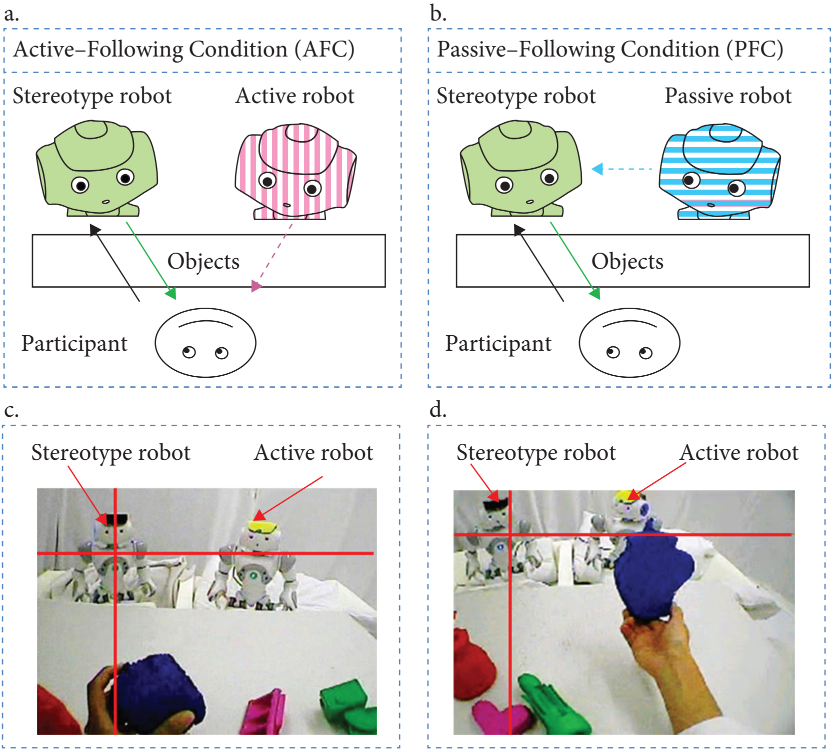

Figure 3. An overview of the experimental design: graph (a) is the active-following condition (AFC) wherein the human teacher was interacting with one stereotype robot and one active robot, graph (c) is a snapshot of this condition from the teacher's first-person view and the cross-hair in the image indicated where the teacher was looking at; graph (b) shows the passive-following condition (PFC) which includes one stereotype robot and one passive robot and (d) is a snapshot of this condition. Details about the behaviors of different robot agents are presented in the experimental design section

The robots' behaviors in the task were contingent on the human's attention in real-time as both robots "knew" where the human teacher was looking and then responded with their own behaviors accordingly. For example, when the user switched his attention to look at an object, the robot agents would follow the human teacher to gaze at the same target ROI as well; if the teacher initiated a face look toward one of the robots, the robot would gaze back at the human's face. By manipulating the robots' gaze-contingent control, we created three different types of robot learners to perform three different types of cooperative gaze behaviors.

1. **The stereotype robot** followed exactly the participant's gaze direction by looking at the same ROI that the participant attended to. When the participant started to look at the robot's face, the stereotype robot looked back to the participant's

face; when the participant looked away from the robot's face and switched to a target object, the robot also followed this gaze switch to look at the same target object; when the participant generated a face look to the other robot, the robot showed stereotypical cooperative gazing behavior by not changing its eye gaze and continuing to gaze at the ROI where it was looking at previously.

2. **The active robot**, in addition to following the participant's attentional switches in real time as the stereotype robot, generated additional face looks toward the human teacher to attract the teacher's attention when the teacher was looking at its robot peer (e.g. a stereotype robot). This situation is illustrated in Figure 3a: first the participant started to look at the stereotype robot; when the stereotype robot correctly detected this event, it looked back to the participant's face as a response; meanwhile, the active robot also detected the user's eye gaze switch, so it initialized a face look to the participant. As a result, both the active and stereotype robots looked at the teacher's face at the same moment as shown in Figure 3b.
3. **The passive robot** followed the same general gaze following strategies as the above stereotype robot. However, when human participant's eye gaze moved to the other robot peer, the passive robot imitated this behavior by looking at its robot peer as well – a passive cooperative gazing behavior. Thus, both the human participant and the passive robot looked at the other robot at the same time. This exact situation is shown in Figure 3c with a snapshot from the human's view shown in Figure 3d.

Based on these three types of learners, two experimental conditions were created. The active-following condition (AFC) included one stereotype robot learner and one active robot learner (Figure 3a and 3c); and the passive-following condition (PFC) includes one stereotype robot and one passive robot (Figure 3b and 3d). Thus, the same stereotype robot appeared in both conditions but was paired with different learning partners: either a more active or a more passive robot peer. This design allows us to directly compare the active and the passive gaze following behaviors of the other robot peer and to see how the human teachers adjusted their own behaviors in response to different situations of multi-robot interaction.

Figure 4 shows examples of gaze data from the two conditions, which reflected the dynamics of joint gaze activities between the two robots and the human participant. The focus of the present study is to understand momentary and dynamic gaze behaviors and speech acts from the human participant interacting with the two robots at the same time. In the active-following condition shown in Figure 4a, a human agent was interacting with an active and a stereotype robot. The first data stream is the ROI stream from the human's eye gaze, indicating which object or robot the human participant was attending to (e.g. gazing at one of the four objects

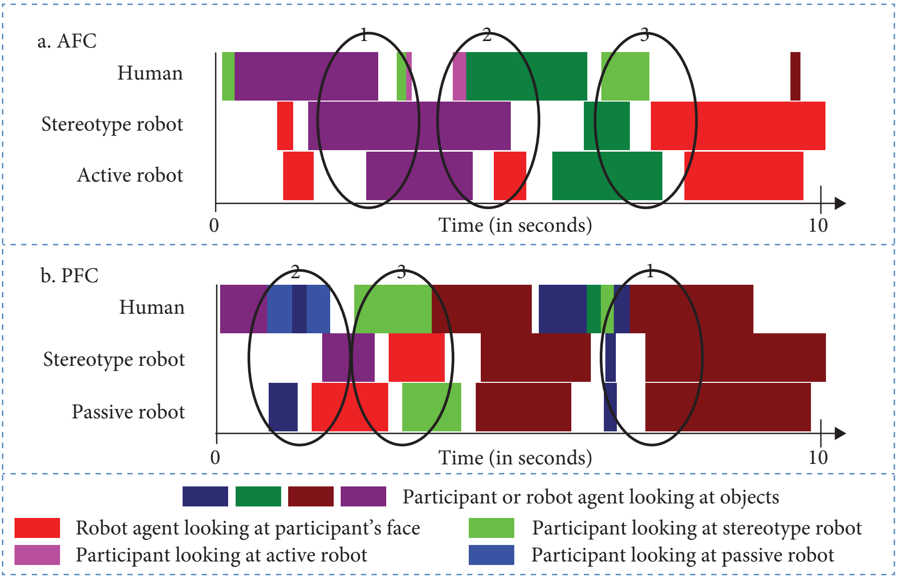

Figure 4. (a) The three data streams are derived from ROI gaze data of the user and the two robots (one stereotype robot and one active robot) in the active-following condition (AFC). (b) The three streams are derived from ROI gaze data of the participant, the stereotype robot and the passive robot in the passive-following condition (PFC). Details of the three labeled moments in each figure are explained in the experimental design section

or on one of the two robots). The other two data streams represent where the two robot looked. Three interactive behavioral patterns from those data streams (labeled from left to right on the top) are highlighted to illustrate the types of patterns we investigated using our multimodal real-time platform: (1) joint attention: all three agents visually attended to the same object (colored in magenta); (2) the human participant gazed at the active robot, the active robot responded by looking back at the human while the stereotype continued to look at the object; and (3) the human user gazed at the stereotype robot, the stereotype robot looked back at the human while the active robot also initialized a face look toward the human.

In the passive-following condition wherein a human agent was interacting with a passive robot and a stereotype robot, Figure 4b shows the three temporal gaze streams derived from eye tracking data from the participant (top), the stereotype robot (middle) and the passive robot (bottom). Three highlighted interaction moments are (1) joint attention: all three agents visually attend to the same object (colored in maroon); (2) face look at the passive robot: the human user gazed at the passive robot, the passive robot responded by looking back to the human while the stereotype robot kept its gaze on the object previously attended to; and (3) face look at the stereotype robot: the human user gazed at the stereotype robot, the stereotype robot looked back to the human and the passive robot joined the human to also gaze at the stereotype robot.

2.3 Hypotheses

Our main goal was to investigate whether different cooperative gazing behaviors between the two conditions have effects on both the teacher's behaviors and the whole group's interaction dynamics. Consequently, we developed five hypotheses:

Hypothesis 1. Participants will be sensitive to the different cooperative gazing behaviors demonstrated by different robot learners within each condition. Participants will look more at the active robot's face and generate fewer face looks to the stereotype robot and the passive robot. When multiple human participants are communicating with each other or interacting with a single humanoid robot, they use eye gaze to monitor their partner's attention allocation and react in real-time (Vertegaal et al. 2001; Rehm & André 2005; Mutlu et al. 2009). So we predict that in our experiment, the face looks generated by the users to one particular robot learner will be proportional to how often that robot looks towards the user's face. Thus, since the active robot was designed to initiate more face looks as its cooperative gazing behavior, the teachers will also look at the active robot more often.

Hypothesis 2. Participants will spend about the same amount of time looking at the stereotype robot in both conditions. The amount of attention focused on the stereotype robot will not be influenced by who its robot peer is between the two conditions. There are three possible outcomes: (1) No influence – participants will spend the same amount of time looking at the stereotype robot in each of the two conditions; (2) Facilitative effect – participants will generate more face looks to the stereotype robot in the active-following condition compared to the passive-following condition. If the active robot elicits more face looks from the participant, then participants may not only spend more time looking at the active robot but also generate more looks to its robot peer; (3) Negative effect – participants will look less to the stereotype robot in the active condition compared to the passive-following condition. Since the active robot will attract more attention from the participants, they will naturally look less toward the stereotype robot; and when the passive robot generates more looks to the stereotype robot, this behavior may encourage the human teacher to look more to the stereotype robot too. Among the three possible effects, we believe that as long as the stereotype robot has the same reactive behavioral strategy in both conditions, participants should also behave consistently toward the same type of learner across conditions.

Hypothesis 3. Participants will pay more attention to the two robot learners and look less at the objects in the active-following condition compared to the passive-following condition. Following our first and second hypothesis, we predict that since the active robot will attract more attention from the user, the overall proportion of time spent looking at the robots will be larger than in the passive-following condition. Consequently, the user will look less at the objects.

Hypothesis 4. Participants will generate more speech utterances in the active-following condition than in the passive-following condition. Utterances from the speaker are shaped collaboratively by the interactors as relevant and meaningful actions in the shared environment (Goodwin 2007). Speech behavior and eye movement are often closely coupled in human-human interaction and human-robot interaction (Kendon 1967; Staudte & Crocker 2009). In both dyadic and group conversions, human participants tend to divert eye gaze when they start to speak, and engage in more mutual gaze when they are about to terminate their turns in conservation (Vertegaal et al. 2001). This led us to predict that the amount of utterances generated by the users will be proportional to the amount of mutual face looks during interaction. Following the previous hypotheses, the teachers will generate more speech acts in the active-following condition than the passive-following condition since they are engaged in more mutual gaze moments in the active-following condition.

Hypothesis 5. Participants will show different gaze shifting patterns around naming moments between the two conditions. Here, we refer to the moments when the participants vocally pronounced the object names as "naming events" or "naming moments". When people describe a visual object, they tend to fixate on the target about one second before actually referring to it (Meyer et al. 1998; Griffin & Bock 2000). In addition, gaze alternation is considered a major form of joint attention: looking at a target object with interspersed glances to the social partner (Bard et al. 2008). In our learning task, the interaction was centered on teaching the object names to the robot learners. Therefore, it is pivotal to see whether we are able to observe this gaze-shifting pattern between the named object and the robot learners in our experiments and whether there will be differences between the two conditions.

2.4 Experimental procedure

The procedure was the same in both conditions. The participants were given four novel single-colored objects in each trial and each novel object was given an artificial two-syllable name (for example, gasser or kaki). The participants were asked to memorize the object names beforehand, and during the trial they were instructed to introduce these novel objects to both robot learners. Each trial lasted for about three minutes. There were three teaching trials in total after which an experimenter signaled the participants to stop and exchange the objects at the end of each trial. 10 undergraduate students at Indiana University were recruited and participated for the active-following condition and 11 for the passive-following condition (an additional 5 participants were excluded due to technical issues). After eliminating some trials where the proportions of valid eye-tracking data were exceptionally low (for example, 40% valid data compared to around 80% to 90% for the retained trials), the data set consisted of 28 trials in the active-following condition and 32 trials in the passive-following condition.

2.5 Data collection

Multimodal data were recorded during the experiment to examine how each agent responded to the others at the micro-behavioral level. As shown in Figure 2, eye movements from human participants were tracked and coded as 6 ROIs in the real-time image processing system (four objects and two robot agents). These frame-by-frame eye gaze data were recorded as sequential data streams with categorical values indicating the ROIs and then were further processed to detect attentional switches from one ROI to another. In addition, we also recorded the participant's speech during the experiment. An endpoint detection algorithm based on speech silence was implemented to segment a speech stream into spoken utterances. Each utterance may contain one or several words. Spoken utterances were transcribed by three human coders.

2.6 Validation of the gaze-contingent interaction system

To our best knowledge, the system presented here is one of the few gaze-contingent systems based on real-time eye tracking (Yoshikawa et al. 2006; Yu et al. 2012; Zhao et al. 2012). Due to the challenge of implementing and debugging a real-time system, validating the implementation is a necessary step. We first analyzed the gaze data from the robots collected in the experiment to ensure that the system was running smoothly and the robot learners were executing the correct cooperative gazing behaviors as designed. The amount of time the robots spent looking at different ROIs (excluding when the robots were executing their head turns) was normalized by the total experimental time to derive proportions. The results (see Table 1) show that in both conditions, both robots were either engaged in mutual gaze with the participants or looking at the objects for more than 80% of the duration of the experiments.

We also calculated the successful following rates of both robots in each condition. Successful following was defined as following the participant's gaze shift within a window of 2 seconds from the shift onset. The lag of 2 seconds was chosen for the

Table 1. Robot attention in the interaction

Proportion of total time	Active-following condition		Passive-following condition	
	stereotype robot	active robot	stereotype robot	passive robot
Looking at objects	71.55%	50.87%	73.93%	65.17%
Looking at the human's face	9.37%	30.34%	7.29%	4.04%
Looking at the other robot	0.00%	0.00%	0.00%	12.64%
sum	80.91%	81.21%	81.22%	81.85%

two reasons: (1) as mentioned in system implementation, there was a 657ms system lag even if a robot decided to follow immediately; and (2) there were situations where the target object or participant's face was not in the robot's view, and the robot had to conduct a visual search by turning his head to locate the target. Based on the 2-second lag, both robots in each condition were able to follow the participant's eye movements correctly for 60–70% out of all the gaze shift instances generated by the participants. As described in the implementation section, the robots would not follow the participant's attention when the participant just briefly glimpsed at an object and then switched attention to somewhere else. This is a sensible decision given the dynamic nature of a human's attention system. Even in human-human interactions, we will not be able to follow every single gaze shift from the other person.

Both the human participants and the robot learners were all actively engaged in the interaction and exhibited highly dynamic multimodal behaviors. Moreover, the different robot learners successfully executed looking behaviors according to the experimental designs. In the next section, we focus on how the human teacher's attention and speech were influenced by the behaviors from the two robots, both jointly and individually.

3. Results

In this section, we will first report the results derived from the gaze data, followed by the results from the speech acts. Next, we will integrate gaze and speech data to analyze dynamic multimodal patterns.

3.1 Eye movements

We first examined how long and how frequently participants looked at the two robots and objects in the interaction. Table 3 shows mean gaze duration and frequency of looks. The mean durations of participants' face looks at the two robots were consistent across the two experimental conditions. This shows that when participants decided to check one of the robots' faces, independent of the cooperative behavioral patterns of the robot, and independent of the particular experimental contexts, they always briefly looked at the robot's face for around 400ms before switching their attention back to an object or the other robot. Future studies in different task contexts will be needed to further test this. If we are able to find a similar face looking duration in other studies, this face-looking duration can be used as a reliable micro-level metric of human social attention in building human-robot interactions.

Table 2. Eye movement results on different measures

	Active-following condition			**Passive-following condition**		
	objects	**stereotype robot**	**active robot**	**object**	**stereotype robot**	**passive robot**
Mean duration (in seconds)	1.40	0.42	0.45	1.19	0.41	0.38
Frequency of looks (per minute)	35.39	10.75	13.32	43.33	9.35	5.58

We found significant differences in face look frequency across the conditions, showing that participants were indeed sensitive to the different cooperative gaze behaviors. Within the passive-following condition, participants looked at the stereotype robot more than the passive one (t(19) = 2.29, p = 0.03); (2). Within the active-following condition, participants looked at the active robot more than the stereotype robot on average, but it was not statistically significant (t(19) = 1.11, p = 0.29). We also compared the active and passive robots in the two experimental conditions as they shared the same social peer (the stereotype robot). There was a large difference in the frequency of face looks to the two robots (t(19) = 2.29, p = 0.03). Those results support our first hypothesis: at least in the present task, even though participants were teaching two robots as a group, they were sensitive to individual behavioral patterns from different robots and treated them differently.

Given that both experimental conditions had a stereotype robot, we next directly compared the participants' gaze behaviors toward the stereotype robot to test contextual influence when the same type of robot learner was accompanied by the different robot peers. We found no significant difference in the frequency of face looks (t(19) = 0.52, p = 0.61) nor the look durations (t(19) = 0.25, p = 0.80) toward the stereotype robot across the two conditions, which supported our second hypothesis. Thus, even though participants treated the two robots differently within each experimental condition, they treated the same robot type consistently across the two conditions. Moreover, we found that there was no significant difference in the duration of looks to the objects between the two conditions ($t(19) = 0.56$, $p = 0.58$). On average, the participants in the active-following condition generated fewer object looks compared with those in the passive-following condition, but the difference was not significant ($t(19) = 1.25$, $p = 0.23$). Thus, our third hypothesis was not well supported.

To summarize the results so far, there were no differences in looking durations from the participants toward the robots and objects. Moreover, there were no differences in their looking behaviors toward the stereotype robot in the two experimental conditions. But the participants in the active-following condition generated significantly more frequent face looks to the active robot compared to the passive robot in the passive-following condition. In the active-following condition, face looks toward the stereotype robot led to a consequential look from the active robot to the participant's face, which in turn may elicit the participants to switch their attention from the stereotype robot to the active robot instead. As a result, there were more face looks toward the active robot. In the passive-following condition, when participants attended to the stereotype robot, the passive robot also looked at its social peer to show its own interest toward the stereotype robot. Even though this behavior did not make participants look more or longer toward the stereotype robot, it did make them look less toward the passive robot.

3.2 Speech acts

On average, participants in the two conditions produced a similar number of words per minute ($M_{active} = 129.75$, $M_{passive} = 123.73$; $t(19) = 0.25$, $p = 0.80$) and their overall vocabulary size used in the experiment did not differ ($M_{active} = 180.40$, $M_{passive} = 176.36$; $t(19) = 0.16$, $p = 0.88$). In addition, the participants in both conditions generated a similar number of words per spoken utterance ($M_{active} = 5.22$; $M_{passive} = 5.97$, $t(19) = 0.95$, $p = 0.35$). However, the human teachers did generate fewer spoken utterances in the passive-following condition ($M_{passive} = 20.71$) compared to the active condition ($M_{active} = 26.45$; $t(19) = 2.36$, $p = 0.03$). This confirmed our fourth hypothesis: the different cooperative gazing behaviors demonstrated by the active robot and the passive robot have influenced the overall group dynamics by eliciting more speech utterances from the teachers in the active-following condition.

Since the task was to teach object names to the two robots, we were interested in how well the human teachers named objects for the two robot learners: in which condition the human participants presented better teaching behaviors by naming the objects more frequently and providing cues to help learners form correct word-object mappings. The participants generated 6.76 naming utterances per minute in the active condition and 4.98 naming utterances per minute in the passive condition, but this difference was not significant (t_{naming} (19) = 1.21, p = 0.24). We defined the moments when the participants produced object names as naming events. The similar numbers of naming events in the two conditions reflect only the quantity of naming instances in the interaction; the quality of naming instances is probably more critical for effective teaching, which is the focus of the multimodal data analysis reported next.

3.3 Attention dynamics around naming moments

When naming an object, the human teachers may attend to the named object most of the time; alternatively, they may look toward one or both of the two robots to check whether their attention was on the named object. Such gaze alternation patterns between the objects and social partners are essential for creating the joint attention moments that facilitate learning in human-human interaction (Carpenter et al. 1998; Tomasello et al. 2005; Yu et al. 2005; Meltzoff et al. 2009). Consequently, we focused on the moments right before and after a naming utterance, and measured the dynamics of participants' visual attention as a way to integrate speech data with gaze data.

There are four types of attentional shifts between different gaze ROIs in total: (1) shifts among different objects (e.g. from the named object to other objects); (2) shifts between objects and the stereotype robot; (3) shifts between objects and the active robot in the active-following condition, or between objects and the passive robot in the passive-following condition; (4) shifts between the two different robot learners. We then took an approach to analyzing temporal dynamics of attention switches which has been used in psycholinguistic studies to capture temporal profiles across a related class of events (Allopenna et al. 1998; Yu et al. 2009). Such profiles enable us to discern important moments within a trajectory and compare temporal trends across trajectories. To generate the temporal trajectories for the participants' attentional shifts prior to the naming events, we first aligned all the naming events by their onsets. Then, a 1000-msec moving window with 200-msec moving step was used. Within every window, we calculated the number of attentional switches from each participant, and then derived the mean value averaged across all participants. For example, in Figure 5, for the first data point of the brown line in subplot (a), we calculated the average frequency of attentional switches between different objects in AFC from 6.40 seconds to 5.40 seconds before the onset of naming events and placed the data marker at the center of this time window which is –5.90 seconds on the x-axis. For the next window which was 6.20 seconds to 5.20 seconds before the onset of naming events, the same calculation was carried out again to derive the second data point in the same trajectory.[1] We applied this calculation to all of the four types of attentional switch from moments prior to and following naming events in both conditions.

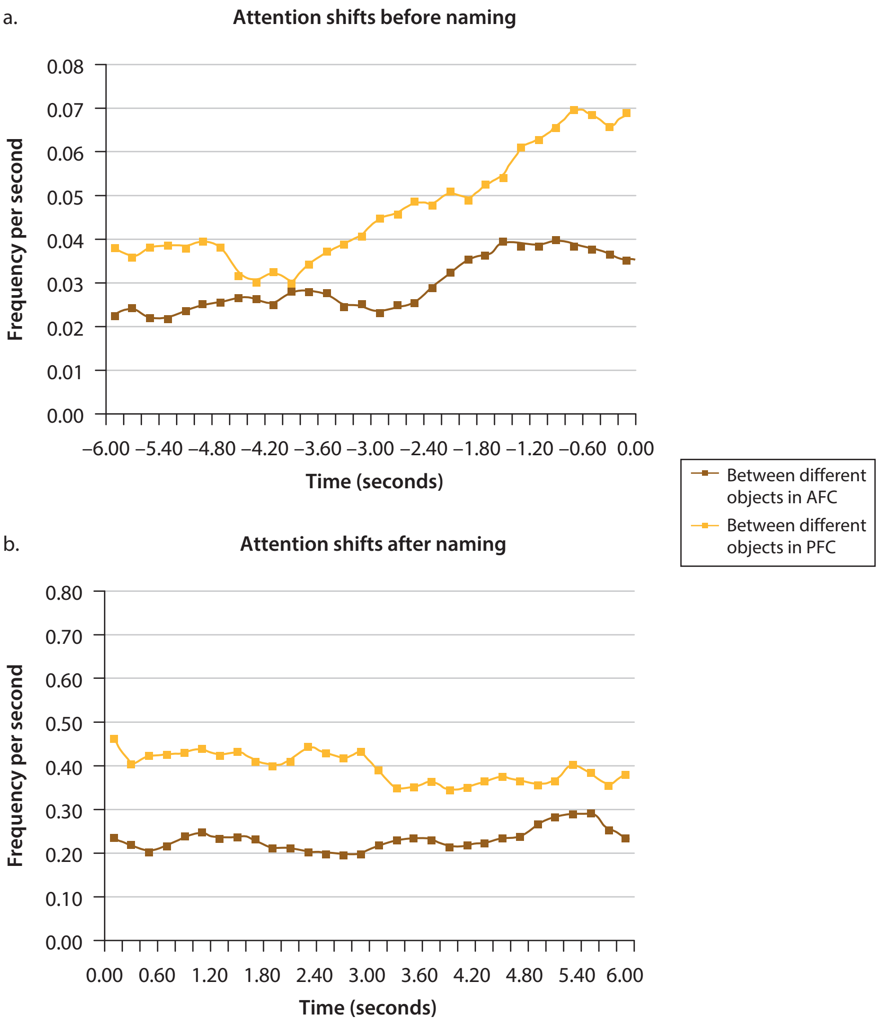

Figure 5. The participants' attentional shifts between different objects before and after naming events in both the active-following condition (AFC) and the passive-following condition (PFC). The subplot (a) depicts the trajectory before naming events; and subplot (b) describes the moments after naming events. All the naming moments were aligned by their onsets in (a) and by their offsets in (b). The detailed procedure for deriving the trajectories was explained in the text above

In Figure 5, the two subplots (a, b) showed the trajectories of attentional shifts between different objects before and after naming events. We tested the differences between the active-following condition and the passive-following condition by conducting two 2 (two between-subjects conditions) × 30 (frequency values from 30 moving time windows, the repeated measures factor) mixed-effects analyses of variance on the frequencies of gaze shifts both before and after the naming moments. There were significantly more attentional switches between different objects in the passive-following condition than in the active-following condition ($F_{before}(1,59) = 127.69$, $p < 0.0001$; $F_{after}(1,59) = 350.52$, $p < 0.0001$). In well-documented literature on the coupling between human eye movement and speech production (Meyer et al. 1998; Griffin & Bock 2000), people generated increased looks toward an object when mentioning or referencing it by its name in speech. Around naming events, gaze cues that provide accurate word-object associations are meaningful signals that can facililate learning for the robots. However, in the passive-following condition, the participants frequently switched their attention between the named object and other objects. This behavioral pattern can be confusing for learners to infer which attended object was the target that the teacher was referring to at that moment. Thus, compared with the passive-following condition, human teachers in the active-following condition provided better teaching signals with fewer attention switches between the target and irrelevant objects.

We further categorized the attentional switches according to named objects and non-named objects in each naming instance, and plotted the trajectories of attentional switches between the named objects and the two robot agents in both conditions. Figure 6 shows that in the active-following condition, the participants switched their attention more often between the target object and the active robot than the stereotype robot. A 2 (two robot learners) × 30 (frequency values from 30 moving time windows, the repeated measures factor) mixed-effects analysis of variance revealed that the pattern is statistically significant ($F_{before}(1,59) = 35.15$, $p < 0.0001$; $F_{after}(1,59) = 21.04$, $p = 0.0001$). Such gaze alternations created more joint attention moments between the participants and the active robot. This suggests that the participants cared more about the active robot's attentional state during naming than the stereotype robot. In the passive-following condition, participants switched their attention significantly more often between the target object and the stereotype robot than the passive robot before and after naming moments ($F_{before}(1,59) = 83.05$, $p < 0.0001$; $F_{after}(1,59) = 9.5$, $p < 0.005$). Taken together, human teachers consistently paid more attention toward the more active robot learner (the active robot in AFC and the stereotype robot in PFC) when they were referring to an object verbally by its name.

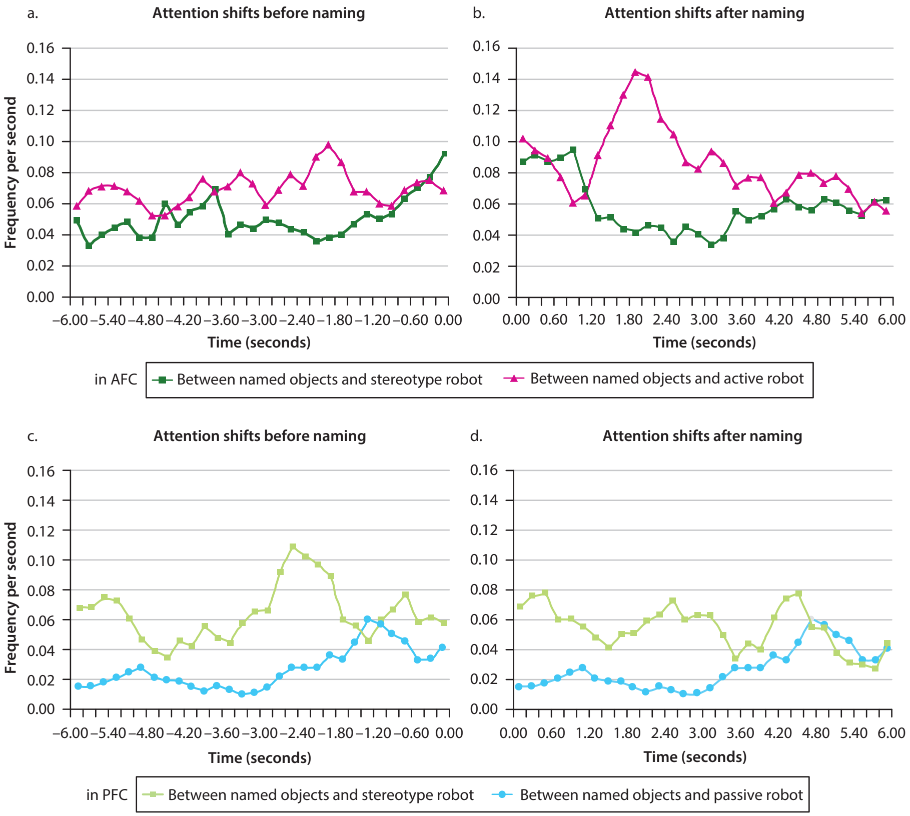

Figure 6. The attentional shifts before and after naming events between the targeted objects and the two robot agents in the active-following condition (AFC) and the passive following condition (PFC). The subplots (a, c) describe the trajectories before naming events and (b, d) are for after naming events. The frequencies of attentional switches were calculated with the same method as Figure 5 by taking a 1-second window at 200ms moving step along the time line. The detailed procedure for deriving the trajectories was explained in the main text

In summary, although there was no significant difference in the overall numbers of naming events between the two experimental conditions, a closer look at the multimodal data between speech and gaze data revealed different micro-level multimodal dynamics around naming moments which confirms our fifth hypothesis. Overall, the participants in the active-following condition paid more attention to the two robots (and primarily to the active robot) and provided more meaningful teaching behaviors in the active-following condition compared to participants in the passive-following condition. This suggests that actively eliciting the human teacher's attention in a multiparty interaction is beneficial to engaging the teacher in the interaction.

4. General discussions

In this paper, we focus on investigating a human multi-robot interaction scenario wherein human participants were addressing multiple robots at the same time by exploring the functionalities of gaze cue provided by robot listeners. Four of our five original hypotheses are well supported by behavioral data. Participants looked at the active robot more frequently compared to the passive robot and generated more speech utterances in the active following condition. The stereotype robot was treated consistently across the two conditions even though it was paired with a different robot peer. Furthermore, we not only found that the human participants were highly sensitive and responsive to different cooperative gaze behaviors, but also provided evidence that the participants were showing more meaningful teaching behaviors in the active-following condition. For the third hypothesis, we did observe that on average, the users generated fewer looks at the objects in the active-following condition, but the difference was not statistically significant.

4.1 Gaze cue in human multi-robot interaction

Due to the importance of gaze cues in human-human communication (Argyle & Cook 1976; Knapp et al. 2013), numerous studies have investigated the effectiveness of eye movement behaviors in human-robot interaction (Imai et al. 2003; Rich et al. 2009). During face-to-face interaction between a human subject and a single robot agent, participants are highly sensitive to the robot's gaze cues and tend to direct and shift their own gaze responsively in the same way as when they are interacting with other humans (Yu et al. 2010a). In Muhl and Nagai (2007), all the human participants responded immediately when the robot agent looked away from the participant's face and averted its attention at random moments. All this suggests that a robotic interactor needs to show appropriate responsive behaviors that may depend on different stages and contexts of conversation and interaction. Nakano and Nishida (2005) classified the user's eye movements into different categories and proposed different gaze strategies for each scenario. For example, in the scenario that the user has not paid attention to either the agent or the target display, in order to keep the interaction natural, the agent itself also needed to show disengagement signals accordingly.

When multiple human participants are engaged in the same interaction with a robot agent, the eye gaze cue from the robot still has a large impact in influencing and constructing the group communication. The study by Rehm and André (2005) showed that when two human participants were interacting with a virtual agent, not only did they accept this agent as a conversational partner but also paid more attention to the virtual agent than the other human participant.

In Trafton et al. (2008), the users rated the robot's behavior as more natural when the robot waited around 500ms after the speakers ended their turns then switched attention as opposed to responding immediately. Gaze cues from a robot agent can even regulate key moments and shape specific roles in human participants during multiparty conversations (Mutlu et al. 2009) and give precise direction to other members among a group of human participants (Kirchner et al. 2011). But we have not found reported results exploring situations when human users are interacting with multiple robots at the same time. We did not know whether the users would be sensitive to the robots' cooperative gaze behaviors and how they would dynamically adjust their attention allocation between different robot agents in real-time.

By implementing a real-time gaze contingent multi-robot interactive platform, we created three distinct types of robot learners showing different cooperative gaze behaviors in two experimental conditions (see Figure 3) to answer the above questions. The results revealed that different cooperative gaze behaviors can significantly influence the participants' gaze and speech acts. The active cooperative gazing behavior has attracted the most visual attention out of all three robot learners. And the passive-following gaze behavior did not succeed in directing more attention to its robot peer by imitating the participant's behavior to look toward the stereotype robot. This behavior has only made the participants pay significantly less attention to the passive robot itself compared to the other two robot agents.

In addition, as for the same type of robot agent between the two different conditions, the human participants paid a similar amount of attention to the stereotype robot even when it was paired with different robot peers between the two conditions. It shows that as long as the robot learners are programmed with the same behavioral strategy, the responses generated by the participants will not be largely influenced by the other robot's gazing behavior within the shared environment. Human participants treat the same robot agent consistently in different social contexts. These findings have provided valuable heuristics in constructing a social agent's behavior during group interaction.

4.2 Micro-level mutual reflexivity

In multiparty communications, the entire group will need to control, evaluate and integrate information conveyed by each participant to maintain common ground (Bales 1950). The key to mediating group interaction is for each agent to consider the subsequent behaviors from all the participants in the conversation and coordinate with them by conveying the correct behavioral cues. This is described as "Mutual Reflexivity" between speakers and listeners "within interaction participants treat their co-participants as reflexive actors" (Goodwin 2007). As a result, the actions of the listeners will considerably influence the speakers' ways of constructing the entire conversation. In the process of achieving such mutual reflexivity, eye movement is often used as a prominent source to monitor each other's momentary attention and regulate group communication (Kendon 1967; Vertegaal et al. 2000; Knapp et al. 2013). Here we provide empirical evidence derived from analyzing micro-level gaze data around naming speech acts, showing that the same phenomenon also holds in human multi-robot interaction.

During the word learning task, the human participants tried to teach object names to the two robot learners by repeatedly naming the objects and describing their features in detail. Thus, we extracted all the naming moments and calculated the attentional switches between different gaze ROIs that led up to, during and after the naming events. The teachers in the passive-following condition were frequently switching their visual attention between different objects around naming moments. To look at non-target objects during naming utterances would greatly confuse the learners on the account of which objects the teachers were referring to. Thus, the teachers in the active-following condition were showing better teaching behaviors. And they were checking the active robot significantly more often which has created more joint attention between the participants and the active robot. This suggested that the participants cared more about where the active robot was looking around naming utterances. In a shared environment, a group of more active learners will elicit more attention from human participants especially during key teaching moments.

5. Conclusion

There are two main contributions of this paper. First, we designed and implemented a real-time gaze contingent multi-robot platform that allows a human participant to interact with two robots performing different reactive behaviors in a shared environment. The participant's first-person view video and eye movement were recorded and processed in real time, so that the participant's gaze direction at each moment was sent to both robots as signals to trigger their responsive eye movements. With a word learning task, the participants as a language teacher were able to freely interact with the robot learners, engage their attention by switching eye gaze, speaking, manipulating the objects, gesturing and generating other nonverbal bodily cues.

Second, we present evidence that in human multi-robot interaction, the participants were highly sensitive to all the robot agents' gaze behaviors even when they were not directly looking at them. With the gaze contingent multi-robot platform, we created three different types of robot learners by choosing three different types of cooperative gazing behaviors. There were two experimental conditions: the active-following condition includes one stereotype robot and one active robot; and in the passive-following condition, the teachers were interacting with the same stereotype robot and a passive robot. We found that the teachers' responsive behaviors were significantly influenced by the different cooperative gazing behaviors. The teachers even showed more meaningful teaching behaviors in the active-following condition. Thus, active robot learners were more likely to engage human teachers to provide more and better teaching signals.

Due to the complexities of multiparty interaction, there are many topics left to be explored in human multi-robot interaction, such as how should the robot behave as an informant, or as a participant being directly addressed and taking turns to speak. With this platform, we are moving toward a more complete understanding of how human participants coordinate their own actions to different behaviors performed by multiple robot agents. This effort will lead to building a more naturalistic and effective participant model for social interactive robots and ultimately to building social robots to interact as a group with human users.

Acknowledgements

This research was supported by National Science Foundation Grant BCS 0924248. We thank Damian Fricker for recruiting participants and collecting data and Hongwei Shen for co-developing the image processing software. We are also grateful to two anonymous reviewers and Alexa Romberg for their insightful feedback and suggestions on this manuscript.

Notes

1. Other windows sizes were also considered; since we were deriving the frequency of gaze shifts, the numerical values as well as the overall temporal trends of the trajectories with other window sizes were very similar to the results presented here.

References

Allopenna, P.D., Magnuson, J.S., & Tanenhaus, M.K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439. DOI: 10.1006/jmla.1997.2558

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Oxford, England: Cambridge University Press.

Balch, T. (2002). Taxonomies of multirobot task and reward. In *Robot Teams: From Diversity to Polymorphism* (pp. 23–35). Natick, MA: A K Peters/CRC Press.

Bales, R.F. (1950). A set of categories for the analysis of small group interaction. *American Sociological Review*, *15*(2), 257–263. DOI: 10.2307/2086790

Bard, K.A., & Leavens, D.A. (2008). Socio-emotional factors in the development of joint attention in human and ape infants. In *Learning From Animals? Examining the Nature of Human Uniqueness* (pp. 89–104). London: Psychology Press.

Bennewitz, M., Faber, F., Joho, D., Schreiber, M., & Behnke, S. (2005). Towards a humanoid museum guide robot that interacts with multiple persons. In *Humanoid Robots, 2005. 5th IEEE-RAS International Conference* (pp. 418–423). IEEE.

Bratman, M.E. (1992). Shared cooperative activity. *The Philosophical Review*, *101*(2), 327–341.

Cakmak, M., Srinivasa, S.S., Lee, M.K., Kiesler, S., & Forlizzi, J. (2011). Using spatial and temporal contrast for fluent robot-human hand-overs. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 489–496). ACM.

Cao, Y.U., Fukunaga, A.S., & Kahng, A. (1997). Cooperative mobile robotics: Antecedents and directions. *Autonomous Robots*, *4*, 1–23. DOI: 10.1023/A:1008855018923

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4)1–143. DOI: 10.2307/1166214

Casper, J., & Murphy, R.R. (2003). Human-robot interactions during the robot-assisted urban search and rescue response at the world trade center. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *33*(3), 367–385. DOI: 10.1109/TSMCB.2003.811794

Clark, H.H., & Carlson, B.T. (1982). Hearers and speech acts. *Language*, *58*, 332–373.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 679–704.

Demiris, Y. (2007). Prediction of intent in robotics and multi-agent systems. *Cognitive Processing*, *8*(3), 151–158.

Dudek, G., Jenkin, M., & Milios, E. (2002). A taxonomy of multi-robot systems. In T. Balch & L.E. Parker (Eds.), *Robot teams* (pp. 3–22). Natick, MA: A K Peters.

Duffy, B.R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*(3), 177–190. DOI: 10.1016/S0921-8890(02)00374-3

Farinelli, A., Iocchi, L., & Nardi, D. (2004). Multirobot systems: A classification focused on coordination. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(5), 2015–2028.

Frischen, A., Bayliss, A.P., & Tipper, S.P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*(4), 694.

Goffman, E. (1981). *Forms of talk*. Philadelphia, PA: University of Pennsylvania Press.

Goodrich, M.A., & Schultz, A.C. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, *1*(3), 203–275. DOI: 10.1561/1100000005

Goodwin, C. (2007). Interactive footing. *Studies in Interactional Sociolinguistics*, *24*, 16.

Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., & Maisonnier, B. (2009). Mechatronic design of NAO humanoid. In *IEEE International Conference on Robotics and Automation* (pp. 769–774). IEEE.

Griffin, Z.M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*(4), 274–279. DOI: 10.1111/1467-9280.00255

Isaacs, E.A., & Tang, J.C. (1994). What video can and cannot do for collaboration: A case study. *Multimedia Systems*, *2*(2), 63–73. DOI: 10.1007/BF01274181

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22–63.

Kirchner, N., Alempijevic, A., & Dissanayake, G. (2011). Nonverbal robot-group interaction using an imitated gaze cue. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (pp. 497–504). ACM.

Kitano, H., Tadokoro, S., Noda, I., Matsubara, H., Takahashi, T., Shinjou, A., & Shimada, S. (1999). Robocup rescue: Search and rescue in large-scale disasters as a domain for autonomous agents research. In *IEEE International Conference on Systems, Man, and Cybernetics*, (Vol. 6, pp. 739–743). IEEE.

Knapp, M., Hall, J., & Horgan, T. (2013). *Nonverbal communication in human interaction*. (*8th ed.*). Boston: Wadsworth/Cengage Learning.

Imai, M., Ono, T., & Ishiguro, H. (2003). Physical relation and expression: Joint attention for human-robot interaction. *IEEE Transactions on Industrial Electronics, 50*(4), 636–643. DOI: 10.1109/TIE.2003.814769

Le Meur, O., Ninassi, A., Le Callet, P., & Barba, D. (2010). Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric. *Signal Processing: Image Communication*, *25*(7), 547–558. DOI: 10.1016/j.image.2010.05.006

Matsusaka, Y., Fujie, S., & Kobayashi, T. (2001). Modeling of conversational strategy for the robot participating in the group conversation. In *Proceedings of European Conference on Speech Communication and Technology*, (Vol. 1, pp. 2173–2176).

McLurkin, J., Lynch, A., Rixner, S., Barr, T., Chou, A., Foster, K., & Bilstein, S. (2010). A low-cost multi-robot system for research, teaching, and outreach. In *Proceedings of the Tenth International Symposium on Distributed Autonomous Robotic Systems* (pp. 597–609). Springer Berlin Heidelberg.

Meltzoff, A.N., Kuhl, P.K., Movellan, J., & Sejnowski, T.J. (2009). Foundations for a new science of learning. *Science*, *325*(5938), 284–288.

Meyer, A.S., Sleiderink, A.M., & Levelt, W.J. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition*, *66*(2), B25-B33. DOI: 10.1016/S0010-0277(98)00009-2

Modi, P.J., Shen, W.M., Tambe, M., & Yokoo, M. (2005). ADOPT: Asynchronous distributed constraint optimization with quality guarantees. *Artificial Intelligence*, *161*(1), 149–180.

Muhl, C., & Nagai, Y. (2007). Does disturbance discourage people from communicating with a robot? In *16th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 1137–1142). IEEE.

Mutlu, B. (2009). *Designing gaze behavior for humanlike robots.* Doctoral dissertation, Northwestern University.

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., & Ishiguro, H. (2012). Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems*, *1*(2), 12.

Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., & Hagita, N. (2009). Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (pp. 61–68).

Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., & Hagita, N. (2009). Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In *Proceedings of 4th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 69–76).

Nakano, Y.I., & Nishida, T. (2005). Awareness of perceived world and conversational engagement by conversational agents. In *Proceedings AISB 2005. Symposium Conversational Informatics for Supporting Social Intelligence and Interaction-Situational and Environmental Information Enforcing Involvement* (pp. 128–134).

Parker, L.E. (1998). ALLIANCE: An architecture for fault tolerant multirobot cooperation. *IEEE Transactions on Robotics and Automation*, *14*(2), 220–240. DOI: 10.1109/70.681242

Parker, L.E. (2008). Distributed intelligence: Overview of the field and its application in multi-robot systems. *Journal of Physical Agents*, *2*(1), 5–14.

Posner, M.I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. DOI: 10.1080/00335558008248231

Rehm, M., & André, E. (2005). Where do they look? Gaze behaviors of multiple users interacting with an embodied conversational agent. In *Proceedings of the 5th International Workshop on Intelligent Virtual Agents* (pp. 241–252). Panayiotopoulos T., Gratch J., Aylett R., Ballin D., Olivier P., Rist T. (Eds.). Springer Berlin: Heidelberg.

Rich, C., Ponsler, B., Holroyd, A., & Sidner, C.L. (2010). Recognizing engagement in human-robot interaction. In *Proceedings of the 5th ACM/IEEE International Conference on Human Robot Interaction* (pp. 375–382).

Rich, C., & Sidner, C.L. (2009). Robots and avatars as hosts, advisors, companions, and jesters. *AI Magazine*, *30*(1), 29.

Rubenstein, M., Ahler, C., & Nagpal, R. (2012). Kilobot: A low cost scalable robot system for collective behaviors. In *IEEE International Conference on Robotics and Automation* (pp. 3293–3298). IEEE.

Sacks, H., Schegloff, E.A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, *50*(4), 696–735. DOI: 10.2307/412243

Searle, J.R. (1976). A classification of illocutionary acts. *Language in Society*, *5*(1), 1–23. DOI: 10.1017/S0047404500006837

Smith, L.B., Yu, C., & Pereira, A.F. (2010). Not your mother's view: The dynamics of toddler visual experience. *Developmental Science*, *14*(1), 9–17. DOI: 10.1111/j.1467-7687.2009.00947.x

Staudte, M., & Crocker, M.W. (2009). Visual attention in spoken human-robot interaction. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction* (pp. 77–84). DOI: 10.1145/1514095.1514111

Tapus, A., Mataric, M.J., & Scassellati, B. (2007). Socially assistive robotics. [Grand Challenges in Robotics]. *IEEE Robotics Automation Magazine*, *14*(1), 35–42.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–690.

Trafton, J.G., Bugajska, M.D., Fransen, B.R., & Ratwani, R.M. (2008). Integrating vision and audition within a cognitive architecture to track conversations. In *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction* (pp. 201–208).

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 301–308). ACM.

Vertegaal, R., van der Veer, G., & Vons, H. (2000). Effects of gaze on multiparty mediated communication. In *Proceedings of Graphics Interface* (pp. 95–102). Montreal, Canada: Morgan Kaufmann Publishers.

Yoshikawa, Y., Shinozawa, K., Ishiguro, H., Hagita, N., & Miyamoto, T. (2006). Responsive robot gaze to interaction partner. In *Proceedings of Robotics: Science and Systems*.

Yu, C., Ballard, D.H., & Aslin, R.N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, *29*(6), 961–1005. DOI: 10.1207/s15516709cog0000_40

Yu, C., Schermerhorn, P., & Scheutz, M. (2012). Adaptive eye gaze patterns in interactions with human and artificial agents. *ACM Transactions on Interactive Intelligent Systems*, *1*(2), 13: 1–13: 25.

Yu, C., Scheutz, M., & Schermerhorn, P. (2010a). Investigating multimodal real-time patterns of joint attention in an HRI word learning task. In *Proceedings of the 2010. 5th ACM/IEEE International Conference on Human-Robot Interaction* (pp. 309–316). IEEE.

Yu, C., & Smith, L.B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244–262. DOI: 10.1016/j.cognition.2012.06.016

Yu, C., Smith, L.B., Shen, H., Pereira, A.F., & Smith, T. (2009). Active information selection: Visual attention through the hands. *IEEE Transactions on Autonomous Mental Development*, *1*(2), 141–151. DOI: 10.1109/TAMD.2009.2031513

Yu, C., Smith, T., Hidaka, S., Scheutz, M., & Smith, L. (2010b). A data-driven paradigm to understand multimodal communication in human-human and human-robot interaction. *Advances in Intelligent Data Analysis IX* (pp. 232–244). Cohen P., Adams N., Berthold M. (Eds.). Springer Berlin: Heidelberg.

Zhao, Q., Yuan, X., Tu, D., & Lu, J. (2012). Multi-initialized states referred work parameter calibration for gaze tracking human-robot interaction. *International Journal of Advanced Robotic Systems*, *9*(75). DOI: 10.5772/50891

Authors' addresses

Tian Xu
Computer Science Department,
Indiana University,
150 S. Woodlawn Ave.,
Bloomington, IN, 47405

Cognitive Science Department,
Indiana University,
1101 E 10th St.,
Bloomington, IN, 47405

Hui Zhang
Pervasive Technology Institute,
Indiana University,
IT414H
Indianapolis, IN, 46202

Chen Yu
Computer Science Department,
Indiana University,
150 S. Woodlawn Ave.,
Bloomington, IN, 47405

Cognitive Science Department,
Indiana University,
1101 E 10th St.,
Bloomington, IN, 47405

Psychological and Brain Sciences Department,
Indiana University,
1101 E 10th St.,
Bloomington, IN, 47405

Authors' biographical notes

Tian Xu (Linger) is currently a Ph.D. candidate in the Computer Science Department and the Cognitive Science Department at Indiana University Bloomington. By grounding the process of "mind-meeting" and coordination between embodied conversational agents in multimodal social signals, she is interested in utilizing robotic platforms and novel data mining techniques to explore the mystery of efficient human communication and learning during social interaction. She obtained her bachelor degree in Software Engineering at Nanjing University in 2009 and her master degree in Computer Science at Indiana University Bloomington in 2012.

Hui Zhang is a principal researcher in the Pervasive Technology Institute at the Indiana University at Bloomington, with an adjunct Assistant Professor appointment in the IU Cognitive Science Program. His research spans a range of topics including scientific data visualization, visual analysis with applications across domain sciences, data intensive analysis with a focus on the development of scalable techniques that are efficient for advanced computing infrastructures, and big data visual analytics techniques that combine scalable machine computation with interactive visualizations.

Chen Yu is Associate Professor in the Department of Psychological and Brain Sciences, Cognitive Science Program, and School of Informatics at Indiana University Bloomington. He holds his Ph.D. degree in computer science from University of Rochester. His research focuses on using head-mounted eye tracking, computational modeling and data mining techniques to understand mechanisms of coordinated behavior, early word learning, statistical learning, selective attention, perception-action coupling and social interaction. Yu received the 2003 David Marr best paper award from the Cognitive Science Society and the 2008 early career award from the International Society on Infant Studies.