Consistency in Physical and On-screen Action Improves Perceptions of Telepresence Robots

David Sirkin and Wendy Ju
Center for Design Research
Stanford University, Stanford, CA
{sirkin, wendyju}@cdr.stanford.edu

ABSTRACT
Does augmented movement capability improve people's experiences with telepresent meeting participants? We performed two web-based studies featuring videos of a telepresence robot. In the first study (N=164), participants observed clips of typical conversational gestures performed a) on a stationary screen only, b) with an actuated screen moving in physical space, or c) both on-screen and in-space. In the second study (N=103), participants viewed scenario videos depicting two people interacting with a remote collaborator through a telepresence robot, whose distant actions were a) visible on the screen only, or b) accompanied by local physical motion. These studies suggest that synchronized on-screen and in-space gestures significantly improved viewers' interpretation of the action compared to on-screen or in-space gestures alone, and that in-space gestures positively influenced perceptions of both local and remote participants.

Categories and Subject Descriptors
H.5.m **[Information Interfaces and Presentation (e.g., HCI)]**: Miscellaneous.

General Terms
Design, Experimentation.

Keywords
Telepresence robotics, embodied interaction, videoconferencing.

1. THE PROXY-IN-PROXY PROBLEM
Embodied proxies are increasingly used by remote workers to participate in the activities of a central workplace. Embodied proxy systems combine a live video representation of the remote worker with a local physical platform, often with human-body-like proportions. This setup enables remote workers to communicate naturally with their distant, collocated peers, have a similar presence (at least in terms of body size and location) to their collaborators, and often allows them to move, or be moved, around the central workplace to engage in day-to-day, informal interactions [18].

However, combining a video display of a remote worker on an articulating base creates a confusing juxtaposition, an issue we call the *proxy-in-proxy* problem. The worker actually has two representations in the central workplace: that of the audiovisual feed, and that of the physical platform. These two distinct channels are likely to portray inconsistent non-verbal facial or gestural cues, even if that is merely the lack of complementary action between them. We know that in face-to-face interactions, inconsistencies in verbal and non-verbal cues are often interpreted as a sign of deceit [17] and mistrust, and cause increased cognitive load [10]. Video-based interactions can also influence observers' abilities to interpret these cues, such as when comparing videos that include facial expressions alone to those that include the entire body [8]. So the question arises: how do inconsistencies between *on-screen* and *in-space* behaviors affect people's interpretations of non-verbal gestures and their perceptions of the remote worker's proxy?

HRI'12, March 5–8, 2012, Boston, Massachusetts, USA.

Figure 1. Our embodied proxy displays the remote participant's on-screen expressions and gestures in physical space.

In this paper, we present two studies that compare people's subjective interpretations and responses to single-channel gestures (either a visible on-screen expression alone, or a physical in-space motion alone) versus coordinated gestures (where both on-screen and in-space representations move in concert). The studies focus on gestural aspects of video-based communication, including facial expressions, head motion and orientation, and posture. Results suggest the importance of creating consistency across the remote worker's various proxy representations, and the great benefit that coordinated action can have, even for the most familiar behaviors. A better understanding of the issues caused by the proxy-in-proxy problem could help robot designers develop technologies to address and resolve inconsistencies and their effects.

2. BACKGROUND
2.1 The Embodied Proxy Metaphor
Unlike earlier telepresence designs, which employed the metaphor of the *avatar* (such as the on-screen avatars in BodyChat [27] or physical avatars like Geminoid [23]) or the *portal* (such as EuroPARC's Portholes [7], Berkeley's MultiView [22] or Cisco's Telepresence System), systems that employ the embodied proxy metaphor commonly have a one-to-one relationship of person to proxy, human-body-like proportions, and a near-life-size display of the remote person's face and upper body. Some, like Sun Micro-systems' Porta-Person [29] and MIT's RoCo [3] feature a flat screen computer monitor system on a robotic head/neck mechanism that sits on a desk or chair. Others, like Paulos' PRoP [24], Microsoft Research's Embodied Social Proxies [26], Willow Garage's Texai [19] and Anybots' QB host a live video on a large

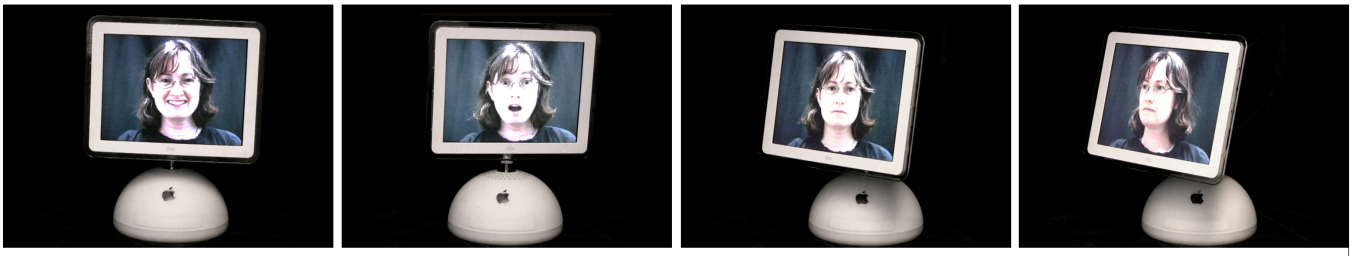

Figure 2. The Study 1 video prototypes showed brief video clips of on-screen, in-space or combined on-screen/in-space gestures. Shown here are: on-screen laugher, on-screen surprise, in-space look to one side, and combined on-screen/in-space look to one side.

flat screen mounted on a remotely steerable base. MIT's MeBot [1] is sized for desktop use, but extends the metaphor to include a three degree-of-freedom screen mounted on a mobile base with articulating arms, which track the operator's motions.

2.2 Challenges in Embodied Representation

Although each embodied proxy platform design has its own unique characteristics and issues, people bring common shared expectations about how to interpret embodied social cues to each [13]. Casting the video telepresence feed into an embodied form has significant ramifications for how on-site workers perceive the proxy and, in turn, the remote worker.

Designers of telepresence systems might assume that a lack of motion in the embodied platform would provide a neutral stage on which to perform on-screen actions. However, as Groom, *et al*. [12] note, "people expect bodies not only to serve as decorations suggesting identity, but also as functional units intended to interact with the environment and to communicate information." We use embodied non-verbal communications such as gestures, body movements, posture, visual orientation, and spatial behavior in concert with our verbal communication to signal our attention, express emotions, convey attitudes, and encourage turn-taking [2], and numerous studies suggest that we (perhaps subconsciously) prefer that our technological counterparts follow suit. People respond positively towards agents that display consistency in verbal and non-verbal cues compared to those with mismatched cues [14]. We prefer that agents use gaze and gesture to provide contextual grounding for the agent and user's shared experience to those that do not [4]. We read intentions from the seemingly unintentional non-verbal cues made by robots [20]. Researchers have found that rigid and frozen postures are a common "deception cue" [13], and habituated social responses to such cues could affect the perception on-site workers have of their remote counterparts. Hence, there are reasons to doubt the neutrality of an unmoving embodied platform, or an unmoving video image that features a large image of a person's head.

2.3 Consistency in Physical & On-screen Action

Recent experiments found that people prefer expressive robotic motions to a static platform [1]. We propose that *consistency* between physical and on-screen action is critical to this preference.

Related work on embodied agents and personality has defined consistency as accordance between voice and body cues [21]. Consistency can thus be seen as one cue supporting another. For gesturing embodied proxies, the remote worker's on-screen expressions and voice (if present), and the platform's physical movements, provide these cues. Such proxies can represent either in-space motions alone, on-screen actions alone, or a combination of the two. We therefore chose to investigate individual aspects of gestural expression, and how they contribute to positive interactions, under these three conditions.

3. RESEARCH METHODOLOGY

In order to explore how people react broadly to proxy-in-proxy designs, we ran our experiments using online video prototypes and crowdsourced participants from Amazon's Mechanical Turk (MTurk) service [16]. MTurk is an online marketplace where *requesters*—people or organizations who typically have small tasks to be performed—connect with *workers* who have the time and attention to perform them for a small payment.

Video prototypes are short movies that demonstrate how an interactive technology would perform. There are a number of practical reasons to employ video prototypes rather than in-person trials. They permit us to find the most salient design factors prior to building a fully functional system; to tune the proxy's motions to be as subtle or obvious, coarse or refined as required; to reproduce precise timings between the remote actor's on-screen gestures and the proxy's motions; to better control the study, by exposing participants to the same, consistent stimulus; and to access a more diverse audience than possible using local participants. To ensure the quality of responses [11], we tracked each submission time, to confirm that responses did not occur before the video completed, and that total task time was not notably below average; we reverse coded particular questions and filtered for straight-line responses; we assigned unique identifiers to match each worker to a location, and confirmed that IP addresses originated in the expected regions.

Another consideration is that online responses may differ from those of in-person trials. Powers, *et al*. [25] found that a remote projected robot could be used to study many critical social processes, including engagement. And studies comparing real-world evaluations of interactive prototypes with web-based video prototypes in the field of human-robot interactions by Kidd [15] and Woods, *et al*. [28] found that results from video studies tend to be consistent with in-person studies, although the simulated studies might not point out every salient factor that may be present in a real-world setting.

4. STUDY 1: ACTIONS IN ISOLATION

4.1 Video Prototypes

This study featured two sets of 27 video clips of different interactive gestures, with each set being performed by a male or a female actor, on a stationed telepresence proxy (see Figs. 1 and 2). Each video clip was approximately five seconds in duration, and depicted a first-person view of the proxy. The proxy was constructed of an iMac G4 computer with hemispherical base and 15-inch screen, connected to by an articulating "neck." The neck allowed three types of motion: pan and lift at the hemispherical base, and vertical screen tilt at the top of the neck-screen connection. The proxy's screen motions were puppeted so as to approximate the movements of a human head, and in particular, to reproduce the movements displayed by the actor, whose image was shown as a head-and-shoulders gesture against a black curtain background. The clips had no audio.

Each video clip presented one of nine behaviors that might be observed during a typical video interaction:

Agree (Nod "Yes")	Laughter	Look Down at Table
Disagree (Shake "No")	Look to One Side	Confusion
Surprise	Lean In to Look Close	Think Carefully

Each of these nine gestures had three variants:

On-screen The screen/frame remained still and only the actor's visible facial expressions and gestures changed.

In-space The screen/frame was actuated but the actor's facial expressions and gestures remained neutral.

On-screen/In-space The screen/frame was actuated and the actor's facial expressions and gestures changed.

4.2 Hypothesis

Our proxy design allowed us to isolate gestures enacted by the remote participant (visible on the proxy's screen), gestures made by the screen itself (which occurred in physical space) and the combination of the two. We could then explore the proxy-in-proxy problem through the following hypothesis:

H1 Consistency between on-screen and in-space action improves observers' comprehension of the message that the remote participant is expressing when compared to:

a. On-screen action alone, without corresponding physical motions.

b. In-space action alone, without corresponding facial expressions.

Consistency refers to coordinated gestural cues between physical and on-screen channels, which occur in concert with one another.

4.3 Study Design

The study was a within-subjects design: each participant was shown video prototypes of all nine different behaviors, with each behavior enacted using all three types of motion.

4.3.1 Procedure

We recruited 55 MTurk participants in each of three regions: the United States, India and the rest of the world, and collected data over four days. This was done to isolate any regional differences, as well as to account for the time-dependent nature of a globally administered study. Participants viewed a brief instructional page that described who we were, and what the study was about. The average time to complete the study was 18 minutes. Participants were paid US$2 each, which equates to about $6 per hour.

4.3.2 Video Questionnaire

Each participant was then shown the series of 27 video clips in random order. Even-numbered participants viewed a set with a male on-screen actor, and odd-numbered participants viewed a set with a female actor. Each clip appeared on its own webpage, and was accompanied by four questions. The first question provided a menu of nine alternative interpretations of the behavior that appeared in the video:

1. Which choice best describes what the remote participant is communicating?
   ○"I'm surprised" ○"Let me look closer" ○"I disagree"
   ○"That was funny" ○"Let's look over there" ○"My feelings are hurt"
   ○"What's down there?" ○"I agree" ○"I'm thinking"
2. How confident are you in your interpretation of this message?
3. How strongly did you respond to the message that you interpreted?
4. Rate the observed actions based on the following 3 parameters.

We ran a pilot study that revealed that the *confusion* and *think carefully* behaviors appeared and were rated very similar to most participants. As a result, responses to both of those clips were mapped to the same "I'm thinking" menu choice.

The next two questions on each page used a 7-point Likert scale to ask participants about their confidence in their interpretation, (where 1="not sure at all" and 7="absolutely sure"), and about how strongly they responded to the message (1="barely at all" and 7="very strongly"). The last question asked them to rate the observed action along three further dimensions (with 1="unnatural" and 7="natural," then 1="unfamiliar" and 7="familiar," and finally, 1="unintentional" and 7="intentional").

4.3.3 Coding of Responses

Each response to the menu of alternative interpretations was coded as either correct or incorrect, so that a guess would have a 1-in-9 (about 11%) chance of being correct. Our assignment of what made a response correct was based on our communicative intent when we created the behavior for each video (prior to collecting data), as well as comparison with the most frequently cited response (the mode) for that behavior among all of the participants' responses. These two bases agreed in every case.

4.3.4 Participants

We had 164 complete responses, with most of the rest of the world participants located in Europe. Nearly all respondents reported their gender and age, with 65% being male, and with ages ranging from 18 to 63 (M=29.3, SD=9.15).

4.4 Results

4.4.1 Participant Interpretation of Gestures

To determine whether participants' interpretations of the behaviors differed significantly from chance alone, we compared responses with the expected value from guessing. For the on-screen/in-space condition, z ranged from 6.5 to 28.2 and for the on-screen only condition, z ranged from 6.6 to 25.9, with all gestures identified as significantly above chance at $p<.001$. For the in-space only condition, six of the nine gestures were identified as significantly above chance (z ranged from 5.629 to 14.432) at $p<.001$, with *confusion* (z=2.2) having $p<.01$, *laughter* (z=1.6) having $p<.05$ and *think carefully* (z=0.7) having $p<.25$.

To determine whether participants' interpretations of the behaviors shown differed from each other, we performed a log linear regression; the model assessed the impact that the type of motion, the particular gesture, and the gender of the actor, had on interpretations. To find significant differences, we used one-way ANOVA with interpretation as a dependent variable, and gesture

Table 1: Observers' subjective ratings for interpretation, for both individual gestures and the type of motion*.

	Interpretation				
	On-screen & In-space	**On-screen**	**In-space**	***M***	***SD***
Shake "No"	0.80	0.82	0.65	0.76 a	0.09
Nod "Yes"	0.80	0.84	0.62	0.75 a	0.12
Surprise	0.78	0.74	0.44	0.65 b	0.18
Laughter	0.86	0.80	0.16	0.61 b	0.39
Look Side	0.71	0.60	0.47	0.59 b,c	0.12
Lean In	0.74	0.66	0.36	0.58 b,c	0.20
Look Down	0.63	0.51	0.32	0.49 d	0.15
Confused	0.59	0.50	0.18	0.42 e	0.22
Thinking	0.35	0.36	0.13	0.28 f	0.13
M	0.69 a	0.65 b	0.37 c		
SD	0.15	0.17	0.19		

*Differing subscripts indicate significant differences.

and type of motion as the fixed factors. Responses from the three study regions followed very similar patterns, but had somewhat different values. Summary statistics for all regions are shown in Table 1, which includes significant differences at $p<.05$ for gesture and $p<.001$ for type of motion, per Tukey post-hoc. Example data from the United States is shown in Fig. 3.

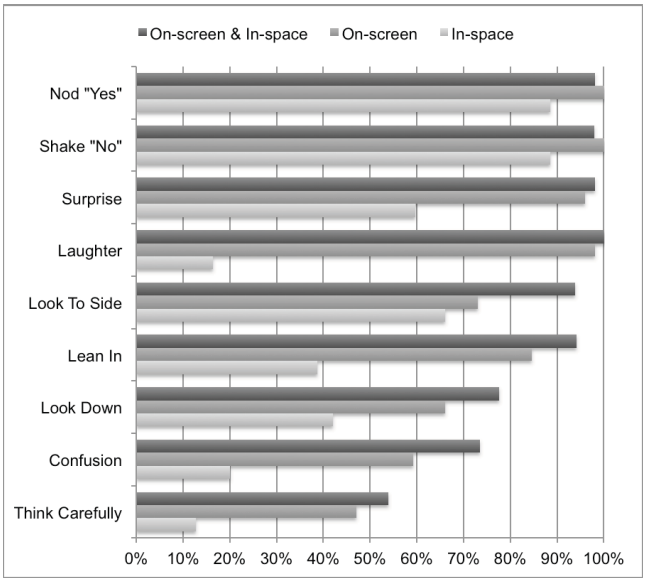

Figure 3: The percentage of responses that agreed with the intended message, for participants in the United States.

We found that the type of motion had a significant effect on participants' interpretation of the behaviors that they saw ($\chi^2(2)=395.7$, $p<.001$). The in-space condition had substantially lower accuracy across the board compared to the other two types of motion, each of which notably included the actor's on-screen expressions. Because this can mask significant differences between the other two types of motion (which concern us most), we performed a follow-up analysis excluding the in-space condition, focusing only on the difference between on-screen and on-screen/in-space conditions. We found that the type of motion still had a significant effect on interpretation ($\chi^2(2)=15.0$, $p<.001$). The on-screen/in-space condition was the better performer in most cases, with the amount of improvement dependent upon the *category* of gesture (which we cover in the discussion).

We also found that certain gestures (of the nine shown: such as laughter, agree, disagree, and so on) were more recognizable than others, with gesture having a significant effect on participants' interpretation ($\chi^2(8)=345.9$, $p<.001$). Post-hoc analysis, including only the on-screen and on-screen/in-space conditions, found that gesture still had a significant effect ($\chi^2(8)=239.9$, $p<.001$).

The interaction between the type of motion and gesture was also significant ($\chi^2(10)=94.4$, $p<.001$). This implies that certain types of proxy motion (say, in-space movement) influenced perceptions of particular gestures (such as the actor's glance to the right) more so than they influenced perceptions of other gestures (such as a nod or a head shake). There was no significant effect due to the on-screen actor's gender, which agrees with our expectation that it was the combination of motion and gesture, rather than the persona of the remote worker, which influenced responses.

4.4.2 Confidence and Strength of Response

Observers' confidence in their interpretations, and the strength of their response, followed the same pattern as the interpretations of the intended message. This agreement was with respect to both the type of motion and the particular gesture, meaning that participants actually were more correct when they thought that they were, and in those cases, the message had a stronger impact. Summary statistics are shown in Table 2, including significant differences per Tukey post-hoc at $p<.05$ for gesture. For type of motion, $p<.001$ for confidence and $p<.05$ for strength of response.

Table 2: Observers' subjective ratings for both confidence and strength of response, for individual gestures and type of motion.

Confidence

	On-screen & In-space	On-screen	In-space	*M**	*SD*
Shake "No"	6.33	6.19	5.23	5.92 a	0.60
Nod "Yes"	6.10	6.08	5.11	5.76 a,b	0.57
Surprise	6.07	5.91	4.35	5.45 b,c	0.95
Laughter	6.28	6.13	3.86	5.42 c	1.36
Look Side	5.52	5.03	4.25	5.26 c	0.93
Lean In	5.86	5.74	4.18	4.93 d	0.64
Look Down	5.46	4.79	4.34	4.86 d	0.57
Confused	5.12	5.14	4.23	4.83 d	0.52
Thinking	4.90	4.90	4.07	4.62 e	0.48
M	5.74 g	5.55 h	4.40 i		
SD	0.51	0.57	0.46		

Differing subscripts indicate significant differences at *$p<.05$ and **$p<.001$.

Strength of Response

	On-screen & In-space	On-screen	In-space	*M*	*SD*
Shake "No"	5.87	5.78	4.77	5.47 a	0.61
Nod "Yes"	5.69	5.64	4.63	5.32 a,b	0.60
Surprise	5.77	5.44	4.40	5.20 a,c	0.72
Laughter	5.89	5.88	3.82	5.19 a,d	1.19
Look Side	4.92	4.93	4.26	5.13 b,c,d	0.78
Lean In	5.57	5.59	4.23	4.71 e	0.38
Look Down	5.14	4.48	4.23	4.62 e	0.47
Confused	4.91	4.81	4.24	4.83 e	0.37
Thinking	4.65	4.72	4.02	4.47 f	0.38
M	5.38 g	5.25 h	4.29 i		
SD	0.45	0.49	0.29		

Differing subscripts indicate significant differences at $p<.05$.

4.5 Discussion

Our online study showed that consistency between the remote actor's facial expressions and gestures and the proxy's physical motions resulted in improved understanding of the behavior portrayed, higher confidence levels, and stronger responses, confirming both H1a and H1b.

But the data also indicate a more nuanced interpretation. One participant stated it this way: "Looking back, I feel that for 'simple' messages (yes, no, etc.), the movement didn't do much, and maybe even got in the way. For more complicated messages it seemed to make more difference." Fig. 3 supports this comment, and suggests even more. Notice that gestures such as *nod "yes"* or *shake "no"* were very well recognized across the conditions, even the in-space only condition. Even without a complementary facial expression, the proxy's motion gave participants the right idea. In contrast, gestures such as *laughter* and *surprise* were not as well recognized in the in-space condition. They seemed to require another cue—such as a smile—to place them in the proper context. In this case, the addition of physical motion did not substantially improve the clarity of the message. We call these two categories of gesture "intentionally communicative" and "unanticipated response." The first is sufficiently communicated

through in-space motion alone, while the second is sufficiently communicated through on-screen expressions alone.

Gestures such as *lean in to look close*, *look to one side*, and *look down at table* serve mostly to indicate the remote actor's focus of attention. The addition of motion to each of these significantly improved interpretations. Again this makes sense, as motion toward one direction tends to draw attention that way. These represent another category of gesture, "focus of attention," best communicated through both on-screen and in-space channels.

Confusion and *think carefully* reflect a fourth category of gesture, "thoughtful, internal states." These gestures were the least well-recognized overall, but accuracy improved noticeably when both on-screen expressions and in-space motions were combined.

5. STUDY 2: TEAM SCENARIO

In Study 1, we identified discernable and meaningful behaviors, then asked observers to respond to them individually. As proxies generally support communication within working groups, findings drawn from behaviors performed in relative isolation only take us so far. To extend our understanding of how proxy motions are perceived in the more situated context of practical use, we ran a follow-up study depicting a collaborative, problem-solving team interaction. Our focus thus shifts from recognizing the affordances and components of proxy gesture to interpreting the social relationships that alternative forms of gesture imply and support.

5.1 Relational Measures

To evaluate the effects that the proxy-in-proxy setup had on the perceived interactions between meeting participants, we chose a set of measures that characterize the remote worker's level of participation, degree of engagement, and relative social status during a distributed team meeting. We based our video scenarios and the subsequent analysis on an adaptation of Dillard's relational message scale [6], which focuses on how people interpret the *relationships* between communicators. The scale describes judgments according to nine relational factors, which include: 1) immediacy, the degree to which someone actively engages another; 2) affect, which suggests warmth, interest and attraction; 3) similarity/depth, which indicates familiarity, personalism and friendliness; 4) receptivity/trust, the degree to which someone expresses or seeks trust; 5) composure, or how relaxed versus tense someone appears; 6) formality, a demeanor of being responsive and disclosive; 7) dominance, the extent to which someone tries to persuade another or control the conversation; 8) equality, being treated as, or treating someone else, as an equal; and 9) involvement, an indication of interest or detachment.

5.2 Video Prototypes

The study featured four video clips of a distributed team during an active design session. The team consisted of three members: two were collocated, and one was remote and communicated through a proxy. The proxy from Study 1 was enhanced to provide remote robotic control. The iMac G4's screen was actuated by three DC motors and a cable drive system to move the neck and screen to positions controlled using a remote interface. Screen motions were controlled gesturally, through the orientation of a hand-held Wii remote, so that larger movements of the remote produced more rapid movements of the screen. Pilot trials also revealed the need for arm-based gestures, so we added a Lynxmotion AL5D five degree-of-freedom robotic arm to provide deictic as well as other symbolic gestures critical to interactive team activities [5].

Figure 4. The Study 2 video prototypes depicted a design collaboration scenario (panel 1). Half of the study conditions showed on-screen and in-space proxy gestures (panel 2); the other half, showed on-screen proxy gestures only (panel 3).

The video scenario shows a third-person view of a remote teammate (Eric) asking an on-site design collaborator (Becky) for assistance revising the design of a hand-held remote control (see Fig. 4 for the physical setup). A brief discussion ensues about how to make the remote work for a wider range of hand sizes, and another local participant is called over for further design support. After a brief period, the three check back in with each other, review the designs they had developed, and choose one that resolves the original problem. The clip is one and a half minutes long, and includes an audio track of the actors' conversation.

5.3 Hypotheses

The team scenario allowed us to examine how people interpret aspects of social interaction, such as friendliness, composure and equality, under different proxy operating modes. Extending our findings from Study 1, we formed two hypotheses:

H1 Consistency between on-screen and in-space action improves observers' interpretations of the remote team member's role (better recognition), when compared to on-screen action alone.

We expect consistency between actions to favorably influence observers' judgments of the indices that have a strong relation to nonverbal, physical expression. These include immediacy, which is enhanced by greater activity, intensity and enthusiasm;

similarity/depth, which is influenced by displays of friendship (a wave or handshake), body orientation and mirroring postures; dominance, which is represented by persuasive or controlling actions (pointing or gazing at someone); and involvement, which includes attentional cues (concentration or distraction). Given the Study 1 finding that in-space motion clarifies focus of attention cues, we expect involvement to be a strong component of H1.

H2 Consistency between on-screen and in-space action heightens observers' awareness of the remote participant's role (strength of effect), when compared to on-screen action alone.

Here, we expect whether or not the proxy physically gestures to influence observers' judgments of those indices that reveal the remote participant's role as a leader or follower in the conversation: in particular, dominance and equality between teammates. Enacting each role with embodied movements should make that role more apparent. For example, the remote participant might be evaluated as a more assertive leader when his proxy is in motion, compared to when it is not.

5.4 Study Design

We recorded four variations of the video scenario. Each followed the same script, but varied one of two factors, yielding a 2x2 study design. The first factor was the type of proxy gesture, which included on-screen expressions and gestures only, or combined on-screen and in-space action. We did not investigate the in-space only condition for this study, as it under-performed the other two in all cases for Study 1. The second factor was the leadership role of the characters. In two of the four conditions, the remote teammate initiates the meeting, asks the questions and calls over the third participant, giving him a more dominant role in the meeting. In the other two conditions, the local designer takes all of these actions, giving her the more dominant role.

During the course of the video, the remote participant exhibits many non-verbal cues: he turns to face each local collaborator, peers closely at a prototype, compares hand sizes, and jolts in surprise. Actions such as pointing out the location and size of particular design features on a prototype, rapping on a table to get someone's attention, and waving to request a turn to speak, which are only visible on the proxy's screen in the on-screen condition, use the robotic arm in the on-screen/in-space condition.

5.4.1 Procedure

We recruited 110 MTurk participants globally, and collected data over two days. Participants viewed instructions and questionnaire pages similar to Study 1, only adjusted for the different video and greater number of statements. The average time to complete the study was 10 minutes. Participants were paid US$1 each, which equates to about the same $6 per hour as Study 1.

5.4.2 Video Questionnaire

Each participant was shown a single video clip of the team interaction scenario, which was selected from the four conditions based on the participant's order of arrival at the study website. The clip appeared on a single webpage accompanied by two sets of 33 statements, oriented in two columns. The subject of one set was the remote team member, and the subject of the other was the local teammate. Statements were adapted from the original messaging scale statements, and included, for example, "Eric was intensely involved in the conversation" (immediacy), "Becky tried to control the interaction" (dominance), and "Eric considered them equals" (equality). Each statement was followed by a seven-point Likert scale for responses, with "strongly disagree" and "strongly agree" as their endpoints.

5.4.3 Coding of Responses

Responses were grouped by their message scale categories, and each group was then averaged, to produce nine response indices.

5.4.4 Participants

We had 103 complete sets of responses. A reverse-IP lookup revealed that 36% of respondents were located in India, 32% were in the United States, 6% were in Canada, and the remaining 26% were distributed among 13 other countries, with anywhere from 1-4% of responses from any single country. Ninety percent of respondents reported their gender and age, with 54% being male, and with ages ranging from 18 to 59 (M=29.4, SD=9.55).

5.5 Results

We conducted two-way between-groups ANOVAs—one for each of the nine indices—to explore the impact of the type of motion and team member role on observers' interpretations of the social dynamics of our design scenario. We used a power transformation with λ=0.8 in order to stabilize variance across conditions. We found a statistically significant main effect of combined proxy-in-proxy movement, as compared to the on-screen only condition, for several measures at p<.05 (see Table 3).

Table 3: Significant differences in perceptions of remote and on-site participants due to the type of proxy motion.

Variable	Condition	*M*	*SD*	F-Ratio	η_p^2
Remote participant friendliness	On-screen	4.53	1.03	F(1,99)=5.46	.05
	On-screen/In-space	**5.05**	1.24		
Remote participant dominance	On-screen	**3.97**	1.14	F(1,99)=4.00	.04
	On-screen/In-space	3.48	1.42		
Remote participant involvement	On-screen	4.98	1.28	F(1,99)=4.67	.05
	On-screen/In-space	**5.50**	1.37		
On-site participant equality	On-screen	4.49	1.24	F(1,99)=3.93	.04
	On-screen/In-space	**4.95**	1.50		

Regardless of the remote participant's role as leader or follower, when the proxy moved in concert with his expressions and gestures, he was perceived to be more friendly, less dominant and more involved. Similarly, regardless of the on-site design collaborator's role as leader or follower, when the proxy moved in concert with the remote participant's behaviors, she was rated as being more equal in stature relative to him.

Table 4: Significant differences in perceptions of remote and on-site participants due to team role.

Variable	Condition	*M*	*SD*	F-Ratio	η_p^2
Remote participant composure	Remote led	**5.32**	1.33	F(1,99)=5.71	.05
	On-site led	4.72	1.33		
Remote participant involvement	Remote led	**5.49**	1.33	F(1,99)=4.22	.04
	On-site led	5.00	1.32		
On-site participant composure	Remote led	**5.29**	1.30	F(1,99)=5.19	.05
	On-site led	4.71	1.25		
On-site participant dominance	Remote led	3.12	1.32	F(1,99)=6.38	.06
	On-site led	**3.78**	1.34		
On-site participant equality	Remote led	**4.98**	1.46	F(1,99)=4.57	.04
	On-site led	4.47	1.27		

We also found a statistically significant main effect of the role assumed by the remote teammate at p<.05 (see Table 4). When

the remote participant assumed a leadership role, regardless of the proxy's motion condition, he was perceived to be more composed and involved, and his on-site teammate was rated as being more composed and equal in stature. When the remote participant assumed a follower role, regardless of the proxy's motion condition, the on-site teammate was perceived as being more dominant.

5.6 Discussion

The addition of physical proxy motion favorably influenced three of the four expected relational measures, including similarity/ depth, dominance and (notably) involvement, providing good support for H1. While effect sizes were small—likely due to the subtle distinctions between behaviors that were portrayed across conditions—the remote participant's being seen as more involved when he was able to physically gesture makes sense: during face-to-face interactions, when someone is animated and moves about, he also tends to feel, and be perceived as, more involved. Similarly, when the remote participant led the interaction, he was perceived as having the characteristics that should accompany that role, particularly greater involvement. As with physical gesture, this is reasonable, since one typically feels and appears more involved when leading a discussion than when following.

We found mixed results regarding H2, the expected influence of the type of proxy motion on the strength of the remote participant's perceived role. This would suggest a magnifying effect on dominance and equality based on whether the proxy was in motion or not. The remote participant *was* perceived to be more dominant when his proxy was capable of movement, but he was not perceived differently regarding equality. Proxy motion *did* influence perceptions of the on-site teammate as being more equal in stature, which provides a further degree of support.

Perhaps most interesting is the influence that proxy motion and participant role had on perceptions of the on-site teammate. That is, when the *remote* participant either displayed proxy motion or led the discussion, the *on-site* teammate was viewed as being more equal in stature. One interpretation of this finding is that participating in a conference through standard, static video portal creates the perception of inequality between colleagues. This inequality may cause remote participants to appear higher, or perhaps lower, in stature than local teammates, depending on the situation. For example, when Eric led the discussion, the static platform may have caused him to appear dictatorial, making requests and issuing instructions from another place, behind his screen. After seeing Eric lead the discussion, one participant commented, "I thought Eric sounded a little rude; it seemed like he tried to hurry Becky at one point, which affected my attitude and answers." Re-introducing physical motion for the remote teammate may ameliorate that perception, and put the colleagues back on more even terms relative to one another.

In their written comments, most participants had a distinct point-of-view in their interpretations of proxy motion and collaborators' roles ("I simply found Becky 'professional' and Eric's distractedness 'unprofessional.'"), but some had difficulty discerning emotions, particularly during the stationary proxy conditions. For example, "Some of the emotions are quite subtle." and "I imagine it's hard to convey a feeling of warmth and sincere interest when communicating in this fashion." Such comments support the Study 1 finding that thoughtful, internal states can be difficult to interpret.

6. DESIGN IMPLICATIONS

In these studies, we found that combined on-screen and in-space action work best for particular categories of gesture ("focus of attention" and "thoughtful, internal states"), but that on-screen only action works best for others ("intentionally communicative"). This presents a challenge for the design of control interfaces, because it is most unnatural to feel and act out behaviors in the moment, and at the same time, control a proxy's gestures that represent those behaviors. A further complication is that gestures that tend to be most unanticipated, such as *laughter* or *surprise*, are those that would most benefit from proxy motion. Users should not have to say, press a "surprise" button when they are (or sense that they are about to be) surprised.

One approach to address this dilemma is to track the remote worker's movements and only pass through to the proxy those that cross a motion threshold that indicates say, a change in focus of attention. For example, a remote participant's head turns to either side would be ignored for a brief period—the *dwell* time—after which they most likely represent shifts in attention, and would be mirrored on the proxy (although this will introduce slight delay in the response). In addition, we can extend motion tracking to include face tracking, and use both movement and expression data to determine the user's category of gesture, then actuate the proxy or not, as appropriate to that category. For example, a quick movement upward and back from the torso is more likely to represent surprise when it is accompanied by raised eyebrows, opened eyes and a rounded mouth. Another approach would be to generate the proxy's behaviors based on user-indicated or detected situational context, as one would for autonomous agents [27].

7. LIMITATIONS AND FUTURE WORK

Evaluating video prototypes in an online study differs from first-hand experience with tangible devices in physical proximity. For example, Adalgeirsson and Breazeal evaluated the *experiential* concerns of interacting with proxies and in-the-moment responses to working with them. To that end, their use of physical, in-person environments provided the most appropriate context. Our studies address the *cognitive* concerns of how people perceive and interpret differences between particular behaviors and roles. Video prototypes, with their consistent reproduction of subtle variations in expression, therefore provide an appropriate context for evaluation. For just this reason, photographs and videos of faces and gestures have been used to evaluate the perception of emotions [9] for decades.

Neither of our studies exhaustively examined the effects of proxy behavior across all possible gestures or scenarios. Follow-up studies could compare whether larger or smaller motions, or alternative facial expressions and gestures, would be more or less effective than those we used here. Or they could look at how people from different cultures interpret the same motions or expressions. Our findings may be influenced by the actors that we chose to appear in video scenes. Alternative characters from different cultures, or with different levels of expressiveness, may affect how observers respond to gestures and team roles, and hence produce different results. Likewise, participants with more or less experience with mediated communication may have different expectations, and therefore, differ in their abilities to compare alternative approaches. Our use of an iMac G4 as a component of the platform may affect generalizability, as it had a familiar appearance, limited range of motion, and was mechanically difficult to actuate. On the other hand, other embodied proxy designs will impose their own motion constraints and influence perceptions in their own ways. In the future, we hope to work with other proxies to test how perceptions of on-screen and in-space gestures vary across configurations. We will also examine the

influence of longer exposure time, raising the issue of whether familiarity and acclimation lead to better understanding.

Finally, we considered inconsistency to be a lack of complementary cues: that is, an on-screen *nod "yes"* either supported by a physical nod or not. An alternative approach would actively contradict the content presented between channels, so an on-screen *nod "yes"* might be paired with an in-space *shake "no."*

8. CONCLUSION

Study 1 found that consistency between physical and on-screen action improved understanding of the messages that remote participants communicated. Study 2 extended this finding to a dynamic social context, and found that the addition of proxy motion also improved measures of perceived collaboration: not just for remotely connected participants, but for their local, physically co-present colleagues as well.

Clearly, the lack of gesture by the remote worker or embodied platform is not interpreted neutrally. Some degree of motion can significantly improve the clarity of communication and the confidence in understanding it, as well as improve perceptions of the friendliness and involvement of remote and on-site colleagues. While the quality of communication through an embodied proxy may never be quite the same as an in-person interaction, our two studies indicate that providing consistency between physical and on-screen cues can enhance the social interactions characteristic of distributed work.

9. ACKNOWLEDGMENTS

The authors thank Eric Kent, Rebecca Currano and Samson Phan, who were core members of the research team, and the HPI–Stanford Design Thinking Research Program for funding support.

10. REFERENCES

[1] Adalgeirsson, S. and Breazeal, C. MeBot: A robotic platform for socially embodied telepresence. *Proc. HRI 2010*, ACM Press, 15-22.

[2] Argyle, M. *Bodily Communication.* London: Methuen, 1988.

[3] Breazeal, C., Wang, A. and Picard, R. Experiments with a robotic computer: Body, affect and cognition interactions. *Proc. CHI 2007*, ACM Press, 153-160.

[4] Cassell, J. and Thorisson, K. The power of a nod and a glass: Envelope vs. emotional feedback in animated conversational agents. *Jnl. Applied A.I.* 13, 4 (2000), 519-538.

[5] Clark, H. Pointing and placing. In S. Kita (Ed.) *Pointing: Where Language, Culture and Cognition Meet*. Hillsdale NJ: Erlbaum (2003), 243-268.

[6] Dillard, J., Solomon, D. and Palmer, M. Structuring the concept of relational communication. *Communication Monographs* 66 (1999), 49-65.

[7] Dourish, P. and Bly, S. Portholes: Supporting awareness in a distributed work group. *Proc. CHI 1992*, ACM Press, 541-547.

[8] Ekman, P. and Friesen, W. Detecting deception from the body or face. *Jnl. Pers. and Soc. Psych.* 29 (1974), 288-298.

[9] Ekman, P. and Friesen, W. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Palo Alto CA: Consulting Psychologists Press, 1974.

[10] Fiske, S. and Taylor S. *Social Cognition*. New York NY: McGraw-Hill, 1991.

[11] Feng, D., Besana, S. and Zajac, R. Acquiring high quality non-expert knowledge from on-demand workforce. *Proc. People's Web 2009*, ACL and AFNLP, 51-56.

[12] Groom, V., Nass, C., Chen, T., Nielsen, A., Scarborough, J. and Robles, R. Evaluating the effects of behavioral realism in embodied agents. *Int. Jnl. Human-Computer St*. 67, 10 (2009), 842-849.

[13] Hall, E. *The Silent Language*. New York NY: Doubleday, 1959.

[14] Isbister, K. and Nass, C. Consistency of personality in interactive characters: Verbal cues, non-verbal cues and user characteristics. *Int. Jnl. Human-Computer St.* 53 (2000), 251-267.

[15] Kidd, C. *Sociable Robots: The Role of Presence and Task in Human-Robot Interaction*. MIT Thesis, 2003.

[16] Kittur, A., Chi, E. and Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proc. CHI 2008*, ACM Press, 453-456.

[17] Kraut, R. Verbal and nonverbal cues in the perception of lying. *Jnl. Pers. and Soc. Psych.* 36 (1978), 380-391.

[18] Kraut, R., Fish, R., Root, R. and Chalfonte, B. Informal communication in organizations: Form, function, and technology. In R. Baecker (Ed.) *Readings in Groupware and CSCW: Assisting Human-Human Collaboration*. San Francisco CA: Morgan Kaufman (1990), 130-144.

[19] Lee, M. and Takayama, L. Now, I have a body: Uses and social norms for mobile remote presence in the workplace. *Proc. CHI 2011*, ACM Press.

[20] Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H. and Hagita, N. Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. *Proc. HRI 2009*, ACM Press, 69-76.

[21] Nass, C. and Brave, S. *Wired for Speech*. Cambridge, MA: MIT Press, 2005.

[22] Nguyen, D. and Canny, J. Multiview: Improving trust in group video conferencing through spatial faithfulness. *Proc. CHI 2007*, ACM Press, 1465-1474.

[23] Nishio, S., Ishiguro, H., and Hagita, N. Geminoid: Teleoperated android of an existing person. In P. Filho (Ed.) *Humanoid Robots: New Developments*. Vienna: I-Tech (2007), 343-352.

[24] Paulos, E. and Canny, J. PRoP: Personal roving presence. *Proc. CHI 1998*, ACM Press, 296-303.

[25] Powers, A., Kiesler, S., Fussell, S. and Torrey, C. Comparing a computer agent with a humanoid robot. *Proc. HRI 2007*, ACM Press, 145-152.

[26] Venolia, G., Tang, J., Cervantes, R., Bly, S., Robertson, G., Lee, B. and Inkpen, K. Embodied social proxy: Mediating interpersonal connection in hub-and-satellite teams. *Proc. CHI 2010*, ACM Press, 1049-1058.

[27] Vilhjálmsson, H. and Cassell, J. BodyChat: Autonomous communicative behaviors in avatars. *Proc. AGENTS 1998*, ACM Press, 269-276.

[28] Woods, S., Walters, M., Koay, K. and Dautenhahn, K. Methodological issues in HRI: A comparison of live and video-based methods in robot to human approach direction trials. *Proc. ROMAN 2006*, IEEE, 51-58.

[29] Yankelovich, N., Simpson, N., Kaplan, J. and Provino, J. Porta-Person: Telepresence for the connected conference room. *Ext. Abstracts CHI 2007*, ACM Press, 2789-2794.