Coordination of Communication: Effects of Shared Visual Context on Collaborative Work

Susan R. Fussell, Robert E. Kraut, and Jane Siegel

Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA, 15213 U.S.A.
susan.fussell@cmu.edu

ABSTRACT

We outline some of the benefits of shared visual information for collaborative repair tasks and report on a study comparing collaborative performance on a manual task by workers and helpers who are located side-by-side or connected via audio-video or audio-only links. Results show that the dyads complete the task more quickly and accurately when helpers are co-located than when they are connected via an audio link. However, they didn't achieve similar efficiency gains when they communicated through an audio/video link. These results demonstrate the value of a shared visual work space, but raise questions about the adequacy of current video communication technology for implementing it.

Keywords

Computer-supported collaborative work, video mediated communication, conversational analysis, wearable computers, empirical studies

INTRODUCTION

Technologies that provide visual information to people collaborating at a distance have been available for decades. Yet, to date there has been no consensus on the effects of video on the quality of an interaction or the success of a task. In this paper we focus on the role of video in remote collaborative repair of physical objects.

Early research on the effects of video suggested that while adding an audio channel to any other medium substantially improves communication, adding video to the audio adds little or no further improvement [see 20 for a review]. In contrast, more recent research shows that in at least some cases, having a shared visual environment improves communication [2, 9, 10, 17, 18].

These inconsistencies in results are not surprising given the diversity of research paradigms used to investigate video. Studies have used a variety of techniques to share visual environments—placing people together, connecting computer screens and connecting them by video conferencing—and among those using video, quality (e.g., size, resolution, or frame rate) have differed substantially. In addition, studies have provided different types of visual information (e.g., head shots of the other participants vs. views of the work environment) and they have used a wide variety of collaborative tasks (e.g., brainstorming, problem-solving, matching or assembling objects, etc.).

In this paper we aim to clarify the role of visual information in one type of computer-supported cooperative work—collaborative repair of complex devices. Research has documented the value of remote expert assistance when field workers are troubleshooting and repairing complex equipment [10, 13, 15]. In the past, audio connections have been the primary means of communication between workers and remote experts. Today, walkie-talkies and mobile phones are being supplemented by mobile maintenance systems using video, on-line manuals, and other sources of information [1, 16]. By considering the ways in which specific types of visual information influence collaborative work, we will argue, designers should be better able to develop systems that meet the needs of collaborative workers.

In the remainder of this paper we first delineate possible benefits of visual information for task-oriented conversations. Specifically, we consider the ways in which the presence of visual information facilitates *grounding,* or the development of mutual understanding between conversational participants. Next, we present a study that aims to test empirically the value of shared visual information in a complex collaborative task—bicycle repair—and to examine how properties of media affect conversation and task performance. We conclude with a discussion of design recommendations for video systems to support remote collaborative work.

Grounding in conversation

Interpersonal communication is demonstrably more efficient when people share greater amounts of *common ground*—mutual knowledge, beliefs, goals, attitudes, etc. [3, 4]. People may have common ground prior to an interaction if they are members of the same group or population. In addition, they construct and expand their common ground over the course of the interaction on the basis of *linguistic co-presence* (because they are privy to the same utterances) and/or *physical co-presence* (when they inhabit the same physical setting). The term

CSCW'00, December 2-6, 2000, Philadelphia, PA.

grounding refers to the interactive process by which communicators exchange evidence about what they do or do not understand over the course of a conversation, as they accrue common ground.

To successfully ground their utterances, communicators must perform a number of conversational subtasks, three of which we focus upon here: (a) they must identify what their partners are attending to, in order to determine whether an object is part of their joint focus of attention; (b) they must monitor their partner's level of comprehension, so that they may expand or clarify their utterances if necessary; and (c) they must strive for efficiency in message formulation by constructing their utterances in accordance with Gricean norms for informativeness, brevity, and the like [7].

Physical co-presence as a resource for grounding

Physical co-presence provides a number of more-or-less independent sources of visual information that can be used to in the process of grounding utterances. In Figure 1, we consider four of these sources— participants' heads and faces, participants' bodies and actions, task objects, and work context—in terms of their benefits for the three grounding subtasks mentioned above. Due to space limitations this table is necessarily sketchy and incomplete but it serves to highlight three key points regarding the role of shared visual space in conversational grounding: First, although views of others' faces provide general information about direction of attention, additional information on participants behaviors and/or the array of task objects is needed to understand what a person is actually focusing upon.

Second, although head movements and facial expressions (e.g., nods, frowns) can be used to monitor general levels of comprehension, the prior literature suggests that this information doesn't improve comprehension much [19]. In contrast, people's behaviors or changes in the task status of objects can provide more detailed information about what is understood or misunderstood. If a speaker says, "pick up the wrench," but the listener picks up a screwdriver, the speaker can see not only that the message was misunderstood but that the problem lies in the identification of the wrench. This finer-grained understanding of others' comprehension should enable speakers to more precisely tailor their messages to the needs of the listener.

Third, the ability to view others' heads and faces does little to improve conversational efficiency, by which we mean the ease and brevity of referring expressions and other utterances. Other sources of visual information are required for gestural references (e.g., pointing) and deictic utterances (e.g., "that one").

Creating virtual physical co-presence

Each of the sources of visual information shown in Figure 1 may be supported to greater or lesser degrees in video systems designed to provide "virtual" physical co-presence.

Traditional "talking heads" video conferencing systems provide head position and gaze information. Views of the task objects can be presented from head mounted display cameras, which show the scene as viewed from the participant [e.g., 10] or from stationary cameras focused upon the task [e.g., 11]. Stationary cameras at different distances and with different fields of view can be used to provide visual information on the wider task environment. Choices among these video configurations can be expected to impact conversational grounding and task performance [2,18].

Although it might be helpful for remote collaborators if a video system were to make all of these sources of visual information available, bandwidth limitations make such a system unfeasible. One approach to this problem suggested by Gaver et al. [6] is to provide multiple video feeds and allow participants to switch between them as they choose. Such an approach is problematic in that equipment requirements may be needlessly high. In addition, Gaver et al. found that the ability to switch between video feeds made it difficult for participants to identify what elements of the visual environment were shared.

Our approach is instead to try to identify the critical elements of visual space for collaborative physical tasks

	Type of Visual Information			
Grounding Subtasks	Participants' heads and faces	Participants' bodies and actions	Shared task objects	Shared work context
Establish joint focus of attention	Eye gaze and head position can be used to establish others' general area of attention	Body position and activities can be used to establish others' general area of attention	Constrain possible foci of attention	Constrain possible foci of attention; disambiguate off-task attention (e.g., disruptions)
Monitor comprehension	Facial expressions and nonverbal behaviors can be used to infer level of comprehension	Appropriateness of actions can be used to infer comprehension, clarify misunderstandings	Changes in state of objects can be used to infer comprehension, clarify misunderstandings	
Conversational efficiency		Gestures can be used to refer to task objects	Pronouns can be used to refer to visually shared task objects	Environment can help constrain domain of conversation

Figure 1. Benefits of four types of visual information for three grounding subtasks.

and to design video systems such that they support these critical elements. Our assumption is that the usefulness of a video system for remote collaborative work will depend on the extent to which the video configuration makes available to collaborators those visual cues to common ground that are important in side-by-side collaborative physical tasks. Given that the importance of different visual cues may depend on the nature of the task (e.g., brain-storming vs. object construction), we focus in the next section on the role of visual information in collaborative remote repair.

Shared Visual Space in Collaborative Remote Repair

Collaborative repair tasks of the type we are considering here are characterized by the presence of one or more work objects (e.g., piece of complex equipment, vehicles) that undergo changes in state as workers perform physical actions upon them. Such tasks are also frequently characterized by the presence of other task-related physical objects such as tools, replacement parts, and the like. Remote assistance consists in large part of helping workers diagnose a problem and in instructing them how to perform unfamiliar operations. In the process, experts must help the worker identify the correct tools and parts and monitor the worker's comprehension and task status to fine-tune repair advice to the current state of the worker and of the task.

Consideration of these dynamics and the functions for visual information outlined in Figure 1 suggests that collaborative remote repair could be improved if systems could be augmented with two specific types of shared visual information: First, views of the task object and supporting tools and parts should facilitate a remote helper's ability to infer the worker's focus of attention, to monitor the worker's comprehension through observed changes in the status of the object and adjust assistance accordingly, and to refer quickly and easily to task elements using short-hand expressions and pronouns such as "this one." Second, views of the worker's actions should likewise facilitate the monitoring of attention and comprehension and should enable workers to use pointing and other gestures to refer to task objects efficiently [5, 11].

Previous Research

Despite these hints that incorporating shared video will be valuable for collaborative, manual tasks performed at a distance, there is as yet no convincing research demonstrating its value. Many previous studies of the value of video telephony are not directly applicable to this setting because the video telephony systems used a "talking heads" model, in which the cameras broadcast pictures of the people in conversation rather than the task they are working on.

There is some indirect support for the importance of shared views of the task object. Gaver et al. [6] found that when collaborators were working on a shared object, they spent most of their time looking at the video feed of that object rather than at each other's faces or the wider context. Nardi et al. [13] found that nurses monitored video feeds of surgeons' operating procedures to anticipate what instruments and supplies they would need next, reducing the need for explicit communication.

Kuzuoka and colleagues [11, 12] compared instructional conversations (an expert teaching a novice how to use a complex piece of machinery) across face-to-face and a number of shared video conditions. Although statistical tests are not reported, it appears that it took dyads in the mediated condition longer to complete the task than those in the face-face condition. No audio-only comparison group was included, so it is not clear whether the addition of video benefited performance.

In a previous study [10] we attempted to build on previous research by comparing performance in a collaborative bicycle repair task in audio-video versus audio-only media conditions. The video system was configured such that a remote helper could view a worker's activities and task objects through the use of a video camera mounted on the worker's head. The system thus provided partial information about the worker's actions (what his/her hands were doing) and partial information about the objects in the task environment (those in his/her immediate visual field), but no information on worker's facial expressions or the wider task context. Our hypothesis was that this video system would capture enough of the essential elements of actual physical co-presence to improve performance over the audio-only condition.

Contrary to our expectations, however, pairs with shared visual context were neither faster nor more accurate than pairs who communicated via audio only. However, the video technology used in this research may not have had enough fidelity on numerous dimensions to provide a fair test of the proposition that shared visual context improves collaborative task performance. Moreover, in this study, a small group of experts participated multiple times, and over time they may have scripted their responses to the workers independently of what they saw the workers doing. Most importantly, the study did not contain a control condition in which participants worked side-by-side.

The Current Study

The current study builds on our previous work by using a combination of experimental manipulation and explor-atory conversational analysis to address two interrelated sets of issues: the effects of communications media on task performance and on the strategies collaborators use to ground their utterances during repair dialogues.

We improve on our previous research design by comparing audio-video and audio-only collaboration to side-by-side collaboration, in which workers and helpers are physically co-present. In addition, we compare collaborations in which helpers are experts in bicycle repair to those in which helpers are, like the workers, novices. Finally, we use a within-group experimental design, in which each pair conducts tasks under all communication conditions, to control for the effects of

individual differences in skill and conversational style. Previous studies using between-subjects designs have found large differences in communicative style between pairs of communicators that might mask media effects.

Hypotheses

We use quantitative and qualitative data analysis to investigate three sets of hypotheses:

Task performance. We predict that performance will be best in the side-by-side condition, because the quality of the shared visual context is maximized, and poorest in the audio-only condition, due to the lack of shared visual context. Performance in the video condition should be intermediate, because the video technology supports some but not all of the benefits of actual physical co-presence.

Conversational grounding. Ease of conversational grounding, as indicated by message length, number of conversational turns, and use of deictic expressions, should be easiest in the side-by-side condition, and hardest in the audio condition. The extent to which performance in the video condition approaches that of the side-by-side condition is predicted to be mediated by the extent to which collaborators are able to use the video technology to facilitate grounding.

Helper expertise. We anticipate that the effects of media condition on task performance and conversational grounding will be mediated by the expertise of the helper. Based on previous research on grounding in referential communication tasks [e.g., 8], we hypothesize that shared visual context will be more important when the collaborative task is new for both parties than when helpers already possess an extensive vocabulary for discussing bicycles.

Study Design

Unskilled workers performed three repair tasks on a ten speed bicycle with the assistance of either an expert or novice helper. Pairs performed one task in each of three media conditions: (a) *side-by-side*: worker and helper worked in the same room, (b) *audio-video:* workers were connected by full-duplex audio plus video link to remote helpers, such that the video feed showed the worker's local activities; and (c) *audio-only*: workers were connected to remote helpers by full-duplex audio only. The experimental design was an incomplete factorial, in which participants were randomly assigned to task/treatment orders.

Participants

Workers consisted of 25 Carnegie Mellon University undergraduate and graduate students (68% male). They received $10 for participation and competed for a $20 bonus for the one with the fastest completion time and best task performance. A total of 12 helpers provided advice and guidance to subjects during their experimental sessions. Three were bicycle repair experts with professional experience; the other nine were novices who had limited prior bicycle repair expertise. The novice helpers had participated in the study as a worker and were also shown a tutorial videotape showing correct procedures. Helpers were paid $10 per session for their participation.

***Figure 2*:** Worker wearing collaborative system.

Apparatus

Each worker donned the apparatus shown in Figure 2, a head-worn mount where we attached various display and audio/video telecommunications devices. The devices included a sports caster style Radio Shack 49 MHz microphone, headphones, and a tiny Virtual Vision VGA (640x480 pixel resolution) monitor mounted in front of the right eye, with optics that placed the image directly in front of the eye. Workers viewed the shared on-line repair manual on this display, navigating with a remote control mouse.

Workers also wore a small CCD camera mounted on the head mount just above their left eye. In the video condition, both the worker and helper could see the output from the camera on their screens and output from a camera focused on the face and upper torso of the remote helper, using Intel's Proshare video conferencing technology[1]. The worker's camera saw approximately what the worker was pointing his or her head at. A view from the video condition is shown in Figure 3.

Figure 3. Display for the video conditions

Shared repair manual. An on-line bicycle repair manual with brief instructions and illustrations was created by

[1] Because Proshare introduces an audio delay, we by-passed Proshare for the audio.

subdividing each of the three main tasks into eight to ten component subtasks. One subtask was explained on each manual page through text and diagrams. In the video and side-by-side conditions, workers and helpers had the same view on their displays including the repair manual. Both worker and helper could control the cursor and flip pages. In the audio-only condition the manuals were not yolked.

Procedure

Participants reported to a room to complete consent forms and pretests. Workers were then taken to the experimental lab to put on the head-worn display shown in Figure 2 and a fanny pack containing components and controls. The head-worn display was fitted on the participant's head and adjusted by the experimenter so that it was comfortable and so that the camera tracked the worker's gaze. Participants were given an eye test to ensure they could read the text on the head-mounted display, instructed on how to navigate through the on-line manual, and then given a practice task.

The experiment was run with one experimenter in the same room as the worker, behind a computer used for real-time coding of communication behavior. The helper and worker could communicate at will, but the helper had the following rules to follow: (1) answer any question asked, (2) try to give the best answer, (3) if the worker was quiet for one minute, ask if he or she was doing all right, and (4) offer help or advice if the worker was doing something incorrectly.

Following completion of each task, participants were asked a set of questions by the experimenter about their experience during that task. At the completion of the three tasks, participants completed a test of bicycle repair knowledge and a questionnaire about their experience and were debriefed and dismissed.

Measures

Four sets of dependent measures were collected: survey data, performance measures, real-time observations of the interaction, and audio/video logs.

Survey data. Surveys were used to collect background information on participants' basic demographic data (e.g., age, gender) and prior computer, bicycle and related abilities. Participants also completed a ten-item bicycle repair knowledge questionnaire and a post-session questionnaire in which they rated their experiences using the mobile system (e.g., visibility of the workspace, ease of communication).

Performance measures. Measures of task performance include number of tasks completed, task completion time, and repair quality. To assess repair quality, both experimenter and the session helper rated the worker's repair against a checklist, assessing such details as whether the saddle was level to the ground and whether the brake anchor was set correctly.

Real-time coding. Two trained observers rated work quality and helper and worker communication in real-time using a 5-point scale ranging from 1 (poor) to 5 (well). Ratings were made at the subtask level (as described under "shared repair manual" above).

Video and audio recordings. Video and audio recordings of the sessions were the basis for verbatim transcripts and more detailed, post-experimental coding of the communication. For this coding, a selection of 3 subtasks from each of the three main tasks was chosen. One coder reviewed all the video recordings from the video-mediated condition and noted all events in the video that pertained to usage of shared visual space (e.g.,. pointing; orienting the head camera to bring an object into shared view).

Conversational Coding

To examine the relationship between media conditions and task dialogues, we developed a coding system that captured the primary purpose of each utterance. Each utterance was classified as either a question, answer, or statement in one of the following content categories:

Procedural: Instructions furthering task completion (e.g., "You might want to tighten the bolts just a little bit more.").

Task Status: State of the task or objects within the task (e.g., "The brake pads are on pretty tight;" "The wheel is in the fork.").

*Referential***:** Utterances pertaining to the identification or location of task objects. (e.g., "The straddle cable is the thing in this diagram;" "What's an anchor plate?")

*Internal State***:** Intentions, knowledge, emotions, etc. (e.g., "I don't understand what you're talking about;" "Do you know how a quick release lever works?").

*Acknowledgements***:** Feedback that message is heard/understood (e.g., "ok," "uh huh").

Other: Nontask and uncodable communication.

Two independent coders classified each utterance; agreement was better than 90% and disagreements were resolved through discussion.

RESULTS

We present the results in three parts: First we examine the effects of communication media on task performance; then, we examine the results of the real-time coding of work and communication quality, and finally we examine the relationship between communications media and discourse characteristics.

Task Performance

To see whether visual information aided a helper-worker pair in repairing the bicycle, we compared the two communications that used visual information with the audio-only condition, in a repeated measures analysis of variance (ANOVA) that included the expertise of the helper as a between-pairs factor. Because the tasks differed in difficulty, scores were standardized by task prior to this and all other analyses reported below.

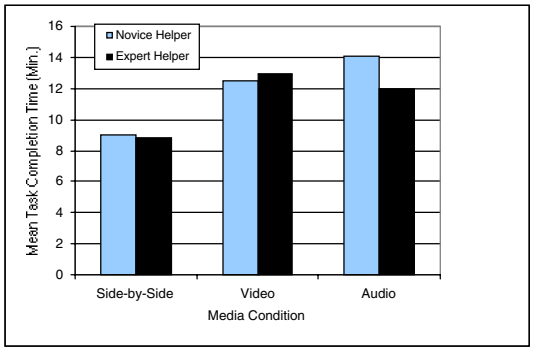

Figure 4. Completion time by media condition

As shown in Figure 4, completion times differed significantly across media conditions ($F[2,46] = 14.20$, $p < .001$). Pairs in the side-by-side condition completed tasks about 4 minutes faster than pairs in the audio-only and video conditions (*p*s < .001), which is approximately a 25% reduction in work time. Surprisingly, neither the expertise of the helper nor the expertise by communication medium interaction approached significance.

Work and Communication Quality

Real-time observers' ratings of work quality, helper communication quality, and worker communication quality varied across communications media (see Figure 5). Ratings were examined in a condition by helper expertise repeated measures ANOVAs. Rated work quality was slightly higher in the side-by-side condition than the mediated conditions, which did not differ from one another ($F[2,46] = 2.74$, $p = .08$). There was no significant effect of helper expertise and no significant interactions.

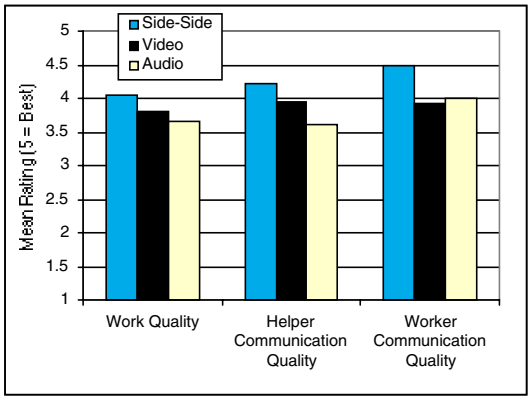

Figure 5. Mean rated work and communication quality by media condition.

Helper and worker communication quality ratings also differed across media conditions (helper: $F[2,46] = 4.21$, $p<.05$; worker: $F[2,46] = 9.17$. $p<.001$). Post-hoc tests indicate helper communication was rated significantly better in the side-by-side than in the audio condition, and worker communication quality was rated significantly better in the side-by-side condition than in the video or audio conditions (ps < .001). There were no main effects or interactions for helper expertise.

Ratings of work quality were correlated with completion time ($r = -.34$, $p <.01$). Work quality was also correlated with rated helper and worker communication quality (rs =.46 and .39, respectively, $p < .001$). Surprisingly, helper and worker quality ratings were virtually uncorrelated ($r = -.05$). In the next section we examine in more detail the aspects of communication that may be influencing work quality.

Conversational Analysis

First, we summarize the properties of messages as a function of media condition, then, we examine in more detail sequential relationships between utterance types as a function of condition, quality of work, and, for the video-mediated conditions, characteristics of the video feed.

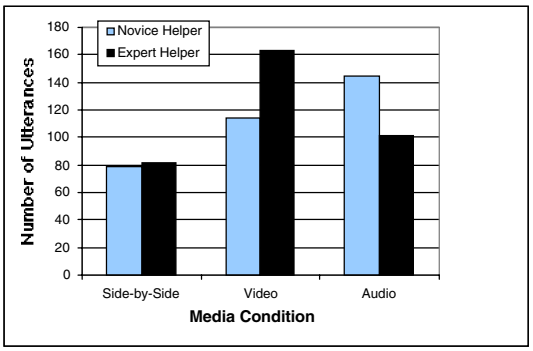

Figure 6. Mean number of utterances per task by media condition.

Message characteristics. As shown in Figure 6, dialogues were significantly more efficient in the side-by-side condition than in the mediated conditions, which did not differ from one another ($F[2,46] = 6.45$, $p < .005$). There was no main effect of helper expertise but a significant expertise by media condition interaction ($F[2,46] = 3.79$, $p < .05$). Post-hoc tests indicated that this was due to longer conversations between expert helpers and workers in the video condition.

Figure 7 shows the mean percentage of message units of each type per pair and task. Here, we have collapsed over statements, questions, and answers but the pattern is very similar when the data is further broken down. As can be seen, acknowledgements, descriptions of task status, and procedural instructions comprised the majority of utterances.

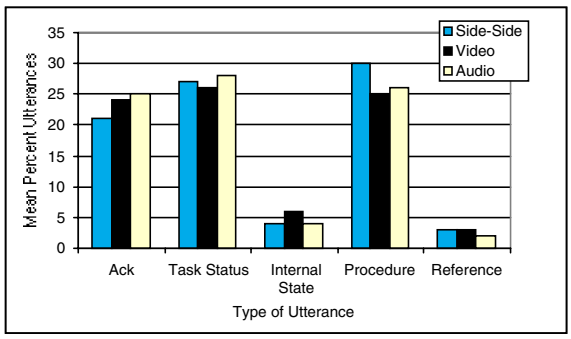

Figure 7. Mean percentage of utterances by content type and media condition.

Differences were examined in media condition by helper expertise ANOVAs of the same form used in the previous analyses. Pairs used significantly fewer acknowledgements when performing the task side-by-side ($F[2, 46] = 5.23$, $p < .01$). Pairs' references to internal states also differed significantly across media conditions ($F[2, 46] = 3.73$, $p < .05$). Post-hoc tests indicated that references to internal states were more frequent in the video condition than either side-by-side or audio. There was a trend for differences in percentages of procedural statements ($p = .11$), with procedural statements more frequent in the side-by-side than the other two conditions ($ps < .05$). There were no significant differences between conditions in percentages of task status utterances or references to task objects ($ps < .20$).

Qualitative analysis of dialogues.

To better understand the role of shared visual space in collaborative maintenance dialogues, we looked more closely at utterances in two of the coding categories described earlier: references to task objects, the brevity of which can be considered a measure of conversational efficiency, and messages about participants' internal states, which can be considered one form of attention and comprehension monitoring.

Reference. References to task objects comprised a small but critical proportion of overall messages in each dialogue—objects had to be identified before instructions for working with them could be given. Qualitative examination of the conversational exchanges through which participants established the identity of objects suggests that while the number of such sequences might have been fairly constant across conditions, the form of the referring expressions differed as a function of the presence of shared visual information. A representative sample of dialogues focused on identifying the bicycle's derailleur in each condition is shown in Figure 8. These dialogues illustrate several points:

First, in the side-by-side condition, in which participants' behaviors and task objects are visually shared, both helper and worker can refer quickly and easily to these objects using gestures and deictic expressions (e.g., "this one"). In the audio condition, lengthy descriptive sequences were typically required Mean durations of messages referring to task objects by media condition are shown in Figure 9.

Second, in the video condition, task objects were often not visually shared until the worker explicitly maneuvered the camera to bring them into the helper's field of view. Once this joint focus of attention was established, workers in the video condition, like those in the side-by-side condition, could use deictic expressions to refer to the objects.

Third, in the video condition, although objects were visually shared, helpers' physical behaviors were not. Hence, helpers were unable to use gestures to refer to task objects within a shared visual space, and occasionally expressed frustration with this situation (e.g., "If I could point to it, it's right there").

Side-by-Side Condition	Video Condition	Audio Condition
W: But what exactly is the derailleur?, the derailleur, whatever. H: Is this thing. W: Ok.	W: I'm not exactly sure what is a front whatever derailleur. H: Derailleur. It will be hanging off probably to the left side of the bicycle. It's ah W: OK H: Yeah, yeah W: That? [shows part with camera] H: That's it, right there.	W: Well what's the derailleur then? H: The derailleur is the piece with the other half of the clamp on it. W: The piece with the other half of the clamp on it? I'm confused …. H: Oh I bet the derailleur is hanging off the bike somewhere W: ok.
H: The derailleur is actually hanging down on this side W: Uh huh, over here. H: Right there.	H: What are you looking for? The derailleur itself? W: Yeah H: It's connected to the bike frame. It's already there. .. … H: Do you see it hanging? W: This? [shows part with camera] H: Yeah, that's the derailleur.	H: The derailleur itself is handing down by its cable. W: Oh ok. H: Off the left hand side of the bike. W: Yeah ok. I see it now.
H: And this is the front derailleur W: Ok.	W: What's derailleur? H: Derailleur is just a little mechanical thing that changes the ah chain from the small ring in the front to the large ring in the front. W: Ok it's just this one, is that right? [shows part with camera] H: Uh yeah.	H: The derailleur has I guess there is gonna be—there should be I think two bolts and a clamp that looks sort of like an elongated "c". W: Yeah, on the table. H: and then the derailleur also has a clamp that looks sort of like a "c".

Figure 8. Representative dialogues identifying the derailleur across media conditions (W = worker, H = helper).

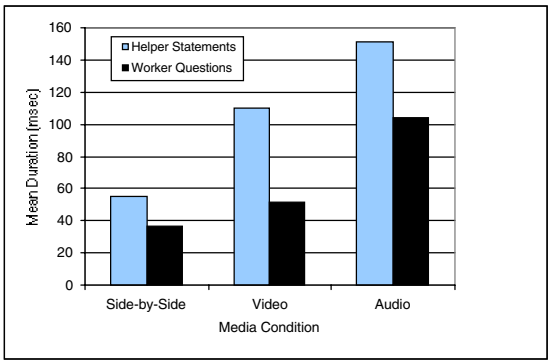

Figure 9. Mean duration of references to task object by media condition and participant role.

Figure 10 shows the percent of references to task objects containing the deictic terms *this*, *these*, or *those* across media conditions. (the term *that* had to be excluded from the analysis due to its multiple discourse functions in addition to deixis). These results suggest that while for workers in the video condition there is a sense of a shared visual space in which deictic references are mutually meaningful, this is not the case for the helpers. Consequently, a portion of the dialogues in the video condition were devoted to clarifying the meaning of deictic references, as in the following exchange:

W: Whoa! [Shows part with camera]

H: What?

W: Look at this.

H: Look at what?

W: You see how warped that is?

We return to this issue in the Discussion section.

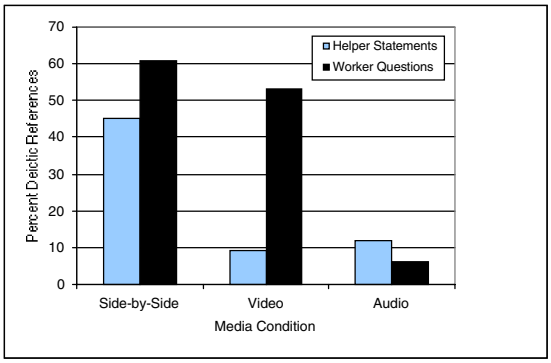

Figure 10. Percentage of deictic references to task objects by media condition and participant role.

Internal state. Our coding category of internal states included messages concerning participants' knowledge (e.g., "are you familiar with the quick release levers?" "I think I misunderstood something") in addition to those concerning helpers' or workers' fields of view. To distinguish between these types of utterances, we calculated the proportion of messages in each condition that contained the word *see.* In the side-by-side condition, in which participants were mutually aware of their shared visual space, only 17% of the utterances included the word *see.* In contrast, 37% of those in the audio condition and 46% of those in the video condition included *see*.

In the video condition, many uses of *see* occurred in the context of workers' queries about helpers' views (e.g., "Can you see the table?", "Can you see what I'm doing?", "See where I'm pointing up here?"). Helpers also volunteered information about their field of view (e.g., "I can't quite see the derailleur cage."). This use of *see* to clarify shared visual space was virtually nonexistent in the audio condition.

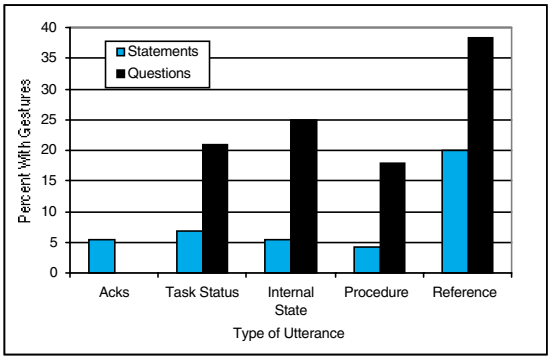

Figure 11. Percentage of worker messages in the video condition with associated gestures.

Using the shared visual space. To better understand how participants made use of the shared visual space created by our video technology, we examined the relationships between message types and behaviors that relied upon video (e.g., pointing to an object, moving the camera to focus on an object). Figure 11 shows the percentage of time messages in the video condition were accompanied by worker behaviors that relied upon the video feed. Video-related behaviors were more frequent during questions and acts of reference, suggesting indirectly that pairs made use of the video's potential to create a shared visual space.

DISCUSSION

In summary, this research shows that collaborative repair tasks can be performed more efficiently with a physically co-present helper than with a remote helper. It also shows that the visual cues made available by communications technologies can impact how collaborators ground their utterances. We consider the implications of these findings below.

Conversational Grounding

Conversation was more efficient and rated higher in quality in the side-by-side condition. Content analyses suggest that one reason this might be so is that straightforward procedural instructions comprise a higher proportion of utterances in the side-by-side condition. In the mediated conditions, not only are dialogues longer but their focus shifts slightly but significantly—more speaking

turns are devoted to acknowledging partners' messages and, in the video condition, to messages about internal state.

One reason we may not have found larger differences in dialogue structure across media conditions stems from a limitation of our conversational coding system—we used an utterance-based definition of a conversational turn. It appears, however, that collaborative repair dialogues do not always follow a conventional turn-taking structure; rather, workers can make behavioral responses to helpers' utterances, as in the following example:

H: No, down a little more.

W: [movement]

H: Down a little more.

W: [movement]

H: Right there.

We are recoding a subset of the data to include these nonverbal turns by the worker to further understand how communicators make use of shared visual space.

Contrary to our expectations, helper expertise had no effect on task performance, nor did it interact with media condition. This may be due to experts' use of technical terms which were unfamiliar to workers and thus required extra time to define and ground. In future studies we will include expert workers as well as expert helpers to examine the hypothesis that shared visual space will be less important when collaborators share specialized vocabularies.

Limitations to Video-Mediated Visual Space

The qualitative analysis of repair dialogues and the use of "pointing" gestures by workers in the video condition suggest that workers and helpers try to make use of shared visual information when it is available. Why, then, were video-mediated dialogues less efficient than side-by-side ones? Four considerations may help answer this question: First, workers' queries about video-linked helpers' fields of view suggest that participants had difficulty establishing what visual information was shared.

Second, the video system did not provide the full array of visual cues present in the side-by-side condition. For video-linked helpers, the extent of the repair scene visible at one time was limited to a view from workers' head-mounted displays. Thus, objects were sometimes outside the view of the camera (e.g., on a work table or the other side of the bicycle).

Third, in the video-mediated condition the worker's face was not visible. Side-by-side helpers may have glanced at workers' faces to monitor attention and comprehension and thus been better able to fine-tune their messages to the worker's current needs.

Finally, in the current video system, the worker's view of the helper was limited to his or her head plus upper torso. As a result, helpers could not use gesture to refer efficiently to task objects. Other research [11, 12] suggests that video-mediated communication will be more similar to side-by-side communication when remote helpers' gestures are made available to the worker through overlay on the video feed.

Implications for Video System Design

Our findings and the discussion above suggest four recommendations for the design of future video-based systems to support collaborative remote repair:

Provide workers with better feedback on what is in the helper's field of view, to clarify what is in the shared visual space.

Provide helpers with a wider field of view, thereby increasing the shared visual space.

Provide helpers with feedback on the worker's attentional focus.

Provide support for helper gestures within the shared visual space.

CONCLUSION

We have argued that shared visual space is essential for collaborative repair because it facilitates conversational grounding, that there are a number of different ways in which visual information can facilitate grounding, and that the suitability of specific video configurations for supporting remote collaboration will depend on the extent to which the configurations capture the essential elements of shared visual space. The system we tested in the current study goes only part of the way towards creating "virtual" physical co-presence but the guidelines we suggest should help future video system designers come closer to this goal.

ACKNOWLEDGEMENTS

This study was conducted with support from the National Science Foundation Grants #9022511 and #9980013. We thank Elise Nawrocki, Tom Pope, Leslie Setlock, and Mei Wang for their assistance.

REFERENCES

1. Bass, L., Kasabach, C., Martin, R., Siewiorek, D., Smailagic, A., & Stivoric, J. (1997). The design of a wearable computer, *CHI Proceedings*, Atlanta, Georgia, 139-146.
2. Clark, H. H. & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, R. M. Levine, & S. D. Teasley (Eds.). *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: APA
3. Clark, H. H. & Marshall, C. E. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10-63). Cambridge: Cambridge University Press.
4. Clark, H. & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*, 1-39
5. Daly-Jones, O., Monk, A. & Watts, L. (1998). Some advantages of video conferencing over high-quality audio conferencing: fluency and awareness of

attentional focus. *International Journal of Human-Computer Studies,* **49**, 21-58.

6. Gaver, W., Sellen, A., Heath, C., & Luff, P. One is not enough: Multiple views in a media space. *Interchi '93* (335-341). NY: ACM Press.
7. Grice, H. Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics* (Vol. 3, pp. 41-58). New York: Academic Press.
8. Isaacs, E., & Clark, H. H. (1987). References in conversation between experts and novices. *J. of Experimental Psychology: General*, **116**, 26-37.
9. Karsenty, L. (1999). Cooperative work and shared visual context: An empirical study of comprehension problems and in side-by-side and remote help dialogues. *Human-Computer Interaction,* **14**, 283-315.
10. Kraut, R. E., Miller, M. D., & Siegel, J. (1996) Collaboration in performance of physical tasks: Effects on outcomes and communication, *Proceedings of CSCW'96* (57-66). NY: ACM.
11. Kuzuoka, H. (1992). Spatial workspace collaboration: A Sharedview video support system for remote collaboration capability. *Proceedings of CHI'92* (533-540). NY: ACM.
12. Kuzuoka, H., Kosuge, T., & Tanaka, K.. (1994) GestureCam: A video communication system for sympathetic remote collaboration, *Proceedings of CSCW 94* (35-43). NY: ACM Press.
13. Nardi, B., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S. & Sclabassi, R. (1993). Turning away from talking heads: The use of video-as-data in neurosurgery. *Proceedings of Interchi '93* (327-334). NY: ACM Press.
14. Orr, J. (1989). Sharing knowledge, celebrating identity: War stories and community memory among service technicians. In D. S. Middleton and D. Edwards (Eds.), *Collective Remembering: Memory in Society*. London: Sage Publications.
15. Sachs, P. (1995). Transforming work: Collaboration, learning, and design. *Communications of the ACM*, **38**(9), 36-44.
16. Smailagic, A. & Siewiorek, D. (1996) Modalities of interaction with CMU wearable computers, *IEEE Personal Communications*, 3(1), 14-25.
17. Veinott, E., Olson, J. Olson, G. & Fu, X. (1999). Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. *Proceedings of CHI'99* (302-309). NY: ACM Press.
18. Whittaker, S. & Geelhoed, E. (1993). Shared workspaces: How do they work and when are they useful? *Int'l J. Man-Machine Studies*, **39**, 813-842.
19. Whittaker, S. & O'Conaill, B. (1997). The role of vision in face-to-face and mediated communication. In K. Finn, A. Sellen & S Wilbur (Eds.) *Video-mediated communication.* Mahwah, NJ: Erlbaum.
20. Williams, E. (1977). Experimental comparisons of face-to-face and mediated communication: A review. *Psychological Bulletin*, **84**, 963-976.