HUMAN–COMPUTER INTERACTION, 2010, Volume 25, pp. 261–287

ISSN: 0737-0024 print / 1532-7051 online
DOI: 10.1080/07370024201050014 6

Taylor & Francis
Taylor & Francis Group

Providing Dynamic Visual Information for Collaborative Tasks: Experiments With Automatic Camera Control

Jeremy Birnholtz,[1,2] Abhishek Ranjan,[2] and Ravin Balakrishnan[2]
[1]Cornell University
[2]University of Toronto, Canada

One possibility presented by novel communication technologies is the ability for remotely located experts to provide guidance to others who are performing difficult technical tasks in the real world, such as medical procedures or engine repair. In these scenarios, video views and other visual information seem likely to be useful in the ongoing negotiation of shared understanding, or common ground, but actual results with experimental systems have been mixed. One difficulty in designing these systems is achieving a balance between close-up shots that allow for discussion of detail and wide shots that allow for orientation or establishing a mutual point of focus in a larger space. Achieving this balance can be difficult without disorienting or overloading task participants. In this article we present results from two experiments involving three automated camera control systems for remote repair tasks. Results show that a system providing both detailed and overview information was superior to systems providing only one or the other in terms of performance but that some participants preferred the detail-only system.

1. 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 example is the consulta-

Jeremy Birnholtz is interested in improving the usefulness and usability of collaboration technologies through a focus on human attention; he is an Assistant Professor in the department of Communication and Faculty of Computing and Information Science at Cornell University and holds an appointment in the Department of Computer Science at the University of Toronto. **Abhishek Ranjan** is a Senior Software Engineer, specializing in Computer Vision applications, at CognoVision Solutions Inc. **Ravin Balakrishnan** is interested in human–computer interaction, information and communications technology for development, and interactive computer graphics; he is an Associate Professor of Computer Science at the University of Toronto.

CONTENTS

tion 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, might be able to provide assistance in performing a procedure (Ballantyne, 2002; Zuiderent, Ross, Winthereik, & Berg, 2003). In the case of repairing a NASA space-based facility, for example, 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) referred to as "talking about things," where by "things" he means physical objects or artifacts being discussed and used in performing a task. When engaged in these conversations, it can be useful for the remote expert (the "helper") to have a visual image of the workspace where the task is actually being performed by the "worker" (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 shared understanding, or common ground (Clark, 1992; Clark & Brennan, 1991).

1.1. 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 similar visual information and can refer to this information in the grounding process.

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—using tasks such as bicycle repair, puzzles, and toy robot 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 "lexically complex," that is, they are difficult to describe or distinguish. Examples of lexically complex objects include the similarly colored Tartan plaid patterns used in Gergle et al.'s (2004) puzzle studies. Visual information is less useful, however, 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.

1.2. Establishing a Shared Point of Focus

Fussell et al. (2000) pointed out that one key function of visual information in the grounding process is establishing a joint focus of attention. This serves to ensure that both parties understand precisely which object is being discussed and can then go on to discuss its properties or how it might be manipulated. Using visual information in this way requires that the visual display must (a) contain an object of interest or possible interest to participants and (b) allow for the object to be visually distinguished from others by both the helper and the worker. For the helper to do this, sufficient detail is required to allow the helper to visually inspect the objects and determine which is the correct one. Several systems have been developed to satisfy these constraints.

In their study of participants completing a bicycle repair task, Fussell et al. (2000) experimented with a head-mounted video camera. This camera was mounted on the worker's head such that the camera was trained on whatever the worker was looking at. Despite satisfying both of the aforementioned constraints, however, this system did not yield performance benefits. One likely reason for this is that the camera moved too much. Indeed, the shot changed every time the worker's head moved, even slightly. This was confusing to the helper, who did not know if a change in shot was intended to re-establish joint focus on a new object or was the result of an inadvertent movement of the worker's head. Thus, changing the shot too often can make it difficult to settle on a fixed point of focus.

An approach that addresses this problem is to allow for explicit indication of the shared point of focus within the video frame, either via gesturing with one's hands or drawing on the video (Fussell et al., 2004; Kirk & Fraser, 2006). These approaches, although they do allow for one object to be distinguished from another to establish a joint point of focus, however, may not allow for zooming in and out, thus not affording sufficient detail to determine which object is the correct one when there are several to choose from. In other words, gesturing allows the helper to point at or designate objects to the worker but may not allow for them to be examined in great detail by the helper.

To allow for visual inspection by the helper when necessary, another approach is to allow the helper to select between multiple views or to control a pan-tilt-zoom camera located in the workspace. In theory, this allows the helper to select the video stream that is most useful at any given moment. In practice, however, several studies have suggested that such systems are underused by helpers, who seem to find it easier to adapt their conversation (e.g., by asking more clarification questions) to the visual information already available, even in cases where minor video adjustments could clearly have significant benefit to them (Fussell, Setlock, & Kraut, 2003; Ranjan, Birnholtz, & Balakrishnan, 2006). Thus, dynamically updating visual information is a viable approach, but we cannot rely on the participants themselves to do the updating.

1.3. Monitoring and Awareness

In addition to establishing a shared point of focus, visual information can also be useful in allowing the helper to monitor the worker's progress on the task at hand, as

well as to facilitate awareness of where in the workspace the work is taking place. These two functions have overlapping but distinct requirements (Fussell et al., 2000).

Monitoring progress requires the ability to see what the worker is doing in sufficient detail that corrections can be made if problems are observed. Sometimes this merely means determining that work is taking place in the correct area of the workspace, which can be achieved with a wide overview shot of the entire space (Fussell et al., 2003). At other times, however, it may mean ensuring that detailed work is being done correctly, in which case a close-up may be necessary. Awareness of worker location, on the other hand, typically requires a wide shot of the entire space.

Consider how these requirements are met by existing approaches. A camera mounted on the worker's head allows the helper to focus in detail on whatever it is that the worker is doing. This facilitates detailed monitoring, but fails to provide overview information. In other words, this view is always relative to the worker location, and no location information is provided that allows the helper to situate the view relative to a fixed view or coordinate space (Fussell et al., 2003).

This can be addressed, as discussed earlier, by providing multiple camera views. Multiple views allow for detailed monitoring, plus a wide shot for awareness. Even when users do take advantage of multiple available views, however, it can be difficult to figure out how multiple views fit together (Gaver, Sellen, Heath, & Luff, 1993).[1]

This leaves us with something of a paradox. Studies have shown that visual information is useful for grounding in remote help tasks (Gergle et al., 2004; Karsenty, 1999), but few systems developed for providing visual information in physical tasks have shown tangible performance benefits. This suggests that visual information is useful but that there are logistical design issues that must still be overcome to provide visual information in a useful and meaningful way.

2. THE PRESENT STUDIES

To conduct the experiments described next, we developed and tested a system that addresses many of the aforementioned issues with prior systems and allowed us to further explore theoretical aspects of providing visual information in collaborative remote repair tasks.

Based on the prior findings discussed here, we designed a system that met the following constraints: (a) It did not require the helper or worker to manipulate the camera or select between shots; (b) it provided detailed shots for establishing joint focus and monitoring detailed progress, but also provided wide shots to allow for location awareness; and (c) it provided views that made it easy to determine how different shots fit together relative to the global space.

[1]This is because of the nature of video. Multiple video cameras provide relative views of an essentially undefined global space. This lack of formal definition makes it hard to specify what is in view. For a more detailed discussion of these issues and possible emerging solutions, see Birnholtz et al. (2007).

To meet the first constraint, multiple views could be provided either by attaching a camera to the worker's body, so the shot would change with no effort (i.e., as with the head mounted camera) on the worker's part, or by somehow automating control of camera movement or shot selection. Attaching a camera to the worker's body, however, did not seem to be an appropriate solution, because it could so easily result in camera movement even when a shot change was not desirable. Detaching the camera from the worker, on the other hand, would allow us to change (or not change) shots independent of the worker's behavior. We therefore decided to develop an automated camera control system that followed a set of simple principles for providing useful visual information.

To address the second constraint, we needed two pieces of information: (a) some knowledge of the worker's likely point of visual focus in the workspace, and (b) some way to determine when a close-up would be more useful than a wide shot and vice versa.

Given that remote repair tasks are typically performed by the worker using their hands, we thought that worker hand location would be a reasonable indicator of their current point of focus. This was confirmed via a study using a human-operated camera control system in which hand position was also tracked (Ranjan et al., 2006). To determine worker hand location, we tracked their dominant hand using a Vicon infrared optical motion capture system, using techniques described by Birnholtz et al. (2007). We acknowledge that such infrared tracking would be impractical outside of the laboratory, but we note that this was a simple way for us to achieve this goal using available technology. We expect that computer vision technology will improve such that inexpensive tracking via traditional cameras will be possible.

To know when a shot change would be useful, we divided the workspace into six discrete regions. A fixed close-up shot of each region was programmed into the camera's preset memory, as well as a wide shot that showed all six regions. When the worker's dominant hand moved from one region to another, the shot changed to one of the newly entered region. To avoid confusing or jarring camera movements when the worker's hand moved to a region only very briefly (e.g., as experienced in the head-mounted system described by Fussell et al., 2000), however, we built in a wait time before the shot would change. Trial and error led us to set this delay parameter at 2 sec. Moreover, to allow for the helper to know the location of the newly-entered region, we showed the wide shot for 2 sec each time a new region was entered. When movement was repeated between two regions, however, the wide shot was not shown. For a more detailed discussion of this, see Appendix A.

Finally, to meet the constraint of facilitating easy helper understanding of how the various views fit together, we used a single camera that sat in a fixed location. We put this camera above the worker's shoulder, thus roughly replicating the worker's perspective on the workspace for easy discussion of task components. The single camera meant that the movement of the camera could also be used as a resource by the helper in determining worker location in the space.

We ran two experiments, each using a full-factorial 2×2 within-participants design to compare the performance of pairs of participants on a series of Lego con-

struction tasks at two levels of lexical complexity, and using two systems for providing visual information.

3. EXPERIMENT 1

In Experiment 1, we compared a static scene camera with a detail-plus-overview system.

Following from Gergle's (2004) findings, we expected that visual information would be more useful in performing lexically complex tasks. We included both simple and complex tasks, but mainly expected to see effects for the complex tasks.

3.1. Hypotheses

Performance

First, we were interested in the extent to which the detailed visual information we provided was used in establishing a shared point of focus. As discussed in detail previously, a shared point of focus should facilitate the grounding process (Clark, 1996; Fussell et al., 2000). As a significant component of remote repair tasks involves grounding, faster grounding should be indicated by faster task performance. Detailed visual information should also result in more accurate performance of tasks, because the helper can more easily visually monitor their performance (Clark, 1996; Fussell et al., 2000). Thus, we hypothesized the following:

H1a: Performance will be faster when detailed visual information is provided than when only overview information is provided.

H1b: Participants provided with detailed visual information will make fewer errors than those who are provided only with overview information.

Perceptions of the System

In addition to performance measures, we also wondered about participants' perceptions both of their performance and of the utility of the system. We hypothesized the following:

H2a: Participants will perceive their performance to be better when detailed visual information is provided than when only overview information is provided.

H2b: Participants will find the visual information to be more useful when detailed visual information is provided than when only overview information is provided.

Language Usage

To better understand the detailed nature of participants' interaction and use of the visual information, we also had expectations regarding language usage. First, visual

information in the grounding process can facilitate reduced ambiguity about what specific objects or properties are being discussed (Clark, 1996). This means that participants will have to ask fewer questions to clarify what is being discussed (e.g., "Are you talking about the red one or the blue one?"). Shared focus should allow also allow for more reference (including deictics) to visual information in conversations describing objects (Barnard, May, & Salber, 1996; Gergle et al., 2004), and the need for fewer verbal acknowledgments of what is taking place on screen (Boyle, Anderson, & Newlands, 1994). Moreover, Jackson, Anderson, McEwan, and Mullin (2000) found that less detailed visual information may cause people to use longer, more detailed object descriptions and be generally more cautious. Thus, we hypothesize the following:

H3a: Participants will ask fewer questions of each other when detailed visual information is provided than when only overview information is provided.
H3b: Participants provided with detailed visual information will use fewer references to task components and their locations than those who are provided only with overview information.
H3c: Participants provided with detailed visual information will use more deictic terms than those provided with only overview information.
H3d: Participants provided with detailed visual information will use more acknowledgements of onscreen behavior than those provided with only overview information.

3.2. Methods

Participants

There were 24 participants (6 female) aged 19 to 33 ($M = 26$, $SD = 5$) in Experiment 1. All participants were recruited via posted flyers and e-mail 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 each.

Task and Setup

The experiment task was for the worker to use Lego bricks to construct three multilayer "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 worker also had an LCD video monitor showing what the helper saw. The monitor was located across the table from the worker's seat (see b in Figure 2).

The helper was seated in front of a 20-in. LCD monitor and given a paper map of the workspace indicating which regions the columns were to be built in. Discrete regions were used in this task in order to replicate real-world tasks in which activities must take place in specific locations (e.g., parts of the body in surgery).

FIGURE 1. Sample Lego objects.

Note. **Each of these objects represents one layer of one column.**

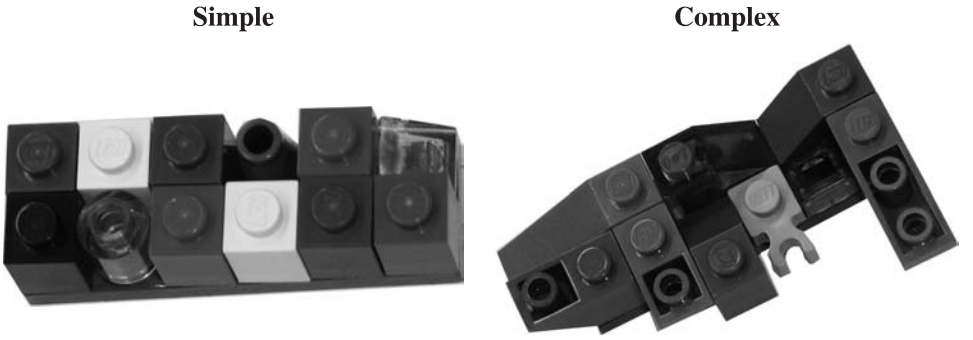
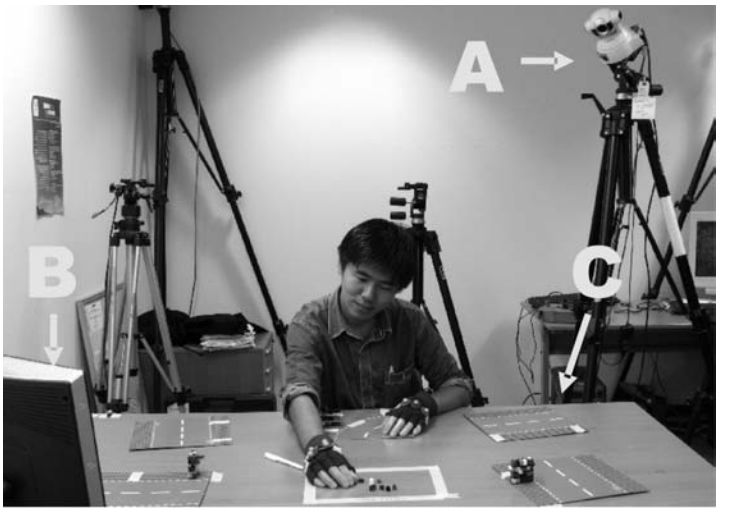


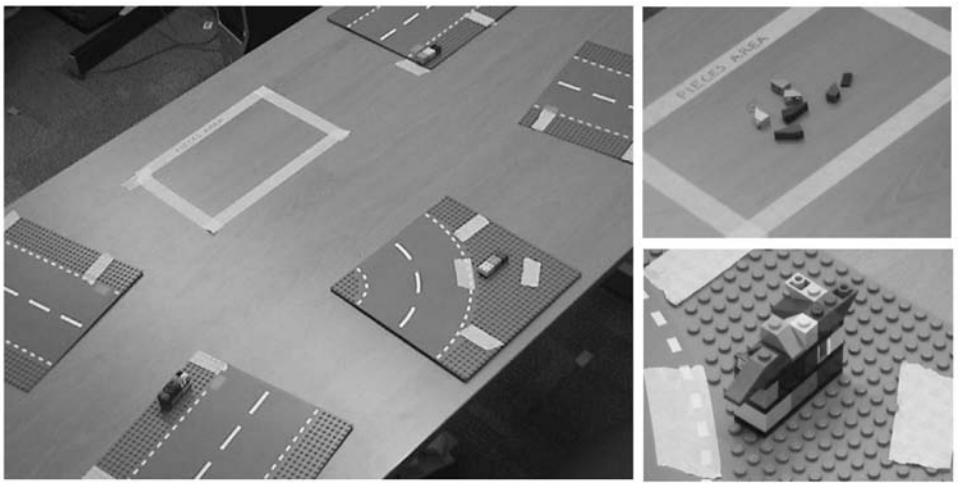
FIGURE 2. Worker's space showing position of the camera (a), the monitor (b), and workspace (c) on the desk.



Each of the columns constructed by the worker consisted of four layers—two involved "identification" tasks and two involved "construction" tasks. The identification tasks are described in detail in Ranjan, Birnholtz, and Balakrishnan (2007) but are not the focus of this article.

In the construction tasks, workers were provided with individual Lego pieces for one layer (all three columns) at a time. This meant between 27 and 36 individual pieces, which were always placed in the "pieces area." The columns were built in separate work regions (see Figure 2). For the lexically simple task, each layer consisted of 9 to 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 (see Figure 1) and therefore harder to describe.

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



Helpers were provided with an exact duplicate of each completed layer, one at a time (i.e., they were given the completed objects shown in Figure 1). The goal was for the helper to instruct the worker in constructing each layer, which included identifying pieces and placing them correctly. Workers 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 they were not permitted to move on until all errors had been corrected.

Experimental Conditions

Participant performance in Experiment 1 was measured in two experimental conditions.

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

Detail Plus Overview Automatic Camera System. We compared performance against the automated detail-plus-overview camera control system described previously and in Appendix A. This system was configured to provide detailed shots of each region, as well as a wide shot of the entire workspace (see Figure 3).

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 completed practice tasks to ensure that they understood the details of the task and how the camera control system worked. When they used the automated system, 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 with the systems.

The order of tasks and conditions was randomized according to a Latin Square design. 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.

Data Analysis

All sessions were video recorded for analysis. All sessions of Experiment 1 were fully transcribed, with one exception due to technical problems with the recording. The complex tasks from each were then 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 spatial and temporal diectics, as well as deictic pronouns referring to objects. This coding scheme includes labels for statements that cannot be classified in any of the stated categories for each attribute. Deictic personal pronouns (e.g., "you") were not included because they did not refer to the task, so could distort the results. Each transcript was coded by at least one of two independent coders, with 15% of them coded by both coders. Agreement between coders on those coded by both coders was better than 90% and disagreements were resolved via discussion.

Individual questionnaire items from the postexperiment questionnaires were aggregated into four constructs (see Appendix B). Each construct consisted of two to five related items. Cronbach's alpha for these constructs ranged between .7 and .9, which is considered adequate for social science research (Nunally, 1978). Confirmatory factor analyses were also performed to ensure that all items loaded onto a single factor (DeVellis, 2003). For system perception results, only the helper questionnaires were used because the workers did not use the system directly.

To statistically analyze performance, word count, and questionnaire data, we used a series of repeated measures analysis of variance models in which the variable of interest was entered as a two-level within-subjects factor, and condition order was a between-subjects factor. In these models, the main effect for experimental condition is

the key point of comparison, whereas a main effect of complexity can be interpreted as a manipulation check. Where there is an interaction between complexity and condition, separate analyses are reported for simple and complex tasks. Where there are interaction effects with the order in which conditions were experienced, these are reported and discussed. Unless noted otherwise, all upcoming analyses use this model specification.

3.3. Results

Manipulation Checks

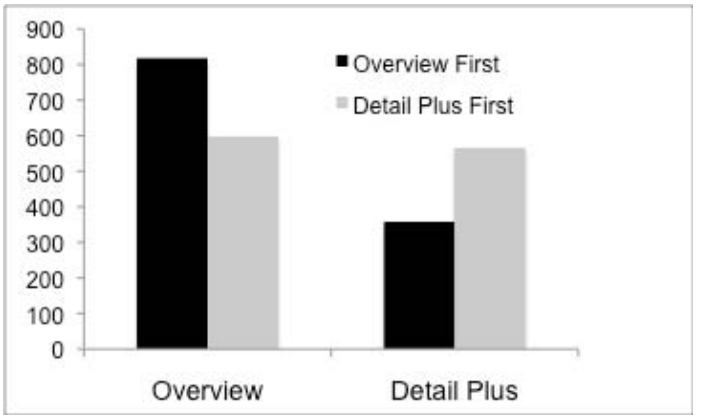
To ensure the validity of our experimental manipulations, we first compared performance times for simple and complex tasks overall and found that simple tasks (M = 292.67, SD = 52.88) were completed significantly faster than complex tasks (M = 585.47, SD = 173.52), $F(1, 11) = 38.2$, $p < .001$. We next assessed whether participants perceived a difference between the two camera control conditions. In the post-experiment questionnaire, we asked participants several questions (see Appendix B) about whether they could see information in sufficient detail to complete the task, which is an indicator of whether the manipulation was noticed. Helpers agreed much more strongly with questionnaire items reflecting their ability to examine objects in sufficient detail in the detail-plus-overview camera condition (M = 5.68, SD = 1.31) as compared with the overview-only condition (M = 2.19, SD = 1.18), $F(1, 9) = 48.24$, $p < .001$.

Performance

We were first interested in participant performance. Hypothesis 1a posited that performance would be faster when detail and overview were provided than when only overview information was provided. Camera control condition clearly had an impact on performance of both simple and complex tasks, but there was a significant interaction between condition and difficulty, $F(1, 10) = 22.00$, $p < .01$. For simple tasks, there was main effect of condition, but it was not in the hypothesized direction. The detail-plus-overview system (M = 319.00 sec, SD = 53.12) was slower than the overview-only system (M = 266.33 sec, SD = 53.12), $F(1, 10) = 10.72$, $p < .01$.

For complex tasks, the difference was as hypothesized. Participants completed tasks significantly faster in the detail-plus-overview condition (M = 462.53 sec, SD = 132.39) than in the overview only condition (M = 708.41 sec, SD = 295.01), $F(1, 10) = 1.03$, $p < .01$. The interaction term (Order × Camera Condition) was statistically significant, $F(1, 10) = 9.86$, $p < .05$. This interaction, however, supports our hypothesis (see Figure 4) that the detailed shots will be used as a resource in grounding. The figure shows that the performance difference between conditions for people who used the detail plus overview condition first was much smaller ($M_{Detail\ plus\ overview}$ = 598.67, SD = 170.70; $M_{Overview}$ = 566.88, SD = 152.00) than the difference between conditions for people who used the overview-only system first ($M_{Detail\ plus\ overview}$ = 818.16, SD = 365.28; $M_{Overview}$ = 358.19, SD = 50.33). This suggests that those who used the detail-plus-over-

FIGURE 4. Interaction effect between performance time and condition order.



view system first were able to use the system to negotiate common ground quickly and then draw on this common ground in doing a similar task using the overview-only system. Those who used the overview-only system first, on the other hand, had no such resource to draw on and it took them longer to negotiate common ground initially. In other words, our system was particularly useful for those encountering the task for the first time.

What is further interesting in Figure 4 is that performance using our system by those who had used the overview-only system first was actually the fastest. This result suggests that those who spent more initial time negotiating common ground may have developed a fast way to refer to task components verbally. This was then possibly enhanced by having detailed visual information, hence their faster performance. It is also possible, of course, that they did not use the visual information provided by the detailed system and that their verbal system of referring to components was simply faster.

Hypothesis 1b was about errors, which are another way to measure performance effectiveness. Participants made a total of seven mistakes that were corrected only when pointed out by the experimenter. Six of these seven were in the overview-only condition, suggesting further that the detailed-plus-overview information improved participant performance.

Perceptions of the System

Hypotheses 2a and 2b referred to participant perceptions via questionnaire. All results use a 7-point Likert scale, with higher scores indicating greater agreement, as presented in Figure 5.

Hypothesis 2a stated that participants would feel better about their performance when detailed visual information was provided than when only overview information was provided. As the table shows, the data support this hypothesis (M_{detail} = 5.97, SD =

FIGURE 5. Comparison of helpers' questionnaire assessment of the two experimental conditions in Experiment 1.

	Overview		Detail Plus Overview	
Variable	*M*	*SD*	*M*	*SD*
Pair Performance*	5.70	.69	5.97	.41
Individual Performance**	5.02	1.04	5.59	.71
Ability to see details**	2.59	1.18	5.68	1.31
Utility of video view**	3.21	1.38	5.06	1.24
Awareness of Partner Location	5.89	.90	5.84	.74
Difficulty of Learning	5.50	1.16	6.05	.88

Note. $N = 11$. All items used 7-point Likert scales.
Asterisks indicate statistically significant mean differences as follows: *$p < .1$. **$p < .05$.

.41; $M_{Overview} = 5.70$, $SD = .69$) by a small but statistically significant margin, $F(1, 10) = 2.05$, $p < .01$.

Hypothesis 2b stated that participants would find the visual information to be more useful when detailed visual information was provided than when only overview information was provided. Again, the table shows that the data support this hypothesis as well ($M_{Detail} = 5.06$, $SD = 1.24$; $M_{Overview} = 3.21$, $SD = 1.38$), $F(1, 10) = 1.91$, $p < .01$. Note also that the utility of the overview is rated 3.21, which is below the neutral point on the scale. This result suggests that participants generally did not find the overview information useful, whereas they did generally find the video from the detail-plus-overview system to be useful.

Language Usage

To better understand the differences just observed, we examined the transcripts. Here we focus specifically on the transcripts of the complex tasks, as these are where the most interesting differences were observed in terms of grounding behavior.

First, Hypothesis 3a stated that participants would ask fewer questions of each other when detailed visual information was provided than they will when only overview information was provided. As Figure 6 shows, this hypothesis was supported. In the detail-plus-overview condition, participants asked a mean of 16.45 questions of each other ($SD = 9.95$), as contrasted with 42.27 ($SD = 20.99$) in the overview-only condition, $F(1, 9) = 20.48$, $p < .01$. There was a significant interaction between condition order and number of questions, $F(1, 9) = 13.13$, $p < .01$. This interaction follows the same pattern seen earlier in the discussion of performance times. Those who used the detail-plus-overview system first asked far fewer questions ($M = 23.2$, $SD = 10.96$) in their first attempt at the task than those who used the overview system first ($M = 54.17$, $SD = 19.46$). This indicates that the detail-plus-overview system was useful in grounding when participants first encountered the task. When the second tasks are compared, those who used the overview-only system first asked fewer questions ($M = 10.8$, $SD = 4.3$) when using the detail-plus-overview condition than did participants who used the detail-

FIGURE 6. Mean frequencies of utterance types in completing the complex tasks in Experiment 1.

	Overview		Detail Plus Overview	
Variable	*M*	*SD*	*M*	*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
Deictic Pronouns	21.45	13.14	22.45	20.94

Note. N = 11 groups.
Asterisks indicate statistically significant mean differences as follows: *$p < .05$. **$p < .01$.

plus-overview condition first, when they used the overview only condition (M = 28.00, SD = 12.70).

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.

Next, Hypothesis 3b stated that participants provided with detailed visual information would make fewer references to task components and their locations than those who are provided only with overview information. This hypothesis was supported for both of these variables. For piece references, participants in the detail-plus-overview condition used fewer statements that described pieces (M = 68.36, SD = 30.7) than did those in the overview-only condition (M = 38.55, SD = 8.41),

$F(1, 9) = 15.40$, $p < .01$. Again, the interaction between condition order and piece references was statistically significant, and the pattern was the same as described previously, $F(1, 9) = 12.83$, $p < .01$.

For statements about piece location, fewer statements were used by participants in the detail-plus-overview condition ($M = 45.18$, $SD = 12.42$) than by those in the overview-only condition ($M = 75.55$, $SD = 22.85$), $F(1, 9) = 19.40$, $p < .01$. The interaction term was also significant and followed the same pattern, $F(1, 9) = 8.24$, $p < .05$.

Hypothesis 3c stated that participants provided with detailed visual information would use more deictic terms than those provided with only overview information. This hypothesis was not supported by these data, as can be seen in Figure 5.

Hypothesis 3d stated that participants provided with detailed visual information would use more acknowledgments of behavior than those provided with only overview information. This hypothesis was supported by the data ($M_{Detail} = 15.36$, $SD = 6.77$; $M_{Overview} = 7.18$, $SD = 8.75$), $F(1, 9) = 9.17$, $p < .05$. The interaction term was not significant.

4. EXPERIMENT 2

In Experiment 2, we compared the detail-plus-overview system with one providing only detailed information. In addition to detailed visual information for establishing a shared point of focus, providing overview information allows for monitoring and awareness. In assessing our approach to camera control, we wanted to be sure we were providing appropriate overview information in addition to the detailed information. This was the focus of Experiment 2.

4.1. Hypotheses

Performance

We were first interested in whether providing overview information contributed to pair performance. As previously noted, overview information should allow helpers to track where in the workspace the work was taking place. The ability to do this should facilitate grounding with regard to task location, and we expected this to impact performance time:

H4: Participants provided with detailed plus overview information will perform faster than those provided with only detailed information.

Perceptions of the System

We also expected information detail to impact participants' perception of the utility of the visual information provided and of their performance.

H5a: Participants provided with detailed plus overview information will feel better about their performance than those who are provided with only detailed information.

H5b: Participants provided with detailed plus overview information will perceive the video to be more useful than those who are provided only with detailed information.

Language Usage

As in Experiment 1, we were interested in language use differences.

We expected differences in the number of questions being asked and verbally answered because questions and statements would be one way that location in the workspace could be determined (e.g., "Which region are you in?"). We therefore hypothesized the following:

H6a: Participants provided with detailed plus overview information will ask fewer questions than those provided only with detailed information.
H6b: Participants provided with detailed plus overview information will use fewer statements than those provided only with detailed information.

4.2. Methods

Participants

There were 32 participants (15 female) aged 19 to 29 ($M = 22$, $SD = 2$) in Experiment 2. All participants were recruited via posted flyers and e-mail 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 each.

Task and Setup

The task and basic setup in Experiment 2 were identical to those used in Experiment 1, except that participants completed only construction (and not identification) tasks.

Experimental Conditions

In Experiment 2, we compared the detail-plus-overview camera system used in Experiment 1 with a detail-only automated system.

Procedure

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 experiment. 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.

Data Analysis

All sessions were video recorded for analysis. All sessions of both experiments were fully transcribed, with two exceptions due to technical problems with the recording equipment. The complex tasks from the transcripts were then coded exactly as in Experiment 1. Quantitative data were analyzed as previously described.

4.3. Results

Manipulation Checks

To check the validity of our manipulation, we first looked at the number of discrete camera movements in each condition, as one of our reasons for a region-based tracking system was that it would result in more stable shots and fewer camera movements. There was a clear difference between conditions, with a mean of 734.50 (SD = 308.71) discrete movements in the detail-only condition, and 134.31 (SD = 55.30) in the detail-plus-overview condition, $F(1, 11) = 55.35$, $p < .001$.

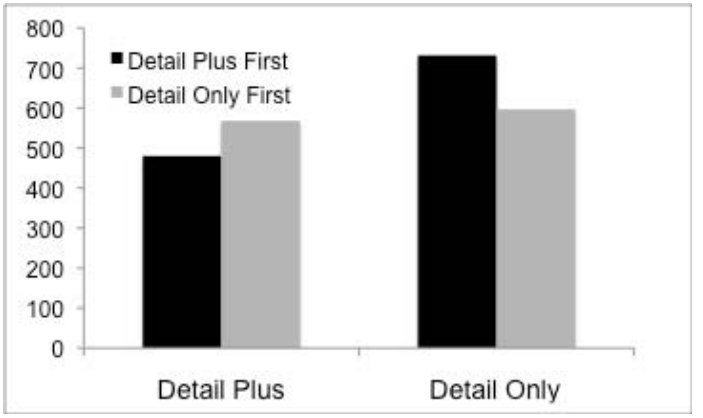
We also looked at participants' perceived ability to tell where in the workspace their partner was working, as this was one of the intended differences between the experimental conditions. Despite the clear technical difference in the number of shot changes, however, there was not a statistically significant difference between conditions on this dimension ($M_{Detail\ plus\ overview} = 5.78$, $SD = .88$; $M_{Detail\ only} = 5.90$, $SD = .84$). There are several possible reasons for this, however, as discussed next, and we do not believe this to be a debilitating threat to the validity of the manipulation. We then checked for a difference in participants' perceived ability to see detail in the two conditions. As both conditions allowed for seeing detail, we did not expect there to be a significant difference, and this was the case ($M_{Detail\ plus\ overview} = 5.66$, $SD = 1.15$; $M_{Detail\ only} = 5.25$, $SD = 1.42$).

Performance

Hypothesis 4 stated that participants provided with detailed plus overview information would perform faster than those provided with only detailed information. For the simple tasks, Hypothesis 4 was not supported. On average, participants in the detail-plus-overview condition (M = 346.81 sec, SD = 94.46) completed the tasks in about the same amount of time as those using the detail-only system (M = 348.88 sec, SD = 126.59), $F(1, 14) = .01$, $p > .05$. There was a statistically significant interaction between condition order and condition, $F(1, 14) = 13.51$, $p < .01$. Participants consistently performed better in their second attempt at the task (regardless of condition) than they did in the first, suggesting that prior experience was a stronger predictor of performance than the availability of visual information for simple tasks.

For the complex tasks, Hypothesis 4 was supported by the data, as participants using the detail-plus-overview system completed the task in 524.06 sec on average (SD = 166.82), as compared with those using the detail-only system (M = 664.12 sec, SD = 228.89), $F(1, 14) = 24.43$, $p < .001$. As Figure 7 shows, there was also a significant interac-

FIGURE 7. Interaction effect for performance times with condition order.



tion between performance time and task order, $F(1, 14) = 15.43$, $p < .01$. As in Experiment 1, the difference in performance between conditions for those who did the detail-plus-overview condition first ($M_{Detail\ plus\ overview} = 480.25$ sec, $SD = 217.25$; $M_{detail\ only} = 731.63$ sec, $SD = 266.02$) was greater than the difference in performance between conditions for those who did the detail-only condition first ($M_{Detail\ plus\ overview} = 567.88$ sec, $SD = 89.71$; $M_{detail\ only} = 596.62$ sec, $SD = 176.30$). Moreover, the detail-only system was substantially worse for those who used the detail-plus-overview system first. This suggests that the detail-plus-overview system provided them with useful information that they became accustomed to, and this was missing when they moved on to the detail-only system.

Perceptions of the System

Next, Hypothesis 5a stated that participants provided with detail-plus-overview information would feel better about their performance than those provided with only detailed information. Support for this hypothesis is weak but nonetheless worth exploring. There was a marginally significant difference between the two camera control conditions ($M_{Detail\ Plus\ Overview} = 6.36$, $SD = .43$; $M_{Detail\ Only} = 6.04$, $SD = 1.10$), $F(1, 13) = 3.62$, $p < .1$. This suggests that participants felt slightly better about their performance when using the detail-plus-overview system. There was also a marginally significant interaction between perceived performance and condition order, $F(1, 13) = 3.26$, $p < .1$. Looking carefully at this interaction shows that those who used the detail-plus-overview system first had a larger difference between their perceived performance in the two conditions (difference in means = –.71), as compared with those who used the detail-only system (difference in means = .14). This suggests that those who used the detail-plus-overview system first felt worse about their performance when they moved on to the detail-only system, whereas those who used the detail-only system felt about the same about their performance when they moved on to the detail-plus-overview system. Thus, people who learned to do the task using the detail-plus-overview system seemed to like having that information.

Hypothesis 5b stated that participants provided with detailed plus overview information would perceive the video to be more useful than those who are provided only with detailed information. This hypothesis was not supported by the data. Rather, those using the detail-only system ($M = 5.52$, $SD = .85$) rated that system as more useful than the detail-plus-overview system ($M = 4.94$, $SD = 1.04$) by a marginally significant amount, $F(1, 14) = 3.29$, $p < .1$.

Language Usage

Hypothesis 6a predicted that participants provided with detail-plus-overview information would ask fewer questions than those provided only with detailed information. The data provide limited support for this hypothesis, as the number of questions using the detail-plus-overview system ($M = 16.15$, $SD = 10.62$) was less than the number in the detail-only condition ($M = 26.38$, $SD = 13.64$) by a marginally significant amount, $F(1, 11) = 4.21$, $p < .1$.

Hypothesis 6b stated that participants provided with detailed plus overview information would use fewer statements than those provided only with detailed information. This hypothesis was not supported by the data, as there was no statistically significant difference between the two conditions ($M_{Detail\ plus\ overview} = 125.38$, $SD = 33.43$; $M_{Detail\ only} = 112.77$, $SD = 27.08$), $F(1, 11) = .83$, $p > .1$.

Given that both conditions in Experiment 2 provided participants with detailed information, we did not expect the elements of discussion of detailed elements (piece selection and placement) to differ significantly. Nonetheless, we tested for differences in these categories. As Figure 8 shows, however, there were no statistically significant differences.

5. DISCUSSION

5.1. Theoretical Implications

From a theoretical standpoint these results have several implications for our understanding of the role of visual information in the grounding process. We have

FIGURE 8. Mean frequencies of utterance types in completing the complex tasks in Experiment 2.

	Detail Plus Overview		Detail Only	
Variable	*M*	*SD*	*M*	*SD*
Statements	125.38	33.43	112.77	27.08
Questions*	16.15	10.62	26.38	13.64
Piece References	44.62	14.97	43.00	9.51
Piece Position	47.77	15.43	61.69	21.22
Acknowledgment of Behavior	9.08	10.15	10.38	9.23

Note. N = 14 groups.
Asterisks indicate statistically significant mean differences as follows: *$p < .1$.

shown that visual information can be useful in the performance of lexically complex remote repair tasks using physical objects in the real world (i.e., not an on-screen task). Moreover, we have shown that having some detailed visual information was clearly more useful in this task than having only overview information. This suggests that participants used the information both to establish a joint focus of attention and to monitor detailed progress on tasks. These findings were supported by differences in performance time, error frequency, word usage, and perceptions of the utility of the visual information. Thus, the grounding process was improved by the presence of visual information but only when this information was sufficiently detailed as to allow for differentiation of specific task components.

This is distinct from the lexically simple tasks, where the availability of detailed visual information actually worsened performance by a slight but nonetheless statistically significant margin. In these cases, participants used the visual information when it was not necessary to do so; they could more easily have described the pieces verbally and ignored the video. They may also have been distracted by the information. This largely confirms Gergle's (2006) finding that visual information is most useful in tasks that are lexically complex. We extend these findings by suggesting preliminarily that visual information can also impede or constrain the grounding process when it is not necessary.

We also cannot ignore the interaction effect between performance times and condition order in Experiment 1. The time data suggest that participants used the visual information in their initial negotiation of common ground but did not necessarily refer to this information later. Although more detailed transcript analysis is necessary to see precisely how the visual information was used in different phases of the task, such an interpretation would be consistent with Newlands, Anderson, and Mullin's (2003) findings that people adapt their communication strategy to available information. Our findings would extend this idea in that they seemed to stick with these strategies, even when more information became available.

We were also interested in the extent to which overview information was useful in combination with detailed information. Fussell et al. (2003) and others have suggested that visual information can also be useful in monitoring task progress and maintaining awareness of what is going on in the workspace, and we believed that the combination of detailed with overview information would be more useful than detailed information alone. To test this we compared our detail-plus-overview camera control system to one that provided only detailed information. Here, the performance time results supported our hypothesis, but there were few differences in terms of participants' perceptions of system utility, and word usage in performing the tasks.

We believe there are two possible reasons for this. One is the structure of the task, which consisted mostly of piece identification and placement (detailed work), and less frequent movement between regions for which the overview information would be more useful. Thus, the overview information was useful less frequently than was the detailed information. It is possible that the performance time difference captures the utility of the detail-plus-overview approach but that there simply were not enough instances of region change to be captured by the coding and word count mea-

sures from the transcripts. It is also possible that the relatively small number of regions (six, including the pieces area) meant that tracking the worker's location was a "lexically simple" task component. In other words, it may be that participants were able to track and identify the regions verbally and did not always need video of them. This interpretation is supported somewhat by the lack of difference on the 'ability to know partner location' questionnaire item.

A second possible reason is that there are other attributes of our designs that influenced the results. We compared two specific design solutions for automated camera control, but these are obviously not the only possible designs for approaching this problem. It is possible, for example, that a region-based, detail-only system could have had different results, as could a detail-plus-overview system that had a different time delay before changing shots. More experiments are needed to fully tease apart the relationships between these variables and their impact on design and theory.

Moving forward, we believe there is a need to theoretically account for the lexical simplicity or complexity of a task, and the extent to which visual information is likely to help, rather than hinder, the grounding process.

5.2. Implications for Design

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, Oh, Fussell, Blum, & Yang, 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 such indicators (e.g., hand position, head position, speech parsing) are inherently imperfect, however. Nonetheless, our results in Experiment 1 clearly show that hand tracking was a useful way to discern what visual information would be useful to the helper, in that it provided a substantial performance benefit over less-detailed information. Thus, one key design implication is that hand tracking is a useful indicator of worker attention in a task using physical components in the real world. This would obviously be different in a task that did not involve work with the hands, or where the helper needed to monitor multiple detailed activities beyond the immediate area where the helper is working.

We noted earlier that one of the advantages of using hand tracking is that, as contrasted with a head-mounted camera, our camera was not physically attached to the worker's body. This means that worker body movement need not necessarily result in camera movement. This disconnection, or "decoupling," can be exploited to avoid jarring shot changes that might result from quick or inadvertent worker movements. In our detail-plus-overview system, we approached this two ways: (a) by implementing a 2-sec delay following any rapid worker hand movement before the corresponding camera movement occurred, and (b) by using workspace regions for tracking, rather than moving the camera following every hand movement. This latter feature allowed for more stable shots (i.e., shots that moved less frequently due to small hand move-

ments). This approach had clear benefits over providing no detailed information in Experiment 1.

In Experiment 2, we compared the same detail-plus-overview system to one that provided only detailed information, and did so without the 2-sec delay or regional tracking. As previously noted, there was a performance benefit here for the detail-plus-overview system. This was only supported by the performance time data, however, and not by the other performance and word count measures. From a design standpoint, there are a few possible reasons for this. One is that the relatively constrained task space meant that there simply were not enough jarring or distracting movements of the camera to really make a difference. Another is that simply removing the camera from the worker's body resulted in sufficient decoupling so as not to be too distracting, in that the speed of camera movement and actual camera location are then under system control (and subject to constraints such as the speed of servo motors, etc.) rather than subject to whimsical movements of the worker's body.

5.3. Limitations and 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, and S. R. Fussell (personal communication) reported that there were few differences in her studies between those performed with "professional helpers" and those using ordinary student participants.

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.

5.4. Future Work

These results point to several directions for future research both on understanding what visual information is useful in collaborative remote repair tasks, and on automating the provision of this information in different contexts.

One clear next step involves further experiments to understand the balance between the role of detailed and overview information in the grounding process, and when each is desirable. Our results suggest that both are useful, but we cannot discern how much of each is desirable. It seems that detailed information was more useful than overview information, but this may also have been an artifact of our task specifications. Additional experiments in which shot-by-shot analyses of transcripts and the use of specific visual elements would be useful in advancing our understanding of this.

Another area for additional work is in understanding the nature of the relationship, or coupling, between worker attention focus or body movement and camera movement. Results here experimented with decoupling on both spatial (i.e., using regions) and temporal (i.e., with a delay parameter) dimensions. Our design, however, did not allow us to tease these dimensions apart to understand them better. Additional experiments could utilize systems that differ on these dimensions to better understand the utility and significance of each approach.

NOTES

Acknowledgments. We thank John Hancock, Xiang Cao, Clarissa Mak, Serena Kao, Tom Byuen, and members of the Dynamic Graphics Project Lab for their assistance with this research.

Authors' Present Addresses. Jeremy Birnholtz, Department of Communication, Faculty of Computing and Information Science, Cornell University, 310 Kennedy Hall, Ithaca, NY 14853. E-mail: Jpb277@cornell.edu. Abhishek Ranjan, Department of Computer Science, University of Toronto, 10 King's College Circle, Toronto, Ontario, Canada. E-mail: aranjan@dgp.toronto.edu. Ravin Balakrishnan, Department of Computer Science, University of Toronto, 10 King's College Circle, Toronto, Ontario, Canada. E-mail: ravin@dgp.toronto.edu.

HCI Editorial Record. Received June 2, 2008. Revisions received April 30, 2009, and December 11, 2009. Accepted by Steve Whittaker. Final manuscript received April 12, 2010.

REFERENCES

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

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

Birnholtz, J., Ranjan, A., Balakrishnan, R. (2007, October 7–9). *Using motion tracking data to augment video recordings in experimental social science research.* Paper presented at the Third International Conference on E-Social Science, Ann Arbor, MI.

Boyle, E. A., Anderson, A. H., & Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language and Speech, 37*, 1–20.

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 Press.

Clark, H. H. (1996). *Using language*. New York, NY: 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.

DeSanctis, G., & Monge, P. (1998). Communication processes for virtual organizations. *Journal of Computer Mediated Communication, 3*(4). doi:10.1111/j.1083-6101.1998.tb00083.x

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

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

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 the ACM 2003 Conference on Human Factors in Computing Systems*. New York, NY: ACM.

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 the ACM 1993 Conference on Human Factors in Computing Systems*. New York, NY: ACM.

Gergle, D. (2006). *The value of shared visual information for task-oriented collaboration*. Unpublished doctoral dissertation, 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,* 491–517.

Jackson, M., Anderson, A. H., McEwan, R., & Mullin, J. (2000) Impact of video frame rate on communicative behaviour in two and four party groups. *Proceedings of the ACM 2000 Conference on Computer Supported Cooperative Work*. New York, NY: ACM.

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 the ACM 2006 Conference on Human Factors in Computing Systems*. New York, NY: ACM.

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 shared view video support system for remote collaboration capability. *Proceedings of the SIGCHI 1992 Conference on Human Factors in Computing Systems*. New York, NY: ACM.

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 ACM 1993 Conference on Human Factors in Computing Systems*. New York, NY: ACM.

Newlands, A., Anderson, A. H., & Mullin, J. (2003). Adapting communicative strategies to computer-mediated communication: An analysis of task performance and dialogue structure. *Applied Cognitive Psychology, 17*, 325–348.

Nunally, J. C. (1978). *Psychometric theory*. New York, NY: 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 the 7th International Conference on Multimodal Interfaces*. New York, NY: ACM.

Ranjan, A., Birnholtz, J. P., & Balakrishnan, R. (2006). An exploratory analysis of partner action and camera control in a video-mediated collaborative task. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. New York, NY: ACM.

Ranjan, A., Birnholtz, J., & Balakrishnan, R. (2007). Dynamic shared visual spaces: Experimenting with automatic camera control in a remote repair task. *Proceedings of the ACM Conference on Human Factors in Computing Systems*. New York, NY: ACM.

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.

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 (Ranjan et al., 2006).

APPENDIX B. QUESTIONNAIRE ITEMS

FIGURE B-1. Questionnaire Items

Item	Variable
My partner and I completed this task effectively	Perceived Performance
My partner and I completed this task faster than most people could.	Perceived Performance
My partner and I communicated well in completing this task.	Perceived Performance
This task would have gone more smoothly if my partner and I were in the same place.	Perceived Performance (Reverse-coded)
My partner was effective in doing what he/she needed to do for us to complete this task.	Perceived Performance
I was able to examine objects in great detail.	Ability to Examine Objects in Detail
I was able to tell all of the Lego pieces apart by looking at the video screen.	Ability to Examine Objects in Detail
I relied primarily on the video view and not conversation with my partner, to tell pieces apart.	Video Utility
Most of the time, I saw exactly what I wanted on the video screen.	Video Utility
I had no trouble seeing what I needed to see in completing this task.	Video Utility
When the camera changed shots, it usually changed to something I wanted to see	Video Utility
It was hard to tell where in the workspace my partner was working	Partner Awareness
I could usually tell what my partner was doing.	Partner Awareness
I usually knew where in the workspace my partner was working	Partner Awareness
There were times when I had no idea what my partner was doing.	Partner Awareness (reverse coded)