Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters

Sean Andrist,[1] **Michael Gleicher,**[2] **Bilge Mutlu**[2]
(1) Microsoft Research, Redmond, WA, USA
(2) Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI, USA
sandrist@microsoft.com; gleicher@cs.wisc.edu; bilge@cs.wisc.edu

Figure 1. Left: A user wears eye-tracking glasses to collaboratively assemble a sandwich with a virtual character. Middle: The virtual character produces gaze cues to relevant task objects. Right: A user interacting with the virtual character in head-mounted virtual reality.

ABSTRACT
Successful collaboration relies on the coordination and alignment of communicative cues. In this paper, we present mechanisms of *bidirectional gaze*—the coordinated production and detection of gaze cues—by which a virtual character can coordinate its gaze cues with those of its human user. We implement these mechanisms in a hybrid stochastic/heuristic model synthesized from data collected in human-human interactions. In three lab studies wherein a virtual character instructs participants in a sandwich-making task, we demonstrate how bidirectional gaze can lead to positive outcomes in error rate, completion time, and the agent's ability to produce quick, effective nonverbal references. The first study involved an on-screen agent and the participant wearing eye-tracking glasses. The second study demonstrates that these positive outcomes can be achieved using head-pose estimation in place of full eye tracking. The third study demonstrates that these effects also transfer into virtual-reality interactions.

ACM Classification Keywords
H.5.2 **Information Interfaces and Presentation**: User Interfaces—*evaluation/methodology, user-centered design*

Author Keywords
Bidirectional gaze; gaze coordination; interactive gaze; dyadic gaze; joint attention; embodied agents; verbal referencing

CHI 2017, May 06 - 11, 2017, Denver, CO, USA.

ACM 978-1-4503-4655-9/17/05$15.00
DOI: http://dx.doi.org/10.1145/3025453.3026033

INTRODUCTION
When people interact they use a number of verbal and nonverbal communication mechanisms to coordinate. Gaze is a particularly important cue in both directions; people use it to indicate their attention as well as sense the attention of others. For example, an instructor might observe the gaze of their student to see that they are looking in the wrong place or are seeking help and then use their own gaze to capture the student's attention in order to guide it to the correct place. Such bidirectional gaze mechanisms can improve coordination in interaction by correcting potential failures before they occur in a subtle way, avoiding interruptions in the flow of activity.

Interfaces utilizing virtual embodied agents hold great promise for situated interaction in domains such as work training, occupational therapy, rehabilitation, counseling, retail, education, entertainment, and more. To build rich, immersive, and fluent interactive experiences with agents in these settings, we must build models that they can use to coordinate their actions and behaviors with their users and shared objects in the task environment. In this paper, we present techniques that improve the quality of human-agent interactive experiences through the use of *bidirectional gaze*—the coordinated production and responsiveness to social gaze cues—and demonstrate that these techniques indeed achieve positive interaction outcomes.

Bidirectional gaze is particularly important when people collaborate over a shared visual space [11, 13, 15, 45]. Coordinated gazing allows conversational participants to monitor their interlocutor for understanding, regulate the amount of mutual gaze and averted gaze, quickly pass and receive the conversational floor, disambiguate verbal references early in their production, and so on. Virtual agents currently lack sophisticated models that would allow them to engage in similar

coordination when interacting with people. The opportunity exists for agents to use their embodiments to express social gaze cues and through the use of gaze tracking technologies detect and interpret the user's gaze cues. However, these gaze cues must be simultaneously expressed by the agent and interpreted from the user in an interactive, dynamic process.

The primary contribution of our paper is a bidirectional gaze model that enables an agent to interpret the gaze of its user and generate its own gaze to effectively communicate coordinative behaviors. We utilize data collected in a previous human-subjects experiment to inform the design of a stochastic finite-state-machine model with heuristic rules that can respond in real time to streams of signals, including eye-tracker information, to drive the animation of a character's gaze. A user study ($n = 32$) utilizing an on-screen agent system validates this model, showing that bidirectional gaze behaviors can lead to positive collaborative outcomes in error rate, completion time, subjective quality, and the ability of the agent to produce quick references that require little verbal disambiguation.

In order to make the model more practical, we also explore removing the need for accurate spatial gaze tracking. We provide methods that use easily obtained head tracking information as a proxy for detailed gaze information and validate that this approximation achieves desired results in a second empirical study ($n = 12$). This ability to use head-tracking data to achieve bidirectional gaze effects also makes our approach practical for virtual reality applications. We demonstrate this in a third version of our system using an Oculus Rift head-mounted display, and confirm that we can achieve the desirable effects of bidirectional gaze in a third user study ($n = 20$).

BACKGROUND

Previous research on gaze behavior has characterized it as a key mechanism for communication and coordination in human interactions and a potential resource for creating natural and effective interactive experiences involving embodied agents.

Gaze in Coordination and Collaboration

Human interactions involve individuals drawing on social exchanges to coordinate their actions toward changing their environment [13, 14, 46]. In these interactions, communicators seek to establish *common ground* [17, 13, 46] by exchanging information and to obtain *situation awareness* by monitoring visual information available in the environment[19, 16, 22]. This awareness enables them to predict breakdowns in coordination and engage in *repair* [26], using language to re-establish common ground, such as a teacher seeing that a student appears confused and offering clarification.

Gaze cues facilitate both the process of establishing common ground and the process of engaging in repair. Conversational partners monitor each others' gaze and engage in shared gaze to indicate attention to and understanding of references to objects [16, 5, 21, 9]. Breakdowns in understanding or need for more information by listeners can be judged based on whether or not their attention is directed toward referents [5]. When breakdowns do occur, gaze cues of the speaker serve to rapidly disambiguate references [24]. The gaze patterns of a partner can also help predict ensuing task actions [50] and cognitive processes such as language comprehension [48].

This continuous process of grounding and repair is facilitated by *gaze coordination*, the emergent coupling of gaze patterns between conversational partners [43]. This coordination signals how well speakers and listeners achieve visual common ground [41, 5] and predicts conversational outcomes such as listener comprehension [42]. Gaze coordination is an efficient way to facilitate collaboration, reducing the cost of language production for coordination and repair [17, 8], particularly for rapid communication of spatial information [36].

Previous research on human communication has examined how people engage in gaze coordination, such as the timings of when people look toward objects to which they or their partners verbally refer [23, 33, 48]. These investigations are generally one-sided, looking at each person's gaze in isolation, and do not capture the intricate coordinative patterns in which partners' gaze behaviors interact. Previous work has also investigated *gaze alignment*, exploring the extent to which conversational partners gaze toward the same targets at various time offsets [41, 5]. This paper extends the human communication work with a computational model of gaze coordination and alignment applicable to embodied agents in HCI.

Gaze in Coordination with Embodied Agents

Previous work has extensively explored how embodied agents can use the production of gaze cues for a number of social functions, e.g., signaling attention [40, 38], spatial referencing [10, 3], and action coordination [7, 34]. However, much less work has considered the use of gaze as an input for agent interaction. Previous studies have explored how embodied agents can monitor the gaze of their users and seek to establish shared attention by aligning their gaze with those of their users [51]; by selectively using gaze shifts, head motion, and voice depending on user orientation [27]; and by dynamically engaging in mutual and averted gaze [6]. Prior research also includes studies of gaze coordination between humans and robots. Skantze et al. [47] investigated user interactions with a robot that employed coordination mechanisms such as joint attention, turn-taking, and action monitoring, finding that these mechanisms facilitated reference disambiguation. Similarly, Mehlmann et al. [32] devised a heuristics-based gaze model for conversational grounding that allowed a robot to utilize user gaze in order to disambiguate user speech, establish joint attention, and regulate turn-taking.

Another line of work among these studies involved enabling embodied agents to monitor user gaze for signals of communication or task breakdowns and to offer repair. These studies included the development of a spoken dialogue system that monitored user gaze patterns to determine the need for reference resolution [12]; a robot system that monitored user gaze as they performed an assembly task and proactively provide task assistance [44]; an intelligent tutoring system that used gaze direction to determine whether or not the student lost attention and prompted the student to pay attention to the tutoring [18]; and a robot system that monitored user gaze and task actions to predict hesitation or breakdowns in understanding and offer additional task information [49]. Assessments of

these systems have demonstrated that, by monitoring, inferring user states based on, and responding to gaze, agents can improve coordination with their users.

Although previous work has extensively studied human gaze coordination and explored integrating some of its mechanisms into the design of embodied agents, humanlike gaze coordination with such agents is still an unrealized goal. This paper seeks to address this knowledge gap by presenting a model of bidirectional gaze, allowing virtual characters to express gaze and react to user gaze in a coordinated fashion in order to create richer and more natural interactive experiences. Our work extends prior results on the use of coordinated gaze behaviors between humans and embodied agents by providing a model constructed directly from human data, implementing the behaviors using new technologies, and providing a unique methodology for evaluation.

MODELING BIDIRECTIONAL GAZE

Our full model of bidirectional gaze is comprised of two major components: a stochastic finite-state-machine with statistical parameters of what to gaze toward, independent of the user, and a heuristic rule-based component on what to do in response to the user's gaze. In this section we first describe a baseline descriptive model of human bidirectional gaze behaviors from previous work. We then present our work to extend the previous analyses into a generative model by turning descriptive statistics into stochastic parameters to a finite-state-machine. To add gaze responsiveness, we also performed extra analysis on the existing data and developed several heuristics to inform the agent on what to do in response to user gaze. At the end of this section we describe a virtual agent system which we developed to utilize the bidirectional gaze model in a physical instruction-based task. We also describe two additional versions of the system that we developed to utilize low-cost head tracking and to situate the interaction in virtual reality.

Human-Human Data & Descriptive Model

The general class of tasks that we focus on in this paper includes physical collaborations in which one participant provides instructions to the other participant. One of the basic building blocks of physical collaborations is the *reference-action sequence*: one participant makes a verbal reference to a physical object and the other participant performs an action on that object. Previous work in psychology, particularly recent work by Andrist et al. [1], has analyzed the high-level structure of reference-action sequences and provided a qualitative understanding of gaze coordination within such sequences. Our model development builds on this work, using data obtained by Andrist et al. [1] as well as the descriptive model resulting from their analysis. The paragraphs below will briefly describe the task and setting in which this data was collected in order to provide the reader with context on our model.

Following Andrist et al. [1], the model scenario we utilize throughout this paper is a sandwich-building task—representing a real-world scenario that most non-experts should be familiar with. In this task, one participant—the *instructor*—provides verbal instructions to the other participant—the *worker*—on what sandwich ingredients, arranged on a table between the participants, to physically move onto some nearby bread. The instructor is directed to request a single ingredient at a time, not to physically touch any of the ingredients, and to communicate using speech and gaze alone. The worker is responsible for all physical actions. The task is complete when fifteen items of the instructor's choosing have been moved from the task space where ingredients are located to the target bread. All ingredients are made of toy fabric and may have multiple instances, such as two different tomato slices. This task consists of fifteen discrete reference-action sequences in which ambiguous verbal references may appear, e.g., a reference to "the cheese" when there are both swiss cheese and cheddar cheese in the source area.

The data obtained by Andrist et al. [1] was acquired from 13 previously unacquainted dyads. Both participants wore mobile eye-tracking glasses. Each interaction was divided into a set of reference-action sequences which were further divided into four discrete phases. These phases include:

- *pre-reference*—time before verbal reference has been made.
- *reference*—time during the verbal request for a specific sandwich ingredient.
- *monitor*—time directly after the verbal reference when the instructor monitors the worker for understanding.
- *action*—time during the worker's successful action of moving the ingredient to the target bread.

The most pertinent analysis conducted by Andrist et al. [1] explored the alignment between gaze behaviors of interacting participants throughout each reference-action sequence. Gaze alignments were computed in each of the four phases, shifting the instructor's and worker's gaze streams at different offsets in relation to each other. The peak of the line graph for each of the four phases in Figure 2 represents the optimal time lag at that phase. Positive lags (peaks on the right side of the graph) put the instructor ahead of the worker, indicating that the instructor is "driving" the gaze patterns, while negative lags (peaks on the left side) indicate that the worker is driving.

This analysis revealed a general rise and fall in alignment throughout a sequence, as well as a back-and-forth pattern of which participant was leading the interaction in terms of their gaze behavior. These observations shed light on the role of gaze in "mixed initiative" conversations [37]. Early in an interaction sequence, the instructor drives the interaction by finding and referencing an object that they wish the worker to find and act upon. Then the worker takes initiative by searching for and acting upon the referent while the instructor monitors for potential breakdowns. In this paper, we utilize these observations in order to construct a model that allows an agent to synthesize similar behavior, as described next.

Our Model

The analysis by Andrist et al. [1] was conducted from the perspective of the instructor, which we also focus on for this paper as a role that a virtual character could effectively take on. Their analysis suggests that the agent should shift from leading with its gaze during *pre-reference* and *reference* phases (producing gaze behaviors to which the user is expected to respond) to following the user's gaze during the *monitor* and

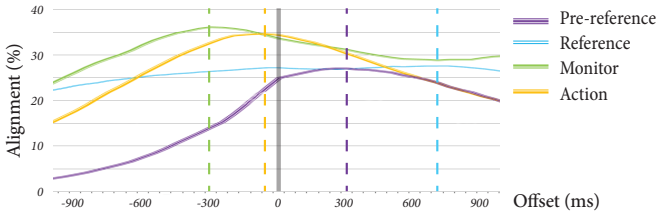

Figure 2. Adapted from Andrist et al. [1]. Peaks of gaze alignment within each phase, which occur in *pre-reference* and *reference* when the instructor's gaze is shifted ahead of the worker's gaze and in *monitor* and *action* phases when the instructor's gaze is shifted behind the worker's.

action phases (gazing in response to the detected gaze behaviors of the user). Thus, our model of bidirectional gaze has two major components: a stochastic component with statistical parameters of what to gaze at when, independent of the user, and a heuristic rule-based component on what to gaze at in response to the user during the *monitor* and *action* phases.

Producing Gaze with a Stochastic Finite-State-Machine

The agent's world is broken down into five categories of gaze targets relevant to the task: the *referent*, objects *ambiguous* to the referent, the *user*, the action target (always the *bread* for this scenario), and all *other* task-relevant objects (the ingredients that are not the referent or ambiguous to the referent).

At the highest level, the model traverses through the separate phases of a reference-action sequence according to the agent's speech and the user's actions. The *pre-reference* phase is entered at the conclusion of the user's action in the previous reference-action sequence. The *reference* phase is entered once the agent starts producing the verbal reference. Once the agent finishes speaking the reference, the *monitor* phase begins. This may involve responding to a user's request for clarification, either explicitly via speech recognition or implicitly from gaze via the heuristic part of the model described below. The *action* phase begins when the worker begins the relevant action, in this case grabbing the appropriate ingredient. Grabbing the wrong ingredient is considered an error, and the agent remains in the *monitor* phase, instructing the user to place that ingredient back and locate the correct one, providing additional description to help clarify. If the correct item is grabbed, the *action* phase lasts until the user brings it to the target bread location. Once that motion is complete, the next reference-action phase begins.

Within each phase, a stochastic finite-state machine is employed to determine which targets to gaze at and for how long.

Mean Gaze Fixation Length (seconds)

Phase	Referent	Ambiguous	Other	User	Bread
Pre-reference	0.85	0.45	0.35	0.65	0
Reference	1.1	0.5	0.45	0.6	0
Monitor	1.2	0.6	0.47	1.7	0
Action	0	0	0.66	0.6	0.86

Gaze Shift Probabilities

Phase	to Referent	to Ambiguous	to Other	to User	to Bread
Pre-reference	0.4	0	0.57	0.03	0
Reference	0.48	0	0.41	0.11	0
Monitor	0.49	0.02	0.34	0.15	0
Action	0	0	0.65	0.11	0.24

Table 1. Top: Mean gaze lengths to targets within each phase. Bottom: Probabilities of gazing toward targets within each phase.

This state machine includes five states, one for each category of possible gaze targets (*referent*, *ambiguous*, *user*, *bread*, and *other*). Each state is associated with a Gaussian distribution with mean and standard deviation values derived from the existing data (Table 1 top). Values sampled from these distributions served as gaze-length values when producing a gaze shift to that target. Transitions among states (realized as a gaze shift to the associated target) are dictated by the probabilities of gazing toward each type of target within the current phase (also derived from the existing data) (Table 1 bottom). When one gaze fixation is completed, the next gaze target is determined through a weighted random sample of the probabilities of gazing at the other targets. For states containing multiple discrete instances of possible locations, e.g., the "other task-relevant objects" state, one of these instances is randomly selected as the next gaze target.

Responding to Gaze with Heuristic Rules

In addition to the stochastic state-machine, which runs at all times during the interaction, the full model for responsive phases (*monitor* and *action*) also includes heuristically defined triggers of what the agent should do in response to user gaze. These heuristics override events in the stochastic state-machine part of the model, which is active at all other times.

To derive these heuristics, we examined phases from the existing human-human data where the instructor seemed to be "following" the gaze of the worker, rather than leading, looking for potentially interesting events or triggers in the worker's gaze in order to discover what the instructor is likely to do in the seconds immediately following that event. For example, when the worker looks at an ambiguous object, what is the instructor likely to look at next? Figure 3 (left) depicts the likelihood, over all collected data, of what the instructor is looking at over time following the event of a worker gazing to a wrong ambiguous ingredient, e.g., the light green lettuce instead of the dark green spinach. As can be observed, the likelihood that the instructor gazes toward the correct referent goes up and that of gazing toward other ingredients goes down, likely as a means to drive attention toward the correct object. Similarly in Figure 3 (right), when the worker looks at the correct referent, the likelihood of the instructor looking towards the referent again goes up over the next second. We examined a variety of such phenomena in our data in order to synthesize the following heuristic rules.

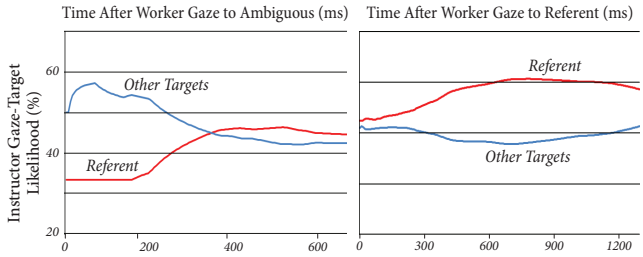

Figure 3. Gaze triggers informing the heuristic model component. Left: Likelihood of instructor gaze to the referent goes up over time following the worker's gaze to an ambiguous item. Right: Likelihood of instructor gaze to the referent goes up following the worker's gaze to the referent.

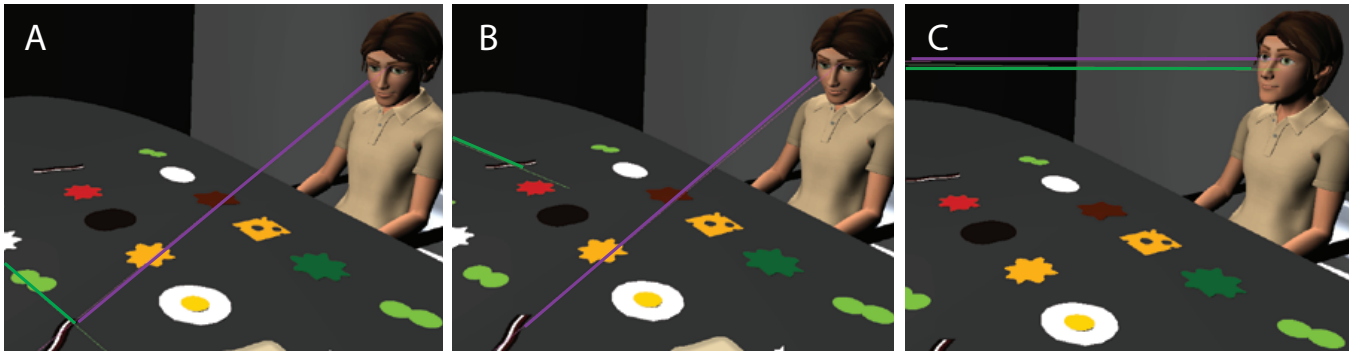

Figure 4. Heuristics in the *monitor* phase. User and agent gaze are shown in green and purple, respectively. The physical task space is replicated in the agent's virtual world for illustrative purposes. (A) Joint attention following to the referent. (B) Shifting joint attention from an ambiguous item to the referent. (C) Mutual gaze in response to agent-directed gaze.

In the *monitor* phase (Figure 4):

1. *Joint attention following* — When the user gazes toward the referent, the agent also gazes toward the referent.
2. *Shifting joint attention* — When the user gazes toward an object that is ambiguous with the referent, the agent gazes toward the referent. If the user is gazing toward the ambiguous object for more than one second without fixating on the referent, the agent gazes toward the user and preemptively offers a verbal refinement.
3. *Mutual gaze* — When the user gazes toward the agent, the agent gazes back toward the user. If the user has not made the correct action within two seconds of gazing toward the agent, the agent preemptively offers a verbal refinement.

In the *action* phase:

1. *Tracking action intent* — When the user gazes toward the target bread, the agent also gazes toward the target bread.
2. *Mutual gaze and shifting joint attention* — When the user gazes toward the agent, the agent gazes back to the user and then to the target.
3. *Joint attention following* — When the user gazes toward any ingredient, the agent gazes toward that ingredient.

System Design

We implemented an interactive virtual agent system on top of the Unity game engine. In order to integrate the bidirectional gaze model, additional system components were required. First, a model of gaze motions is necessary to actually execute shifts in gaze from one target to another. The gaze model developed by Andrist et al. [3] was utilized for this purpose.

Throughout interaction, the agent requires real-time access to the human interlocutor's point-of-regard. In the on-screen interaction version of the system (Figure 1 left), the user wears SMI eye-tracking glasses to provide the agent with a constant stream of gaze points within the glasses' front-facing camera view. To classify these gaze points in terms of the actual object being looked at, a system of augmented reality (AR) tags are used to convert camera-space points into real-world points. The ArUco library [20] was used for the detection of tags, providing the camera-space corner points of any and all tags detected in the camera's view (10 fps). Given these corner points, a nearby gaze tracker point-of-regard, the known real-world dimensions of the tag, and the assumption that the gaze point falls on the same plane as the tag, the Jacobi iterative method is used to solve for the homography between camera-space and real-space. This produces the real-world coordinates of the gaze point-of-regard. Tags are arranged on a table to create a grid of 18 locations. Using this system, the agent is provided with real-time access to the grid location being looked at by the human, which the agent can then associate with a particular item given its internal task representation. Two AR tags are also placed around the agent on the frame of the display to detect when the user looks toward the agent.

To make successful reference utterances, the speaker needs some form of feedback from the addressee. Despite the best efforts of speakers, there will inevitably be breakdowns—misunderstandings or miscommunication—that can impede ongoing progress of the interaction or lead to breakdowns in the future [52]. The preemptive verbal refinements provided by the heuristic portion of the model allow the agent to engage in repair in a natural and efficient way. In addition to these gaze-triggered refinements, the overall system enables the agent to respond to explicit verbal requests for refinement.

We developed a second version of the system by replacing the expensive eye-tracking technology with a lower-cost head-tracking solution. This system used a Kinect to track the head pose of the user. A virtual ray is extended forward from the point between the eyes and intersected with the task space. The bidirectional gaze model is relaxed to treat head pose as a gaze "cone" rather than a precise gaze point. For example, a reference gaze is detected when the user's head pose is directed toward the referent or to any objects within one grid cell of the referent in the task space. A third version of the system was subsequently developed to target head-mounted virtual reality, specifically the Oculus Rift. This implementation also utilizes the relaxed head-tracking version of the model, making use of the Rift's built-in head tracking as a proxy for gaze direction.

All three systems include speech recognition and task tracking modules. Speech recognition for verbal clarification requests is performed using a microphone and Microsoft Speech. Task tracking in the first two implementations utilizes a ceiling camera focused on the task space. As items are moved, AR tags become revealed, which communicates to the agent that an action has been performed. This action can be compared to the agent's current task model to check if the correct action has been performed, and if not, to issue a verbal repair to correct

the error. In VR, task tracking is accomplished by gazing with the head toward an ingredient and tapping a button to indicate ingredient selections, whereupon the selected ingredient (if correct) is animated to move to the action destination.

STUDY 1: ON-SCREEN INTERACTION

We next sought to evaluate the ability of our model of bidirectional gaze to improve a virtual character's ability to interact with people in natural, engaging, and effective ways. We chose the same interaction scenario that was previously used in data collection for the model—learning how to make a sandwich.

Study Design: The user study followed a 2×2 within-participants design. The independent variables included whether or not the agent *produced* gaze motions and whether or not the agent *responded to* user gaze with its own gaze and verbal clarifications. Considering both variables separately enabled us to explore the relative effectiveness of a virtual character reacting to user gaze as input and/or producing gaze as output to facilitate interactive experiences. Each participant interacted with the virtual character system four times, making one sandwich in each condition in a randomized order.

Hypotheses: Our evaluation of the bidirectional gaze model was guided by three central hypotheses, focusing on the production of referential gaze behavior, the responsiveness to user gaze, and the benefits of doing both simultaneously.

Hypothesis 1—A virtual character that *produces* gaze according to the bidirectional gaze model will improve user interaction over one that statically gazes toward the user.

More specifically, we predict that producing gaze cues will enable the agent to utilize faster, ambiguous referencing without breakdowns or requests for repair (e.g., clarification). This prediction is supported by prior research that has found that a virtual agent mimicking user behaviors more successfully directs user attention, resulting in faster task completion and fewer errors [4].

Hypothesis 2—A virtual character that *responds to* user gaze according to the bidirectional gaze model will improve user interaction over one that does not respond to user gaze.

By tracking and responding to user gaze with its own gaze and speech, the virtual agent will be able to anticipate breakdowns before they occur or before a repair request is made. This hypothesis is supported by previous work that has demonstrated that gaze behaviors precede physical actions [31, 25] and characters that respond to user gaze improve learning gains in tutoring scenarios [18].

Hypothesis 3—A virtual character that both *produces* gaze and *responds to* user gaze will further improve user interaction over only producing gaze or only responding to user gaze.

Previous work has shown that coordinated gaze between pairs of people improves performance in visual search tasks [36] and in answering comprehension questions [41].

Procedure: Following informed consent, the experimenter provided the participant with high-level instructions and calibrated the eye-tracking glasses. A virtual character named "Jason" was introduced, which then instructed participants on how to assemble four sandwiches with different ingredients. The order of the sandwiches, as well as their assignment to condition, was randomized. Each sandwich required 12 ingredients, at least four of which were inherently ambiguous, e.g., the character asked for "cheese" when both Swiss cheese and cheddar cheese were present. Following each sandwich, the participants completed a questionnaire about their experience with that version of the agent and took a quiz that measured their recall of the ingredients of the sandwich. Following the completion of all sandwiches, the participant filled out a demographic questionnaire and received $5 USD for compensation.

Measures: All studies included the same core set of objective, behavioral, and subjective measures to assess the quality and effectiveness of a character's use of bidirectional gaze.

Objective measures: Completion time and number of errors (i.e., incorrect actions taken) were measured both within each reference-action sequence and across the making of the sandwich. We also counted the number of verbal requests the participant made for clarification. Finally, participants completed a recall quiz to list the ingredients used for each sandwich.

Behavioral measures: We recorded participant gaze to extract behavioral measures, including the amount of shared and mutual gaze—percentage of time that the agent and user looked toward the same object and toward each other, respectively—as measures of behavioral synchrony. The percentage of time that users looked toward ambiguous objects not relevant to the interaction served as an indicator of confusion. Finally, we measured the time from the onset of the character's verbal reference to when the user fixated on the referent.

Subjective measures: Participants completed a questionnaire about their impressions of the agent after finishing each condition. Using an exploratory-confirmatory factor analysis, four scales were constructed, each comprised of several seven-point-rating-scale questions of the form, "I believe that this version of Jason was...:"

1. *Task competence*: "engaged in the task," "dedicated to the task," "active part of task," "helpful" (Cronbach's $\alpha = .88$);
2. *Cognitive abilities*: "sensitive to my needs," "intelligent," "an expert" (Cronbach's $\alpha = .87$);
3. *Expressiveness*: "lively," "expressive," "excited to help me," "humanlike in behavior" (Cronbach's $\alpha = .91$);
4. *Visual attentiveness*: "watchful," "attentive," "observant" (Cronbach's $\alpha = .91$).

Participants: Forty participants from the University of Wisconsin–Madison campus participated in the first user study. Due to occasional system malfunctions, eight were excluded from analysis, resulting in 32 participants (14 females, 18 males) whose ages ranged from 18 to 32 ($M = 22.4$, $SD = 4.0$).

Results

The data collected from the user study were analyzed with a repeated-measures analysis of variance (ANOVA). The statistical model included two independent variables: *gaze-producing* (on or off) and *gaze-responsive* (on or off). Both trial number and the particular sandwich type, e.g., "bacon special" or

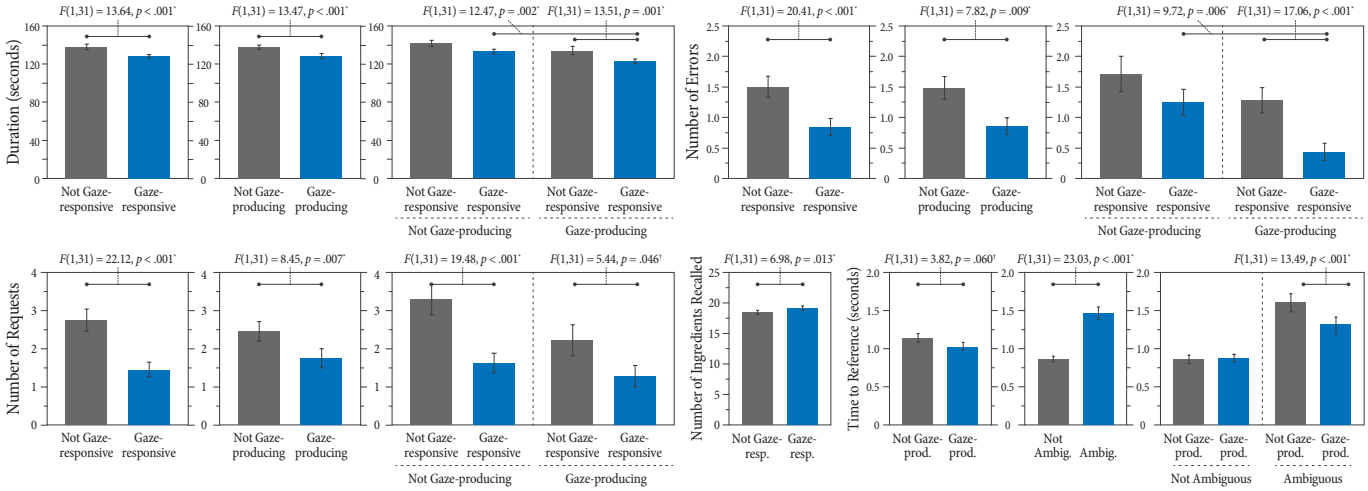

Figure 5. Study 1: results from the objective measures of task duration (seconds), number of errors, number of clarification requests, information recall, and time it took participants to look toward the referent. Test details are provided only for significant (*) and marginal (†) differences based on Bonferroni-corrected alpha levels for multiple comparisons ($\alpha = 0.05$ for H1 and H2 and $\alpha = 0.025$ for H3).

"turkey special" were modeled as control variables in order to control for differences in inherent ambiguity across ingredients. A Bonferroni correction was applied to control for Type I errors in multiple comparisons. All statistical test results are reported in Figures 5, 6, and 7 for ease of reading. We report on the full set of measures only for Study 1 and focus on the more important and illuminating results for Studies 2 and 3.

Objective Results: Our analysis of the task duration measure supported all three hypotheses. Participants completed the task more quickly and with fewer errors when the character responded to user gaze and when it produced gaze cues compared to when it did not engage these mechanisms. The use of both mechanisms had an additive effect; participants completed the character's instructions faster and with fewer errors when it used both mechanisms than when it engaged only one of these mechanisms (Figure 5).

Our analysis of the number of requests participants made for refinement provided support for H1 and H2 and partial support for H3. While gaze responsiveness and gaze production both reduced the number of requests made, the use of both mechanisms further reduced requests only marginally over gaze production alone. The character's use of the gaze-responsive mechanism improved participant recall of the ingredients of the sandwich after the task, while its use of gaze production had no effect, and the combined use of both mechanisms did not further improve recall (Figure 5).

Behavioral Results: Our analysis of the participants' gaze focused on three specific behavioral measurements: *shared gaze* toward objects of interest, gaze toward irrelevant objects (which we call "*ambiguous gaze*"), and *mutual gaze* toward each other. We found higher shared gaze when the agent used the gaze-responsive mechanism, indicating a higher degree of behavioral synchrony. There was also less ambiguous gaze when the character produced gaze, reducing the amount of time spent looking toward task-irrelevant objects. Finally, we found higher mutual gaze when the agent was both gaze-producing and gaze-responsive (Figure 6).

The character's production of gaze marginally reduced the time it took participants to fixate on referent objects, indicating effective engagement of joint attention between the character and its user. The character's use of gaze responsiveness did not affect participant fixation time. Reference ambiguity, i.e., whether or not multiple potential referents were present, delayed the ability of the participants to identify and fixate on the referent. When references were ambiguous, the character's gaze reduced participant fixation time, while it had no effect when the references were not ambiguous (Figure 5).

Subjective Results: We next tested the hypotheses using data from our four subjective measures, including how competent, cognitively able, expressive, and attentive participants rated the character on a seven-point scale. We found support for H1 in ratings of the character's cognitive ability and partial support in ratings of its competence; participants found the gaze-detecting character to be more cognitively able and marginally more competent. Support for H2 was provided by data from the perceived competence and perceived expressiveness measures, and partial support was provided by the perceived awareness measure. The gaze-producing character was rated as more competent and expressive and marginally more attentive. Finally, only the perceived expressiveness measure provided partial support for H3; the use of both the gaze-responsive and gaze-producing mechanisms resulted in higher

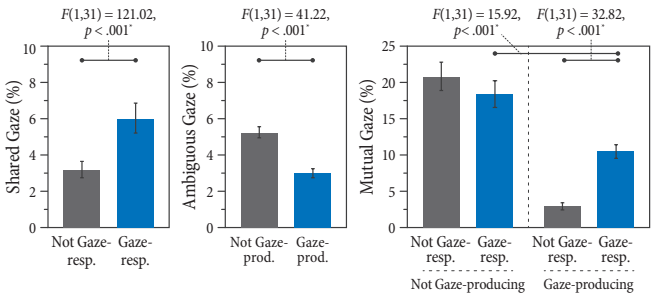

Figure 6. Study 1: results from behavioral measures of shared, ambiguous, and mutual gaze (%). Test details are given for significant (*) and marginal (†) differences based on Bonferroni-corrected alpha levels.

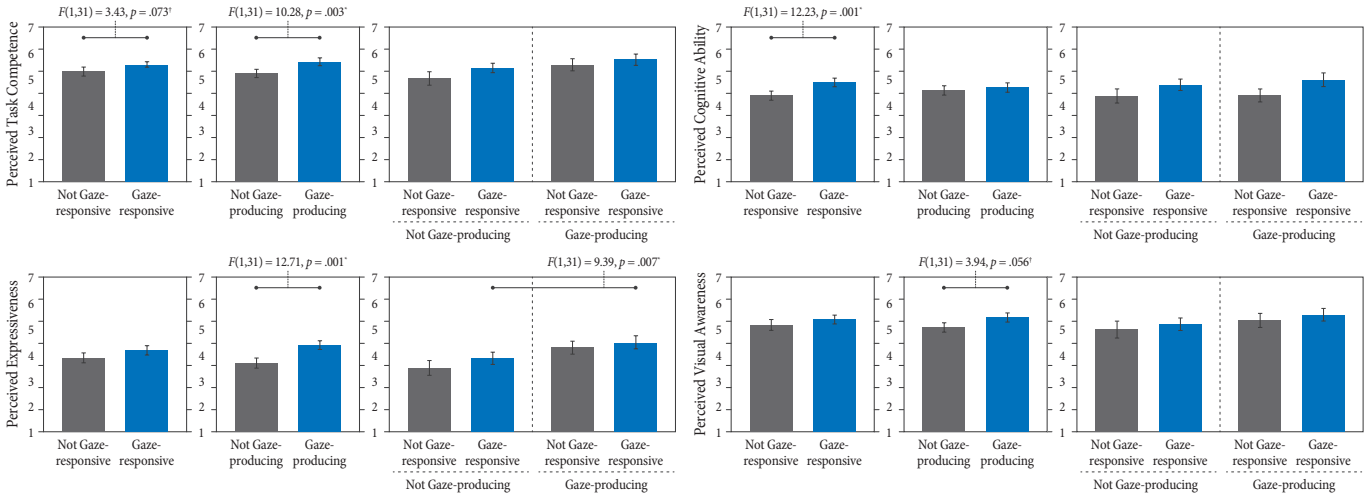

Figure 7. Study 1: results from subjective measures of how competent, cognitively able, expressive, and aware participants found the character to be. Test details are provided only for significant (*) and marginal (†) differences using Bonferroni-corrected $\alpha = 0.05$ for H1 and H2 and $\alpha = 0.025$ for H3.

ratings of expressiveness over the use of gaze-responsive alone but not over the use of gaze-producing alone (Figure 7).

Discussion

The first user study was designed to evaluate the ability of our model of bidirectional gaze to improve a virtual character's ability to interact with people in natural, engaging, and effective ways. We tested conditions in which the agent did or did not produce gaze cues to the task space and in which the agent did or did not respond to the gaze cues of the user. This study demonstrated the benefits of bidirectional gaze in several objective, subjective, and behavioral measures. Participants conducted the collaborative sandwich-task more efficiently when the agent used the full bidirectional gaze model, completing the task faster with fewer errors and less need to ask the agent for clarification. Participants also scored better in a recall quiz and engaged in higher amounts of mutual gaze and shared gaze with the agent, indicating a higher degree of coordination. Subjectively, participants felt that the agent was most expressive when producing gaze, most competent when producing and responding to gaze, and possessed greater cognitive abilities when responding to user gaze.

STUDY 2: HEAD POSE AS GAZE PROXY

In the second study, we investigated whether or not the bidirectional gaze model is effective when full eye-tracking is replaced with low-cost head tracking. This study utilized an identical on-screen agent as in the first study, but with the second version of our system as described in the "System Design" section. We hypothesized that the bidirectional gaze model with head tracking would be more effective than not having any kind of gaze detection (and thus not utilizing the gaze-responsive portion of the model) and that there would be no significant differences in outcomes when moving from the precise eye tracking version of the system to head tracking.

This study included three conditions: not gaze-responsive, gaze-responsive via eye tracking, and gaze-responsive via head tracking. In all conditions, the agent produced gaze cues according to the gaze-producing stochastic finite-state-machine component of the model. The overall procedure was similar to the previous study, with participants assembling three sandwiches (one for each condition) in a random order. We recruited 15 participants, different from those who participated in the first study. After excluding three due to system difficulties, we were left with 12 final participants for analysis (4 females, 8 males), aged 18–23 ($M = 21.2$, $SD = 1.57$). Participants were compensated $5 USD.

Results

Our analysis utilized a repeated-measures analysis of variance (ANOVA) and pairwise comparisons with Bonferroni correction to establish benefits of the gaze-responsive via head tracking method over not gaze-responsive and the equivalence of gaze-responsive via head tracking gaze-responsive via eye tracking. To establish equivalence, we followed guidelines suggested by Julnes & Mohr [29], including a p-value larger than .50. The analysis showed that participants took significantly shorter time to complete the sandwich, made significantly fewer errors, and made marginally fewer requests for clarification when the character used the gaze-responsive method via head tracking compared to not utilizing the gaze-responsive mechanism. On the other hand, gaze-responsive conditions via head tracking and via eye tracking were equivalent across these measures (statistical results reported in Figure 8).

Discussion

The second user study was designed to evaluate whether head tracking could serve as a sufficient proxy for more complex eye

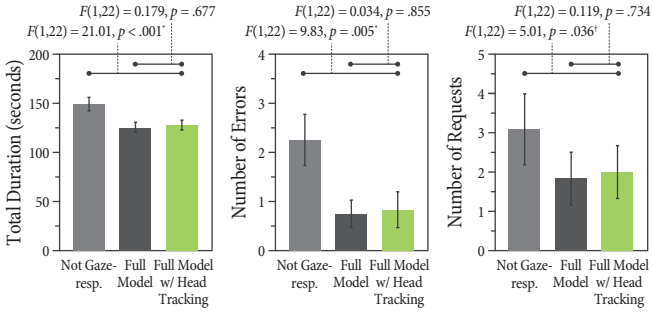

Figure 8. Study 2: data from measures of duration, number of errors, and requests for clarification. Test details are provided for significant (*) and marginal (†) differences (using Bonferroni-corrected $\alpha = 0.025$).

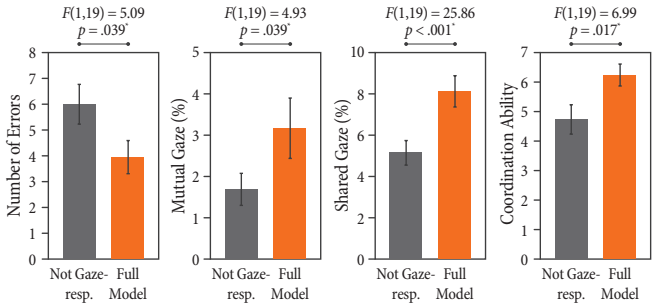

Figure 9. Study 3: results from measures of number of errors, mutual gaze, shared gaze, and perceived coordination ability of the character. Data and test details are provided only for significant (*) differences.

tracking in the bidirectional gaze model. The study confirmed this to be the case in terms of objective task performance. Participants completed the sandwich task more efficiently, faster with fewer errors and requests for clarification, when the agent responded to the participant's head pose rather than when it performed gaze cues unresponsively. We observed no significant differences between eye tracking and head tracking, but a future study with more participants might be able to tease out the small differences.

STUDY 3: VIRTUAL REALITY

Once we confirmed in the second user study that head tracking could serve as an adequate proxy for eye tracking in the bidirectional gaze model, we next asked the following research question: Does the bidirectional gaze model also work when interacting with agents in virtual reality? To answer this question, we designed and executed a third user study in which participants carried out the same sandwich-building task with the same virtual agent as in the previous studies, but this time while wearing the Oculus Rift DK2 headset (Figure 1 right) using the third version of the system described earlier.

This study tested two conditions: gazing responsively using the full bidirectional gaze model and only producing gaze cues with no responsiveness. Based on lessons learned in the first two studies, we constructed an additional subjective scale targeting users' perceptions of the agent's level of *coordination ability*. This scale included the following items on a seven-point scale: "How would you rate Jason's *responsiveness* to your attention and behaviors?" and "How *in sync* were you and Jason?" Twenty participants (8 females, 12 males), aged 18–47 ($M = 23.1$, $SD = 6.81$), experienced both conditions in a random order and received $3 USD.

Results

A repeated-measures ANOVA showed that, while the use of the full model reduced errors, it did not significantly improve task completion time, $F(1,19) = 2.13, p = .16$, or reduce the number of requests for clarification, $F(1,19) = 0.10, p = .75$. Participants established more mutual and shared gaze with the character that used the full model compared to the character that only used the gaze-producing mechanism. Finally, the character that used the full model was rated by the participants as more effective in coordinating its behaviors with them (statistical results reported in Figure 9).

Discussion

The third user study was designed to evaluate the effectiveness of bidirectional gaze for virtual characters in head-mounted virtual reality. This study revealed that an agent using the full bidirectional gaze model, rather than producing gaze in isolation, is able to help participants complete a collaborative task with fewer errors. Task completion time was not significantly reduced in the bidirectional gaze condition, possibly explained by a large amount of variance from differing levels of familiarity with VR technology. However, participants made more mutual gaze and engaged in more shared gaze with the agent when it responded to their gaze, indicating a higher degree of coordination. Participants explicitly indicated that they subjectively felt more coordinated with the agent when it was using bidirectional gaze than when it was not.

GENERAL DISCUSSION

Our first study demonstrated the effectiveness of bidirectional gaze, differentially examining both the gaze-producing and gaze-responsive components of the model. In general, it showed that bidirectional gaze mechanisms are effective in improving collaborative interaction with on-screen characters providing instructions over a physical task space and are even more effective than either of its components alone. The second study demonstrated that relaxing and extending the model to utilize head pose rather than full eye tracking retains much of that effectiveness. The third evaluation demonstrated that the bidirectional gaze model utilizing head tracking is useful in virtual reality collaborations with virtual characters.

People are able to use a range of communication mechanisms to collaborate and correct each other's mistakes and misunderstandings quickly and efficiently. Bidirectional gaze mechanisms similarly enable interactive characters to preemptively offer refinements in a quick and unobtrusive way, responding to subtle nonverbal gaze of the user rather than relying on explicitly spoken questions. People also frequently use ambiguous speech that they easily resolve through context and the use of nonverbal cues, making speech faster and more efficient. Bidirectional gaze enables agents to utilize this power, making potentially ambiguous but faster verbal references along with appropriate gaze production and responsiveness.

Our model is grounded in observations of human-human interactions in order to ensure that it accurately captures human communication mechanisms of bidirectional gaze. Evaluations were driven by hypotheses derived from research on human-human interaction. This human-based modeling allowed us to target human-level competence while making specific predictions about how these gaze mechanisms should improve interactions objectively, subjectively, and behaviorally.

Beyond the statistically significant differences found in many of the experimental comparisons, we would like to highlight the practical implications and applications of our findings. For example, differences in means indicate that error rate is cut by about half when the agent is gaze-responsive compared to when it is not gaze-responsive and approximately half of a point is gained in user perceptions of their cognitive abilities (Study 1). While these gains may individually appear small in

magnitude, they persist as robust effects across different measures and technology implementations. Furthermore, there are a number of potential scenarios beyond toy sandwich building where even small improvements in error rates can have a large impact. In scenarios where these differences are not as impactful, it may be more appropriate to utilize low-cost head tracking rather than fully instrumenting users with gaze trackers. Study 2 demonstrates that much of the positive impact from bidirectional gaze will still be retained in this case.

Limitations & Applicability

The model presented in this paper makes three simplifications that should be addressed in future research. First, it focuses on agents making verbal references to task objects in the form of reference-action sequences. While these sequences are a fundamental building block of collaborative interactions, other elements must be addressed to extend beyond these discrete, well-defined sequences, such as a more general model of turn-taking, coordinating conversational gaze, and providing users with verbal and nonverbal feedback. Previous work has explored agent gaze across these different interaction contexts, such as how agents should utilize gaze in conversations without physical objects or actions in the environment [2].

The second simplification is that the model only applies to the instructor role. The worker role, and roles in other types of collaborations, should be considered in future work. Recent work has explored how a collaborative robot in the worker role might react to user gaze in order to increase efficiency and naturalness of the interaction [28]. Additionally, handling multiparty interactions in which the agent must distribute its gaze to multiple users would also require extending the model.

Third, we only explored one task instance of collaboration—assembling toy sandwiches. While many complex real-world tasks are comprised of the same reference-action sequences we have studied here, the applicability of our model will vary across three categories of scenarios: (1) those that the model would apply directly to, (2) those that would require modification to the model, and (3) those that are fully outside the model's scope, which we discuss in the paragraphs below.

Direct Application: Our approach is directly applicable to instructional tasks in which an agent (physical or virtual) is training a user on how to act on a set of objects (building sandwiches, preparing recipes, assembling furniture, fixing a bicycle, arranging table settings, etc). These tasks all involve making mutual gaze, shifting and responding to joint attention cues, tracking action intent, and so on. Implementations may differ in terms of dialogue handling and object tracking, but the model would still apply. An open question for future work concerns the sensitivity of the precise timing parameters collected from the sandwich-building task.

Requires Modification: The model would require some modification for fully collaborative tasks in which the agent is not only instructing, but also receiving instructions and taking actions. Extensions of the model would account for agent gaze behaviors in more fluid and open-ended roles. This category also includes tasks that do not involve taking physical action on objects, such as tour guiding. Parts of the model are still applicable (mutual gaze, gazing to referents, following joint attention, etc.), but others would need to be adapted, e.g., "reference" and "action" phases might be merged.

Outside the Scope: Casual conversations not situated in an environment of relevant objects or tutoring scenarios where the agent is conveying abstract information are outside the scope of the model. A number of existing models that have been developed in the virtual agent and human-robot interaction literatures [39, 30, 35, 2, 38] can be applied in these situations.

Future Work

Fully validating our models, including establishing their generalizability to other tasks, requires applying them to future scenarios outside of lab settings. Future work should seek to replicate these results across technologies as well as tasks varying in similarity. Another fruitful direction for future research is comparing human-agent interactions utilizing the bidirectional gaze model to a human-human gold standard, which can be achieved by comparing the experimental results directly to the data by Andrist et al. [1] from which the model was derived. However, such comparison will involve many challenges, as the two sets of data differ not only in terms of gaze behavior, but also in the use of unconstrained dialogue, gestures, repair strategies, and so on.

CONCLUSION

In face-to-face conversation, the gaze of one conversational participant is constantly and dynamically shaping, and being shaped by, the gaze of the other participant. Thus, interactive virtual characters stand to benefit from tracking user gaze and the knowledge of these coordinated gaze patterns. Grounded in previous data collected from pairs of collaborating people, we identified a number of subtle features of human gaze coordination, including timings, spatial mappings, and repair strategies. These features were built into a model of *bidirectional gaze* that enables interactive virtual characters to interpret the gaze of their users and generate gaze to effectively communicate coordinated behaviors. This model enables virtual characters to achieve more efficient verbal referencing by signaling attention to the user and to items in the environment appropriately and to infer user state and goals—such as confusion leading to an impending request for repair—from the user's gaze. We also demonstrated that bidirectional gaze is achievable with low-cost head tracking and is an effective mechanism for human-agent interaction situated in virtual reality. Overall, this work extends prior research on the use of coordinated gaze behaviors between humans and embodied agents by providing a model constructed directly from human data, implementing the behaviors using new technologies, and providing a unique evaluation methodology with results that have clear implications for the design of effective human-agent interactions.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation awards 1149970 and 1208632. We would like to thank Ethan Jesse, Faye Golden, Brandon Smith, and Tomislav Pejsa for their help in executing the experiments.

REFERENCES

1. Sean Andrist, Wesley Collier, Michael Gleicher, Bilge Mutlu, and David Shaffer. 2015. Look together: analyzing gaze coordination with epistemic network analysis. *Frontiers in psychology* 6, 1016 (2015), 1–15.
2. Sean Andrist, Bilge Mutlu, and Michael Gleicher. 2013. Conversational gaze aversion for virtual agents. In *Intelligent Virtual Agents*. Springer, 249–262.
3. Sean Andrist, Tomislav Pejsa, Bilge Mutlu, and Michael Gleicher. 2012. Designing effective gaze mechanisms for virtual agents. In *Proc. of CHI*. ACM, 705–714.
4. Gérard Bailly, Stephan Raidt, and Frédéric Elisei. 2010. Gaze, conversational agents and face-to-face communication. *Speech Communication* 52, 6 (2010), 598–612.
5. Ellen Gurman Bard, Robin Hill, Manabu Arai, and ME Foster. 2009. Referring and gaze alignment: Accessibility is alive and well in situated dialogue. In *Proc. of CogSci ('09)*. Cognitive Science Society, 1246–1251.
6. Nikolaus Bee, Johannes Wagner, Elisabeth André, Thurid Vogt, Fred Charles, David Pizzi, and Marc Cavazza. 2010. Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application. In *Proc. of ICML-MLMI ('10)*. ACM, 1–8.
7. Jean-David Boucher, Ugo Pattacini, Amelie Lelong, Gerard Bailly, Frederic Elisei, Sascha Fagel, Peter Ford Dominey, and Jocelyne Ventre-Dominey. 2012. I reach faster when I see you look: Gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in Neurorobotics* 6 (2012).
8. Susan E Brennan, Xin Chen, Christopher A Dickinson, Mark B Neider, and Gregory J Zelinsky. 2008. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition* 106, 3 (2008), 1465–1477.
9. Susan E Brennan, JE Hanna, GJ Zelinsky, and Kelly J Savietta. 2012. Eye gaze cues for coordination in collaborative tasks. In *Proc. of CSCW DUET 2012 Workshop*, Vol. 9.
10. Andrew G Brooks and Cynthia Breazeal. 2006. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proc. of HRI ('06)*. ACM, 297–304.
11. Sarah Brown-Schmidt, Ellen Campana, and Michael K Tanenhaus. 2005. Real-time reference resolution by naïve participants during a task-based unscripted conversation. *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions* (2005), 153–171.
12. Ellen Campana, Jason Baldridge, John Dowding, Beth Ann Hockey, Roger W Remington, and Leland S Stone. 2001. Using eye movements to determine referents in a spoken dialogue system. In *Proceedings of the 2001 Workshop on Perceptive User Interfaces*. ACM, 1–5.
13. Herbert H Clark. 1996. *Using language*. Cambridge university press.
14. Herbert H Clark. 2005. Coordinating with each other in a material world. *Discourse studies* 7, 4-5 (2005), 507–525.
15. Herbert H Clark and Susan E Brennan. 1991. Grounding in communication. *Perspectives on socially shared cognition* 13, 1991 (1991), 127–149.
16. Herbert H Clark and Meredyth A Krych. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language* 50, 1 (2004), 62–81.
17. Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition* 22, 1 (1986), 1–39.
18. Sidney D'Mello, Andrew Olney, Claire Williams, and Patrick Hays. 2012. Gaze tutor: A gaze-reactive intelligent tutoring system. *International Journal of Human-Computer Studies* 70, 5 (2012), 377–398.
19. Mica R Endsley. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37, 1 (1995), 32–64.
20. S Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. 2014. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition* 47, 6 (2014), 2280–2292.
21. Darren Gergle and Alan T Clark. 2011. See what I'm saying?: Using dyadic mobile eye tracking to study collaborative reference. In *Proc. of CSCW ('11)*. ACM, 435–444.
22. Darren Gergle, Robert E Kraut, and Susan R Fussell. 2013. Using visual information for grounding and awareness in collaborative tasks. *Human–Computer Interaction* 28, 1 (2013), 1–39.
23. Zenzi M Griffin. 2004. The eyes are right when the mouth is wrong. *Psychological Science* 15, 12 (2004), 814–821.
24. Joy E Hanna and Susan E Brennan. 2007. Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language* 57, 4 (2007), 596–615.
25. Mary Hayhoe and Dana Ballard. 2005. Eye movements in natural behavior. *Trends in cognitive sciences* 9, 4 (2005), 188–194.
26. Graeme Hirst, Susan McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. 1994. Repairing conversational misunderstandings and non-understandings. *Speech communication* 15, 3 (1994), 213–229.
27. Mohammed Moshiul Hoque and Kaushik Deb. 2012. Robotic system for making eye contact pro-actively with humans. In *Proc. of ICECE ('12)*. IEEE, 125–128.

28. Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory robot control for efficient human-robot collaboration. In *Proc. of HRI ('16)*. IEEE, 83–90.
29. George Julnes and Lawrence B Mohr. 1989. Analysis of no-difference findings in evaluation research. *Evaluation Review* 13, 6 (1989), 628–655.
30. B.J. Lance and S.C. Marsella. 2010. The Expressive Gaze Model: Using Gaze to Express Emotion. *Computer Graphics and Applications, IEEE* 30, 4 (2010), 62–73.
31. Michael Land, Neil Mennie, and Jennifer Rusted. 1999. The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 11 (1999), 1311–1328.
32. Gregor Mehlmann, Markus Häring, Kathrin Janowski, Tobias Baur, Patrick Gebhard, and Elisabeth André. 2014. Exploring a model of gaze for grounding in multimodal HRI. In *Proc. of ICMI ('14)*. ACM, 247–254.
33. Antje Meyer, Femke van der Meulen, and Adrian Brooks. 2004. Eye movements during speech planning: talking about present and remembered objects. *Visual Cognition* 11, 5 (2004), 553–576.
34. AJung Moon, Daniel M Troniak, Brian Gleeson, Matthew KXJ Pan, Minhua Zheng, Benjamin A Blumer, Karon MacLean, and Elizabeth A Croft. 2014. Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing. In *Proc. of HRI ('14)*. ACM, 334–341.
35. Bilge Mutlu, Takayuki Kanda, Jodi Forlizzi, Jessica Hodgins, and Hiroshi Ishiguro. 2012. Conversational gaze mechanisms for humanlike robots. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1, 2 (2012), 12.
36. Mark B Neider, Xin Chen, Christopher A Dickinson, Susan E Brennan, and Gregory J Zelinsky. 2010. Coordinating spatial referencing using shared gaze. *Psychonomic bulletin & review* 17, 5 (2010), 718–724.
37. David G Novick, Brian Hansen, and Karen Ward. 1996. Coordinating turn-taking with gaze. In *Proc. of ICSLP ('96)*, Vol. 3. IEEE, 1888–1891.
38. Tomislav Pejsa, Sean Andrist, Michael Gleicher, and Bilge Mutlu. 2015. Gaze and Attention Management for Embodied Conversational Agents. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 5, 1 (2015), 3.
39. C. Pelachaud and M. Bilvi. 2003. Modelling gaze behavior for conversational agents. In *Intelligent Virtual Agents*. Springer, 93–100.
40. Christopher Peters, Stylianos Asteriadis, and Kostas Karpouzis. 2010. Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces* 3, 1-2 (2010), 119–130.
41. Daniel C Richardson and Rick Dale. 2005. Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science* 29, 6 (2005), 1045–1060.
42. Daniel C Richardson, Rick Dale, and Natasha Z Kirkham. 2007. The art of conversation is coordination common ground and the coupling of eye movements during dialogue. *Psychological science* 18, 5 (2007), 407–413.
43. Daniel C Richardson, Rick Dale, and John M Tomlinson. 2009. Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science* 33, 8 (2009), 1468–1482.
44. Kenji Sakita, Koichi Ogawara, Shinji Murakami, Kentaro Kawamura, and Katsushi Ikeuchi. 2004. Flexible cooperation between human and robot by interpreting human intention from gaze information. In *Proc. of IROS ('04)*, Vol. 1. IEEE, 846–851.
45. Michael F Schober. 1993. Spatial perspective-taking in conversation. *Cognition* 47, 1 (1993), 1–24.
46. Natalie Sebanz, Harold Bekkering, and Günther Knoblich. 2006. Joint action: bodies and minds moving together. *Trends in cognitive sciences* 10, 2 (2006), 70–76.
47. Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication* 65 (2014), 50–66.
48. Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. 1995. Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 5217 (1995), 1632–1634.
49. Cristen Torrey, Aaron Powers, Susan R Fussell, and Sara Kiesler. 2007. Exploring adaptive dialogue based on a robot's awareness of human gaze and task progress. In *Proc. of HRI ('07)*. ACM, 247–254.
50. Weilie Yi and Dana Ballard. 2009. Recognizing behavior in hand-eye coordination patterns. *International Journal of Humanoid Robotics* 6, 03 (2009), 337–359.
51. Yuichiro Yoshikawa, Kazuhiko Shinozawa, Hiroshi Ishiguro, Norihiro Hagita, and Takanori Miyamoto. 2006. Responsive Robot Gaze to Interaction Partner.. In *Proc. of RSS ('06)*.
52. Christopher J Zahn. 1984. A reexamination of conversational repair. *Communications Monographs* 51, 1 (1984), 56–66.