Motion and Attention in a Kinetic Videoconferencing Proxy

David Sirkin[2], Gina Venolia[1], John Tang[1], George Robertson[1], Taemie Kim[3], Kori Inkpen[1], Mara Sedlins[1], Bongshin Lee[1], and Mike Sinclair[1]

[1] Microsoft Research
[2] Stanford University
[3] MIT
sirkin@cdr.stanford.edu,
{ginav,johntang,ggr,kori,a-marase,
bongshin,sinclair}@microsoft.com,
taemie@media.mit.edu

Abstract. Compared to collocated interaction, videoconferencing disrupts the ability to use gaze and gestures to mediate interaction, direct reactions to specific people, and provide a sense of presence for the satellite (i.e., remote) participant. We developed a kinetic videoconferencing proxy with a swiveling display screen to indicate which direction that the satellite participant was looking. Our goal was to compare two alternative motion control conditions, in which the satellite participant directed the display screen's motion either explicitly (aiming the direction of the display with a mouse) or implicitly (with the screen following the satellite participant's head turns). We then explored the effectiveness of this prototype compared to a typical stationary video display in a lab study. We found that both motion conditions resulted in communication patterns that indicate higher engagement in conversation, more accurate responses to the satellite participant's deictic questions (i.e., "What do *you* think?"), and higher user rankings. We also discovered tradeoffs in attention and clarity between explicit versus implicit control, a tension in how motion toward one person can exclude other people, and ways that swiveling motion provides attention awareness, even without direct eye contact.

Keywords: Video-mediated communication, videoconferencing, gaze awareness, proxy, telepresence.

1 Introduction

Attention is fundamental to the flow of face-to-face conversations. Each participant projects cues of what he is paying attention to and other participants interpret these cues to maintain awareness of his locus of attention. This awareness helps them understand his deictic references. Both production and consumption of awareness cues occur at conscious and subconscious levels [1].

Videoconferencing systems disrupt the link between attention projection and attention awareness. They do this in part because they do not faithfully reproduce the

P. Campos et al. (Eds.): INTERACT 2011, Part I, LNCS 6946, pp. 162–180, 2011.

spatial characteristics of gaze, body orientation, and pointing gestures. This disruption is one of the reasons why video-mediated communication is less effective than face-to-face interaction. The lack of a shared physical environment further hampers participants' abilities to use spatial cues to support conversation and direct attention [2, 3]. Videoconferencing configurations that involve multiple people at one site offer multiple plausible loci of attention, increasing the potential for confusion.

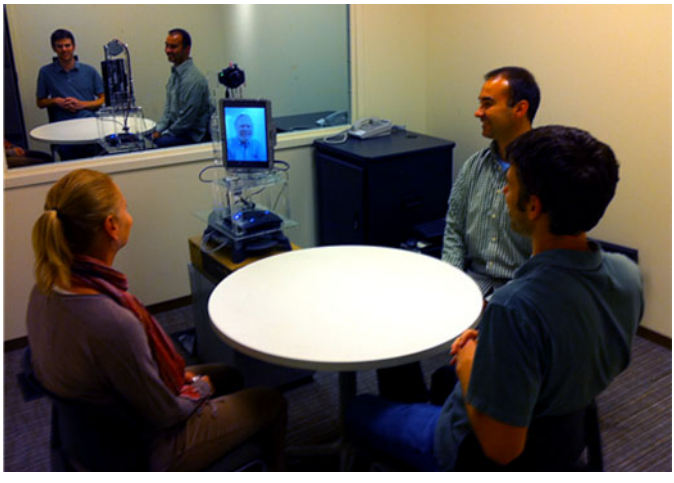

Fig. 1. The experimental setup, showing three collocated participants and the kinetic proxy seated around a table. The proxy was operated by a confederate, and positioned so that its swiveling display approximately matched participant eye-height.

We are particularly interested in videoconferencing systems to support *hub-and-satellite* meetings, where most participants are collocated except for one participant at a satellite location. This satellite is represented in the collocated space by a *proxy* device consisting of a display screen, camera, speaker, and microphone (Fig. 1). The satellite perceives the hub location through streams of audio and video displayed on his computer display.

A study of proxies in everyday use has documented the benefits of a physical representation of the satellite in group interaction [4]. Our own use and studies of proxies in our day-to-day work has led to a design that includes a wide-field-of-view camera that shows most of the meeting room at the hub site (see Fig. 3). The satellite views this panorama of the hub room displayed in a window that is full-screen width across his display. Relative to this window, the satellite's camera is positioned horizontally centered and vertically as close as possible.

This view gives the satellite a good sense of the spatial relationships among the people and objects in the meeting room. He can maintain awareness of the locus of attention for each of the hub participants. Because the camera the satellite views is positioned near the screen representing him, he has a good sense of when a hub participant looks directly at him or gestures toward him.

The reverse, however, is not true. The hub participants have a general sense of whether the satellite is looking left, center or right, but nothing more fine-grained than

that. The video mediation introduces too many invisible parameters, such as the field of view of the wide-angle camera and the size of the satellite's desktop monitor, preventing the hub participants from having a precise sense of the satellite's focus of attention. With the multiple potential foci of attention in a typical meeting (e.g., people, whiteboard drawings, artifacts), breakdowns occur when the hubs cannot determine the satellite's focus of attention.

Moreover, the hub participants do not have a visceral sense of mutual eye contact with the satellite participant. When he looks straight at the camera, *all* of the hubs perceive him to be looking straight at each one of them. We call this the *newscaster effect*[1] [5]. When he looks to the left, *none* of the hubs perceive him to be looking directly at any of them, but instead experience him to be looking over their right shoulder. This disruption of eye gaze awareness adds to making it difficult for the hubs to maintain awareness of the satellite's focus of attention.

A more subtle problem is that the satellite is just not as "present" as his collocated counterparts. We have observed several instances where people are talking in order around the table, e.g., to introduce themselves or report status, and the satellite on his proxy is skipped. We call this the *skip-over effect*. Despite the physical presence of the proxy, we believe there are many factors contributing to this deficit in presence.

We sought to mitigate these problems by physically moving the proxy in response to the satellite's actions. We began with one degree of motion by putting the display of the proxy on a motorized turntable that is controlled by the satellite participant. We called this a *kinetic proxy*.

To explore how kinetic proxy movement may address the newscaster and skip-over effects, we implemented several forms of motion control, and assessed hub participants' perceptions of the differences. We performed a lab study of the prototype, comparing a stationary proxy, a kinetic proxy under the satellite's explicit control by mouse cursor, and a kinetic proxy controlled implicitly by the satellite's head motion. This is the first study to directly compare alternative ways to project attention and evaluate people's responses.

The study's setting was a distributed conference consisting of several collaborative tasks. We used measures of conversation effectiveness such as engagement, naturalness of interaction, cognitive effort and sense of presence.

We found that the kinetic conditions were generally better than the stationary condition, with interesting caveats. For example, screen motion toward one person is more akin to turning one's back (rather than one's head) toward someone else. We also found unexpected and previously undescribed benefits and drawbacks to both means of control of the kinetic proxy. Among these is that implicit control generates more incidental proxy motion, which increases the cognitive effort experience by hub participants. These findings suggest designs for future experimentation in kinetic proxies.

2 Background

Researchers of interpersonal interaction recognize that the nuances of body orientation and non-verbal behavior—some consciously controlled, others not—are

[1] What we call the newscaster effect is more commonly referred to as the Mona Lisa effect. We prefer the former in this context, as the affordances of live audio and video viewed on a digital screen are more directly comparable, and are closer to most people's experiences.

important parts of the messages that people send and receive when communicating in person. Goffman highlighted the difference between the expressions that people *give* and those that they *give off* [6]. The former are verbal signs of the content they want to communicate; the latter are nonverbal and contextual. The former are more conscious; the latter are more subconscious. Hall's study of proxemics formalized that, with cultural differences, the way people orient their bodies toward or away from one another indicates their degree of intent to engage in conversation [7]. Face-to-face is more direct, a 90-degree angle is more casual, and a 180-degree angle is more transitory and disengaged. The design and motion of the kinetic proxy, with its explicit and implicit forms of control, is inspired by these insights.

2.1 Gaze Awareness in Videoconferencing

There is a long series of research prototypes that have tried to improve gaze awareness in videoconferencing. Hydra [8] was a 4-way distributed meeting system that packaged a video camera, small display, microphone, and speaker in self-contained, table-top surrogates dedicated to represent each participant at the meeting. Participants could look toward any of the other participants, and everyone would have an appropriate indication of his or her gaze. A study that compared mediated with same-room conversations found very similar patterns in overall speaking time, speech segment duration, and the distribution of turns taken. Our current study draws upon these same three conversational measures (among others).

More recent videoconferencing studies have also focused on gaze awareness. GAZE-2 [9] used several cameras at each site to capture video of each participant. The system detected which video window was being looked at by tracking each participant's eye movements. Each participant's video view was then shared with the other participants in a rendering of a virtual meeting room. Each video in the virtual room was presented as a flat screen, which was digitally skewed to face the remote participant toward which each participant was looking. These virtual screens rotated in a similar way to our single physical screen.

HP's Halo and Cisco's TelePresence are conference room-scale installations that support individual as well as group meetings. Because all of the participants at each site share the same views of the participants at the other site, correct gaze is not maintained as one moves to different positions within the room. The Virtual Window [10] sought to adjust for motion parallax such as this by tracking participants' head motions and moving a remotely-operated camera to simulate the effect of looking through an opening directly into the remote space.

MultiView [11] preserved gaze and gesture spatial relationships for groups of participants in a two-site (extensible to three) conference. Multiple cameras at each site—one facing each participant—sent multiple video streams to a directional viewscreen at the other site. Positioning was carefully arranged so that every participant at one site had a correct, angle-adjusted view of every participant at the other site, relative to his or her seating locations.

2.2 Sociable Robots

Other research prototypes expressly examine gaze direction in human-robot interaction. Yamazaki *et al.* [12] developed robots with movable heads to support

turn-taking in their communications. These robots engaged humans in one-on-one monologue or simple dialog while orienting their heads toward people or objects of interest. The studies emphasized the coincident timing of robotic gestures with transitional words. Our work also explores how orientation cues can influence interaction, but in a highly collaborative context.

Such robots also act as agents rather than avatars. By representing themselves in an interaction rather than a human other, and by not simultaneously presenting live video of that remote other, they avoid the potential to both complement and contradict an operator's actions. Kinetic proxies take this hybrid approach to providing physical motion as well as onscreen video.

2.3 Kinetic Proxies

A number of embodied telepresence systems have focused on kinetic proxies for hub-and-satellite interactions. PRoP [13] was a series of explorations of mobile, robotic personal stand-ins, composed of a video camera and LCD panel (and later, a small pointer) mounted atop a vertical pole and connected to a drivable base. Due to mobility constraints, PRoP's primary means of directed gaze was through a pan-tilt-zoom camera head, which served as a partial indicator of the operator's focus of attention, much like our proxies. But as we have found, this can be an ambiguous cue, as the camera may not always follow the operator's attention, or agree with his or her gaze. This overall form and interaction experience has recently appeared in commercial telepresence robots, including Willow Garage's Texai [14], Anybots' QB [15], and InTouch Health's Remote Presence [16].

Sun's Porta-Person [17] prototype also addressed the social presence of remote participants through motion, but specifically within "hybrid meetings," which include a mix of conference rooms, remote and local participants. The device included a video camera and display—replaced by a laptop computer in a later design—stereo speakers and microphones, all mounted atop a turntable and positioned on, or alongside, a conference table. Porta-Person and its turntable represent a direct lineage influence on the physical design of our kinetic proxy.

MeBot [18] was a small, desktop proxy with a three degree-of-freedom head that displayed cropped video of the operator's face, mounted to a mobile base with articulating arms. A study found that the proxy displaying motion was more engaging and likable than without motion. The role of motion as an indicator of attention was not evaluated, and since the participant's head motion was tracked (only), alternative forms of control were not compared.

Though it is not a telepresence proxy, the RoCo prototype [19] is relevant because it uses physical motion to influence engagement. It consisted of an LCD screen mounted on a 5 degree-of-freedom robotic "neck" that could rotate, lean and gesture expressively, mirroring the posture of the person standing in front of it. Studies of RoCo demonstrated that the system could create emotional engagement. RoCo, with its gesture mirroring capability, was an inspiration for our implementation of implicit control.

GestureMan's [20] goal was to support a remote operator in projecting his or her intentions in a workspace shared with a human collaborator. Unlike other proxies, it did not support live video of the operator. Instead, it had the ability to orient its own

robot head, body and a pointing arm, which were controlled by tracking the operator's head movements, screen touches and joystick use.

Animatronic Shader Lamps Avatars [21] were another form of kinetic proxy: a life-scale Styrofoam head mounted on a pan-tilt unit, onto which a video feed of the satellite operator's likeness is projected. The system tracked the operator's yaw and pitch head motions and mirrored them on the avatar. An advantage of this approach was that it presented correct focus of attention cues for all of the individuals interacting with the avatar and over a broad range of viewing angles. It has not been systematically studied from a human-factors perspective.

3 Laboratory Study

To test the effectiveness of the kinetic proxy, we conducted a laboratory study to compare it to a typical stationary video display, and to compare explicit and implicit motion control mechanisms. The study sought to explore the satellite's ability to project gaze cues under these alternative conditions, and hub participants' resulting sense of gaze awareness and presence, by including the directional affordances that people enjoy in face-to-face conversation.

3.1 Study Design

We ran the study as a within-subject design to encourage participants to make comparisons that primarily reflect the absence or presence of motion affordances, and their form of user control. Each group of participants experienced all of the following three conditions:

Stationary: The proxy showed no physical motion at all. This condition most closely resembles a traditional video-chat style conference.
Explicit Control: The proxy screen swiveled in response to the satellite participant explicitly selecting the location she wanted to aim her proxy towards. The position of the proxy was directly linked to the position of the mouse cursor over the panorama view (there was no need to click the mouse button).
Implicit Control: The proxy screen swiveled in response to where the satellite was looking, based on automatic tracking of her head motion.

We set out to test two hypotheses about the perception of motion and control of a kinetic proxy:

Hypothesis 1 (H1): Physical motion of the proxy results in greater conversational engagement, improved sense of directional attention, and preferred interactions by hub participants, compared to no motion at all.

By *physical motion*, we mean the physical movement of the proxy within the meeting room (where the hub participants are located). This is in contrast to apparent motion, which might be represented by repositioning a projected image on a stationary screen. For this study, we focused on physical motion of the screen that displayed the satellite participant's video stream.

Hypothesis 2 (H2): Implicit control of the proxy results in more natural interactions, with lower cognitive effort, and greater sense of the satellite participant's reactions, compared to explicit control.

In both cases, the goal is to more closely reflect the way that people interact during a collocated discussion, or at least, a reported improvement over the non-motion conference experience. We used a combination of behavioral and perceptual measures that are described in more detail below.

Procedure

We ran six groups of subjects through the experiment. We counterbalanced the ordering of conditions across groups. All of the groups were composed of three collocated hub participants who were recruited, plus a confederate acting as the satellite. Participants were led to believe that the confederate was an untrained recruit like them. The same confederate participated in all of the groups, so that the kinetic proxy would be operated in a consistent way throughout the experiment.

Each group worked in each condition for approximately 10-15 minutes. Immediately after each condition, participants individually completed a questionnaire that asked them to rate their experience with that condition. After all three conditions were completed, participants individually rated their preference among the conditions on a questionnaire and then participated in a semi-structured group interview. All sessions lasted approximately one hour, and the entire session was recorded using two overhead cameras and microphones in the hub room that captured the team's activity for later analysis.

Tasks

During each condition, the group performed a decision-making task with no right answer [22] that was intended to evoke discussion and interaction within the group. The following three tasks were always performed in the same order.

Task 1: Decide on a local restaurant to visit as a group after the study (hypothetically) that would work for everyone's dietary constraints and interests.
Task 2: Recommend a number of sites or attractions for a first-time visitor to the region, identified as an acquaintance of the satellite participant.
Task 3: Generate a personalized license plate for a well-known regional celebrity figure, whom the group selected from a short list of alternatives.

Participants were instructed at the beginning of the experiment that the members of the group with the best solution to Task 3, as judged by the experimenter, would receive a $20 gift card. (In fact all participants received the gift card.)

Participants

The 18 participants (9 male, 9 female) were recruited from the local region and did not know each other prior to the study. They were given a gratuity for their participation. Participants ranged in age from 20 to 55 years old. Their prior experience with videoconferencing varied from this study being their first exposure, to participating in

conferences on a weekly basis. While individual groups had an uneven makeup, every group had both genders (the confederate was female).

3.2 Experiment Setup

Turntable Kinetic Proxy

The satellite participant and hub group were located in adjacent rooms, and communicated through a videoconferencing proxy that we built for the study (see Fig. 2). A 12" Tablet PC supported in the portrait orientation displayed head-and-shoulders video of the satellite participant. The tablet was mounted atop an 8" turntable, which the satellite could remotely position within ±90°, to directly face any of the hub participants in the room. The frame, constructed of ¼" sheet acrylic to minimize its visual appearance, positioned the display approximately at eye level to the seated hub participants (see Fig. 1). The proxy and hub participants were evenly distributed around a 3' round conference table.

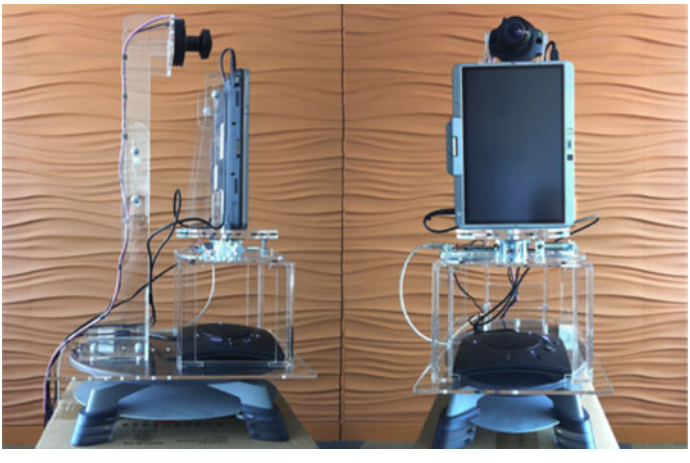

Fig. 2. Side and front views of the kinetic proxy. At its top is a fixed-position wide-angle video camera. Below the camera is a 12" tablet PC, which shows video of the satellite participant. The tablet is mounted atop a remotely-operated turntable which, in turn, is mounted atop a hutch that holds a videoconference speakerphone.

The proxy also included a fixed-position Axis 212 wide-angle camera. The view from this camera was displayed across the entire width of the satellite's 30" monitor (see Fig. 3). This configuration allowed the satellite to see all of the hub participants and their positions around the table, as well as the top edge of the Tablet PC, to confirm that it was oriented as she expected. The satellite's screen also displayed the video directly from the tablet's integrated webcam, but the confederate preferred to focus on the larger, wide-screen image, both to more directly engage the hub individuals and because it provided sufficient feedback of the screen's orientation.

All audio and video communication was over wired and wireless LAN, while the position control stream for controlling the motion of the proxy was carried by USB cable between the two rooms.

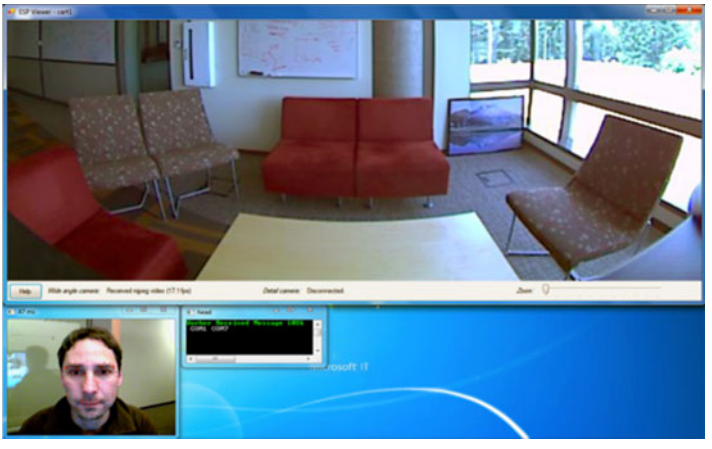

Fig. 3. View of the satellite's interface. At the top is the view coming from the fixed wide-angle camera. The lower left window provides feedback of the image the satellite is projecting on his or her proxy. It also provides feedback that the head-tracker has a good representational model of the satellite's head.

Explicit and Implicit Control

For the explicit control condition, the satellite participant moved her mouse to place the cursor at a particular spot on the client's widescreen view of the hub's workspace. Doing so sent a command which rotated the proxy's screen to face that location in the room. The client program updated the desired 'go-to' position approximately 30 times per second. Since the proxy's turntable was only capable of rotation in the horizontal plane, we only tracked the horizontal component of the cursor's position.

The satellite's wide display meant that she had to turn her head to see the hub participants to her left and right. For the implicit control condition, we tracked this head motion using an in-house, webcam-based software head tracker. This mode also updated the proxy position approximately 30 times per second. Similar to the explicit control condition, we used only the horizontal component of head rotation to orient the proxy's screen.

In both cases, proxy motion was calibrated so that the satellite's mouse or head motions mapped directly to the intended positions around the conference table.

3.3 Measurements

To measure the effects of the experimental conditions, we collected both objective and subjective data during each videoconference session. The objective measures included a tally of responses to deictic prompts and sociometric data that was captured by sensor badges that all of the participants wore around their necks [23].

The subjective measures included the individual questionnaire after each session that focused on the quality of interaction with the satellite member during that condition, the final questionnaire that probed participants' perceptions and preferences about group communication across all three conditions, and the semi-structured group interview, where we discussed these issues in greater detail.

Sociometric Badges

Each participant wore a sociometric badge that consists of several sensors, including an internal microphone, accelerometer, and infrared emitter and detector, which store their data on a removable Micro-SD card. The sensors are all housed in a lightweight black plastic case that is similar in shape, but slightly smaller in size, than a deck of playing cards. One badge was worn around each participant's neck on a lanyard that positioned it about mid-chest.

The microphone recorded the wearer's speech amplitude and time codes [23], and was the only sensor used in this study. Badge data from each day's sessions were downloaded, combined and analyzed using scripts that had been developed earlier. These scripts measured speaking time, speech energy, speech-segment length, and turn-taking per participant and condition.

Five badges were used in total: one worn by each hub participant, one worn by the satellite, and one placed on the proxy. The badge worn by the satellite was not used in the analysis, as the one on the proxy produced a duplicate but cleaner audio signal.

Deictic Prompts

Interspersed through each session, the confederate satellite participant would directly address one of the hub group members with a question of the form "What do *you* think?" or "What type of food do *you* like?" Often, this *first* question was followed with similar *second* or *third* questions to the other two members about their opinions. Responses to these questions were tallied during subsequent video analysis, paying attention to whether it was a first, second, or third question in a series. This distinction is important, because there are three potential respondents to a first question, while there are two for a second question, and only one for a third. Because of this increasing likelihood of responding correctly, we only analyzed responses for the first question according to the following protocol:

Correct Response: The intended person responded immediately.
Correct Confirmation: The intended person checked or confirmed whether he or she was being addressed before responding (e.g., "Do you mean *me*?").
Multiple with Correct: More than one person responded immediately. The group that responded included the intended person.
Incorrect Response: Someone other than the intended person responded immediately.
No Response: None of the participants responded to the question. An example is several seconds of silence, followed by a new thread in the conversation.

Two other codes are possible: 1) someone other than the intended person checked or confirmed whether he or she was being addressed before responding and 2) more than

one person responded immediately but this did not include the intended person. Neither of these categories was present in our data.

The number of deictic prompts varied from session to session, depending on the flow of the conversation or the personality of the participants, but when summed over the study, each condition had 12 first-question prompts and responses.

Questionnaires and Post-Study Semi-Structured Interview

The questionnaire after each condition included nine Likert-style statements and a single brief written response (see Table 1). Each had a 7-point scale with "Strongly Disagree" or "Strongly Agree" as their endpoints. The written response question invited participants to, "Please comment on your experience interacting with the remote person in this session."

The final questionnaire included three questions that asked which of the three conditions the participant felt the group communicated best with the satellite participant, as well as which condition was most preferred and least preferred.

Following the last of the three conditions, study administrators explained the nature of the study, revealed the confederate's role in the study, and discussed with the group reflections on the experience. The group interview provided an opportunity to interactively engage participants about their responses, and to follow-up on particular comments. The debrief sessions were video recorded and reviewed after the study.

4 Results

4.1 Conversational Engagement

Our goal was to measure the influence of the kinetic proxy on conversation effectiveness. For this comparison, we included sociometric measures such as speaking time and energy, segment length, and turn-taking. Prior work demonstrates how these measures can characterize interaction [24], as well as establishes desirable values and directionality for the measures in face-to-face interaction [7, 25, 26, 27].

In order to assess participants' level of engagement in the conversation, we measured their communication behavior using the sociometric badges. A within-subject analysis of this data detected several significant differences between the conditions, which suggest that the different configurations did have an impact on participants' engagement.

Speaking Time

Greater speaking time suggests a higher level of activity in a conversation [25]. We found a significant difference in the percentage of time that people spoke in each of the conditions ($F_{2,46}$=3.62, p=0.03). Participants in the kinetic conditions on average spoke for a larger percentage of time (Explicit: 18.7% and Implicit: 17.8%) than in the Stationary condition (16.6%). Posthoc pairwise comparisons showed a significant difference between the Stationary condition and the Explicit condition (p=0.02), and a marginally significant difference between Stationary and Implicit (p=0.09), but no difference was detected between Explicit and Implicit (p=0.29).

Speech Energy

Speech energy is the variance in speech volume. Higher speech energy often correlates to the perceived excitement of speakers [26] and can be used to indicate their level of activity in a conversation [28]. We found a significant difference in the variance in speech energy in each of the conditions ($F_{2,46}$=3.99, p=0.03). Participants spoke with higher speech energy in the Explicit (0.110 units) and Implicit conditions (0.107 units), than the Stationary condition (0.103 units). Posthoc pairwise comparisons showed a significant difference between the Stationary condition and the Explicit condition (p=0.008), and between Stationary and Implicit (p=0.04), but no difference between Explicit and Implicit (p=0.44).

Speech Segment Length

Speech segment length can indicate level of attentiveness and engagement in a conversation [25]. Speech segment lengths are shorter when there are more interjections such as "Oh," "Uh-huh," or "Wow" and when there are more frequent turn transitions. More interjections and turn-transitions may show that the listeners are more attentive to or engaged with a main speaker. Hence calculating the average segment length of all types of speech (such as interjections, interruptions, or full turns) allows us to estimate the attentiveness of the conversation. We found a significant difference in the length of speech segments in each of the conditions ($F_{2,46}$=10.53, p<0.001). The average speech segment length was longest in the Explicit condition (0.80 sec), followed by the Implicit condition (0.75 sec), while the Stationary condition had the shortest speech segments on average (0.72 sec). Posthoc pairwise comparisons revealed that the Explicit condition had significantly longer speech segments than both the Implicit condition (p=0.01) and the Stationary condition (p<0.001), but no difference between Implicit and Stationary (p=0.13).

Number of Turns per Second

Conversation turn-taking can indicate level of activity in a conversation. The level of interaction among the group members can be estimated by the frequency of turn-taking per second [24]. We found that the number of speech segments per unit of time was significantly different across the conditions ($F_{2,46}$=4.8, p=0.01), with the Explicit and Implicit conditions having more turns per second (0.77 turns/sec and 0.75 turns/sec, respectively) than the Stationary condition (0.63 turns/sec). Posthoc pairwise comparisons showed significant differences between Explicit and Stationary (p=0.009), and Implicit and Stationary (p=0.01), but no difference between Explicit and Implicit (p=0.72).

Turn-Taking with the Satellite Participant

We also examined turn-taking in relation to the satellite participant to see if turn-taking to and from the satellite participant was affected by condition. More turn-taking with the satellite participant indicates more active involvement of the satellite participant in the conversation. We found a significant difference ($F_{2,34}$=6.4, p=0.005) with hub participants having more conversational turns after or overlapping the satellite participant in the Explicit condition (0.09 turns/sec) and the Implicit condition

(0.08 turns/sec) than in the Stationary condition (0.07 turns/sec). Posthoc pairwise comparisons showed that both Explicit and Implicit had significantly more conversational turns with the satellite participant than the Stationary condition (p=0.007 and p=0.006, respectively), but no difference between Explicit and Implicit (p=0.90).

Relationship between Measures

When compared to the static condition, implicit and explicit conditions were associated with both longer speech segment length and number of turns. For tasks where the goal is to share a fixed amount of information, total speaking time may be somewhat constant among groups, so the number of turns depends on how information is shared in each turn. For open-ended tasks such as ours, total speaking time is highly dependent on how comfortable participants feel, how many ideas come up, etc. Groups could have more turns (to propose ideas) as well as longer speech segments (to elaborate on them). This interpretation is confirmed by the differences in speaking time. Prior studies do not reveal a general trend: Angura [29] shares the negative relationship that we found, whereas Mutlu *et al.* [30] shows a positive relationship.

4.2 Directing Attention

Responding appropriately when being addressed is important for fluid, natural interaction. For the deictic prompts in our study, we examined whether the intended person responded appropriately (see Fig. 4). Both motion conditions (Explicit and Implicit) had the highest number of correct responses, where 100% of the time (twelve instances in each condition), the correct person responded (although in two cases, others in the group also responded, indicating some ambiguity). Additionally, in the Implicit condition, one of the correct responses first sought confirmation. The Stationary condition was the most problematic. In four of the twelve instances, either the incorrect person responded, or no one responded. Examining just the number of correct responses, we found a significant difference across the conditions (Kruskal Wallis Test, χ^2=6.049, df=2, p=0.049), with Explicit and Implicit having more correct responses than the Stationary condition.

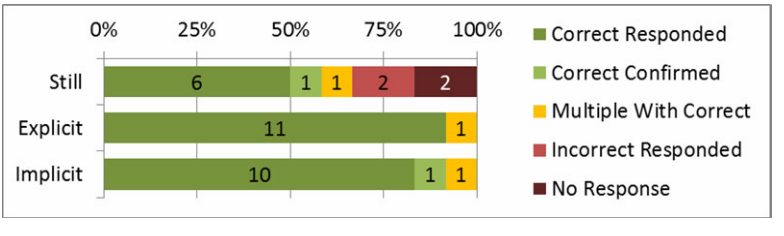

Fig. 4. Results of the confederate's deictic prompts, as determined by observing the responses in the session videos

4.3 Reactions to the Proxy

Table 1 shows the nine Likert questions participants were asked at the end of each session. The responses revealed that our participants felt positively about their experience

interacting with the satellite, in all of the conditions. They felt that they worked together as a team (average rating 6.6 out of 7) and that they communicated well with the remote person (6.4). They also indicated that it was natural to talk with the remote participant (5.6), and that it was comfortable (6.2) and not fatiguing (6.1) to interact with her.

Table 1. Mean post-task questionnaire results on a normalized* 7-point scale from most negative (1) to most positive (7). *Indicates that the scale has been inverted to match the other questions, where higher numbers indicate positive responses. ($^{+}p<.05$, $^{\pm}p<.01$)

Question	Stationary	Explicit	Implicit
1. Worked together as a team	6.8	6.4	6.6
2. Communicated with the remote person	6.4	6.2	6.6
3. Interaction with the remote participant was very comfortable	6.2	6.2	6.2
4. Was fatiguing to talk to the remote participant*	6.1	6.0	6.1
5. Felt natural to talk with the remote participant	5.7	5.4	5.6
6. Did not have a good sense of the remote participant's reactions*	5.8	6.1	5.9
7. Often confused about whom the remote participant was talking to*±	5.1	5.8	5.7
8. Could easily tell when the remote participant was talking directly to me±	4.7	5.8	5.2
9. Sense of the remote participant '*being there*'+	6.2	5.4	5.9

In terms of confusion about whom the remote participant was talking to, or being able to easily tell whom the remote person was talking to, ratings were lowest (more confusing) in the Stationary condition, and highest (less confusing) in the Explicit and Implicit conditions; however, the differences across conditions were not statistically significant (Friedman Test, p=0.054 and p=0.083) (see Fig. 5).

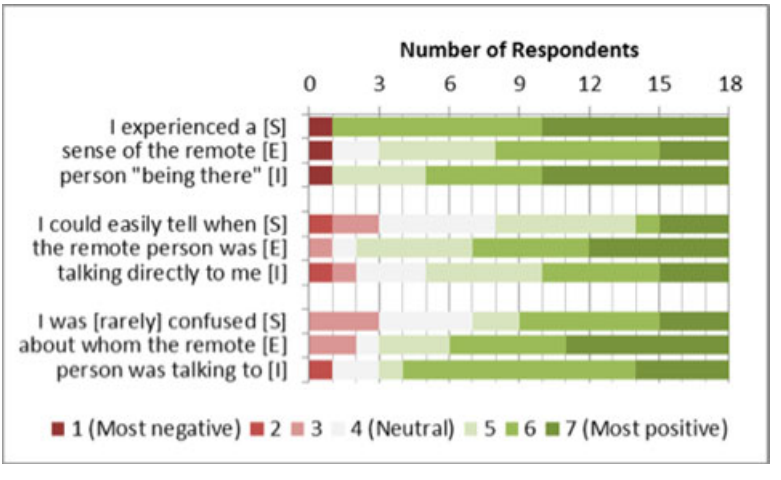

Fig. 5. Results of three post-session 7-point Likert questions: [S] = Stationary condition, [E] = Explicit, and [I] = Implicit. Responses to the third question are presented inverted to match the scale of the other two.

Interestingly, participants rated the sense of presence for the remote participant higher in the Stationary condition (6.2) than the Explicit (5.4) and Implicit (5.9) conditions (Friedman Test, p=0.003) (see Fig. 5). Pairwise comparisons showed a significant difference between the Stationary and Explicit conditions (p<0.032).

4.4 Overall Preference

At the end of the session, all participants were asked to indicate "In which of the three sessions did you and the group communicate best with the remote participant?" Four indicated the Stationary condition, five indicated the Explicit condition, and seven indicated the Implicit condition (with two stating no preference). Overall, participants felt that the kinetic conditions were more effective by 3:1 over Stationary.

4.5 Satellite's Perspective

Since the satellite in the lab study was a confederate, she was able to reflect on her experiences across all groups and conditions. She found that in the Explicit condition, she became more consciously aware of who she was directing her attention to than in the Implicit or Stationary conditions. Each movement of the mouse was made in order to either demonstrate listening to a particular person or direct speech toward someone. But over time, intentionally using the mouse in this way became more like a natural extension of her nonverbal communication and movements became more automatic.

In the Implicit condition, she sometimes found herself intentionally "driving" the proxy with her head, rather than moving naturally and trusting the proxy to follow her actions (perhaps due to some technical difficulties with the prototype). She also noticed that, unintentional head movements sometimes distracted the hub participants, making her a bit self-conscious about the way she moved her head.

Overall, the satellite had a slight preference for the Explicit condition. Each movement was meaningful, and participants seemed to pay attention to each movement and interpret its intent correctly. The proxy movement also added communicative value compared to the Stationary condition. She felt that her "voice" was amplified by the motion, and having explicit control over this was the most comfortable for her.

5 Discussion

Our lab study results showed that motion in the kinetic conditions performed better than the Stationary condition in terms of conversational engagement, accurate responses to deictic prompts, and trends in user ranking, confirming H1. A participant illustrated these results with the comment, "The motorized action brought the remote person to life." Hub participants were able to perceive the satellite's attention in motion through the swiveling of the display.

However, we also discovered some tradeoffs with motion. The swivel motion of the display could clearly communicate focus of attention ("Rotating LCD made it more clear who remote participant was talking to."). But, the more general locus of attention suffered, especially because swiveling toward one hub participant often meant that another hub participant was left looking at the edge of the display screen ("…when the remote person was talking to other people, I couldn't see her and I felt

excluded."). Given the flat surface of the display screen, swiveling the display was more narrowly directional than the physical affordance of head and body orientation. We expect that these tradeoffs are the main reason why participants rated the motion conditions as lower in a sense of 'being there.' Thus there was a tension in motion as it directed attention towards some hub participants but excluded others.

Rotating the display also introduced delay in the conversation, especially when the satellite had to explicitly control the aiming of the display ("…and then kind of like these awkward interruptions, every time she turned."). Some also found the motion to be distracting ("I felt like whenever she turned we all kind of like stopped talking for a second.").

Furthermore, aiming the display in the direction of the satellite's gaze did not address the eye contact issue. Especially in the Implicit condition, where turning the head was used to aim the display, the combination of swiveling the screen and having the head aimed off center from the camera combined to disrupt a true sense of eye contact ("Seemed as if she was looking over my shoulder.").

Our lab study also illustrated the tradeoffs between explicit and implicit control of motion. Explicit control showed the intent of the satellite more clearly, and was preferred by the satellite confederate, but incurred a delay in operating the interface to aim the display. Some evidence for this delay is found in the longer speech segment measure for the Explicit condition compared to both Implicit and Stationary conditions. This measure may reflect protracted speech while the satellite is simultaneously aiming the direction of the proxy. Or, explicitly aiming the display may leave the satellite's "gaze" aimed at a conversation partner longer than natural, protracting the partner's speech turn until the display turns away.

Implicit control tried to lessen the delay in moving the screen and reduce the cognitive burden on the satellite in aiming the display. However, it also added more ambiguity in the intention of the satellite, and the increased amount of motion exposes the hub to more of the negatives of motion (i.e., distraction, noise). Head motion in the physical world can be communicative (e.g., turning toward someone to elicit their response) or incidental (e.g., a side effect of not being able to remain completely still). But the kinetic display motion generated by the turntable was largely interpreted as communicative. Consequently, implicit control caused incidental head movement to be perceived as communicative movement, leading to more of a sense of distraction. In this way, implicit control transferred cognitive effort from the satellite to the hub.

We also expected that hubs would perceive implicit control as more natural than explicit, as more of the satellite's gestures would be available to them, but we found mixed indications in the questionnaire responses. While our results are equivocal about H2, we have a richer understanding of the tradeoffs between explicit and implicit control.

It is interesting that measures of behavior (sociometric turn-taking, deictic prompt responses) were more demonstrative than perceptual questionnaire responses. Some participants reported not even noticing that the display remained stationary during that condition. The behavioral measures showed that participants reacted to the motion conditions even though their perceptual rankings do not show statistically significant differences. Taken together, these results show that people's mechanisms of attention awareness may operate at a subconscious level, as has also been seen in other research [31].

6 Conclusion and Future Work

Returning the two practical problems with our proxy that prompted this exploration, do we believe that the kinetic proxy will address the skip-over and newscaster effects? Regarding the skip-over effect, we certainly believe the attention projection provided by motion will give the hub participants enough awareness of the satellite's attention to include her in the conversation. We look forward to deploying kinetic proxies into everyday usage to gain more experience with that. Regarding the newscaster effect, we set out to improve gaze awareness, in the tradition of the research prototypes reviewed earlier that have attempted to do so. Kinetic motion did provide a physical sense of gaze direction. However, it did not achieve eye contact, as the satellite's turning head turned the kinetic proxy display, but also caused her gaze to be directed off angle from the camera.

Our study shows that despite not achieving true eye contact, the kinetic proxy does project the satellite's attention focus so that hub participants could have engaging conversations and correctly respond to deictic requests. Our experiences with the lab study have led us to explore teasing apart attention awareness from gaze awareness.

Eye contact and gaze awareness are mechanisms used for attention projection and awareness when face-to-face. There has been a long series of research prototypes that have indicated how difficult it is to re-create eye contact and correctly convey gaze awareness in videoconferencing. But there are other mechanisms for conveying attention awareness without having to recreate eye contact. While Fels and Weiss [32] have begun to explore this space, we see more opportunities to support attention awareness without relying strictly on eye contact and gaze awareness.

We set out to explore using motion to improve interaction with the satellite participant. We discovered that motion helps, but has some tradeoffs. Swiveling a flat display screen toward one hub participant often excludes other participants, which can diminish the sense of presence of the satellite. Plus, rotating the visual mass of a display incurs lag and some found it to be distracting. Furthermore, swiveling the display did not succeed in improving eye contact.

Based on our study results, we would like to explore designs that leverage the benefits of physical motion, but avoid the exclusion of turning away from participants. Since swiveling the display still did not create true eye contact, perhaps there are ways to use a physical pointer, like a weather vane, to indicate attention projection while keeping the flat display stationary, so all hub participants maintain visual contact with the satellite. Alternatively, it would be interesting to explore convex displays, rather than the flat display screen, which might afford a wider range of directing a satellite's gaze while not 'turning her back' on some participants.

By distinguishing between gaze awareness and attention awareness, what we learned in our lab study generalizes beyond the particular turntable proxy that we examined. The motion of our turntable proxy did provide a stronger sense of attention projection and awareness, even though it did not offer true eye contact.

Complementary to the approaches of re-creating eye contact in videoconferencing systems, we should also explore ways of providing attention projection and awareness which may not depend on gaze awareness. This approach may open up options that are mechanically simpler, more abstract, and perhaps more diverse than previous approaches for creating engagement through videoconferencing solutions.

The current study focused on ameliorating the social asymmetries particular to hub-and-satellite teams. We have had meetings with multiple satellites in attendance by static proxy. In our experience, they have proceeded in much the same way—with comparable improvements in social integration but shortcomings in newscaster and skip-over effects—as meetings with single satellites. As we construct further prototypes, we hope to explore how interaction quality may differ (such as proxy-to-proxy conversations) through the use of multiple kinetic proxies.

References

1. Argyle, M.: Bodily Communication. Methuen, New York (1988)
2. Heath, C., Luff, P.: Disembodied conduct: Communication through video in a multi-media office environment. In: Proc. CHI 1991, pp. 99–103. ACM Press, New York (1991)
3. Gaver, W.: The affordances of media spaces for collaboration. In: Proc. CSCW 1992, pp. 17–24. ACM Press, New York (1992)
4. Venolia, G., Tang, J., Cervantes, R., Bly, S., Robertson, G., Lee, B., Inkpen, K.: Embodied social proxy: Mediating interpersonal connection in hub-and-satellite teams. In: Proc. CHI 2010, pp. 1049–1058. ACM Press, New York (2010)
5. Vishwanath, D., Girshick, A., Banks, M.: Why pictures look right when viewed from the wrong place. Nature Neuroscience 8, 1401–1410 (2005)
6. Goffman, E.: The Presentation of Self in Everyday Life. Doubleday Anchor, Garden City (1959)
7. Hall, E.: A system for the notation of proxemic behavior. American Anthropologist 65, 1003–1026 (1963)
8. Sellen, A.: Speech patterns in video-mediated conversations. In: Proc. CHI 1992, pp. 49–59. ACM Press, New York (1992)
9. Vertegaal, R., Weevers, I., Sohn, C., Cheung, C.: Gaze-2: Conveying eye contact in group video conferencing using eye-controlled camera detection. In: Proc. CHI 2003, pp. 521–528. ACM Press, New York (2003)
10. Gaver, W., Smets, G., Overbeeke, K.: A virtual window on media space. In: Proc. CHI 1995, pp. 257–264. ACM Press, New York (1995)
11. Nguyen, D., Canny, J.: MultiView: Spatially faithful group video conferencing. In: Proc. CHI 2005, pp. 799–808. ACM Press, New York (2005)
12. Yamazaki, K., Yamazaki, A., Okada, M., Kuno, Y., Kobayashi, Y., Hoshi, Y., Pitsch, K., Luff, P., von Lehn, D., Heath, C.: Revealing Gauguin: Engaging visitors in robot guide's explanation in an art museum. In: Proc. CHI 2009, pp. 1437–1446. ACM Press, New York (2009)
13. Paulos, E., Canny, J.: PRoP: Personal roving presence. In: Proc. CHI 1998, pp. 296–303. ACM Press, New York (1998)
14. Lee, M., Takayama, L.: Now, I have a body: Uses and social norms for mobile remote presence in the workspace. In: Proc CHI 2011. ACM Press, New York (2011)
15. Anybots, `http://www.anybots.com/`
16. InTouch Health, `http://www.intouchhealth.com/`
17. Yankelovich, N., Simpson, N., Kaplan, J., Provino, J.: Porta-Person: Telepresence for the connected conference room. In: Ext. Abstracts CHI 2007, pp. 2789–2794. ACM Press, New York (2007)
18. Adalgeirsson, S., Breazeal, C.: MeBot: A robotic platform for socially embodied telepresence. In: Proc. HRI 2010, pp. 15–22. ACM Press, New York (2010)

19. Breazeal, C., Wang, A., Picard, R.: Experiments with a robotic computer: Body, affect and cognition interactions. In: Proc. HRI 2007, pp. 153–160. ACM Press, New York (2007)
20. Kuzuoka, H., Kosaka, J., Yamazaki, K., Suga, Y., Suga, Y., Yamazaki, A., Luff, P., Heath, C.: Mediating dual ecologies. In: Proc. CSCW 2004, pp. 477–486. ACM Press, New York (2004)
21. Lincoln, P., Welch, G., Nashel, A., Ilie, A., State, A., Fuchs, H.: Animatronic shader lamps avatars. In: Proc. ISMAR 2009, pp. 423–432. IEEE, Los Alamitos (2009)
22. McGrath, J.E.: Groups: Interaction and Performance. Prentice-Hall, Inc., Englewood Cliffs (1984)
23. Olguín Olguín, D., Waber, B., Kim, T., Mohan, A., Ara, K., Pentland, A.: Sensible organizations: Technology and methodology for automatically measuring organizational behavior. IEEE Transactions on Systems, Man, and Cybernetics 39(1), 43–55 (2009)
24. Kim, T., Chang, A., Holland, L., Pentland, A.: Meeting mediator: Enhancing group collaboration using sociometric feedback. In: Proc. CSCW 2008, pp. 457–466. ACM Press, New York (2008)
25. Curhan, J., Pentland, A.: Thin slices of negotiation: Predicting outcomes from conversational dynamics within the first 5 minutes. Journal of Applied Psychology 92, 802–811 (2007)
26. Hung, H., Gatica-Perez, D.: Estimating cohesion in small groups using audio-visual nonverbal behavior. Transactions on Multimedia 6(12), 563–575 (2010)
27. O'Conaill, B., Whittaker, S., Wilbur, S.: Conversations over video conferences: An evaluation of the spoken aspects of video-mediated communication. Human Computer Interaction 8(4), 389–428 (1993)
28. Vertegaal, R., Ding, Y.: Explaining effects of eye gaze on mediated group conversations: Amount or synchronization? In: Proc. CSCW 2002, pp. 41–48. ACM Press, New York (2002)
29. Angura, X.: Robust speaker diarization for meetings. PhD dissertation, Department of Signal Theory and Communications, Universitat Politecnica de Catalunya, Barcelona (2006), http://xavieranguera.com/phdthesis/node47.html
30. Mutlu, B., Shiwa, T., Kanda, T., Ishiguro, H., Hagita, N.: Footing in human-robot conversations: How robots might shape participant roles using gaze cues. In: Proc. HRI 2009, pp. 61–68. ACM Press, New York (2009)
31. Sato, W., Okada, T., Toichi, M.: Attentional shift by gaze is triggered without awareness. Experimental Brain Research 183(1), 87–94 (2007)
32. Fels, D., Weiss, T.: Toward determining an attention getting device for improving interaction during video-mediated communication. Computers in Human Behaviour 16(2), 99–122 (2000)