HUMAN-COMPUTER INTERACTION, 2003, Volume 18, pp. 13–49

Visual Information as a Conversational Resource in Collaborative Physical Tasks

Robert E. Kraut, Susan R. Fussell, and **Jane Siegel**
Carnegie Mellon University

ABSTRACT

In this article we consider the ways in which visual information is used as a conversational resource in the accomplishment of collaborative physical tasks. We focus on the role of visual information in maintaining task awareness and in achieving mutual understanding in conversation. We first describe the theoretical framework we use to analyze the role of visual information in physical collaboration. Then, we present two experiments that vary the amount and quality of the visual information available to participants during a collaborative bicycle repair task. We examine the effects of this visual information on performance and on conversational strategies. We conclude with a general discussion of how situational awareness and conversational grounding are achieved in collaborative repair and with some design considerations for systems to support remote collaborative repair.

Robert Kraut is a social psychologist with an interest in communications and the social impact of computing; he is a professor in the Human Computer Interaction Institute at Carnegie Mellon University. **Susan Fussell** is a social and cognitive psychologist with interests in face-to-face and computer-mediated communication; she is a system scientist in the Human Computer Interaction Institute at Carnegie Mellon University. **Jane Siegel** is a psychologist with interests in interpersonal communication and wearable computers; she is a senior system scientist in the Human Computer Interaction Institute at Carnegie Mellon University.

CONTENTS

1. INTRODUCTION

As the workforce becomes increasingly distributed across space and time and increasingly mobile, the need to collaborate with remote partners to ac-

complish collaborative tasks has increased substantially. The field of computer-supported cooperative work (CSCW) both invents and studies these human–machine systems and has made significant progress in describing, understanding, and improving them. Despite this progress, however, most systems to date (e.g., desktop video conferencing, electronic mail, audio teleconferencing) are designed to support group activities that can be performed without reference to the external spatial environment (e.g., decision making). Development of systems to support collaborative tasks involving physical objects has been much slower.

In this article, we consider the ways that visual information is used as a conversational resource in "collaborative physical tasks," tasks in which two or more individuals work together to perform actions on concrete objects in the three-dimensional world. Such tasks play an important role in many domains, including education, design, industry, and medicine. For example, an expert might guide a worker's performance of emergency repairs to an aircraft, a group of students might collaborate to build a science project, or a medical team might work together to save a patient's life.

Observational studies of physical collaboration suggest that people's speech and actions are intricately related to the position and dynamics of objects, other people, and ongoing activities in the environment (e.g., Flor, 1998; Ford, 1999; Goodwin, 1996; Kuzuoka & Shoji, 1994; Tang, 1991). Conversations during collaborative physical tasks typically focus on the identification of target objects, descriptions of actions to be performed on those targets, and confirmation that the actions have been performed successfully. During the course of the task, the objects themselves may undergo changes in state as people perform actions upon them (e.g., a piece of complex equipment may undergo repair) or as the result of outside forces (e.g., a patient might start hemorrhaging).

The performance of collaborative physical tasks requires substantial coordination among participants' actions and talk. As we discuss in detail later, in face-to-face settings much of this coordination is managed through the use of visual information. Visual information plays at least two interrelated roles. First, visual information helps people maintain up-to-date mental models or *situational awareness* of the state of the task and others' activities. This awareness can help them plan what to say or do next and to coordinate their utterances and actions with those of their partners. Second, visual information can help people communicate about the task, by aiding *conversational grounding*, or the development of mutual understanding between conversational participants.

In face-to-face settings, people use a variety of visual cues to achieve situational awareness and conversational grounding, including views of others' faces, bodies, and actions; views of the task objects; and views of the environment. This diversity of visual cues presents a challenge for designers of systems

to support the remote accomplishment of collaborative physical tasks. It is rarely feasible due to bandwidth and other limitations for a system to provide all sources of visual information. Our approach is instead to try to identify the critical elements of visual space for collaborative physical tasks and to design video systems that support these critical elements. Our assumption is that the usefulness of a video system for remote collaborative work depends on the extent to which the video configuration makes the same visual cues available to collaborators that they use when performing the task when co-located.

In the remainder of this introduction, we first present the theoretical framework guiding our empirical work on shared visual spaces. We focus on a single type of physical task—collaborative bicycle repair—and describe how visual information can be used to maintain situational awareness and ground conversations during this task. We then describe how features of technologies might make them more or less suitable for supporting remote collaboration on our bicycle repair task, and we outline our general hypotheses. In the second and third sections of the article, we describe two experiments that aim to test empirically the value of shared visual information in the bicycle repair task by examining how properties of media affect task performance and conversation. Some results from these studies have been reported elsewhere (Fussell, Kraut, & Siegel, 2000; Kraut, Miller, & Siegel, 1996). In this article, we focus on integrating and extending previous findings with respect to ways situational awareness and conversational grounding are achieved by using different media. We conclude with a general discussion that includes design recommendations for video and other systems to support distributed physical work.

1.1. Conversation and Coordination in Collaborative Repair

Collaborative physical tasks can vary along a number of dimensions, including number of participants, temporal dynamics, and the like. The task on which we focus here, a bicycle repair task, falls within a general class of "mentoring" collaborative physical tasks, in which one person directly manipulates objects with the guidance or one or more other people, frequently who have greater expertise about the task. In our bicycle repair task, one person, whom we call the "worker," uses tools to repair a bicycle. A second person, whom we call the "helper," provides guidance to the worker during the course of the repairs but does not actually manipulate the bike, tools, or parts. The relation between helper and task is thus similar to a teacher guiding a student's lab project, advice from a call-in help desk, or the like.

The collaborative bicycle repair task is one form of complex coordination problem (Clark, 1996; Malone & Crowston, 1994): For helpers to provide useful assistance, they must determine what help is needed, when to provide

the help, how to phrase their messages of assistance such that the worker understands them, and whether the message has been understood as intended. That is, assistance must be coordinated not only with the worker's utterances but also with his or her actions and the current state of the task.

Consider the following fragment from a conversation in which a helper is telling a novice worker how to attach a bicycle saddle to its seat post using clamps.

Helper: Now you want to fit the rails of the seat into that groove.
Worker: I see. How can it fit into?
Helper: You might want to unscrew those nuts a bit.
Worker: Oh—okay.
Helper: It will give you a little more room.

To have this dialogue, the helper needs to overcome several challenges. One challenge is for the helper to identify what the worker is attending to, to determine whether an object is part of the joint focus of attention. The helper's use of the definite article in "the rails" and "the seat" and deictic adjectives in "that groove" and "those nuts" depends on knowing the worker's focus of attention, to be assured that he was referring to the rails, seat, grooves, and nuts the helper was attending to. A second challenge is to make sure that the worker understands a prior utterance before continuing the conversation. In this example, the worker verbally indicated understanding with phrases like, "I see" or "Ok." The helper could also infer understanding because he could see that the worker had indeed started to loosen the nuts. Finally, the helper needs to comply with Gricean norms, for example, informativeness and brevity (Grice, 1975). In this case, he does so by using deictic references (e.g., "that groove") along with pointing. Helpers can meet these challenges through processes of situation awareness and conversational grounding.

Situation Awareness

For help to be effective, workers must receive it when they need it and when their preconditions for taking advantage of it have been met (e.g., when they are paying attention, when they are not overloaded with other information). To determine what help is needed and when to provide it, helpers must maintain an ongoing awareness of what their collaborators are doing, the status of the task, and the environment (cf. Orr, 1996). Endsley (1995) used the term *situational awareness* for people's mental models of complex, dynamic environments.

In collaborative physical tasks, people must maintain awareness both of the state of task objects and of one another's activities. In the bicycle repair task, helpers can use their awareness of the state of the bicycle—what repairs

have been made thus far, with what level of success—to determine what information to present next. Instructions for each new step in the repair process can be coordinated with the completion of the previous step. In addition, helpers can use their awareness of what the worker is currently doing—what actions he or she is performing, with what tools and parts and with what success—to determine if clarifications or expansions of the instructions are required. If, for example, the worker is using the wrong tool, the helper can interject a comment to correct this (e.g., "no, not that wrench, the larger wrench").

Conversational Grounding

In addition to timing assistance appropriately, helpers need to ensure that their messages are properly understood; that is, that they become part of the common ground between helper and worker. *Common ground* refers to mutual knowledge, beliefs, goals, attitudes, and the like shared by partners in a communication (Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986). In conversation, each participant's messages build on previously established common ground. New contributions are presented and then "grounded" through an acceptance phase. In this some cases, contributions may be grounded immediately by an acknowledgment ("uh huh," "okay"). In other cases, sequences of questions, repairs, clarifications, and the like may be required before grounding is established (Jefferson, 1972; Sacks, Schegloff, & Jefferson, 1974). The term *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 (Clark & Brennan, 1991).

Research has shown that communication is more efficient when people share greater amounts of common ground. Clark and Marshall identified three primary sources for common ground: First, people may have common ground prior to an interaction if they are members of the same group or population (e.g., Fussell & Krauss, 1992; Isaacs & Clark, 1987). For example, if both helpers and workers are familiar with bicycle tools, they can refer to them using their names (e.g., "take the Allen wrench"); when the worker does not share the helper's tool expertise, longer descriptive phrases may be necessary to identify the tool. Second, people can 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). Finally, people can share common ground due to *physical co-presence*—when they inhabit the same physical setting (Clark & Marshall, 1981). Physical co-presence provides multiple resources for awareness and conversational grounding (e.g., sights, smells, touch). In the next section we focus on one of the most important of these resources, visual information.

1.2. Visual Cues in Maintaining Situational Awareness and Grounding Conversations

When all parties to the interaction are co-present—located at the same place at the same time—they share a rich visual space. People can monitor one another's facial expressions, watch each other's actions, and jointly observe task objects and the environment. This shared visual space can facilitate task awareness and conversational grounding (e.g., Daly-Jones, Monk, & Watts, 1998). Helpers in our bicycle repair task can monitor workers' facial expressions, the worker's actions, and changes in the state of the bicycle; thus, they can formulate and time their instructions to ongoing changes in the workers' need for assistance. In the excerpt we gave previously, the helper used his observation that the worker was finished with a step to time his next instructions ("Now …"), and he used his view of the bicycle to determine that the nuts needed to be unscrewed. Similarly, because workers can view helpers' actions and hand movements, helpers can use pointing gestures and deictic expressions (e.g., "that one") to refer quickly and efficiently to task objects. In the preceding excerpt, the helper used a combination of pointing and a deictic expression, "those nuts," to refer effectively to the nuts in question.

Sources of Visual Information

Physical co-presence provides a number of more or less independent sources of visual information. These sources vary in terms of their importance for maintaining awareness and grounding conversation. A challenge, for both theoretical development and technology design, is to understand how people use specific types of visual evidence for specific collaborative purposes. The approach we take to this challenge is a decompositional one, in which we strive to specify the components of physical collaboration that rely on visual information, to identify the types of visual cues each of these components requires, and to understand how affordances or properties of specific technologies provide or fail to provide these visual cues (see Kraut, Fussell, Brennan & Siegel, in press, for a full discussion of this approach).

Our decompositional approach is illustrated in Figure 1. In this figure we consider four sources of visual information—participants' heads and faces, participants' bodies and actions, the focal task objects, and the work environment or context—in terms of their benefits for several aspects of situation awareness and conversational grounding: (a) monitoring task status, (b) monitoring people's actions, (c) identifying what one's partner is attending to, (d) communicating efficiently, and (e) monitoring one's partner's level of comprehension. The figure is intended to illustrate rather than define our approach; future research is needed to fully specify the rows and columns of this figure.

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

	Type of Visual Information			
Collaborative Process	**Participants' Heads and Faces**	**Participants' Bodies and Actions**	**Task Objects**	**Work Context**
Monitor task status	N/A	Inferences about intended changes to task objects can be made from actions.	Changes to task objects can be directly observed.	Activities and objects in the environment that may affect task status can be observed.
Monitor people's actions	Gaze direction can be used to infer intended actions.	Body position and actions can be directly observed.	Changes to task objects can be used to infer what others have done.	Traces of others' actions may be present in the environment.
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).
Create efficient messages	Gaze can be used as a pointing gesture.	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.
Monitor comprehension	Facial expressions and nonverbal behaviors can be used to infer level of comprehension.	Appropriateness of actions can be used to infer comprehension and clarify misunderstandings.	Appropriateness of actions can be used to infer comprehension and clarify misunderstandings.	Appropriateness of actions can be used to infer comprehension and clarify misunderstandings.

When two people are working side-by-side, they have all four sources of visual information easily available. To assess the other's focus of attention, they can monitor each other's facial expressions and body orientations vis-à-vis task objects. Facial expressions and visible actions vis-à-vis the task provide evidence of whether someone understands an instruction. Knowledge of the physical environment constrains what objects are likely to be talked about, making both production and understanding of reference easier. Because all participants are able to interact with the physical work, all can point and use gestures along with deictic expressions to refer efficiently to objects. If, however, the participants have to work together at a distance, they must communication through some type of telecommunications, which substantially limits the type of visual information that can be shared. In the next section we consider the effects of technology on shared visual space.

1.3. Effects of Communications Media on Shared Visual Space

Although it might be helpful for remote collaborators if a video system were to make all sources of visual information available, bandwidth limitations make such a system unfeasible. One approach to this problem, suggested by Gaver, Sellen, Heath, and Luff (1993), 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 impractically high. In addition, Gaver et al. found that the ability to switch between video feeds made it difficult for participants to identify which elements of the visual environment were shared.

An alternative approach is to determine the key visual information used in collaborative physical tasks and to design or implement technologies to provide this information. As Clark and Brennan (1991) have discussed, specific features or "affordances" of communications media can affect the ease and methods by which conversationalists maintain task awareness and achieve common ground. Clark and Brennan focused on decomposing features of different classes of communications technologies (e.g., telephone, e-mail, video conferencing). Here, we focus our discussion on the types of visual information that different systems make available.

Currently, the majority of video systems provide only a subset of the visual cues available when people are co-present. AT&T's earliest PicturePhone® had arrangements to show documents and other small objects (Noll, 1992), but such systems are the exception rather than the rule—most video conferencing systems train their camera on the people in a meeting and provide views only of facial expressions and, in some cases, upper body movements. These sorts of "talking heads" systems provide only the types of information listed in the second column of Figure 1. They provide almost no

support for situational awareness and limited support for conversational grounding (except in certain circumstances, for instance, when communicators are not speaking in their native language; e.g., Veinott, Olson, Olson, & Fu, 1999).

Other camera arrangements can be used to provide the three other types of visual information listed in Figure 1 (i.e., participants' bodies and actions, focal task objects, work context). For example, views of task objects can be presented from stationary cameras focused on the task. Stationary cameras at different distances and with different fields of view can be used to provide visual information on the wider task environment. Head-mounted cameras can show a detailed view of the scene as viewed by the person wearing the camera, including the objects in his or her field of view.

Choices among these video configurations can be expected to affect situational awareness and conversational grounding and, as a result, also affect task performance. In the extreme, for example, when one person is giving another instructions over the telephone, no shared visual information is available. In this case, participants have to rely on language to signal and monitor focus of attention and comprehension. As a result, they are likely to be far more explicit in their descriptions of the objects they are working on, the instructions they are giving, the changing state of the task, and their own level of understanding than if they were side-by-side (Beattie & Barnard, 1979). Because "talking heads" video systems provide few visual cues to task objects and work environment, these systems are unlikely to reduce the need for explicitness found in audio-only systems.

Video communication systems that provide a view of the work area are likely to be more useful in supporting situational assessment and conversational grounding. Recent research has shown that sharing a two-dimensional visual space improves instruction in computer-based tasks (Karsenty, 1999). Other research has suggested the value of workspace-oriented video systems for three-dimensional tasks. Gaver et al. (1993), for example, 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. (1993) 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 (Kuzuoka, 1992; Kuzuoka, Kosuge, & Tanaka, 1994) found that experts could teach novices how to use a complex piece of machinery via a number of shared video systems, although the instructional dialogues were longer than those in side-by-side settings. These studies suggest the potential importance of shared views of the workspace for conversations during collaborative physical tasks. The studies discussed here expand on this research by analyzing in much greater detail how task conversations are

shaped by the types of visual information that a communication medium makes available.

1.4. The Studies: Overview and General Hypotheses

In this pair of studies, we examine the value of a head-mounted video system that provides remote helpers with a view of what the worker is looking at and a portion of the surrounding environment. With respect to the framework we presented in Figure 1, our system provides no visual access to the workers' heads and faces, a view of the worker's hands and actions, and partial views of task objects and the work environment (when these are in the worker's field of view). This system is compared with an audio-only condition in which helpers cannot see the work area and, in Experiment 2, to a side-by-side condition in which helpers and workers are co-located. Two general sets of hypotheses are examined, one concerning task performance and one concerning communication (specific hypotheses are described in greater detail in the introduction to each experiment).

Task Performance

Our first hypothesis is that this video system will improve a pair's ability to perform the collaborative task over an audio-only system because the visual cues provided by the system will improve situational awareness and conversational grounding. At the same time, because the video system does not provide all the visual cues of physical co-presence, we do not expect performance of pairs using the video system to match that of pairs working side-by-side.

Communication

The anticipated effectiveness of the video versus side-by-side and audio-only conditions is hypothesized to stem from the ways people communicate about the task in the different conditions. More specifically, we hypothesize that helpers will use visual information to help them time their assistance, communicate this assistance effectively, and monitor comprehension. Workers, in turn, are hypothesized to be less explicit about what they are doing and when they need help, because they know the helper can see what they are doing.

In Experiment 1, we examine these hypotheses by using a between-subject comparison of audio-only and audio–video conditions. In Experiment 2, we build on the findings of our first study by using a within-subjects design comparing audio-only, audio–video, and side-by-side conditions.

2. EXPERIMENT 1

In Experiment 1, we compared people's performance on the bicycle repair tasks when working alone and when working with a helper who could guide them through the repair process. When they worked with the helper, they were connected or not with a video link between the worker and helper so that the helper could see what the worker was doing. We also varied the quality of audio connection between the worker and helper (either full- or half-duplex audio). The experimental design was an incomplete factorial, in which we compared, first, solo performance to collaborative performance and, second, three different technology configurations to support the collaborative pairs: (a) full-duplex audio alone, (b) full-duplex audio plus video, and (c) half-duplex audio plus video. Study participants were randomly assigned to a single treatment. Because this article is concerned with the role of visual information on conversation and performance, we concentrate here on the collaborative pairs communicating with full-duplex audio, with versus without a video capability. We examined three hypotheses (H):

H1: *Performance.* Performance in the video condition was expected to be better than that in the audio-only condition because pairs could use visual information to help coordinate their activities.

H2: *Timing and content of helper instructions.* We predicted that helpers would provide better instruction in the video condition, because the shared visual space would enable the helper to maintain awareness of the worker's behaviors and changes in task objects and use this awareness to time precisely when to give instructions and which instructions to give. The workers' activities can be monitored to figure out which instructions are most appropriate and to determine when clarifications are necessary.

H3: *Explicitness of worker descriptions.* When there is no shared visual space, we predict that workers will explicitly describe what they are doing and the status of the task, because the helper has no other way of knowing. In contrast, workers in the video condition, because they are aware that the helper can view their activities, are anticipated to be less explicit in describing their behaviors and task status.

2.1. Method

Apparatus

Each worker wore a head-worn mount where we attached various display and audio–video telecommunications devices (see Figure 2). The devices in-

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

cluded a sports caster–style Radio Shack 49 MHz microphone, headphones, and a tiny Virtual Vision VGA (640 × 480 pixel resolution) monitor mounted in front of the right eye, with optics that placed the image directly in front of the eye. 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 the bottom right corner of Figure 3.

An online bicycle repair manual with brief instructions and illustrations was created by subdividing each of the three main tasks into 8 to 10 component subtasks, such as attaching clamps to the bike saddle. 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 yoked. Workers viewed the shared online repair manual on their head-worn display, navigating with a remote control mouse.

Participants and Procedure

Participants consisted of 60 Carnegie Mellon University undergraduate students (69% male), who received a candy bar for their participation and competed for a $20 bonus for the fastest completion time and highest quality

1. Because Proshare introduces an audio delay, we bypassed Proshare for the audio.

Figure 3. **Display for the video conditions.**

task performance. Two bicycle repair experts also participated in the study. They were trained for the specific repairs they would advise about during this experiment. They were paid for their participation.

When workers arrived at the laboratory, they 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, were instructed on how to navigate through the online manual, and then were 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 to follow these rules: (a) answer any question asked, (b) try to give the best answer, (c) if the worker was quiet for 1 min, ask if he or she was doing all right, and (d) offer help or advice if the worker was doing something incorrectly.

Measures

Three sets of dependent measures were collected: performance measures, real-time observations of the interaction, and audio–video logs.

Performance measures. Measures of task performance included 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 by using a 5-point scale ranging from 1 (*poor*) to 5 (*well*). They rated each subtask as described previously in describing the shared repair manual.

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

Conversational Coding

To examine how media changed coordination strategies, a single experimenter coded experts' and workers' communication behavior in real-time, during the study, using the content categories shown in Figure 4.

These codes represent more than 90% of speech activity during a session. Interrater reliability analyses of the real-time codes based on recoding of three videotapes by five judges show Cohen's kappas in the mid-40s for the set of six codes. As one would expect from coding done in real time, judgments are substantially more reliable than chance but contain a substantial amount of error. The practical consequence of both the small number of differentiations made and their relatively low reliabilities is that we were able to observe only gross differences in the communication behavior by technology condition.

2.2 Results

We present the results in two parts, first examining the effects of collaboration and communication media on measures of task performance and then examining the effects of media on conversational grounding patterns among the collaborative pairs.

Performance

Workers performed substantially better on the three repair tasks with collaborative help. Average time to complete the tasks with a remote expert was half as long as in the solo condition: 7.5 versus 16.5 min, respectively, $t(54) = 4.54$, $p < .001$ (XXX-tailed), $d = .XX$. In addition, the quality of the repairs they completed was superior when they had assistance than when they worked alone:

Figure 4. **Conversational coding system used in Experiment 1.**

Message Type	Definition
Worker questions	Worker questions about the task or technology
Worker descriptions	Worker descriptions of the state of the task or technology
Worker acknowledgments	Workers' indications that their partner's messages had been heard or understood (e.g., "mhm" or "ok")
Expert questions	Expert questions about the worker's state (e.g., "Do you have that step done?")
Expert help	Expert's instructions on how to perform the task (e.g., "Insert the bolt")
Expert acknowledgments	Experts' indications that their partner's messages had been heard or understood. (e.g., "mhm " or "ok ")

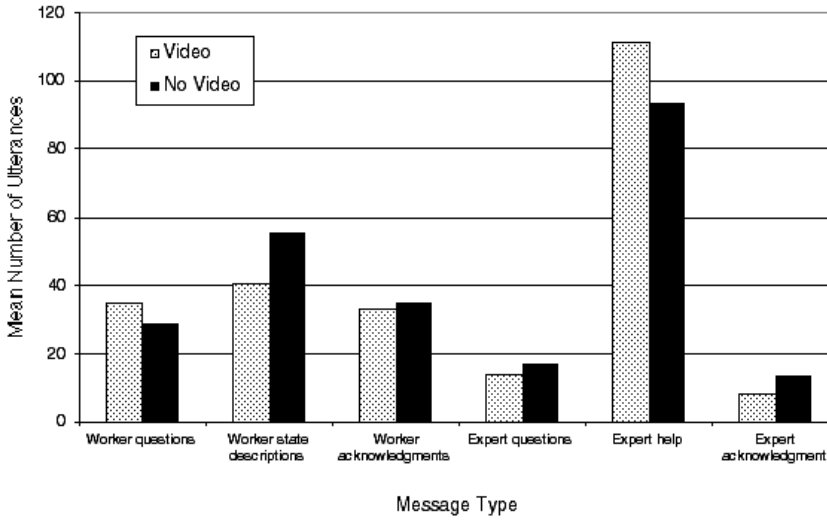
79% of the quality points for the collaborative condition versus 51% for the solo condition, $t(54) = 6.50$, $p < .001$ (XXX-tailed), $d = .XX$. Although having access to an expert dramatically improved performance, having better tools for communication with the expert did not improve the number of tasks completed, the average time per completed task, or performance quality: for all three dependent variables, $F(2, 42) < 1$, $p > .5$. In particular, neither video (the comparison of the full-duplex video condition with the full-duplex no-video condition) nor full-duplex audio (the comparison of the full-duplex video condition with the half-duplex video condition) helped workers perform more tasks, perform tasks more quickly, or perform them better.

Conversational Analysis

Although video technology did not change performance, it did influence how workers and helpers talked about the task. We concentrate here on the effects of video (i.e., the contrast of the full-duplex video condition with the full-duplex no-video condition) on experts' ability to give effective help. When video was present, the worker and expert had a similar view of what the worker was doing, on a moment-by-moment basis. Our goal in this section is to understand how this common view changes coordination of conversation.

Figure 5 shows the average number of utterances in the full-duplex video and no-video conditions. A comparison between the two conditions suggests how a shared visual space influences conversation. As we hypothesized, when visual information was not available, workers were more explicit in describing the state of the task ($p < .02$), for example, what tool or part they were holding or what they had just completed. They were also more explicit in describing their personal internal state ($p < .05$), for example, what they were seeing or whether they understood an instruction. Furthermore, helpers with-

Figure 5. **Message type by video condition (Experiment 1).**



out video were more likely to acknowledge worker comments (e.g., to utter "yes" or "uh huh," $p < .003$), as if they needed to be explicit about having heard the workers' questions and descriptions.

We also calculated the conditional probability of one speech act following another. The differences in the probability of workers' descriptions being followed by helper assistance in the video and nonvideo conditions shows that experts treated a worker's description of state differently depending on whether video was available. When video was available, helpers seem to treat workers' descriptions of state as implicit request for help; they followed with help in 56% of the cases. On the other hand, because the descriptions were necessary for simply tracking what workers were doing in the no-video condition, they were followed by help in only 42% of the cases ($p < .001$). In contrast, there was no difference between conditions in helpers' responses to workers' explicit requests for help. Worker questions were followed by help 95% of the time in the video condition and 92% of the time in the audio-only condition ($p = .26$, nonsignificant).

The coding system in Figure 4 describes in a rough way the conversational interactions between workers and helpers, but it does not differentiate some interesting speech behaviors, such as whether a description or help statement was proactive or a reaction to a prior speech act. To examine these phenomena, we used videotapes and transcripts to examine a single subtask—attaching a brake anchor plate to a straddle-cable connecting the two brake pads—in greater detail. We divided workers' state descriptions and helpers' assis-

tance, to distinguish reactive cases, in which the speaker was responding to an explicit request for information from a partner, from proactive cases, in which the speaker initiated the description or help. In addition, we categorized help into cases of instruction, in which the helper was telling the worker what procedural steps to take, or clarification, in which he or she attempted to clarify a previous instruction. As we discussed in the introduction, we predicted that helpers would be more likely to offer proactive help and more likely to spontaneously clarify their messages when they could use video to monitor their partners' behaviors.

Results for the helpers' utterances are shown in Figure 6. As predicted, pairs coordinated differently depending on whether they had video present. When helpers could see the worker, they gave more proactive assistance (i.e., without the worker explicitly asking for it, $p < .03$), presumably because they could see when the worker was having trouble or had completed a step and was ready to move on. For example, in the following fragment, the helper times his assistance to the moment when the worker has finished getting a pair of pliers and has started to apply it to a bolt.

Worker: I'm just going to go get some pliers so I can tighten it on the opposite side. So I'll try out the opposite side also.
Helper: Good enough. (PAUSE) Although, well really, it's, it's the same bolt throughout so you can probably just hold one side and hold the other.

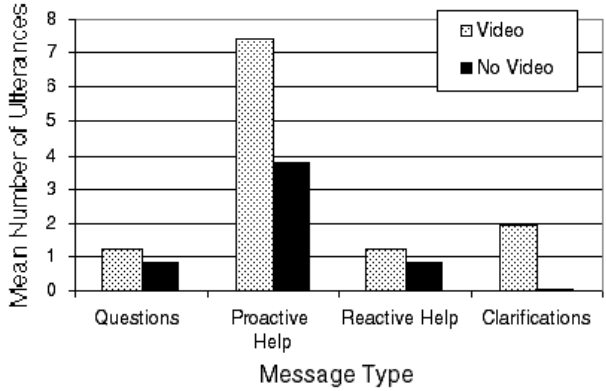
In addition, helpers in the video condition were more likely to clarify and elaborate their prior instructions, presumably because they could better monitor workers' comprehension ($p < .04$).

2.3. Discussion

We hypothesized that the video system would capture enough of the essential elements of actual physical co-presence to improve performance over the audio-only condition. This hypothesis was not supported—contrary to our expectations, 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. We discuss possible limitations of this video system in greater detail in the General Discussion section.

Despite the lack of performance effects, Experiment 1 did provide support for our two communication hypotheses. Workers were less explicit in describing the state of the physical world and what they had accomplished when they

Figure 6. **Mean number of expert messages during the straddle-cable subtask, by presence of video (Experiment 1).**



shared a view of the work environment with their collaborators. When they shared this view, helpers were more likely to offer proactive instruction, basing the instruction they delivered and when they delivered it on a combination of the worker's explicit descriptions and their visual inspection of the worker's behavior. When the shared view was available, helpers were more likely to treat the workers' explicit description of state as an implicit request for assistance.

One limitation of Experiment 1 was that no side-by-side condition was included to serve as a baseline for evaluating the success with which participants used visual information for maintaining task awareness and grounding conversations. In Experiment 2, we add a condition in which worker and helper work side-by-side and have access to a shared visual space.

3. Experiment 2

In Experiment 1, visual information may have influenced task performance less than we predicted because the video communication system we used may have been inadequate. Experiment 2 also incorporates a revised conversational coding scheme and includes additional qualitative analyses of the use of deixis. In addition, we used 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-subject designs have found large differences in communicative style between pairs of communicators that might have masked media effects on performance in Experiment 1.

In Experiment 2, workers performed three repair tasks on a 10-speed bicycle with the assistance of either an expert or a novice helper. Pairs per-

formed one task in each of three media conditions: (a) *side-by-side*, where worker and helper worked in the same room, (b) *audio–video*, where workers were connected by full-duplex audio plus the head-mounted video camera and monitor used in Experiment 1, such that the video feed showed the worker's local activities; and (c) *audio-only*, where 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 or treatment orders.

H1: *Performance.* We predicted that performance would be best in the side-by-side condition, because the quality of the shared visual context is maximized, and because quality is 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. The extent to which performance in the video condition approaches that of the side-by-side condition was predicted to be mediated by the extent to which collaborators were able to use the video technology to facilitate task awareness and grounding.

H2: *Conversational grounding.* We also predicted that 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.

H3: *Deixis.* We predicted that deixis and pointing gestures would be most frequent when pairs worked side-by-side, because helpers and workers shared the same visual space. When helpers can view the same scene as the workers, they can refer quickly and efficiently to task objects, tools, and the like by using short-hand expressions and pronouns such as "this one." We predicted that use of deixis in the video condition would be less frequent than in the side-by-side condition but more frequent than in the audio-only condition because workers who are aware that helpers share their view of the scene can manipulate the visual field such that they too can use deictic terms. Because workers are aware that helpers share their view of the scene, they can manipulate the visual field such that they can use deictic terms.

3.1 Method

Apparatus

The video system used in this study was identical to that in Experiment 1. Full-duplex audio was used in all of the video sessions.

Participants and Procedure

Workers consisted of 25 Carnegie Mellon University undergraduate and graduate students (68% male). 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 workers and were also shown a tutorial videotape illustrating correct procedures. Both workers and helpers received $10 per session for participation and competed for a $20 bonus for the pair with the fastest completion time and best task performance.

The procedure for Experiment 2 was identical to that used in Experiment 1, with the exception that each participant performed one task in each of the three media conditions (side-by-side, audio + video, and audio-only). Trial numbers, repair tasks, and media conditions were counterbalanced across subjects. As in Experiment 1, three sets of dependent measures were collected: performance measures, real-time observations of the interaction, and audio–video logs.

Conversational Coding

To examine the relation between media conditions and task dialogues, we developed a new coding system intended to capture some new distinctions among message types not coded in the system used in Experiment 1. Each utterance was classified as either a question, an answer to a question, or a statement in one of the content categories shown in Figure 7. Two independent coders classified each utterance; agreement was better than 90% and disagreements were resolved through discussion.

3.2 Results

First we examine the effects of communication media on task performance; then, we examine the relation 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 that included the expertise of the helper as a between-pair factor. Completion times differed significantly across media conditions, $F(2, 46) = 14.20$, p

Figure 7. **Conversational coding system used in Experiment 2.**

Message Type	Definition
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, and so forth (e.g., "I don't understand what you're talking about," "Do you know how a quick release lever works?")
Acknowledgments	Feedback that message is heard/understood (e.g., "ok," "uh huh")
Other	Nontask and uncodable communication

$< .001$. Pairs in the side-by-side condition completed tasks about 25% faster than pairs in the audio-only and video conditions ($M = 9$ vs. 13 min, respectively, ps $< .001$). Surprisingly, neither the expertise of the helper nor the expertise by communication medium interaction approached significance. Rated work quality was somewhat higher ($p = .08$) in the side-by-side condition than in the two mediated conditions, which did not differ significantly from one another.

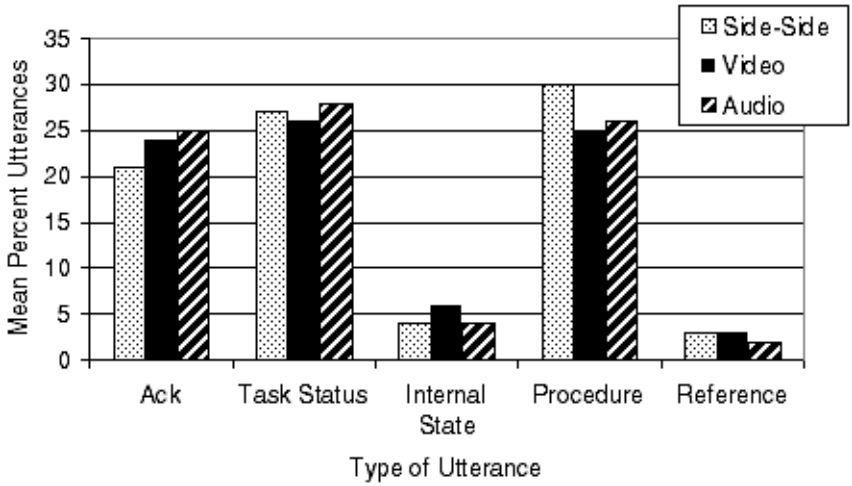
Conversational Analysis

Dialogues were significantly more efficient in the side-by-side condition than in the mediated conditions, which did not differ from one another (mean utterances = 80 for side-by-side, 138 for video, and 123 for audio-only), $F(2, 46) = 6.45$, $p < .005$.

To determine whether use of different media influenced the pattern of conversations in addition to their length, we computed the percentages of utterances in each of our coding categories. The results are shown in Figure 8. Here, we have collapsed over statements, questions, and answers, but the pattern is very similar when the data are further broken down by type of utterance.

Acknowledgments, descriptions of task status, and procedural instructions comprised the majority of utterances. Pairs were less likely to explicitly acknowledge one another's messages when performing the task side-by-side, $F(2, 46) = 5.23$, $p < .01$, probably because they could see if their instructions or comments were acted upon. Pairs' references to internal states also differed significantly across media conditions, $F(2, 46) = 3.73$, $p < .05$. However, contrary to our expectations, references to internal states (e.g., "Do you see that?" "Do you understand?") were more frequent in the video condition than either

Figure 8. **Mean percentage of utterances by content type and media condition (Experiment 2).**



side-by-side or audio conditions. Procedural statements occurred more frequently in the side-by-side condition than in either mediated communication conditions ($ps < .05$), presumably because the overhead of acknowledgments, clarifications, and feedback about internal state took time away from the core task of explaining how to repair the bike and then repairing it. There were no significant differences between conditions in percentages of task status utterances or references to task objects ($ps < .20$).

Qualitative Analyses

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. We also examined how the visual information influenced the maintenance of task awareness.

Reference. References to task objects comprised a small but critical proportion of overall messages in each dialogue—objects had to be identified before workers could complete tasks involving them. Qualitative examination of the conversational exchanges through which participants established the identity of objects suggests that although 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. Figure 9 shows a representative sample of dialogues in each condition in which the worker is attempting to identify the bicycle's derailleur. 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 combining gestures and deictic expressions (e.g., " this thing," "this side," "over here"). In the audio condition, where neither party could see what the other was doing vis-à-vis objects, they had to use lengthy descriptive sequences to describe the objects (see audio sequence 3). This is confirmed by Figures 10 and 11. Figure 10 shows the percentage of references to task objects containing the deictic terms *this*, *these*, or *those* across media conditions. These terms were used much more frequently in the side-by-side condition than the audio one. Figure 11 shows the mean duration of a referring expression across condition. Referring expressions were substantially shorter in the side-by-side condition than in the audio-only condition.

Second, in the video condition, although objects and worker's behaviors were visually shared to some degree, the helpers' physical behaviors were not. Hence, as seen in Figure 10, workers in the video condition used deictic expressions, but helpers did not. In addition, helpers were unable to use gestures to refer to task objects within a shared visual space (and sometimes expressed frustration with this situation, e.g., "If I could point to it, it's right there"). These results suggest that although for workers in the video condition there was a sense of a shared visual space in which deictic references were mutually meaningful, this was 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:

Worker: Whoa! [Shows part with camera]
Helper: What?
Worker: Look at this.
Helper: Look at what?
Worker: You see how warped that is?

Third, in the video condition, the camera showed a more limited field of view than that which the worker could see and was not always aligned with what the worker was attending to. Therefore, task objects were often not visually shared until the worker explicitly maneuvered the camera to bring them into the helper's field of view. In the video condition, but not in the side-by-side one, participants often negotiated what they saw in common. Many messages about internal states consisted of worker queries about what was in the shared field of view (e.g., "Can you see the table?" "See where I'm

Figure 9. **Representative dialogues identifying the derailleur across media conditions (W = worker, H = helper).**

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

Figure 10. **Percentage of deictic references to task objects by media condition and participant role (Experiment 2).**

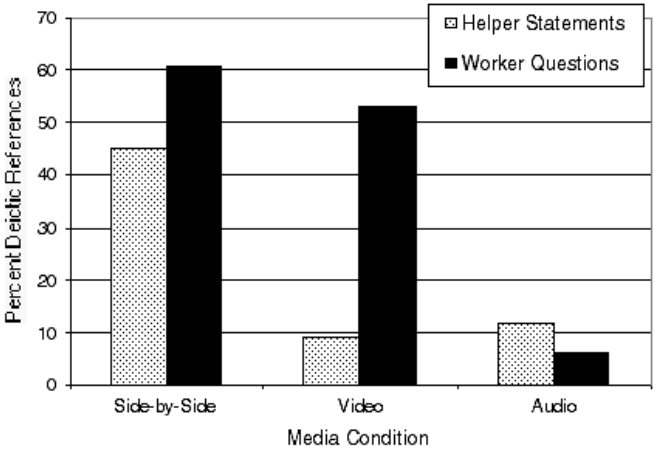
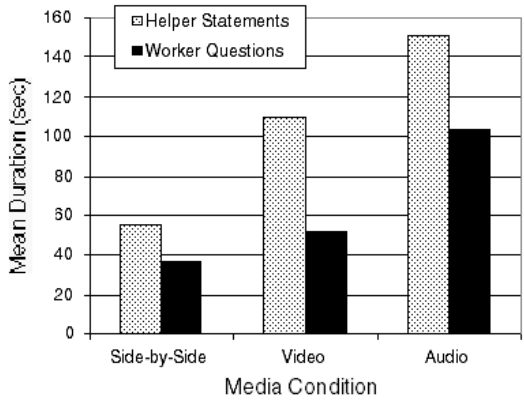


Figure 11. **Mean duration of references to task objects by media condition and participant role (Experiment 2).**



pointing up here?"). Helpers also volunteered information about their field of view (e.g., "I can't quite see the derailleur cage"). 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.

Helper: A little bit lower wouldn't hurt.
Worker: Ok. Is that alright?
Helper: Can't quite see it.

Worker: Um. [Worker moves his head to adjust camera position.] Right here.
Helper: Ah. Yeh, Good.
Worker: Ok.
Helper: That will do it. As long as everything is aligned right we can go ahead and tighten it.

We return to this issue in the Discussion section.

Use of visual information for establishing task awareness. Examples from Figure 12 show that the visual information was used to maintain task awareness as well as to improve conversational efficiency. When the helper could see what the worker had done, he or she would confirm or intervene appropriately. This happened most in the side-by-side condition. As the examples in Figure 12 illustrate, in the side-by-side condition, the helper could observe and intervene to correct a problem (Example 1) or confirm correct behavior (Examples 2 and 3), without explicit description by worker. This also could occur in the video condition, when the pieces were big and in view (Examples 1 and 2). However, when the objects were small or out of view, the helper needed to explicitly query what the worker was doing before being able to intervene (Example 3). In the audio-only condition, helpers needed to rely on the workers' verbal descriptions of the work environment to know if a task had been done well and had no autonomous ability to intervene, based on their own assessment of the state of the task (Examples 1–3).

Using the shared visual space. To better understand how participants used the shared visual space created by our video technology, we examined the relations between message types and behaviors that relied on video (e.g., pointing to an object, moving the camera to focus on an object). Figure 13 shows the percentage of the time messages in the video condition were accompanied by worker gestures that relied on the video feed. Video-related gestures were more frequent during questions and acts of reference, suggesting indirectly that pairs used the video's potential to create a shared visual space.

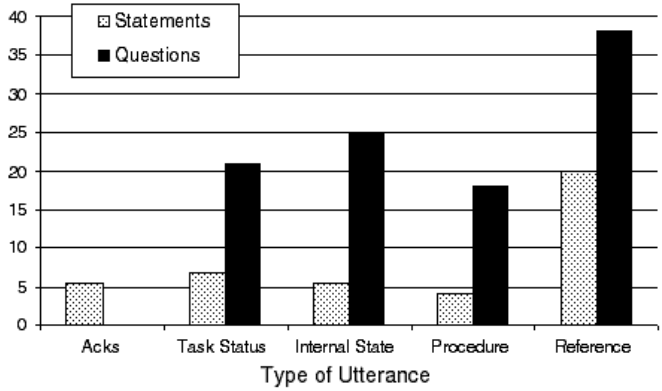
3.3. Discussion

In summary, Experiment 2 provides additional support for the hypothesis that the presence of visual information affects conversational grounding in collaborative physical repair. When pairs worked side-by-side, they performed the task faster and using fewer utterances than when they were remotely linked. Examination of the transcripts shows abundant evidence that participants used visual information in deciding when to converse and what

Figure 12. **Representative dialogues show helper's interventions in correcting a worker error (W = worker, H = helper).**

	Side-by-Side Condition		Video Condition		Audio Condition	
1	W:	Okay- still needs to tighten a little bit more-	H:	Now uh- you need to, you want to angle the seat so that you can actually sit on it.	H:	All right now make sure that's adjusted so that … .You know what it should look like.
	W:	It's funny that the hardest part is the tightening.	W:	[Adjusts seat]	W:	Yeah. I guess that looks okay-
	H:	You're loosening it right now.	H:	It looks like you'd fall forward right now.		
2	H:	just keep tightening and make sure that the seat is parallel to the floor which …	H:	Uh- next go on and adjust it so it's parallel to the bar- the top-	H:	The seat should be parallel with the floor. But if you step back and look at the bike you'll see that it's angled up just slightly.
	W:	Oh- okay.	W:	This bar here? Is that good?	W:	So it should be angled up a little bit?
	H:	Which it is- it's almost parallel	H:	Uh- angle the nose up a little bit more.	H:	Yeh.
			W:	{Adjusts seat}	H:	Just to match the angle of the bike.
			H:	Cool.	W:	Ok.
					H:	Ok. And tighten it down all the way. Ok?
					W:	Alright.
					H:	You got it?
					W:	Yep.
					H:	Ok.
3	H:	And what we need to do now is make sure that it's - the top of the saddle is parallel to the floor.	H:	What are you doing?	W:	It's not perfectly lined up but - it sticks.
			H:	Loosening or tightening or what?	H:	It looks rideable?
	W:	To the floor?	W:	Tightening.	W:	Yeh.
	H:	Yeh. So it looks like its in sort of a comfortable riding position.			H:	Well that's good.
	W:	Ok.				
	H:	It looks like it's there already.				

Figure 13. **Percentage of worker messages in the video condition with associated gestures (Experiment 2).**



to talk about and to effectively construct their conversations once they started talking. The visual information that the helper received over the video system influenced the form of pairs' dialogues but did not improve performance over that in the audio-only condition. We next discuss in greater detail why the video system was unsuccessful in improving performance.

4. GENERAL DISCUSSION

Taken together, our findings show that physical tasks can be performed most efficiently when a helper is physically co-present. Having a remote helper leads to better performance than working alone (Experiment 1), but having a remote helper is not as effective as having a helper working by one's side (Experiment 2). A major benefit of working side-by-side is the extensive shared visual space that this arrangement affords. Both parties working on the task can see task objects, the environment, and their collaborators' behaviors in this environment. The analysis of the talk surrounding these collaborative tasks shows that the visual information was valuable for keeping aware of the changing state of the task, so that they could precisely time their conversational interventions. In addition, partners were able to use the visual information and their ability to gesture in the space to ground the conversations more efficiently.

In the remainder of this Discussion section, we first consider our findings with regard to the role of visual information in communication in greater detail. Then, we consider how affordances of our video technology might account for our pattern of results. We conclude with some general ideas for future research.

4.1. Effects of Visual Information on Communication

Conversation was more efficient—fewer words were required to complete the task—in the side-by-side condition. Content analyses suggest that one reason this might be so is that procedural instructions comprise a higher proportion of utterances in the side-by-side condition; that is, in the side-by-side condition, the helper spent more time telling the worker what to do. In the mediated conditions, not only are dialogues longer but their focus shifted slightly but significantly—more speaking turns are devoted to acknowledging partners' messages and, in the video condition, to messages about internal state.

Despite these findings, however, differences in dialogue structure between conditions were not as large as we had anticipated. One reason we may not have found larger differences stems from some limitations of our conversational coding systems. Specifically, these systems did not fully capture several types of interrelations between speech and action that we have observed in video recordings of the task sessions:

First, our coding systems focused on the type of message content contained in an utterance as opposed to its syntactic form. That is, the coding categories did not distinguish among messages of a given type that did or did not rely on visual information. Our qualitative analyses in Experiment 2 suggested that the form of referring expressions differs depending on the presence or absence of a shared visual environment: When shared visual space is present, pairs could use deictic expressions and gestures to refer quickly and efficiently to task objects. It is likely that similar analyses would show media effects on the form of procedural statements, state descriptions, and other utterance types.

Second, we used an utterance-based definition of a conversational turn. It appears, however, that such reliance on verbal messages may overlook important aspects of how meaning is grounded in collaborative physical tasks. We have observed that behaviors may be alternated with verbal utterances in a turn-taking structure, as in the following example:

Helper: No, down a little more.
Worker: [movement]
Helper: Down a little more.
Worker: [movement]
Helper: Right there.

In our coding systems, the preceding example would be coded as three helper speaking turns with no worker responses between them. We are currently recoding a subset of the data to include these nonverbal turns by the worker to further understand how actions and speech are integrated in collaborative physical tasks.

Third, our coding scheme did not provide a way to encode ongoing worker behaviors. Even when worker behaviors were not integrated with speech in the type of turn-taking structure discussed previously, these behaviors both provided situational awareness cues that influenced when helpers decided to give proactive advice and served as ongoing feedback as to whether an instruction had been understood. In other words, some of the cues that allow experts to make inferences about the timing and content of help are side effects of the worker's task performance, with no explicit communicative intent. A better understanding of the ways experts watch and interpret different types of worker behaviors as they monitor the situation and provide their advice should lead to a greater understanding of what sorts of visual cues should be provided in systems to support remote collaborative repair.

4.2. 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 use shared visual information when it is available. Why, then, were video-mediated dialogues less efficient than side-by-side ones? We consider three sets of explanations for these findings.

Appropriateness of Visual Information

One possible explanation for the lengthier conversations in video as opposed to side-by-side conversations is that our video system did not capture the most important visual elements. As we noted earlier, our video system did not provide the full array of visual cues present in the side-by-side condition (outlined in Figure 1). For example, the remote helper could not view the worker's face, whereas side-by-side helpers may have glanced at workers' faces to monitor attention and comprehension. If this were so, perhaps our video system should be augmented by a camera feed of the worker's face and upper body. However, we consider this explanation unlikely given the failure of many previous studies to show benefits of head-oriented video conferencing systems (see Whittaker & O'Conaill, 1997, for a review of this literature).

A more plausible limitation of our system is that it provided remote helpers with only a partial view of the repair scene. Which objects were in view at a particular time depended on the worker's head position. Thus, objects were sometimes outside the view of the camera (e.g., on a work table or the other side of the bicycle). We are examining the importance of this type of visual information in ongoing studies by providing remote helpers with either the

view from the head-worn camera, a view from a wider angle scene camera, or both views together.

Negotiating Shared Visual Space

A second possible explanation for the longer dialogues in the video condition is that although our system provided the important types of visual information for the task, the costs of obtaining this information were higher than they would be when pairs worked side-by-side. In terms of Clark and Brennan's (1991) framework for analyzing the effects of media on communication, we would argue that although our video systems did provide many of the key types of visual information required for collaborative physical tasks, the costs for obtaining this information were higher than they would be when pairs work side-by-side. It appears that creating shared visual space in the video condition has costs in terms of both conversational efficiency and worker behaviors.

First, workers' queries about video-linked helpers' fields of view suggest that participants had difficulty establishing what visual information was shared and that additional speaking turns were often required to achieve a joint visual focus of attention. As we indicated previously, many messages about internal states consisted of worker queries about what was in the shared field of view, and helpers also volunteered information about their field of view. This use of *see* to clarify shared visual space was virtually nonexistent in the side-by-side and audio conditions.

In addition, as we showed in Figure 9, workers often had to explicitly maneuver the video camera to bring objects into the joint visual field. The additional time required to achieve this joint visual field may explain why conversations via video link were longer than side-by-side conversations. For example, although workers in the video condition could position the camera such that an object could be referred to by deictic expressions not possible in the audio-only condition (e.g., "this one?"), the additional time required to position the camera may have counteracted any benefits in conversational efficiency that the use of deixis permitted. This suggests that the design of systems to support remote collaboration on physical tasks would benefit from a more systematic analysis of the benefits and costs of alternative technologies for providing the same visual information.

Visual Co-presence Versus Physical Co-presence

A final possible explanation for the differences in efficiency between video and side-by-side conversations stems from essential differences between visual co-presence, or sharing a joint view, and full physical co-presence, in which spatial relations among people and task objects are maintained. For ex-

ample, although a view of another person's upper body would enable a remote partner to see that he or she is pointing at something, a spatially consistent view of both partner and task objects is required to determine the target of the pointing gesture. Similarly, other behavioral indicators of attention, such as direction of eye gaze, are difficult to interpret unless the target of that gaze is also in view.

In our research paradigm, the worker and task objects were spatially oriented such that the camera could indicate where in the work scene the worker was looking. Although the system did not provide remote helpers with enough information to interpret workers' pointing hand gestures, workers were able to use their spatial relation with the environment to develop surrogate pointing techniques (e.g., using camera focus to indicate a target item, as we discussed previously).

However, remote helpers had no spatial relation to task objects. Although workers were able to see the helpers' hand gestures, helpers themselves had no mechanisms for pointing to task objects. Given that research suggests the importance of designating objects in collaborative physical tasks (e.g., Barnard, May, & Salber, 1996; Bauer, Kortuem, & Segall, 1999; Bekker, Olson, & Olson, 1995; Bolt, 1980; Kuzuoka & Shoji, 1994), this inability to support helper gesture may be an important limitation to our system. A number of technologies have been developed to permit remote gesture (e.g., remote-controlled laser pointers, overlays of gestures on a view of the remote scene, etc.; see Kuzuoka, 1992, Kuzuoka et al., 1994; Kuzuoka, Oyama, Yamazaki, Suzuki, & Mitsuishi, 2000). We are currently attempting to incorporate remote pointing technology in our video system to assess the value of this capability on collaborative repair tasks.

4.3. Implications for System Design

Our findings and the preceding discussion suggest four recommendations for the design of future systems to support collaborative remote repair:

- Provide people with a wide field of view, including both task objects and the wider environment, so that they can more easily maintain task awareness and ground conversations.
- Clarify what is part of the shared visual space. All parties to the task should have a clear understanding of what one another can see; that is, the contents of the shared visual space should be part of participants' mutual knowledge or common ground.
- Provide mechanisms to allow people to track one another's focus of attention. When people can see where each person is looking, it is easier to establish common ground.

- Provide support for gesture within the shared visual space. Talking about things is most efficient when people can use a combination of deictic expressions and gestures to refer to task objects.

Additional research is needed to clarify the best methods for visually providing these capabilities.

Alternatives to Video

Although we have focused on one type of system to support remote collaboration on physical tasks, namely, a system that provides the same visual information participants use when they perform the task side-by-side, there are alternative ways to approach the remote support of these tasks. Each of the conversational functions provided by visual cues in a side-by-side setting (i.e., attention, comprehension, deixis) could alternatively be provided by other technologies that create a representation of this information rather than conveying it directly. For example, Jie Yang (personal communication, XXDATEXX) has developed a system that indicates through dynamic visual arrows where each person at a meeting is looking at any given time, and Kuzuoka and colleagues (1994, 2000) have designed remote laser pointing systems that overlay a point of light on the intended target rather than showing the helper's hand gesture itself. Similarly, a system might be designed to send messages such as "the worker has picked up the wrench" to a remote helper in lieu of providing a direct video feed of this event.

4.4. Conclusion

We have argued that shared visual space is essential for collaborative repair because it facilitates situational awareness and 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 this study goes only part of the way toward creating "virtual" physical co-presence, but the guidelines we suggest should help future system designers come closer to this goal.

NOTES

Acknowledgments. We thank Elise Nawrocki, Tom Pope, Leslie Setlock, and Mei Wang for help in running the experiment and analyzing data. We also thank the edi-

tors of this special issue and four anonymous reviewers for their very helpful comments on previous versions of this article.

Support. This study was conducted with support from National Science Foundation Grants 9022511 and 9980013.

Authors' Addresses. Robert E. Kraut, Human–Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenues, Pittsburgh, PA 15213. E-mail: robert.kraut@cmu.edu. Susan R. Fussell, Human–Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenues, Pittsburgh, PA 15213. E-mail: susan.fussell@cmu.edu. Jane Siegel, Human–Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Avenues, Pittsburgh, PA 15213. E-mail: jals@cs.cmu.edu.

HCI Editorial Record. First manuscript received April 10, 2001. Revision received November 13, 2001. Accepted by Elizabeth Churchill and Thomas Erickson. Final manuscript received August 18, 2002. — *Editor*

REFERENCES

Barnard, P., May, J., & Salber, D.(1996). Deixis and points of view in media spaces: An empirical gesture. *Behavior and Information Technology, 15*, 37–50.

Bauer, M., Kortuem, G., & Segall, Z. (1999). "Where are you pointing at?" A study of remote collaboration in a wearable video conference system. *Proceedings of the ISWC 99 Third International Symposium on Wearable Computers. City, STATE: IEEE Press.*

Beattie, G. W., & Barnard, P. J. (1979). The temporal structure of natural telephone conversations. *Linguistics, 17,* 213–230.

Bekker, M. M., Olson, J. S., & Olson, G. M. (1995). Analysis of gestures in face-to-face design teams provides guidance for how to use groupware in design. *Symposium on Designing Interactive Systems '95* (pp. 157–166). New York: ACM.

Bolt, R. (1980). "Put-that-there": Voice and gesture at the graphics interface. *Proceedings of the SIGGRAPH 80 Conference on XXXXXXX* (pp. 262–270). New York: ACM.

Clark, H. H. (1996). *Using language.* Cambridge, England: Cambridge University Press.

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: American Psychological Association.

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, England: Cambridge University Press.

Clark, H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22,* 1–39.

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.

Endsley, M. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37,* 32–64.

Flor, N. V. (1998). Side-by-side collaboration: A case study. *International Journal of Human–Computer Studies, 49,* 201–222.

Ford, C. E. (1999). Collaborative construction of task activity: Coordinating multiple resources in a high school physics lab. *Research on Language and Social Interaction, 32,* 369–408.

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62,* 378–391.

Fussell, S. R., Kraut, R. E., & Siegel, J. (2000). Coordination of communication: Effects of shared visual context on collaborative work. *Proceedings of the CSCW 2000 Conference on XXXXXXX* (pp. 21–30). New York: ACM.

Gaver, W., Sellen, A., Heath, C., & Luff, P. (1993). One is not enough: Multiple views in a media space. *Interchi '93* (pp. 335–341). New York: ACM.

Goodwin, C. (1996). Professional vision. *American Anthropologist, 96,* 606–633.

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3, pp. 41–58). New York: Academic.

Isaacs, E., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General, 116,* 26–37.

Jefferson, G. (1972). Side sequences. In D. Sudnow (Ed.), *Studies in social interaction* (pp. 294–338). New York: Free Press.

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.

Kraut, R. E., Fussell, S. R., Brennan, S., & Siegel, J. (in press). A framework for understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work. In P. Hinds & S. Kiesler (Eds.), *Technology and distributed work.* Cambridge, MA: MIT Press.

Kraut, R. E., Miller, M. D., & Siegel, J. (1996). Collaboration in performance of physical tasks: Effects on outcomes and communication. *Proceedings of the CSCW 96 Conference on XXXXXXX* (pp. 57–66). New York: ACM.

Kuzuoka, H. (1992). Spatial workspace collaboration: A Sharedview video support system for remote collaboration capability. *Proceedings of the CHI 92 Conference on XXXXXXX* (pp. 533–540). New York: ACM.

Kuzuoka, H., Kosuge, T., & Tanaka, K. (1994). GestureCam: A video communication system for sympathetic remote collaboration. *Proceedings of the CSCW 94 Conference on XXXXXXX* (pp. 35–43). New York: ACM.

Kuzuoka, H., Oyama, S., Yamazaki, K., Suzuki, K., & Mitsuishi, M. (2000). GestureMan: A mobile robot that embodies a remote instructor's actions. *Proceedings of the CSCW 2000 Conference on XXXXXXX* (pp. 155–162). New York: ACM.

Kuzuoka, H., & Shoji, H. (1994). Results of observational studies of spatial workspace collaboration. *Electronics and Communications in Japan, 77,* 58–68.

Malone, T. W., & Crowston, K. (1994). The interdisciplinary study of coordination. *ACM Computing Surveys, 26,* 87–119.

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 the Interchi 93 Conference on XXXXXXX* (327–334). New York: ACM.

Noll, A. M. (1992, May/June). Anatomy of a failure: Picturephone revisited. *Telecommunications Policy, XX,* 307–316.

Orr, J. (1996). *Talking about machines*. Ithaca, NY: Cornell University Press.

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

Tang, J. C. (1991). Findings from observational studies of collaborative work. *International Journal of Man–Machine Studies, 34,* 143–160.

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 the CHI 99 Conference on XXXXXXX* (pp. 302–309). New York: ACM.

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* (pp. 23–49). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.