Error and Coupling: Extending Common Ground to Improve the Provision of Visual Information for Collaborative Tasks

Abstract

One possibility presented by novel communication technologies is the ability for remotely located experts to provide guidance to others who are performing difficult or technical tasks in the real world. In these scenarios, video views and other visual information have been shown to be useful in the ongoing negotiation of common ground, but their actual impact on performance has been mixed. We argue here that one reason for this is that some means for providing visual information are more error-prone than others. One source of error is "coupling," which we define as the extent to which changes in behavior are tied directly to changes in visual information. We present data from two laboratory experiments comparing three video-mediated communication systems that differ in the degree of coupling they exhibit. Results indicate that the moderately coupled system resulted in superior performance overall, but that participants had a slight subjective preference for the tightly coupled system on two dimensions of assessment.

Introduction

Recent advances in communication and collaboration technologies have allowed groups of geographically distributed individuals to work together in unprecedented ways(DeSanctis & Monge, 1998; Olson & Olson, 2001). One area in which such work is increasingly common is the consultation of remote experts in the performance of repair or construction tasks in the real world (Fussell, Kraut, & Siegel, 2000; Gergle, Kraut, & Fussell, 2004; Kirk & Fraser, 2006; Nardi et al., 1993). Remote expertise can be particularly valuable in cases where constraints on time or distance prevent the expert from physically traveling to the location. In the case of a medical emergency in an isolated location, for example, it may not be possible for a doctor to travel quickly enough to save the patient's life. A remote doctor, however, can provide assistance in performing a procedure (Ballantyne, 2002; Zuiderent, Ross, Winthereik, & Berg, 2003). Or in the case of repairing a NASA space station, it is simply not practical for engineers to travel into space to do repair work themselves. They can, however, provide guidance to the astronauts actually performing the tasks.

These are instances of what Whittaker (2003) refers to as "talking about things," where by "things" he means physical objects or artifacts being discussed and used in performing a task. In particular, it can be useful for the remote expert (the "helper") to have a visual image of the worker's workspace (Fussell et al., 2000; Kraut, Fussell, & Siegel, 2003). This view provides a shared visual space that can be referenced by both parties in their ongoing negotiation of common ground, which is a state of mutual understanding of what is being discussed (Clark, 1992; Clark & Brennan, 1991). In the examples described above, negotiation of common ground might include discussion to ensure that the worker is performing tasks in the proper region and using the correct components.

Clark and Brennan (1991) point out that different communication media have varying attributes that can facilitate or constrain the negotiation of common ground. With regard to video, they note that a shared visual space allows for visibility and cotemporality, meaning that discussion can take place in real time. One problem with this assessment, however, is that the video information provided is assumed to be perfect (or at least adequate). That is, there is an implicit assumption that what is shown in the video view is useful to both the helper and the worker and that it will be used. Research on the performance of collaborative physical tasks using various video systems, however, suggests that video is not always relevant or useful (Fussell, Setlock, & Kraut, 2003). Providing information that *is* useful has proven to be a difficult research challenge both because it is hard to predict what the helper wants to see (Ou, Oh, Fussell, Blum, & Yang, 2005) and because both the helper and worker adapt their physical behavior (Anonymous, 2006) and conversation to the available information (Gergle, 2006; Schober, 1993).

In this paper we present an extension to Clark and Brennan's work that takes this into account. In particular, we present data from two laboratory experiments to introduce the concept of "error" as a measure of how relevant visual information is to task participants and suggest that one source of error is "coupling" – the extent to which change in visual information is coupled to change in participant behavior or focus. These studies compare and evaluate pair performance using three video-mediated communication systems that differ in the degree to which they are coupled to movement by the worker. Results indicate that the moderately coupled system resulted in superior performance overall, but that participants had a slight subjective preference for the tightly coupled system on two dimensions of assessment.

Background and Literature Review

Visual information as a resource for grounding

Common ground refers to a shared understanding of what is being discussed in a conversation with multiple participants, and is achieved through the ongoing negotiation process referred to as "grounding" (Clark, 1992; Clark & Brennan, 1991). In situations where the task involves the identification and manipulation of physical objects, a shared visual context can serve as an important resource in the grounding process (Brennan, 2005; Gergle et al., 2004; Karsenty, 1999; Kraut et al., 2003). By "shared visual context" we mean that all participants have access to the same or very similar visual information and can refer to this information in the grounding process. In particular, these authors point out that shared visual context can be useful for establishing a shared point of focus, and in facilitating mutual awareness.

One example of a task where visual information is particularly useful is the "remote repair" scenario mentioned earlier in which a remote "helper" provides guidance to a "worker" performing a physical task in the real world. Several laboratory studies have been conducted using tasks intended to replicate critical aspects of the remote repair scenario – namely the identification and manipulation of objects – via tasks such as bicycle repair, and toy construction (Fussell et al., 2000; Gergle, 2006; Kirk & Fraser, 2006; Kuzuoka, 1992). These studies suggest that visual information is particularly useful when task components are difficult to describe or distinguish, or "lexically complex." Examples of lexically complex objects include the similarly-colored Tartan plaid patterns used in Gergle, et al.'s puzzle studies (2004).

Visual information is less useful, on the other hand, when objects can be easily and succinctly described verbally (e.g., by saying, "the red one" when there is only one red object) or when the needed visual information is not readily available. There may be cases, for example, where detail is needed, but a video camera is showing a wide shot, or vice versa. In these cases, participants relied primarily on verbal descriptions, even when they had the capacity to control the camera and alter the available visual information (Fussell et al., 2003; Anonymous, 2006). This was the case even when it would have been clearly beneficial to change the view. Moreover, selecting between views can be confusing or disorienting to the helper, who may not understand how they fit together (Gaver, Sellen, Heath, & Luff, 1993).

Error in the provision of visual information

Given this apparent tendency to adapt conversation to available visual information, there is an increased imperative to provide information when it is needed and in a form that is immediately useful (Ou et al., 2005). Failure to do so could result in difficult verbal negotiations and mistakes that could be costly on multiple dimensions. Providing the right information, however, requires a framework for understanding how to identify what information is needed and, perhaps more importantly, an understanding of how to minimize error in this process.

For our relatively simple purposes here, we define error in visual information at any given moment as the difference between what the helper wants to see and what the helper is able to see. When the helper is able to see exactly what she wants, we can say error is zero. As the information becomes less relevant and useful, error increases. This is roughly analogous to Shannon's (1948) notion of "noise," and one can imagine a rigorous information-theoretic definition of error and a similarly rigorous set of calculations to compute it. This, however, raises measurement and modeling issues that are beyond the scope of this paper. For the present discussion of relative error rates, we are concerned with the conceptual notion of error. Its units and exact measures are not important.

This concept allows us to carefully consider a key difference between tasks such as Clark's (1992) collocated shape-matching or Gergle's (2006) on-screen puzzle tasks and real-world physical tasks with a video view provided to a remote helper (Fussell et al., 2000; Fussell et al., 2003). In the collocated and on-screen puzzle tasks, both participants have nearly-identical views of the workspace, and the entire task takes place within the confines of a single screen or shared physical space. Error in this case is therefore arguably quite close to zero. In the real-world tasks, on the other hand, video was used to provide a shared visual context. Error in these cases was introduced, for example, by inopportune or missing camera shots (Fussell et al., 2003; Gergle, 2006). This is likely one reason why visual information resulted in greater performance benefits in the on-screen tasks than in the real-world tasks.

If this is the case, we should expect error to vary according to how visual information is provided and the nature of the task. While we do not have a formal way to measure error (because we cannot know exactly what the helper wants to see at any given moment without asking them, which would distract from performance), we can measure and evaluate several factors likely to correlate with error in the formal sense.

First, we know from Gergle (2006) and Schober (1993) that people adapt their conversations to the visual information that is available. One form of adaptation that would indicate inadequate visual information is relying on the visual information less and asking more verbal questions to designate a shared point of focus or to identify difficult-to-describe task components. In this way, more questions can be said to indicate more error in the visual information stream. We would therefore expect that:

H1: When visual information is less detailed, participants will ask more questions.

A second indicator of error is an overall sense of how useful participants found the visual information in performing the task. While it is not practical to ask this constantly as they perform the task, asking them to reflect on the utility at the conclusion of the task will provide some sense of how useful the information was. We would expect that:

H2: When visual information is less detailed, participants will find the video stream less useful.

Our final indicator of error in visual information is the number of mistakes that participants make. Given that their goal is to complete the task accurately and efficiently, mistakes are a potential indicator that the visual information provided was not adequate. We would therefore expect that:

H3: When visual information is less detailed, participants will make more mistakes.

Coupling

Minimizing error in visual information by automatically providing appropriate and useful views to the helper has proven a difficult research problem, for which there have been several approaches. One is to provide a static wide shot of the entire workspace, under the assumption that whatever the helper wants to see will always be in the camera shot. On the one hand, this is useful in that overview information is provided, but it is not useful when negotiation or discussion of detail is required (Fussell et al., 2003). In these cases, some have experimented with provisions for remote "gesturing" but such functionality is useful only when there is enough detail in the wide shot to see what is being gestured at (Fussell et al., 2004; Kirk & Fraser, 2006).

A second approach is to assume that the helper will often want to see whatever it is that the worker is focused on at any given moment. This led Fussell, et al. (2003) to experiment with mounting a camera on the worker's head. While this was useful in that the helper could always see the worker's current visual focus of attention, this same attribute was also a significant liability in that the worker's head moved far more often and more quickly than was necessary or desirable to give the helper the requisite visual information.

The inadequacy of both of these approaches suggests that reducing error in visual information is not a linear optimization problem – neither very infrequent nor very frequent information updates yielded optimal results. Given individuals' capacity for adaptation in physical behavior (Anonymous, 2006) and conversation (Gergle, 2006), the goal is arguably to minimize error by providing information that is useful most of the time and generally makes it easy for people to adapt. Since the helper cannot reasonably be constantly polled, this requires an indicator of what the helper is likely interested in seeing that can be used in determining what to show.

One way to think about this problem is in terms of coupling, an idea that surfaces repeatedly in discussions of interactions and interrelations between complex systems of people and/or machines (Miller, 1978; Simon, 1996; Weick, 1982). Where interactions within or between systems are frequent, and components affect each other directly and immediately, they are considered to be "tightly coupled." When interactions are less frequent and effects less direct, the systems are "loosely coupled."

In our case, we can describe coupling as the extent to which change in some indicator of likely helper interest correlates with change in the visual information that is being provided. In the case of Fussell, et al.'s (2003) head-mounted camera, coupling is very tight – every move of the head necessarily correlates with a move of the camera. In the case of a static wide-shot, coupling is extremely loose – no indicators correlate with camera movement because there is no camera movement.

One problem with tightly coupled systems is that errors can easily be magnified in unexpected ways, as illustrated in Simon's (1996) discussion of feedback and "feedforward" mechanisms. He notes that predictive (feedforward) information can be destabilizing when erroneous information is taken too seriously, and that systems must be designed to maintain stability. In other words, because interactions between system components in tightly coupled systems are frequent and impact is direct, erroneous information may be acted upon immediately in ways that have much broader consequences. This suggests that tight coupling may not be appropriate in certain error-prone situations, because the broader consequences are potentially so significant. In this way, we can describe the head-mounted camera as a system that is error-prone in that the camera moves more than the helper wants it to (because it is attached to the worker, as described above) and therefore often provides information that the helper does not want. In a sense, it is precisely because of the tight coupling that the system is error-prone.

In this way, coupling and error have an interesting relationship. A system that was too tightly coupled to worker behavior resulted in sufficient error that the system provided few performance benefits (Fussell et al., 2003). At the same time, however, a static camera system which was very loosely coupled to worker behavior, also resulted in few performance benefits. Nonetheless, we know that there is utility in visual information (Clark, 1992; Kraut et al., 2003). This suggests that there is a middle range for coupling that will result in improved performance, and this is the focus of our argument.

In making this argument, we suggest first that some coupling is better than none at all. In other words, providing visual information that is somehow dynamically tied to an indicator of helper's desired information will result in improved performance as compared with static visual information. More specifically, we expect that:

H4: Participants' task performance will be faster when visual information is dynamic than when it is static.

H5: Participants will use fewer words to identify task components when visual information is dynamic than when it is static.

At the same time, we want to avoid coupling that is too tight. Here, we hypothesize that when we compare a moderately coupled system to a more tightly coupled system, the moderately coupled system should result in improved performance. While the actual degree of coupling is, in a sense, arbitrary, we based our system designs on data from our own prior work and that of others, as we describe below. We expect that:

H6: Participants will perform more quickly using the moderately coupled system than with the more tightly coupled system.

H7: Participants will find the moderately coupled visual system more useful than the tightly coupled system.

Methods

Design and Participants

In both experiments, we used full-factorial 2 x 2 within-participants designs to compare the performance of pairs of participants performing a series of Lego construction tasks at two levels of lexical complexity, and using two systems for providing visual information. In Experiment 1, we compared a static (very loosely coupled) visual system with a moderately coupled system described below. In Experiment 2, we compared the same moderately coupled system with a more tightly-coupled system, also described below.

There were 24 participants (6 female) aged 19 – 33 (*M* = 26, *SD* = 5) in Experiment 1 and 32 participants (15 female) aged 19-29 (*M* = 22, *SD*=2) in Experiment 2. All participants were recruited via posted flyers and email notices, and were required to have normal or corrected-to-normal color vision and to use English as their primary language. All were paid $10.

Task and Setup

The overall task was for the worker to use Lego bricks to construct three multi-layer “columns” (see Figure 1) in specifically defined regions of her workspace (see Figure 2), based on instructions from the helper. The worker sat at a table that was divided into six discrete regions (see Figures 2 and 3). Five were used for building objects and the sixth was where the pieces were placed before each trial. The helper was seated in front of a 20” LCD monitor and given a paper map of the workspace indicating which regions the columns were to be built in. Regions were used to replicate real-world tasks in which activities must take place in specific locations (e.g., parts of the body in surgery).

In Experiment 1, each column consisted of four layers – two involved “identification” tasks and two involved “construction” tasks. The identification tasks are described in detail in (Anonymous-b, 2007), but are not the focus of this paper. Participants completed only construction tasks in Experiment 2.

In the construction tasks, workers were provided with individual Lego pieces for one layer (3 columns) at a time. Pieces were always placed in the “pieces area” and the columns were built in separate work regions (see Figure 2). For the simple task, each layer consisted of 10-12 easy-to-describe pieces. In the lexically complex construction task, a similar number of pieces was used, but the pieces were irregular in shape and orientation. Helpers were provided with an exact duplicate of each completed layer, one at a time. The goal was for the helper to instruct the worker in constructing each layer, which included identifying pieces and placing them correctly.

Participants were permitted to move only one piece at a time, and all construction had to be done in place – the entire layer could not be lifted up.

Participants were in the same room, but separated by a divider. They could hear each other and were permitted to talk, but they could not see each other. They indicated to the experimenter when they thought each layer was complete, but were not permitted to move on until all errors had been corrected.

Experimental Conditions

Participant performance was measured in three experimental conditions (two in each experiment). In all conditions, the helper had a 20" LCD monitor to view the video image.

Loosely Coupled/Static Camera System

A camera above the worker's left shoulder provided a fixed wide shot of the entire workspace (see Figure 3). This shot was available throughout the duration of the experiment.

Moderately Coupled Automatic Camera System

Based on data from prior work(Anonymous, 2006), we developed an automated camera system (described in more detail in Appendix A). This system used a single pan-tilt-zoom (PTZ) camera (Sony SNC-RZ30) to provide detailed close-up shots of the six workspace regions, as well as wide-shots to allow the helper to remain aware of where in the workspace the task was taking place. Shots were changed based on the location of the worker's dominant hand, which was tracked using a Vicon optical motion capture system with techniques described in Anonymous-a, (2007).

We elected to use hand tracking for several reasons. First, we knew from our own prior work that, in a remote repair task involving physical effort, worker hand location is a reasonable indicator of what the helper wants to see (Anonymous, 2006). Second, using motion tracking allows us to physically disconnect the camera from the worker in order to avoid the frequent or distracting movements encountered with head-mounted cameras. This means that some “discretion” can be built into the system in determining when camera moves should take place, such as avoiding camera moves when the worker moves her hand to a region only for a second or two. It also means that the camera shots can be region-based. When the worker’s hand moves but stays within a particular region of the workspace, the shot need not change. Our moderately coupled system was based on four simple camera shot transition rules, described in Appendix A. These transition rules were developed iteratively.

Tightly Coupled Automatic Camera System

A single PTZ camera was located above the worker’s shoulder. The camera shot was continuously adjusted based on the position of the worker’s dominant hand in the workspace. Hand position information was gleaned from the motion capture system, as in the previous condition. In this case, however, only close-up shots were used. To the extent possible, the worker’s hand was constantly kept in the center of the shot.

Procedure

Participants were randomly assigned (via coin toss) on arrival to “helper” and “worker” roles, and were shown to their separate workspaces. The task was then explained to them, and they were told that their goal was to complete it as quickly and accurately as possible. Participants then completed practice tasks to ensure that they understood the details of the task and how the camera control system worked.

In the moderately- and loosely-coupled conditions, the basics of system operation were explained. Participants were told that the camera movements were guided by the position of the dominant hand of the worker. They were not given any specific details of the control algorithm, but were required to complete a practice task in each condition to gain experience.

The pieces for the first task were then placed in the pieces region, the helper was given the first model layer, which was an exact duplicate of the object to be constructed by the worker, and the workspace map, and the pair was permitted to begin.

After each condition, the helper and worker both completed questionnaires that evaluated their perceived performance, the utility of the visual information for examining objects and tracking partner location, and the ease of learning to use the system. The questionnaire items were developed for this study and validated by pilot data.

Analysis

All sessions were video recorded for analysis. All sessions of Experiment 1 (with one exception due to technical issues) and a random sample[1] of 7 transcripts from Experiment 2 were fully transcribed and coded using the coding scheme developed by Gergle (2006). In this scheme, each utterance is coded in terms of three attributes: type (e.g., "statement," "question," "answer"), function (e.g., "piece reference," "piece position," or "strategy") and the use of deictics spatially, temporally or with pronouns. Each transcript was coded by at least one of two independent coders, with 15% of them coded by both coders. Agreement between coders was better than 90% and disagreements were resolved via discussion.

Individual questionnaire items from the post-experiment questionnaires were aggregated into 6 constructs. Each construct consisted of 3 – 5 related items. Cronbach's $\propto$ for these constructs ranged between .7 and .9, which is considered adequate for social science research

[1] A subset of transcripts were used in this experiment in order to preliminarily explore the data. There is no reason to believe that the participants in the subset differ in any substantive way from those whose sessions were not analyzed.

(Nunally, 1978). Confirmatory factor analyses were also performed to ensure that all items loaded onto a single factor (DeVellis, 2003).

Results

Error and Performance

Our first goal is to illustrate the notion of "error" in the provision of visual information as a factor that influences the utility and use of that information in task performance.

In H1 we hypothesized that participants would ask more questions of each other in the static condition than in the moderately coupled condition. This is because the visual information in the static condition would serve as a starting point for conversation, but more questions would be needed to clarify details unavailable in the video. As can be seen in Table 2, the data clearly support this hypothesis. Pairs asked significantly more questions in the static condition (M = 42.27, SD = 20.99) than in the moderately coupled condition (M = 16.45, SD = 9.95), t (10) = 3.28, p < .01. This indicates that the visual information did not provide exactly what the helpers wanted or needed to see because they had to ask clarifying questions.

A related indicator is the number of times that visual information was used to acknowledge that behavior was correct. One would expect that when visual information is ambiguous with regard to detail participants would be less likely to verbally acknowledge on-screen behavior (e.g., "Yeah, like that."). The data support this assertion as well, as shown in Table 2. It can be seen that the number of verbal acknowledgements of on-screen behavior was twice as high in the moderately coupled condition (M = 15.36, SD = 8.75) as in the static condition (M = 7.18, SD = 6.77), t (10) = 2.78, p < .05.

In H2 we hypothesized that participants would find the visual information from the static condition to be less useful (which would indicate more error) than visual information in the moderately coupled condition. To assess this hypothesis, we used 3 measures of utility from the questionnaires: overall usefulness of the video, ability to see detail, and ability to see where in the workspace one's partner was working. As can be seen in Table 1, the data support this hypothesis strongly on all but one of these measures. Using seven-point Likert scales anchored by "Strongly Disagree" and "Strongly Agree" with 4 as a neutral point, participants rated the overall utility of the video view to be 2.9 (*SD*=1.2), which indicates negative utility and is less than their rating of the moderately coupled system (*M*=5.2, *SD*=1.2) by a statistically significant margin, t (21) = -6.72, p < .001. Participants also indicated that the static system (*M*=3.1, *SD*=5.9) did not allow them to see detailed aspects of what the worker was focused on when compared with the moderately coupled system (*M*=5.9, *SD*=1.4) again by a statistically significant margin, t (21) = -9.08, p < .001. With regard to the ability to maintain awareness of where in the workspace the worker was, however, responses were nearly identical in the two conditions (M_{Static} = 5.5, SD_{Static} = 1.3; $M_{Moderate\ Coupling}$ = 5.7, $SD_{Moderate\ Coupling}$ = .9), t (21) = -.93, n.s.. While the null hypothesis does not indicate with certainty that there is no difference between these conditions, we believe the responses are sufficiently similar and positive to indicate that the static camera did provide adequate information in this regard.

H3 was about mistakes. Mistakes are another potential indicator of error in that confusion about or misidentification of pieces on the part of the helper indicates that visual information was not likely sufficient. In Experiment 1, participants made a total of 7 mistakes that were corrected only when pointed out by the experimenter. Six of these 7 were in the static condition, suggesting again that there was more error in the provision of visual information in this condition.

To also get a sense of the number of mistakes that participants corrected themselves, we used the logged motion capture data to analyze the number of dominant hand moves to and from the pieces area. A larger number of moves to the pieces area would indicate a larger number of mistakes in the form of misidentified pieces. Even after standardizing the number of moves by dividing by the total number of minutes taken to complete each task, there were more moves to and from the pieces area in the static condition (*M*=4.66, *SD*=3.16) than in the moderately coupled condition (*M*=3.54, *SD*=2.10) by a marginally significant amount (F(1,9)=3.76, p<.1).

Coupling and Performance

Having established that visual information with less error is more useful and effective than visual information with more error, our problem becomes one of minimizing error in the visual information stream. One potential source of error, as discussed above, is coupling – the extent to which change in visual information correlates with change in an indicator of desired information. We generally hypothesized based on prior work that there is a coupling "sweet spot" that exists someplace between too much and too little.

Experiment 1: Moderate Coupling

H4 suggested that performance would be faster using the moderately coupled system than with the static system. H4 was supported in that participants in Experiment 1 completed the complex tasks significantly faster under the automatic camera condition (*M*=462.5s, *SD*=153.4) than under the static camera condition (*M*=680.6s, *SD*=258.6) (t(11)=2.66, p<.05). For the simple tasks, the static camera condition (*M*=250.3s, *SD*=45.6) was significantly faster than the automatic camera condition (*M*=313.9s, *SD*=95.4) (t(11)=-2.47, p<0.05), but by a much smaller margin. The likely reason for this was that the objects in the simple tasks could be described more quickly with words, and trying to use the visual information may have slightly hurt performance. When the task was complex, however, the moderately coupled system improved performance time by over 30%, which is a substantial gain.

One hypothesized reason (stated in H5) for this improvement was that appropriately provisioned visual information should reduce the number of conversational utterances needed to describe the lexically complex task components. When we compare the coded transcripts of the complex tasks, we find that participants used significantly more utterances in referring to specific individual Lego pieces in the moderately coupled condition than in the static condition (M_{Static} = 68.36*, SD*= 30.7*; $M_{Moderately\ Coupled}$* =38.55*, SD*=8.41*),* t (10) = 2.89, p < .05. The same was true for the number of utterances describing the position or orientation of individual pieces (M_{Static} = 75.55*, SD*= 22.85*; $M_{Moderately\ Coupled}$* =45.18*, SD*=12.42), also by a statistically significant margin t(10) = 3.57, p < .01.

For example, one pair clearly used the visual information in determining the proper location of a piece in this example:

Helper: Okay, the darker one and place it on the edge of the black piece on the right side and the smaller side face down.

Worker: (moves the piece)

Helper: Yeah, exactly right, yeah No, no, not on …
Worker: On, on this side here, on the red side?

Helper: Yeah, this side.

This is in contrast to the same pair using the static system, where the information was not as useful. Note how the worker asks a complete question and is not interrupted by the helper, who is not aware of exactly what the worker is doing:

Helper: So it's a dark gray piece and it's upside down. And the triangle piece….

Worker: Sorry, it's upside down?

Helper: Huh?

Worker: You said it's upside down?

Helper: Yeah, there's two gray pieces. There's one with two.

> **Worker**: There's one with one hole on the bottom and two, sort of things sticking out on the top?
>
> **Helper**: Yeah, that's the one you want.
>
> **Worker**: That's the one, okay.

These results suggest strongly that some coupling is useful when compared with no coupling at all, and that our moderately coupled system performed well from an error standpoint.

Experiment 2: Tight Coupling.

H6 suggested that performance in the moderately coupled condition would be faster than the tightly coupled condition. As can be seen in Figure 5, the data support this hypothesis, but only for the complex tasks. Performance time for the simple task was nearly identical in the two conditions ($M_{Moderately\ Coupled}$ = 346.81, *SD* = 94.45; $M_{Tightly\ Coupled}$ = 348.88, *SD*= 31.65) and did not differ by a statistically significant margin, t (15) = -.05, n.s..

For the complex tasks, however, performance in the moderately coupled condition (*M* =524.06, *SD* = 166.82) was significantly faster than the tightly coupled condition (*M* = 664.13, *SD* = 228.89), t (15) = -3.53, p < .01. This suggests that the visual information was more effective when camera movement was moderately coupled to worker hand movement than when it was tightly coupled.

To better understand the nature of this difference, we first turn to H7, which suggested that participants should find the moderately coupled system more useful than the tightly coupled system in the post-task questionnaires. As can be seen in Table 3, the data do not support this hypothesis. There was no statistically significant difference between the two conditions in terms of participants' reported ability to see detail and ability to see where one's partner was working. What is most interesting, however, is that the overall utility for the moderately coupled system (*M* = 4.86, *SD* = 1.17) was rated lower by a slight, but nonetheless statistically significant margin

(*M* = 5.40, *SD* = 1.00), t(31) = -2.08, p < .05. While both systems were rated positively (as they are both above the scale midpoint of 4), it is nonetheless quite interesting that the tightly coupled system was rated to be more useful. This is particularly true given the performance benefit.

One possible reason for this is reflected in participants' ratings of system unpredictability. Here the systems also differed by a small, but also statistically significant margin ($M_{Moderately\ Coupled}$ = 3.76, *SD* = 1.34; $M_{Tightly\ Coupled}$ = 3.06; *SD* = 1.21), t(31) = 2.77, p < .01. This slight difference likely stems from the fact that the moderately coupled system relied on a set of rules (as explained in Appendix A) for changing the visual information, while the tightly coupled system tracked hand motion directly. While the rules for shot change were fairly simple, participants may not have understood them right away, thus rendering the moderately coupled system less predictable. This, in turn, may have influenced their rating of the systems' overall utility. We will return to this issue in the discussion section of the paper.

To further explore these performance and rating differences, we again turned to the transcripts of the experiment sessions, which provided some preliminary answers. We first looked at the overall usage of different statement types in completing complex tasks, as we did in Experiment 1. As can be seen in Table 4, however, there were few differences between conditions. One exception to this was the number of statements used to describe specific pieces, where the two conditions differ marginally significantly (p < .1). We then separated statements by the worker and the helper and considered these separately for all of the categories in Table 4.

When considered separately, there was again a difference between conditions only for the number of utterances used to describe specific pieces. Helpers used, on average, 37.14 (*SD* = 9.72) of these statements in the moderately coupled condition, and 43.29 (*SD* = 8.36) in the tightly coupled condition, t(6) = -1.16, p < .05. This is the only apparent difference between these two conditions in terms of the speech categories described above, and is interesting in several respects.

First, it suggests that the visual information in the tightly coupled condition was less useful for piece identification than for positioning the pieces (because there was no difference in the number of utterances used to describe piece positions). This could be due to the regional nature of the shots in the moderately coupled condition. Rather than constantly trying to keep the worker's hand in the center of the shot, the moderately coupled system used pre-configured shots of each workspace region. This may have meant that the helper was better able to focus on specific pieces because there was less movement, or because the worker could move her hands so as not to block the view without triggering a new shot. It is not entirely clear, however, why this should benefit piece description more than piece position. We discuss this aspect of coupling below, however, as an area for future research.

Even with this explanation, however, an average difference of 6 utterances does not explain the magnitude of the performance time difference. This suggests that particpants may have used the extra time doing other things – not necessarily talking. Some of this time may have been spent asking the workers to move their hands to get shots of the spots they wanted. Some of this time may have been spent simply reacting to a display that was changing much more frequently. If this is the case, however, it would have been subconscious since the two video views were subjectively rated as equally useful overall and the tightly coupled condition was rated more predictable. It is also possible that these factors varied in time in ways that could not be captured using a single questionnaire.

Discussion

While many have considered the role of visual information in the negotiation of common ground by geographically distributed individuals, few have systematically raised the question of how to assess the utility and relevance of visual information. We have presented a preliminary framework for extending Clark and Brennan's (1991) work to include the notions of error and coupling. Error refers to the extent to which visual information is useful and relevant at any given moment. Error relates closely to coupling, which is the extent to which changes in participant behavior result directly and immediately in change to the visual information that is being provided

From a theoretical standpoint, the notion of error allows us to more systematically consider and evaluate the quality of information, which is generally assumed to be perfect according to existing theories. This idea is importantly different from theories of media richness (Daft & Lengel, 1986). Such theories would consider video to be a "rich" medium that necessarily provides more information than a text- or audio-only channel. Without knowledge of what information is needed and the capacity to provide such information, however, even the richest of channels will provide little value, as our results from Experiment 1 clearly illustrate. This problem is exacerbated by the finding in prior work that people tend not to adjust visual information on their own, even when they have the capacity to do so. In other words, allowing participants in Experiment 1 to control the camera in the static condition would not have improved their performance much, since they would not likely have taken advantage of this feature (Fussell et al., 2003; Anonymous, 2006).

Automatically providing visual information, as mentioned earlier, is a difficult problem because it is hard to predict what the helper wants to see at any given moment (Ou et al., 2005). Given that constantly polling the helper is not a practical or desirable option, we must rely on external indicators that provide cues allowing us to approximately discern what information is desired. Experiments with a range of systems (including those described here) have shown that these indicators (e.g., hand position, head position, speech parsing) are inherently imperfect, however, and this imperfection is a key source of error in the process of providing this information, as illustrated by, for example, the problems experienced with a head-mounted camera in prior work. If we assume that all visual information will involve at least some error, our question then becomes one of minimizing that error.

We know from our results presented here and other studies cited above that one source of error is insufficient detail to allow for detailed monitoring of task progress and the establishment of a shared point of focus. This suggests that one way to minimize error is to provide appropriate close-up shots, based on some indicator of what the helper is likely to be focused on. Results from both of our experiments presented here suggest that worker hand position is a good way to approximate the helper's desired focus of attention in a remote repair task. Adjusting camera shots based on worker hand position substantially improved performance and resulted in a system that was perceived by participants to provide useful and relevant information.

At the same time, however, we also know from prior work that too many close-up shots, or moving between close-up shots too quickly can be confusing or jarring to the helper (Fussell et al., 2003; Gaver et al., 1993). The helper may become disoriented or the system may "fall behind" in trying to track motion that is too rapid. Our notion of coupling can aid in addressing these issues. Rather than *constantly* trying to provide optimal visual information (Ou et al., 2005), our results suggest strongly that there is utility in heeding Simon's warning that the cues we use to predict desired information are inherently imperfect. Coupling that is too tight can magnify these imperfections. Moderate levels of coupling mean less sensitivity to imperfection, and visual information that can be more relevant and easier to use.

In our experiments, moderate coupling resulted in improved performance over tight coupling, but this improvement was not reflected in participants' assessment of the system. While both systems were felt to be useful, participants found the moderately coupled system to be slightly less useful and slightly less predictable than the tightly coupled system. This was likely precisely because of the moderate coupling. Not tying visual information change directly to behavior means there will necessarily be times when behavior change does not result in information change. This may be confusing to participants, but this confusion must be weighed against performance benefits from moderate coupling and additional potential confusion from very tightly coupled systems.

It must also be acknowledged that another possible source for this difference in perceived utility has more to do with the details of our moderately coupled system than the notion of coupling itself. It could be, for example, that our system seemed to zoom out in unpredictable ways. While we developed and tested the system iteratively and it showed significant performance benefits in both experiments, it is certainly possible that it could still be improved and this possibility cannot be ruled out.

Limitations/Liabilities

The experimental task has both strengths and weaknesses. Having a consistent set of construction tasks allows for valid comparison across pairs, and the task involves components of many real-world tasks, such as piece selection and placement, and detailed manipulation of physical objects. However, the task is necessarily contrived and relies on a remote helper with limited experience in the task domain. A possible limitation from this is that the helper was relying more heavily on explicit directions than memory, which could impact desired visual information. On the other hand, this limitation is common to many experimental studies.

Also, our task was serial in nature and involved a single focus of worker attention. One could imagine that the worker's hand location would be a less accurate predictor of desired helper focus in a case where there are multiple activities taking place in parallel, or where activity in one region is dependent on information from other regions (e.g., activities in surgery that can take place only when a particular heart rate has been reached, or switchboard repair operations that require knowledge of the state of other circuits). This limitation does not negate these results, but cautions as to the set of domains to which they apply.

Another possible limitation of this work is the effect of the participants having known each other beforehand. It is, of course, possible that participants had a shared vocabulary that would make these results less applicable to pairs of strangers. We considered this and deliberately used abstract, difficult-to-describe Lego pieces and orientations for which participants were unlikely to have a shared language, in order to minimize the effects of the participants' existing relationship.

Future work

First, we have defined an abstract notion of error in the provision of visual information that is conceptually rooted in the much more quantitatively rigorous realm of classic information theory (Shannon, 1948). While we examined several aspects of participant experience that likely correlated with this idea of error, we did not measure error directly. One key area for future research in this area is drawing on this idea in considering ways to more concretely and rigorously conceptualize, measure and minimize error in a range of settings.

Second, we have briefly discussed coupling as a way to consider the relationship between change in the indicators of interesting behavior (that should be visually captured) and the change in the visual information itself. We have used the term “coupling” somewhat loosely, however, in not specifying the precise dimensions on which these changes might be measured. In a sense, our systems can in retrospect be said to have manipulated coupling on two dimensions: time and space. Loosening temporal coupling allowed us to avoid rapid movement by waiting for two seconds after a hand movement before moving the camera. Loosening spatial coupling allowed us to track the hand regionally, rather than constantly try to keep it centered in the video view. Both of these dimensions allowed for more stability in the video view, but we do not know the extent to which each dimension actually had an effect. More research is needed to determine the independent effects of coupling along these two dimensions, and to more broadly map out the space of “coupling” in terms of other possible dimensions (e.g., level-of-detail or zoom, multiple views, etc.).

Third, we have described a task that is serial in nature. It involves a sequence of discrete steps, and only one step is permitted to happen at a time. While these attributes are common to many real-world tasks, they are not universal. More research is needed to determine how these principles apply to scenarios where there are multiple activities occurring in parallel that require attention-shifting by both worker and helper that may be much more difficult to track and anticipate.

References

Anonymous for blind review-a. (2007, October 7-9). Using motion tracking data to augment video recordings in experimental social science research. *Proceedings of Third International Conference on E-Social Science*, Ann Arbor, MI.

Anonymous for blind review-b (2007). Dynamic shared visual spaces: Experimenting with automatic camera control in a remote repair task. *Proceedings of ACM CHI*, San Jose, CA, 1177-1186.

Anonymous for blind review (2006, November 4-8). An exploratory analysis of partner action and camera control in a video-mediated collaborative task. *Proceedings of ACM Conference on Computer Supported Cooperative Work*, Banff, AB, 403-412.

Ballantyne, G. H. (2002). Robotic surgery, telerobotic surgery, telepresence, and telementoring. *Surgical Endoscopy, 16*, 1389-1402.

Brennan, S. (2005). How conversation is shaped by visual and spoken evidence. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world-situated language use: Bridging the language-as-product and language-action traditions* (pp. 95-129). Cambridge, MA: MIT Press.

Clark, H. H. (1992). *Arenas of language use*.Chicago, IL: University of Chicago Pressfield.

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.

Daft, R., & Lengel, R. (1986). Organizational information requirements, media richness and structure design. *Management Science, 32*(5), 554-571.

DeSanctis, G., & Monge, P. (1998). Communication processes for virtual organizations. *Journal of Computer Mediated Communication, 3*(4).

DeVellis, R. F. (2003). *Scale development: Theory and applications*.Thousand Oaks: Sage Publications.

Fussell, S. R., Kraut, R., & Siegel, J. (2000). Coordination of communication: Effects of shared visual context on collaborative work. *Proceedings of ACM Conference on Computer Supported Cooperative Work*, 21-30.

Fussell, S. R., Setlock, L. D., & Kraut, R. E. (2003). Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. *Proceedings of ACM CHI*, 513-520.

Fussell, S. R., Setlock, L. D., Yang, J., Ou, J., Mauer, E., & Kramer, A. D. I. (2004). Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction, 19*, 273-309.

Gaver, W., Sellen, A., Heath, C., & Luff, P. (1993). One is not enough: Multiple views in a media space. *Proceedings of InterCHI Conference*, New York, 335-341.

Gergle, D. (2006). *The value of shared visual information for task-oriented collaboration.* Carnegie Mellon University, Pittsburgh, PA.

Gergle, D., Kraut, R., & Fussell, S. R. (2004). Language efficiency and visual technology: Minimizing collaborative effort with visual information. *Journal of Language and Social Psychology, 23*(4), 491-517.

Karsenty, L. (1999). Cooperative work and shared visual context: An empirical study of comprehension problems in side-by-side and remote help dialogues. *Human-Computer Interaction, 14*, 283-315.

Kirk, D., & Fraser, D. S. (2006). Comparing remote gesture technologies for supporting collaborative physical tasks. *Proceedings of ACM CHI*, Montreal, Canada, 1191-1200.

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction, 18*, 13-49.

Kuzuoka, H. (1992). Spatial workspace collaboration: A sharedview video support system for remote collaboration capability. *Proceedings of ACM CHI*, 533-540.

Miller, J. G. (1978). *Living systems*.New York: McGraw-Hill.

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 ACM CHI*, Amsterdam, The Netherlands, 327-334.

Nunally, J. C. (1978). *Psychometric theory*.New York: McGraw-Hill.

Olson, G. M., & Olson, J. S. (2001). Distance matters. *Human-Computer Interaction, 15*, 139-179.

Ou, J., Oh, L. M., Fussell, S. R., Blum, T., & Yang, J. (2005). Analyzing and predicting focus of attention in remote collaborative tasks. *Proceedings of Proceedings of the 7th International Conference on Multimodal Interfaces*, 116-123.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47*(1), 1-24.

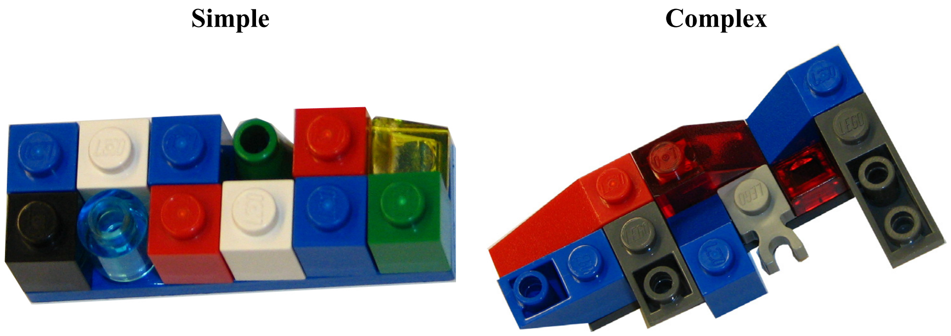
Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379-423, 623-656.

Simon, H. (1996). *The sciences of the artificial*.Cambridge, MA: MIT Press.

Weick, K. (1982). Management of organizational change among loosely coupled elements. In P. Goodman (Ed.), *Change in organizations*.San Francisco, CA: Jossey Bass.

Whittaker, S. (2003). Things to talk about when talking about things. *Human-Computer Interaction, 18*, 149-170.

Zuiderent, T., Ross, B., Winthereik, R., & Berg, M. (2003). Talking about distributed communication and medicine. *Human-Computer Interaction, 18*, 171-180.



Simple

Complex

Figure 1. Sample Lego objects. Each of these objects represents one layer of one column.

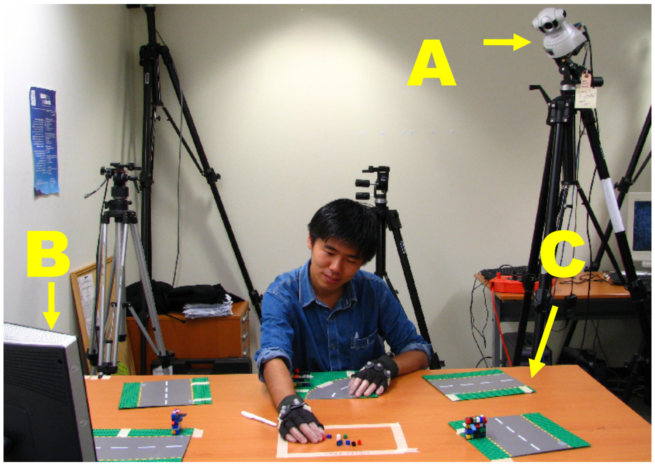


Figure 2. Worker's space showing position of the camera (a), the monitor (b) and workspace (c) on the desk

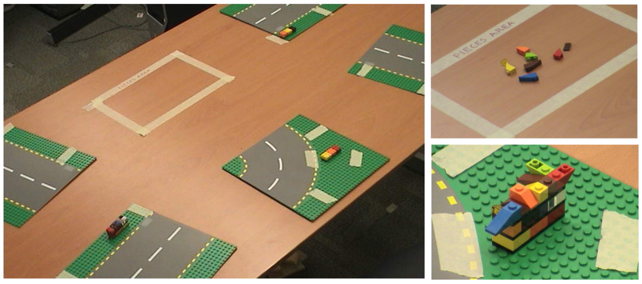


Figure 3. Left: Wide shot of the workspace, Right: Example close-up shots (Top: pieces region, Bottom: work region)

Table 1. Comparison of participants' questionnaire assessment of the two experimental conditions in Experiment 1 (N=24).

	Static/Loosely Coupled Camera		Moderately Coupled Camera	
Variable	Mean	SD	Mean	SD
Pair Performance*	5.8	.6	6.0	.6
Individual Performance**	5.4	1.0	5.7	.7
Ability to see details**	3.1	1.4	5.9	1.4
Utility of video view**	2.9	1.2	5.2	1.2
Awareness of Partner Location	5.5	1.3	5.7	.9
Difficulty of Learning	5.6	1.2	6.0	.7

Notes: Asterisks indicate statistically significant mean differences as follows: * p < .1; ** p < .05. All items used 7-point Likert scales.

Table 2. Mean frequencies of utterance types in completing the complex tasks in Experiment 1 (N = 11 groups)

	Static/Loosely Coupled Camera		Moderately Coupled Camera	
Variable	Mean	SD	Mean	SD
Statements**	110.27	43.11	73.82	15.61
Questions*	42.27	20.99	16.45	9.95
Piece References*	68.36	30.70	38.55	8.41
Piece Position**	75.55	22.85	45.18	12.42
Acknowledgement of Behavior*	7.18	8.75	15.36	6.77

Notes: Asterisks indicate statistically significant mean differences as follows: * p < .05; ** p < .01.

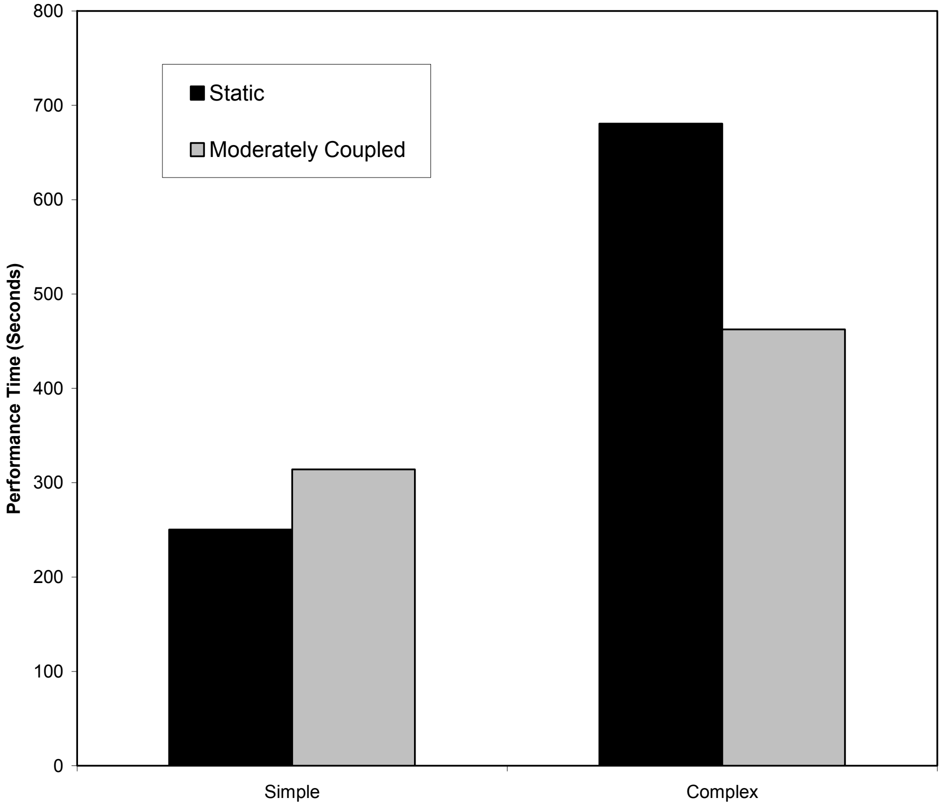


Figure 4. Mean performance time by condition for both task types for Experiment 1.

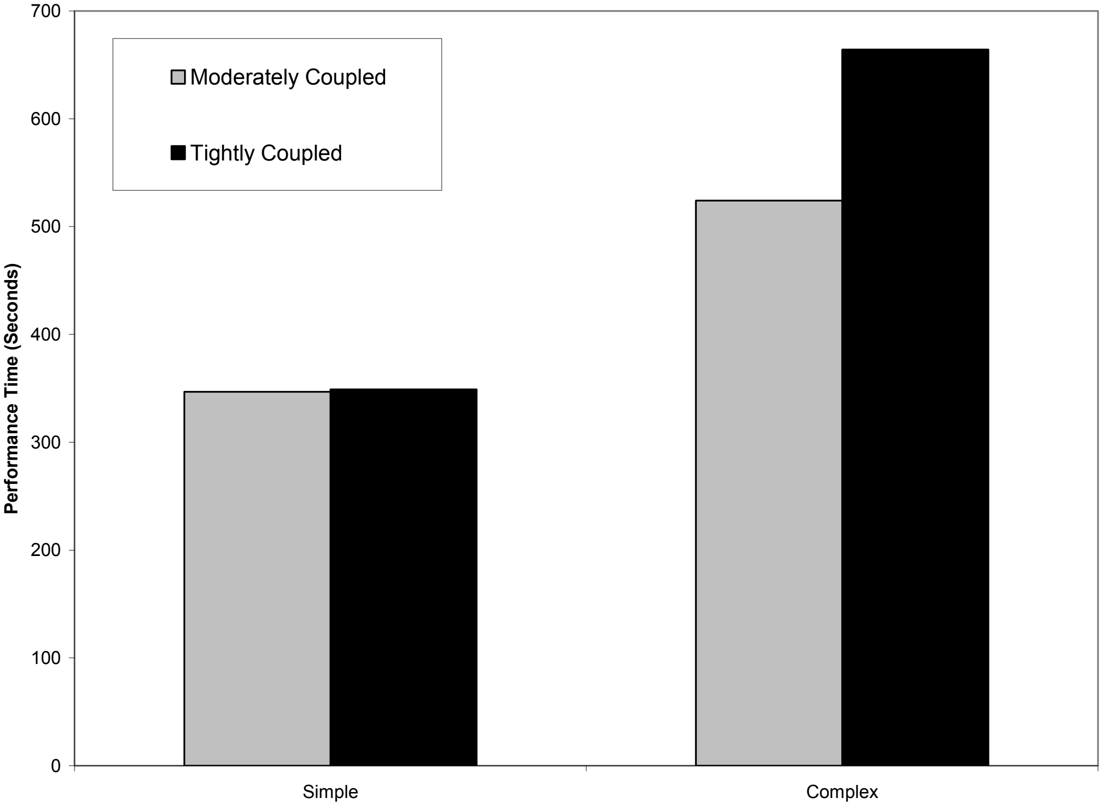


Figure 5. Mean performance time by condition for both task types for Experiment 2.

Table 3. Comparison of participants' questionnaire assessment of the two experimental conditions in Experiment 2 (N=32).

	Tightly Coupled Camera		Moderately Coupled Camera	
Variable	Mean	SD	Mean	SD
Pair Performance	6.09	.86	6.31	.5
Individual Performance	5.69	.65	5.72	.76
Ability to see details	5.25	1.42	5.66	1.15
Utility of video view*	5.40	.97	4.87	1.17
Awareness of Partner Location	5.90	.84	5.78	.88
Difficulty of Learning	2.11	.98	2.16	.88
Unpredictability*	3.06	1.20	3.76	1.34

Notes: Asterisks indicate statistically significant mean differences as follows: * p < .05. All items used 7-point Likert scales.

Table 4. Mean frequencies of utterance types in completing the complex tasks in Experiment 2 (N = 7 groups)

	Moderately Coupled Camera		Tightly Coupled Camera	
Variable	Mean	SD	Mean	SD
Statements	84.00	19.72	96.29	20.97
Questions	25.14	13.28	27.71	11.38
Piece References*	46.71	16.42	53.43	12.01
Piece Position	57.00	17.18	64.71	14.67
Acknowledgement of Behavior	5.29	5.41	4.86	3.39

Notes: Asterisks indicate statistically significant mean differences as follows: * p < .1

Appendix A – Camera Control System Rules

In these rules, the work region location of the worker's dominant hand is called the "current work region," and the previous work region location is the "previous work region." These are both distinct from the "pieces region," which is referred to by this name.

There were four possible movement types and each resulted in a unique system response:

1. *Movement:* The dominant hand enters a "current work region" that is different from the "previous work region."

   System Action: Go to the overview shot.

   Rationale: Moving to a new region meant that the helper was likely to need awareness information about where the worker was now located in the overall space.

2. *Movement*: The dominant hand stays in the "current work region" for at least 3.5 seconds after *Movement 1*.

   System Action: Show close-up of current work region.

   Rationale: Close-up of a work region shown only after it has been selected for construction and to avoid quickly changing views during the region selection process.

3. *Movement:* The dominant hand moves to a "current work region" that is identical to "previous work region" (e.g., returning after a move to the pieces region).

   System Action: Immediately move to close-up of the current work region.

   Rationale: Moving from the pieces area to a work area typically indicated that detailed work was about to occur.

4. *Movement:* The dominant hand moves to the pieces region and stays there for at least 2 seconds.

   System Action: Show close-up shot of the pieces region.

   Rationale: In prior work, most moves to the pieces region were extremely brief and having the camera simply follow the hand was confusing due to quickly changing views. It is only when the hand lingers in the pieces area that a close-up is required. The exact wait time of 2 seconds was decided after several pilot trials and on the basis of data from prior work (Anonymous, 2006).