Dynamic Shared Visual Spaces: Experimenting with Automatic Camera Control in a Remote Repair Task

Abhishek Ranjan[1], Jeremy P. Birnholtz[2], Ravin Balakrishnan[1]

[1] Department of Computer Science
University of Toronto
www.dgp.toronto.edu
aranjan, ravin@dgp.toronto.edu

[2]Knowledge Media Design Institute
University of Toronto
www.kmdi.utoronto.ca
jeremy@kmdi.utoronto.ca

ABSTRACT
We present an experimental study of automatic camera control in the performance of collaborative remote repair tasks using video-mediated communication. Twelve pairs of participants, one "helper" and one "worker," completed a series of Lego puzzle tasks using both a static camera and an automatic camera system that was guided in part by tracking the worker's hand position. Results show substantial performance benefits for the automatic system, particularly for complex tasks. The implications of these results are discussed, along with some lessons for the use of motion tracking as a driver for camera control.

Author Keywords
Camera control, computer-supported cooperative work, collaboration, video mediated communication, video conferencing, motion tracking, empirical studies.

ACM Classification Keywords
H.5.3 Group and Organization Interfaces – *Computer-supported Cooperative Work*

INTRODUCTION

There is a range of settings in which expert assistance may be required by a novice who is completing a complex real-world task. Experts are not always physically close, however, so there is increasing interest in the use of collaboration technologies for tasks such as surgery in remotely located hospitals [2, 21], repair of equipment in remote locations (e.g., aircraft engines), and operation of scientific equipment [8, 17].

In the development of technologies to support these tasks, there is growing evidence to suggest the importance of providing the remote expert (the "helper") with visual information, often in the form of a video view of the workspace where the physical task is being performed by the "worker" [9, 19]. This shared visual context can be used to facilitate the negotiation of "common ground" in the ongoing conversation between the helper and worker [6].

Providing this shared visual context, however, can be difficult when the task involves the detailed manipulation or identification of objects in specific but disparate locations in a work area. In surgery, for example, a detailed activity may occur in multiple areas of a patient's body. In such scenarios, fixed-view "scene cameras" provide a useful overview, but little detail [9], while a camera mounted on the worker's head can provide greater detail, but constrains the helper's view to what the worker is focusing on [10]. While it is possible to simultaneously provide detail and overview by allowing the helper to control the camera or select between multiple shots (a dynamic shared visual space), this has been shown to be potentially distracting, confusing and time-consuming [10, 12].

An alternative approach proposed by Ou et al. [24, 25] is to automate the provision of dynamic visual information by predicting what the helper will want to see. However, Ranjan et al. [27] showed that workers behaved differently under different camera control conditions. This suggests that a purely predictive approach may not be as effective as a hybrid one that attempts basic prediction as well as exploits the expected adaptation by the worker. In this paper, we build on prior work by exploring the basic premise that worker hand position is a reasonable indicator of the helper's desired visual information. We develop an automatic camera control system based on this premise, and provide empirical evidence indicating that it is highly effective in certain types of tasks when compared to a fixed camera.

BACKGROUND AND RELATED WORK

Providing Shared Visual Context

Shared visual context has been shown to play an important role in the completion of a range of collaborative tasks, such as toy robot construction and on-screen puzzles [5, 6, 15, 18]. In particular, prior work has shown that a shared

CHI 2007, April 28–May 3, 2007, San Jose, California, USA.

visual space facilitates the negotiation of common ground, or a level of shared understanding of what is being discussed in a conversation between two or more parties [4]. Fussell et al. [9] point out that, in completing collaborative tasks, people rely on visual cues in the grounding process for monitoring task status, monitoring people's actions, establishing a joint focus of attention, formulating messages and in monitoring the comprehension of their partner.

Video systems necessarily constrain the range of cues that are available to do these things as compared with a face-to-face environment, but have nonetheless been shown to be more useful than audio-only systems in completing collaborative tasks [19]. This is particularly true when the task in question is lexically complex – that is, when it involves elements that are difficult for participants to describe verbally, as was the case with the tartan plaid patterns used in Gergle's puzzle studies [13].

Applying Fussell et al.'s [9] framework, there are two task components in a puzzle-style task for which this visual information could be useful. First is the identification of difficult to describe pieces. Second is in their placement, when placement requires detailed manipulation or difficult-to-describe orientation.

Moreover, work in this area has found that while there is typically not a strong need to use visual cues to monitor partner comprehension, this may be different if some component of the task requires face monitoring [23] or if users do not share linguistic common ground [30]. In most cases, however, video images of the shared workspace are more valuable than images of ones partner's face. Thus, the most valuable cues seem to be those used for monitoring partner actions, task status, and establishing a joint focus of attention.

Static vs. Dynamic Visual Spaces

Shared visual spaces can be either static or dynamic in nature. Static visual spaces provide the helper with a fixed view of the worker's work area, typically via an overhead or over-the-shoulder camera view [10]. Static visual spaces can be effective for monitoring the progress of tasks that take place in a constrained workspace, or that do not require very detailed observation of task actions or joint focus on minute details. The capacity to establish a joint focus of attention can be augmented somewhat via systems that facilitate gesturing [11, 16], but these do not allow for zooming in for detail.

Dynamic visual spaces, on the other hand, provide a range of views to the helper, either via a movable camera [20, 26, 27] or cutting between shots from multiple cameras [10, 12]. Fussell et al. [9] also experimented with head-mounted cameras that were consistently focused wherever the worker was looking. This was useful, but substantially constrained the helper's range of view, and did not result in performance benefits over a static visual space. Ranjan et al. [27] experimented with a user-controlled camera, which was also useful, but participants did not use it to a great extent, even when it would have been useful to do so. It therefore resulted in no performance benefits when compared with a static space.

Fussell et al. [10] had similar results with a multiple camera system that allowed helpers to select among several shots. It was potentially helpful, but underused and resulted in no performance benefits. Gaver [12] also experimented with media spaces providing multiple views, and reported that cutting between views could be confusing and distracting.

Automatic Dynamic Visual Spaces

Much of the work mentioned above suggests that there are some clear benefits to dynamic visual spaces, but helpers generally seem unlikely or unwilling to control the camera or select shots themselves. While there are several possible reasons for this, the important point is that it has been shown repeatedly that they do not do it. This suggests two potential approaches: third party human camera operation and automatic camera operation.

Human operators, when they are aware of the task and adept at camera operation, can be effective [27], but they can also be costly in financial terms and must be trained.

With this in mind, there has been substantial recent interest in automatic operation. The goal in this work is to use some cue or combination of cues to predict the helper's desired focus of visual attention at any given moment, and use that prediction to drive camera operation or shot selection. Prior studies have examined automating camera control in lecture rooms and meeting rooms using speaker tracking, detection and cinematography rules [14, 28]. However, automatic camera control has not been explored in the context of a collaborative task as discussed here.

Ou et al. [24], in a task scenario more closely related to the one explored here, use a combination of speech parsing and gaze detection to develop a preliminary predictive model of desired visual information in an on-screen, PC-based puzzle task. More elaborate models relating speech patterns to desired and actual visual information have also been developed [3, 13].

In addition to speech, worker activity is another likely indicator of the helper's desired visual focus. In observing pairs performing a Lego construction task, Ranjan et al. [27] tracked worker motion in a human-operated camera condition. They found a substantial correlation between the worker's dominant hand location in the workspace and the field of view of the operator's camera shot. At the same time, however, they also found that pairs, consciously or not, used the workspace and shared visual space differently when it was dynamic than they did when it was static. This suggests that predicting the helper's desired focus of visual attention is a slippery problem, in that what the helper wants to see at any moment depends, in part, on what the helper can see at that moment.

THE PRESENT STUDY

First, we explore the extent to which the worker's hand position can be used as a predictor of the helper's desired focus of visual attention in a collaborative remote repair task. Second, we are interested in developing insights for the design of automatic systems that have roots in prediction, but that exploit adaptations in user behavior.

Design

We use a full-factorial 2x2 within-participants design to compare the performance of pairs of participants – a "worker" and a "helper" – performing Lego construction and identification tasks at two levels of complexity, and in two camera control (i.e., visual space) configurations:

Static camera: A camera above the worker's left shoulder provided a wide shot of the entire workspace.

Automatic camera: A single pan-tilt-zoom camera was located above the worker's left shoulder. The camera shot was adjusted (described below) based on the position of the worker's dominant hand.

As with the PC-based puzzle tasks used by Gergle [13], these tasks involve elements common to a range of real-world, collaborative remote repair tasks: piece identification, piece movement, piece manipulation and placement, and verification of correct placement.

Hypotheses

With regard to the effect of camera configuration on task performance, we hypothesized that:

1. Participants would complete all tasks faster with the automatic camera than with the static camera.
2. Participants would make fewer errors in the automatic camera configuration than in the static configuration.
3. The benefit of the automatic camera would be greater for lexically complex tasks than for simple tasks.

We also expected differences in satisfaction with the visual information provided and with system experience overall:

4. Participants would be more satisfied with their performance in the automatic camera configuration.
5. Participants would value the automatic camera more for detailed views of pieces than awareness of partner activity in the workspace.

Based on Ranjan et al.'s [27] observations of behavior changes due to camera movements, we also expected differences in worker behavior:

6. Hand movements towards the camera will be less in the automatic camera configuration.
7. The use of the dominant and non-dominant hand will differ significantly across camera conditions, i.e. participants would adapt their behavior [27] depending on the type of camera control provided.

Participants

24 volunteers (6 female, 18 male) participated in the study, ranging in age from 19 to 33, M = 26, SD = 5. All were required to have normal or corrected-to-normal color vision, and to use English as their primary language of communication. Participants were paid $10, and were recruited via posted flyers and email lists at our university.

Setup and Equipment

The helper and worker were located in the same room, so they could hear each other, but separated by a 5-foot-high partition wall. The worker was seated at a desk (Figure 1) divided into 6 discrete regions. Five of these regions, referred to as "work regions," were marked with green Lego base plates. The sixth, referred to as the "pieces region," was where the unattached pieces were placed, with white markings to define its rectangular boundaries.

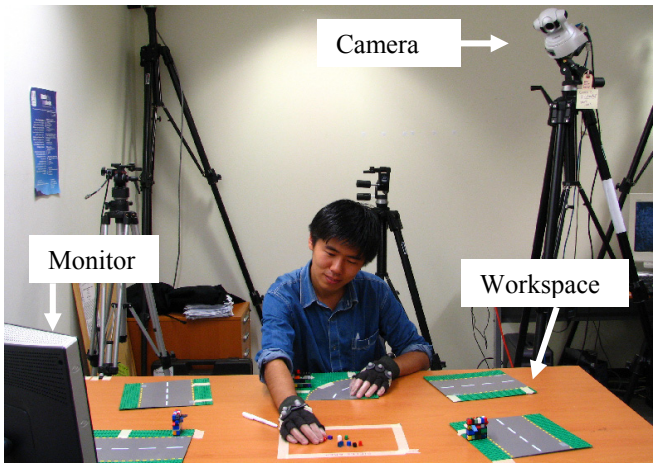


Figure 1. Worker's space showing position of the camera, the monitor and workspace on the desk

Motion Tracking- The workers wore partial-finger gloves (see Figure 1) that had wireless, passive reflective markers attached to them. We tracked the location of these markers with sub-mm precision [1]. Due to very slight shifting of the markers on the gloves themselves, the exact precision of whole-hand tracking was slightly less than this, but still adequate for our purposes

Camera- A Sony SNC-RZ30 pan-tilt-zoom camera was positioned on a tripod 30 cm behind the worker's space, and above the worker's left shoulder. The camera was connected via analog coaxial cable to the worker and helper monitors. The camera was positioned so that it could capture all six regions of the workspace.

Displays- A 20-inch LCD monitor was located 20 cm in front of the worker's desk. It displayed the camera output so that the worker was aware of what the helper could see.

The helper's space consisted of a desk with a 24-inch LCD monitor that displayed the camera output.

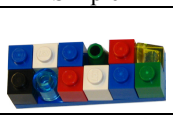
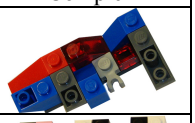
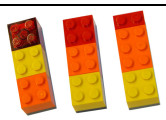
A Sony Mini-DV camcorder was located just outside the worker's space, and recorded all sessions for later analysis.

Task and Materials

The overall task was for the worker to use Lego bricks to construct three four-layer "columns" in specific regions of the workspace, based on instructions from the helper. Helpers were given a paper map of the workspace indicating which regions the columns were to be built in. The columns were built one layer at a time, so a layer in all the columns had to be finished before moving on to the next layer. In order to assess the value of visual information for different tasks, we used two types of tasks in each condition. Two of the layers involved primarily "identification" of difficult-to-describe pieces, while the other two primarily involved "construction," which included detailed placement and manipulation of pieces.

In identification tasks, workers were provided with three similar, but not identical, pre-constructed Lego pieces (see Table 1). Simple identification pieces were composed of three smaller parts. Complex identification pieces were composed of 10-12 smaller parts. Helpers were provided with an exact duplicate of each piece, one at a time. The goal was for the helper to get the worker to pick up the correct piece, and place it in the correct region.

Table 1. Sample Lego column layers

	Simple	Complex
Construction		
Identification		

In construction tasks, workers were provided with several smaller pieces with which to construct the layers of three columns. In the simple construction task, each layer consisted of 10-12 square- or rectangle-shaped pieces. In the 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 here was for the helper to instruct the worker in constructing the next layer of each column, which included identifying pieces and placing them correctly.

Participants were permitted to talk to each other, but 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.

In order to more closely replicate activities (such as the real-world examples mentioned above) where detailed activity must take place in specific, discrete regions of a workspace, workers were not permitted to have more than one unattached piece outside of the pieces area at a time. In other words, construction had to happen in the target region and be completed one piece at a time. It was not acceptable, for example, to lift up the entire column and construct it in the "air space" above the worktable or in the pieces area.

After each camera 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.

Camera Control System

The automatic camera control system was based on data from Ranjan et al. [27] and worked as follows:

The system used two types of camera shots: close-ups of specific regions, and a wide shot of the entire workspace (see Figure 2). There were seven distinct shots that could be selected from: six were close-up views of each of the six regions and one was the overview shot of the workspace. The overview shot was included to allow the helper to see where in the workspace the worker was, to be sure the tasks were taking place in the correct work regions. Close-up shots were included to show detailed views of the construction and pieces as the tasks were underway.

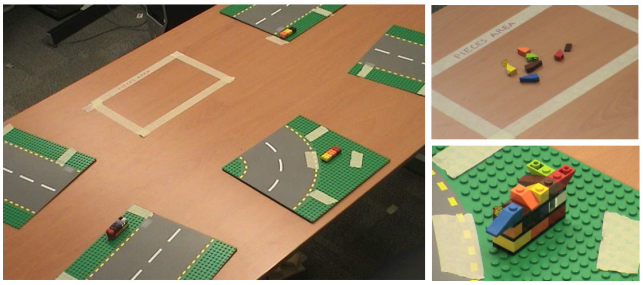


Figure 2. Left: Wide shot of the workspace, Right: Example close-up shots (Top: pieces region, Bottom: work region)

The position of the worker's dominant hand was constantly tracked in 3D using the motion capture system. This information was used in real-time to determine the workspace region in which the worker's hand was located. This, in turn, was used to determine the appropriate camera shot according to the following rules.

In these rules, the current 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, essentially, 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 [27].

Figure 3 shows a state diagram of the automatic camera control. The states represent camera shots and the transitions represent possible movements. These transition rules were developed iteratively, and we experimented with both continuous tracking and discrete, region-based tracking. In the final design, even though the camera moves were guided by continuous movements of the dominant hand, the camera was programmed to make only discrete moves from one preset to another, as opposed to continuously following the hand over the entire workspace. Discrete moves provided stable views of the regions despite significant hand movements inside the region.

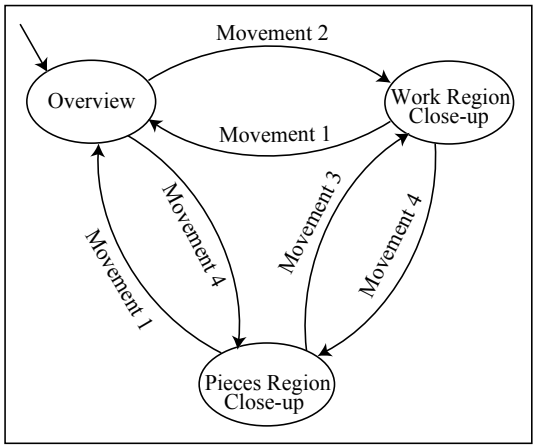


Figure 3. State diagram of the camera control algorithm showing the three camera shots as states and various movements as transitions from one camera shot to another

Procedure

The order of difficulty and camera condition were counterbalanced across all participants.

Participants were randomly assigned (via coin toss) to "helper" and "worker" roles, and were shown to their separate workspaces on arrival. The task was then explained to them, and they were told that their goal was to complete it as quickly and as accurately as possible.

Workers then put on the gloves and participants completed simplified practice identification and construction tasks to ensure that they understood the details of the task.

In the automatic camera condition, the basics of the operation of the system were explained to the participants. They were told that the camera movements were guided by the position of the dominant hand of the worker. They were not given any specific detail of the algorithm controlling the camera. However, as we will discuss later, the participants quickly understood the basic principle behind the automatic camera control, and some consciously made use of this understanding to "manually" control the camera.

The pieces for the first task were then placed in the pieces region, the helper was given the first model block (the duplicate of the piece the worker was to identify or construct, depending on the task) and the workspace map, and the pair was permitted to begin. The completion of each layer, or subtask, was determined first by the participants, who reported to the experimenter when they believed the subtask was complete. If, after examining their work, the experimenter determined that there were no errors, they were permitted to move on to the next subtask. If errors were found, participants were informed that there was at least one error (but not what it was), and required to fix it.

Analysis

Completion Time and Error Analysis

Video of each session was analyzed to track and extract the completion times and the number of errors made. Completion time was defined as the time from start to finish for the complete layer, as reported by the participants. We considered only errors that were in place when the participants reported to the experimenter that they were done. Errors made prior to self-reported completion were not tracked because it was not clear how these should be classified or when one would be considered an error (e.g., if discussed incorrectly or only on final placement).

Where there were errors, the number at the completion of each layer was counted, and the time taken to detect and correct errors was recorded separately.

Motion Capture and Camera Movement Data Analysis

The worker's hand position in 3-D space, along with the camera position and locations of the workspace regions, were recorded once per second for the entire duration of the experiment. All instances of the hands' movements across various regions in the workspace were extracted and

counted. The camera shot selection was also recorded along with the hand positions, so that it could easily be determined whether hand activity was within the camera shot or not.

Questionnaire Data Analysis

Reliability of the questionnaire items was assessed using Cronbach's α, which is a measure of the extent to which a set of scale items can be said to measure the same latent variable [7]. All of the scales used here except one had α values between .7 and .9, which is within the range considered acceptable for well-established scales [22]. The one remaining scale had an α value of .62, which is acceptable for exploratory work. Confirmatory factor analyses indicated that each scale loaded on a single factor.

RESULTS

The study involved two independent task types: identification and construction. Each task had two task complexity levels: simple and complex. Each task was performed under two camera conditions: static and automatic. Two-factor repeated-measures ANOVA models were run separately for the two tasks using task complexity and camera condition as independent variables. Dependent variables were completion time and number of errors.

Participants also filled out questionnaires on completion of each camera control condition. Questionnaire data were analyzed using repeated measures ANOVA models, including each term as a within-participants factor, and participant role (helper or worker) as a between-participants factor to test for interaction effects.

Completion Time

We hypothesized above that the automatic camera condition would result in faster performance for all tasks (Hypothesis 1), but that the benefit would be greater for complex/difficult tasks (Hypothesis 3). For the construction tasks, there was no statistically significant main effect for camera condition on completion time, but a significant interaction was found between camera condition and task difficulty ($F(1,11)=15.41$, $p<0.01$). No significant asymmetric transfer was observed between the two camera conditions.

Paired sample t-tests for completion times showed that participants finished 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) ($t(11)=-2.47$, $p<0.05$). This combination of results supports Hypothesis 3 and suggests that the automatic camera assisted task performance to a greater degree when the task was complex than when it was simple. The left half of Figure 4 shows mean completion times under various conditions for the construction task. The error correction times are shown on top of the bars.

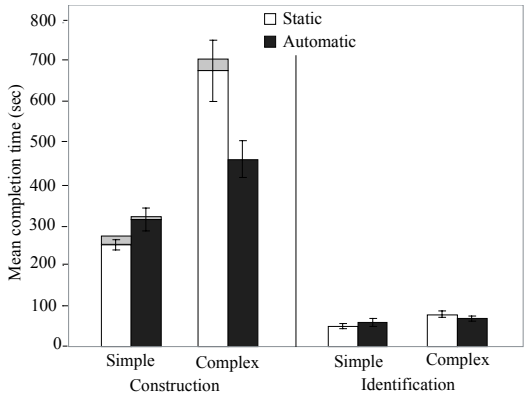


Figure 4. Mean completion time by camera condition for both task types. Error correction times are shown in light gray.

For the identification tasks, there was not a significant main effect for camera condition overall, but there was a significant interaction between task difficulty and camera condition ($F(1, 11)=7.03$, $p < .05$). A trend similar to that in the construction task completion time can be seen here, though paired samples t-tests showed that the result is not statistically significant. It should be noted that identification task completion times are substantially shorter than construction because the task involved fewer discrete steps.

Errors

We were also interested in the errors participants made in performing these tasks, for two reasons. First, a reduced number of errors would suggest that an automatic camera system could be particularly useful in mission-critical settings where errors are costly or fatal [31]. Second, the situations in which participants made errors give us a potentially useful sense of the strengths and weaknesses of both camera conditions.

Only seven errors were detected upon the completion of all subtasks across all pairs of participants, and they were all in the construction task. Six out of seven errors were detected in the static camera condition. This suggests that the automatic camera system enabled participants to perform the tasks more accurately.

This was further reflected in the analysis of the number of dominant hand moves to and from the pieces area, where a larger number of moves in the completion of a task under one camera condition would indicate a larger number 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 camera condition (M=4.66, SD=3.16) than in the automatic camera condition (M=3.54, SD=2.10) ($F(1,9)=3.76$, $p<.1$). These results support Hypothesis 2.

Errors caused by incorrect description or interpretation of color or other piece attributes (e.g., size, shape, markings) are considered piece identification errors. Four out of the six errors detected under the static camera condition were

related to piece identification. This suggests that the additional visual information provided by the automatic camera was particularly useful for focusing on detailed aspects of the task. This is further reflected in the questionnaire results below.

Perceived Performance

Participants evaluated the quality of their performance as a pair, and their individual performance of the tasks. Individuals rated their performance as more effective in the automatic camera than in the static camera condition ($F(1,20)=5.44$, $p<.05$), supporting Hypothesis 4. Moreover, there was a marginally significant interaction between participant role and self-reported individual effectiveness ($F(1,20)=3.95$, $p<.1$). While helpers reported slightly higher performance in the automatic camera condition ($M=5.59$, $SD=.71$) than in the static camera condition ($M=5.02$, $SD=1.04$), there was no such difference for workers.

Somewhat surprisingly, particularly given the performance data presented above, there was only a small and marginally significant difference in perceived pair performance between the two conditions. As can be seen in Table 2, perceived pair performance was slightly higher in the automatic camera condition than in the static camera by a relatively small, but still marginally significant amount ($F(1,20)=3.66$, $p<.1$).

Role of Visual Space

Participants also assessed the utility of both systems, in terms of how useful the video information was in performing the tasks, their ability to examine objects in detail, and their awareness of where in the visual space their partner was working. In all of these cases, workers were assessing the perceived utility of this information to their partners, since they themselves were not relying on the video view.

As Table 2 shows, participants generally did not find the video useful (as the mean rating is below the midpoint on the 7-point scale) in the static camera condition, but did find it to be useful in the automatic camera condition ($F(1,20)=45.86$, $p<.001$). This suggests that there was value in the detailed view provided by the automatic camera condition, but that participants were able to adequately describe things verbally when this view was not available. Combined with the completion time results presented earlier, however, these descriptions seem to have taken longer when the task was complex.

When we consider participants' self-reported ability to examine objects in detail, it is not surprising that they reported that they were substantially less able to do so in the static camera condition than in the automatic camera condition ($F(1,20)=81.04$, $p<.001$).

There was, on the other hand, no statistically significant difference in participants' self-reported ability to know where their partner was in the visual space (or, in the workers' case, their perception of their partner's ability to do so). This supports Hypothesis 5 and suggests that the static camera condition was adequate for providing this information (since both were on the positive end of the Likert scale), and that the main difference between conditions was in participants' ability to examine detailed components of the task objects.

Table 2. Mean ratings and their SD for performance, effectiveness of visual space, and learning under the two camera conditions

	Static Camera		Automatic Camera	
	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.

Ease of Learning

Finally, participants were asked about the ease of learning to use and work with the two systems, where a higher score on this construct indicates an easy to learn system. Again, there was no statistically significant difference between conditions. This, combined with the fact that both mean scores were above the midpoint on the scale, suggests that the automatic camera system was not difficult for participants to learn. It is not surprising that the static camera condition was easy to learn.

User Behavior

We were interested in the extent to which workers' physical movement in the workspace varied across camera control conditions. To do so, we analyzed the motion capture data in which left and right hand positions were tracked for the duration of the experiment. We first examined the vertical height of the worker's hands relative to the workspace. In the static camera condition, holding a piece up towards the camera could be a way to distinguish that piece and provide a sort of primitive 'zoom' capability. If the automatic camera condition was effective, we would expect to see less vertical movement in this condition than in the static camera condition. A repeated-measures ANOVA showed that camera condition had a significant main effect on the worker's mean hand height, with the average hand height lower in the automatic camera condition (M=800mm,

SD=26), than in the static camera condition (*M*=806mm, *SD*=51), ($F(1,11)=9.03$, $p<0.05$). While the difference in means is small (only 6mm), it should be noted that the range of vertical movement is substantially greater in the static camera (Max = 1142mm) than in the automatic camera condition (Max = 664mm). This helps to explain the statistically significant finding and shows that the workers' hands were lifted substantially higher above the workspace in the static camera condition. These results support Hypothesis 6.

We were also interested in user adaptation to the camera control system (Hypothesis 7). We were particularly interested in whether participants used their dominant and non-dominant hands differently in the two camera conditions. While statistical analyses yielded no overall patterns in this regard, one worker did show signs of adaptation and we have analyzed his behavior here.

This participant made 94 dominant hand moves and 31 non-dominant hand moves to the pieces region under the static camera condition, but only 40 dominant-hand moves and 74 non-dominant hand moves under the automatic camera condition. By analyzing the video, we observed that this worker used the dominant hand to keep the camera focused on a particular region by leaving the dominant hand in that region, and using the non-dominant hand to get pieces from the pieces region. This led to more frequent moves of the non-dominant hand to the pieces region. This observation, though not common, has some design implications as we will discuss later.

Not surprisingly, hand type (dominant or non-dominant) had a significant main effect on the number of moves made to the pieces region ($F(1,9)=6.9$, $p<0.05$), with the dominant hand making more moves than the non-dominant hand. Moreover, the amount of movement by the dominant hand relative to the non-dominant one gives us some sense of the reliability of dominant hand movement as an indicator of changes in visual focus.

Camera Performance

In order to evaluate the performance of our automatic camera system in capturing dominant hand activity, we examined the percentage of time the worker's dominant hand was inside the camera view. For all the tasks combined, this percentage was 78.8%, indicating that the visual information about the dominant hand was presented to the helper a reasonable percentage of the time. Further, for complex tasks the dominant hand was in the camera view more often than for simple tasks (see Table 3).

Table 3. Percentage of time the dominant hand was in the camera shot for different tasks

	Simple	Complex
Identification	60.7	70.3
Construction	79.3	83.4

As can be seen in Figure 5, the mean number of times the camera moved to the pieces region for simple construction tasks is less than half the times the dominant hand moved to that region. Since our automatic camera was programmed to follow all trips to the pieces region longer than 2 seconds, the fact that more than half of the trips were not followed shows that those trips were short. On the one hand, the presence of numerous such short trips that were not followed by the camera explains why the percentage of time the dominant hand was in the camera view was lower for simple tasks; on the other hand, it restates our earlier assertion that visual information is not critical for simple tasks. This indicates our camera control system succeeded, at least to some extent, in providing the information only when it was critically needed, which was one of the intents of our initial system design.

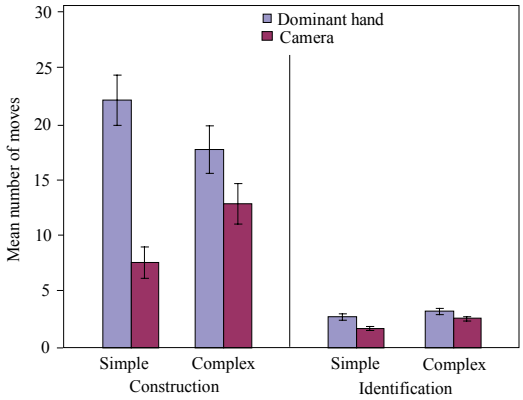


Figure 5. Mean number of moves for the dominant hand and the camera by task complexity for both task types.

DISCUSSION

Implications for Theory

We began this study with the goal of exploring the value of worker hand location as a predictor of the helper's desired focus of visual attention in a collaborative remote repair task. We developed an automatic camera control system that selected and adjusted camera shots based on the location of the worker's dominant hand, and hypothesized that this system would improve pair performance in terms of completion time and the number of errors, with possibly greater benefits for complex tasks.

The results show that our system had a substantial impact on reducing completion time and errors, but the benefits were not seen for both levels of task complexity. Completion times were improved by a statistically significant margin only for complex tasks, but not for simple ones. This partly reinforces Gergle's [13] finding that a shared visual space is more helpful for lexically complex tasks than for simple ones, but suggests further that the shared visual space must provide sufficient detail to allow for monitoring and discussing specific task elements.

Indeed, our questionnaire data suggest that the real value of the automatic camera system lie in the helper's ability to

identify and monitor the placement of detailed task objects. This ability, however, is not unique to our study. Prior systems, such as head-mounted cameras [9] or helper selection between multiple shots [10], have allowed for detailed task monitoring, but did not result in performance benefits. This leaves the question of what it is about our system that yielded the benefits seen here. We believe our use of hand tracking plays a significant role in this story.

Selecting camera shots via hand tracking has two significant benefits over prior systems. First, compared with a head-mounted camera, hand tracking allows for looser coupling [29] of movement to shot change. A head mounted camera can be described as extremely tightly coupled in that the camera necessarily changes focus every time the worker does – even when the changes are rapid or irrelevant (e.g., looking at the clock). This is potentially both intrusive for the worker and distracting for the helper, since the visual information is constantly changing.

Our system allows for the loosening of this relationship on both of these dimensions. Waiting periods can be programmed so that the camera does not follow the worker on very rapid hand moves, and the camera can be restricted to task-centric regions (possibly subject to worker override, if this were desirable) such that the worker's every glance is not taken to indicate a change in focus.

Second, our system requires less effort than those relying on manual operation by the helper or a third party operator. Our participants indicated that the system was easy to learn, and its use required little, if any, conscious effort. A few participants did, however, somewhat adapt their behavior to consciously control the camera.

This brings us to our final point of theoretical interest, which is the extent to which a system allows for and exploits behavior adaptation. Clearly, a head-mounted camera allows for very little adaptation since the worker only has one head, and it must move if focus is to change. Our system, however, allows for adaptation in that hand location is a reasonable predictor of focus, but the hand can also be easily moved to another region to "draw" the camera there, even if hand activity is not required in the new region. Moreover, the non-dominant hand can also be used if camera movement is not desirable, as we saw with some of our participants.

Implications for Practice

On the one hand, full automation of camera control seems theoretically possible by better understanding the visual focus of attention; on the other hand, manual override cannot be avoided in practice for various reasons including the adaptive nature of humans. Various instances of manual override in this study indicate that adaptive systems should provide fluid techniques for manual override.

The integration of low-overhead manual control with an automatic system is a challenging problem. In our study, the worker's dominant hand helped in the integration by serving dual purposes: the visual focus of attention and a cue for explicit manual override. The approach of tracking the objects serving such dual purposes could also be extended to other scenarios. For example, in Gaver et al.'s [12] room layout task, tracking the worker's position could be a potential way to automate the control.

We observed that the static camera was as effective as the automatic camera for simple tasks, and was also efficient in conveying the information about where the task was being performed. This suggests a potential role for static views as a fallback view for automatic systems in case of failures.

One of the reasons previous attempts to create a shared dynamic visual space using head-mounted cameras failed was unstable and shaky views [9]. In this study, special attention was paid to making the views stable in the system via region-based tracking and by introducing pauses at various transitions. This strategy was specifically useful in the simple construction task in which the worker's dominant hand was moving frequently to the pieces area but the camera was not following it tightly. This indicates that automatic systems must make provisions to balance the rate of showing visual information and the rate at which humans can process this information as excessive changes can potentially create a confusing visual space.

Limitations and Future Work

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 it 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 in this area.

Since 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). While this limitation does not negate our results, it 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.

We plan to continue investigating several areas. First, the worker's non-dominant hand was tracked, but the tracking information was not used in the automation. Considering the encouraging results based on the dominant hand only, a better understanding of the role of the non-dominant hand and its incorporation in camera control is one possible research direction. We could also consider incorporating other parameters such as gaze and head position. Finally, we are also interested in exploring the possibility of combining motion detection with other means of predicting desired helper focus, such as speech parsing [13].

ACKNOWLEDGEMENTS

We thank John Hancock, Xiang Cao, members of the Dynamic Graphics Project lab (www.dgp.toronto.edu) and the anonymous reviewers for their comments and suggestions.

REFERENCES

1. http://www.vicon.com.
2. Ballantyne, G.H. (2002). Robotic surgery, telerobotic surgery, telepresence, and telementoring. *Surgical Endoscopy*, *16*(10). 1389-1402.
3. Beun, R.J. and Cremers, A.H.M. (1998). Object reference in a shared domain of conversation. *Pragmatics and Cognition*, *6*(1). 111-142.
4. Clark, H.H. (1992) *Arenas of language use*. University of Chicago Press, Chicago, IL.
5. Clark, H.H. (1996). *Using language*. Cambridge University Press, New York.
6. Clark, H.H. and Brennan, S.E. (1991). Grounding in communication. in Resnick, L.B., Levine, R.M. and Teasley, S.D. eds. *Perspectives on Socially Shared Cognition*, American Psychological Association, Washington, DC, 127-149.
7. DeVellis, R.F. (2003). *Scale development: theory and applications*. Sage Publications, Thousand Oaks.
8. Finholt, T.A. (2003). Collaboratories as a new form of scientific organization. *Economics of Innovation and New Technologies*, *12* (1). 5-25.
9. Fussell, S.R., Kraut, R. and Siegel, J. (2000). Coordination of communication: effects of shared visual context on collaborative work. in *ACM CSCW Conference*, 21-30.
10. Fussell, S.R., Setlock, L.D. and Kraut, R.E. (2003). Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. in *ACM CHI Conference,* 513-520.
11. Fussell, S.R., Setlock, L.D., Yang, J., Ou, J., Mauer, E. and Kramer, A.D.I. (2004). Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction*, *19*(3). 273-309.
12. Gaver, W., Sellen, A., Heath, C. and Luff, P. (1993). One is not enough: multiple views in a media space. in *ACM InterCHI Conference*, 335-341.
13. Gergle, D. (2006). The value of shared visual information for task-oriented collaboration, Ph.D. Thesis, *Human Computer Interaction Institute*, Carnegie Mellon University, Pittsburgh, PA.
14. Inoue, T., Okada, K. and Matsushita, Y. (1995). Learning from TV programs: application of TV presentation to a videoconferencing system. in *ACM UIST Symposium*, 147-154.
15. 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*(3). 283-315.
16. Kirk, D. and Fraser, D.S. (2006). Comparing remote gesture technologies for supporting collaborative physical tasks. in *ACM CHI Conference*, 1191-1200.
17. Kouzes, R., Myers, J.D. and Wulf, W. (1996). Collaboratories: doing science on the internet. *IEEE Computer*, *29* (8). 40-46.
18. Kraut, R. (2003) Applying social psychological theory to the problems of group work. in Carroll, J.M. ed. *HCI Models, Theories and Frameworks*, Morgan Kaufmann, New York, 325-356.
19. Kraut, R.E., Fussell, S.R. and Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction*, *18*(1/2). 13-49.
20. Kuzuoka, H. (1992). Spatial workspace collaboration: a sharedview video support system for remote collaboration capability. in *ACM CHI Conference*, 533-540.
21. Nardi, B., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S. and Sclabassi, R. (1993). Turning away from talking heads: the use of video-as-data in neurosurgery. in *ACM CHI Conference*, 327-334.
22. Nunally, J.C. (1978). *Psychometric theory*. McGraw-Hill, New York.
23. O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G. and Bruce, V. (1996). Comparison of face-to-face and video-mediated interaction. *Interacting with Computers*, *8*(2). 177-192.
24. Ou, J., Oh, L.M., Fussell, S.R., Blum, T. and Yang, J. (2005). Analyzing and predicting focus of attention in remote collaborative tasks. in *ACM ICMI Conference*, 116-123.
25. Ou, J., Oh, L.M., Yang, J. and Fussell, S.R. (2005). Effects of task properties, partner actions, and message content on eye gaze patterns in a collaborative task. in *ACM CHI Conference*, 231-240.
26. Paulos, E. and Canny, J. (1998). PRoP: personal roving presence. in *ACM CHI Conference*, 296-303.
27. Ranjan, A., Birnholtz, J.P. and Balakrishnan, R. (2006). An exploratory analysis of partner action and camera control in a video-mediated collaborative task. in *ACM CSCW Conference,* 403-412.
28. Rui, Y., Gupta, A. and Grudin, J. (2003). Videography for telepresentations. in *ACM CHI Conference*, 457-464.
29. Simon, H. (1996). *The sciences of the artificial*. MIT Press, Cambridge, MA.
30. Veinott, E., Olson, J.S., Olson, G.M. and Fu, X. (1999). Video helps remote work: speakers who need to negotiate common ground benefit from seeing each other. in *ACM CHI Conference, 302-309*.
31. Weick, K. (1998). Organizing for high reliability: processes of collective mindfulness. *Research in Organizational Behavior*, *21*. 81-123.