

An Exploratory Analysis of Partner Action and Camera Control in a Video-Mediated Collaborative Task

Abhishek Ranjan¹, Jeremy Birnholtz², Ravin Balakrishnan¹

¹Department of Computer Science
University of Toronto
www.dgp.toronto.edu
aranjan | ravin@dgp.toronto.edu

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

ABSTRACT

This paper reports on an exploratory experimental study of the relationships between physical movement and desired visual information in the performance of video-mediated collaborative tasks in the real world by geographically distributed groups. Twenty-three pairs of participants (one “helper” and one “worker”) linked only by video and audio participated in a Lego construction task in one of three experimental conditions: a fixed scene camera, a helper-controlled pan-tilt-zoom camera, and a dedicated operator-controlled camera. “Worker” motion was tracked in 3-D space for all three conditions, as were all camera movements. Results suggest performance benefits for the operator-controlled condition, and the relationships between camera position/movement and worker action are explored to generate preliminary theoretical and design implications.

Categories and Subject Descriptors

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

General Terms

Design, Experimentation, Human Factors.

Keywords

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

1. INTRODUCTION

There are 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 proximate, however, so there is increasing interest in the use of collaboration technologies for tasks such as surgery in remotely located hospitals [2, 22], repair of equipment in remote locations (e.g., jet engines, etc.), operation of scientific equipment [7, 18] and others.

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 a video view of the workspace

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference '04, Month 1–2, 2004, City, State, Country.
Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00.

where the physical task is being performed by the “worker” [11, 19]. This shared visual context can then 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 detailed manipulation and identification of objects, while still requiring a higher level overview of what is taking place. Fixed-view “scene cameras” provide a useful overview, but little detail [11], while a camera mounted on the worker’s head provides greater detail, but constrains the helper’s view to what the worker is focused on [10]. While it is possible to provide detail and overview by allowing the user to control the camera or select between multiple shots, this has been shown to be potentially distracting, confusing and time-consuming [10, 12].

An alternative approach proposed by Ou, et al. [26, 27] is to automate the provision of dynamic visual information by predicting what the helper will want to see. In this paper, we build on Ou, et al.’s exploratory work by comparing three camera control conditions and by using high-quality motion tracking technology to track worker motion in three-dimensional space. We will show some evidence pointing to the utility of automated camera control, and identify patterns in behavior that can be used to develop design heuristics for future collaboration technologies.

2. BACKGROUND AND RELATED WORK

2.1 Providing Shared Visual Context

Shared visual context has been shown to play an important role in the completion of a range of collaborative tasks [4, 6, 17, 19]. In particular these authors point out that a shared 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 [5]. In completing collaborative tasks, Fussell, et al. [11] point out that 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]. Moreover, these studies have found that there is typically not a strong need to use visual cues to monitor partner comprehension, though additional work suggests that this may be different if some component of the task requires face monitoring [25] or if users do not share linguistic common ground [32]. In most cases, however, video images of the shared workspace are more valuable than images of one’s 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.

The mechanics of configuring and providing these cues via video, however, remain an open problem. Too broad a range of views or controls for the helper can be distracting and make it difficult to establish a joint focus of attention. At the same time, having too narrow a field of view can make it difficult for the helper to effectively monitor task status and partner action. It is thus unclear how many views should be provided, where cameras should be placed, and how, if at all, they should be controlled.

2.1.1 Number of views and shot selection

A single camera constrains the helper's potential field of view to the range of that camera. Assuming the worker is aware of the helper's view (via a monitor or other awareness indicator [14]), this simplifies the task of establishing shared visual focus because there is easy mutual awareness of what is in this single shared view. However, monitoring of all actions may be difficult in that moving the camera, if possible at all, incurs time and effort.

With this in mind, some researchers have experimented with multiple simultaneous camera feeds. Gaver, et al. [12] found some value in a multi-camera approach, but also found that cutting between camera views could be jarring and did not give remote participants a good sense of how the physical space was actually configured. Moreover, such an approach requires additional bandwidth to stream video feeds and forces the user to select between video sources, which may be disruptive or distracting, and may not result in performance benefits [10].

2.1.2 Camera Placement and Orientation

With a single camera, a major constraint on the range of views available to a helper is camera placement. Prior studies have attempted to increase this range by providing, individually or in some combination, "over the shoulder" views [9] of the overall workspace and a "worker's-eye view" via a camera mounted on the worker's head. In these studies, however, the head-mounted camera was not effective in that it constrained the helper's view to what the worker was looking at. This works well for establishing joint focus within the worker's field of view, but can make it more difficult to establish joint focus outside that area [10, 11].

At the same time, however, a fixed over-the-shoulder shot is also not entirely satisfactory. Such a shot makes it easy to focus outside the worker's immediate field of view, but establishing a joint focus of attention within a wider shot can be difficult because there may be many objects in the frame simultaneously. Systems with on-screen helper gestures [9] can simplify this somewhat, but are limited by the wide shot's level of detail.

An alternative solution would have the camera placed in front of the worker. This has the significant advantage of affording a range of detailed, close-up shots that would not otherwise be possible. While this creates some potential confusion by disrupting the sense of shared orientation, Schober [31] suggests that the pairs he studied quickly adjusted their conversations to accommodate a similar shift. In other words, this will happen as part of the negotiation of grounding within a shared visual space that happens anyway [4]. There may be a slight initial delay, but it should not cause extensive confusion.

2.1.3 Camera Control

We have seen so far that there is value in a single-camera approach, but both close-up and wide shots are desirable. Thus, another approach that has been attempted is allowing the helper to

control a pan-tilt-zoom camera, sometimes mounted on a robotic dolly, at the worker site [16, 20, 28]. This facilitates establishment of a joint focus of attention and monitoring at a fine level of detail, but requires some effort on the part of the helper. Further, the act of manipulating the camera takes time and energy and may disrupt the flow of conversation.

2.2 Automation

One way to address the difficulty of camera operation while maintaining shot flexibility is to automate camera control. Prior studies have examined automating camera control in lecture rooms [30] and meeting rooms using speaker tracking, detection and cinematography rules [15]. However, automated camera control has not been explored in the context of a collaborative task as discussed here.

Ou, et al. [27] suggest that such an approach can ideally be configured to "present the right visual information at the right time" (p. 231). In developing such a system, the key question becomes one of predicting the most valuable information to show the helper [26].

This, in turn, presents the twin problems of: 1) determining what it is that the helper needs to see and when, and 2) finding patterns in behavior that correlate with a desired shot change on the part of the helper.

2.2.1 What does the helper need to see?

Determining what visual information is most valuable to the helper at any given moment is nontrivial. Even in a simple task, there are many locations where attention may be focused, such as the worker's head, hands, or different areas of the workspace. Moreover, there are different levels of detail that may be desired depending on the type of task being completed. Finally, it is difficult to gather consistent preference data from helpers on the type of visual information they need to see.

In recent work, Ou, et al. [26] simplified this problem somewhat by focusing on a PC-based puzzle task. Using a PC-based task has the advantages of constraining helper focus to the dimensions of the computer screen, affording easy tracking of worker mouse movement as a proxy for task-based action, as well as relatively easy tracking of helper gaze (via eye tracking hardware) as a proxy for visual information that is of value at a particular point in time.

This approach is also problematic in that it gives only a limited sense of the relative value of different bits of visual information potentially available to the helper. Because all visual information (e.g., the pieces bay, the workspace and the target puzzle in the Ou, et al. study) were displayed on screen at all times, it is difficult to discern the extent to which a gaze at a particular area is "necessary" for task completion because there is no "cost" (in terms of time) to a quick look at another area of the screen as there might be in, say, panning the camera to change the shot in a single-camera video system.

2.2.2 Finding patterns

There are two primary data streams available for finding patterns that correlate with desired changes in visual information: helper and worker speech, and worker activity. In their study, Ou, et al. [26] found important links between speech patterns, worker mouse activity and helper gaze. For example, the worker's mouse

was in the same screen region as the helper's gaze a substantial fraction of the time.

At the same time, however, these authors acknowledge substantial differences between 3-D tasks in the real world and the 2-D on-screen puzzle task they studied. On screen, the range of worker motion is substantially constrained, there is no occlusion of objects, and the entire screen is visible to both participants at all times (i.e., there is no need to pan, tilt or zoom in for detail).

3. THE PRESENT STUDY

In the present study, we seek to build on prior work by accomplishing two goals: 1) explore the potential usage and value of automated camera control by comparing performance between groups with and without a dedicated camera operator, and 2) explore the nature of worker motion and camera motion in carrying out 3-D tasks.

Though automated camera control via user tracking is our long term goal, we track user behavior in this study only for exploratory purposes. We believe that it is only by better understanding the relationships between user behavior and camera movement that we will be able to develop effective predictive models that will eventually drive camera control.

3.1 Design

In this study, we compare performance between pairs of participants performing four construction tasks of varying complexity using Lego plastic pieces. Participants were randomly assigned on arrival to the "helper" or "worker" condition. The "worker" carried out the construction task, and the "helper" provided guidance. Each pair performed four construction tasks of varying complexity in one of three camera control conditions:

Fixed Scene Camera: A single camera, located directly in front of the worker, was fixed on an overview shot of the worker's workspace. The output was displayed on a 13" video monitor in front of the helper.

Helper-Controlled Camera: A single pan-tilt-zoom camera, located directly in front of the worker, was controlled by the helper and the output was displayed on a 13" video monitor.

Operator-Controlled Camera: A single pan-tilt-zoom camera, located directly in front of the worker, was controlled by a single dedicated operator. The operator was located in the same room as the worker, but could hear both helper and worker via headset. There was no direct interaction permitted between the operator and the worker. The operator was instructed to operate the camera as consistently as possible across pairs of subjects and to use her best judgment in showing the helpers what they needed to see. Most frequently, as we will show later, this involved following the worker's hand back and forth to the pieces area. She had spent several hours over a three week period practicing operation of the camera during pilot and practice sessions, was unaware of the goals of our research, and was the operator for all participants in this condition. The camera output was displayed on a 13" video monitor in front of the helper.

While the first condition is included largely for control purposes, the second two give us some sense of the value of visual information to the helper. When time is taken by the helper or operator to change the shot on the camera, the new information is likely of some value. Shot changes can then be correlated with specific worker motions, which were also being tracked.

3.2 Hypotheses

With regard to the effect of camera control condition on performance, we hypothesized that:

- Adding pan-tilt-zoom functionality to scene cameras would, on the whole, result in improved performance, as measured in terms of performance time, number of errors, and self-reported effectiveness.
- The benefits of camera control would be strongest in the Operator-Controlled Camera condition, because of the disruptive effects of camera operation on helper performance in the Helper-Controlled Camera.

We also expected differences across the three camera control conditions in how the workers moved their hands and how these movements correlated with camera movement.

Based on prior work using two-dimensional mouse tracking, we expected that:

- In general, hand movement in an area will correlate with camera focus on that area.

We further expected that worker action would differ across the camera control conditions. Where camera movement is not permitted, the only way for the pair to establish a visual joint focus of attention is for the worker to point at or move objects up towards the camera. Thus, we expected that:

- In general, there would be more hand movement closer to the camera (and away from the worker's body) in the Fixed-Scene Camera Condition than in the other two conditions.
- Adding pan-tilt-zoom functionality to scene cameras would result in less constrained movements of the worker's head and hands, as measured by comparing the distribution pattern of the worker's hand position and head orientation over the entire workspace during the course of the four tasks.

Finally, we expected that the nature of the task and progress in the task would impact the amount of camera movement we saw. Specifically, we hypothesized that:

- Increased task complexity would result in increased camera movement in the Helper-Controlled Camera condition, due to increased detail and the need for more detailed monitoring .
- Because there would be fewer and fewer pieces to choose from as each task neared completion, and because the object itself would have a more definite form, there should be less camera movement in the Helper-Controlled Camera condition during the final third of a task than in the first third.

3.3 Participants

Forty-six individuals participated in the study, of whom 16 were female and 30 were male. They ranged in age from 19 to 56, with a mean of 24. All were tested for normal color vision and all but one were right-handed. Participants were not compensated directly for participating, but were competing for the chance to win \$25 gift cards awarded to the fastest pair in each of the three camera-control conditions. One pair was unable to complete the experiment in the allotted time, and was withdrawn from the data set. Participants were recruited via posted flyers and various email lists at three universities in a major city in North America.

3.4 Setup and Equipment

The helper and worker were located in separate rooms in our laboratory space. Both wore headsets attached to PC's and were able to speak to and hear each other clearly via a Google Talk connection over a wired Ethernet network.

The worker's space consisted of a desk at which the worker was seated (see Figure 1), that was divided into three distinct areas: the pieces area (25cm wide, to the worker's right), the work area (60cm wide, directly in front of the worker), and the display area (left of worker). The following equipment was used in the worker space:

Motion Tracking- Worker motion was captured utilizing a Vicon motion capture system [1] with five cameras.. The workers wore partial-finger gloves and a baseball cap (see Figure 1) that had wireless passive reflective markers attached to them. These markers allowed for all motions to be tracked in three dimensions with sub-millimeter accuracy. Specifically we tracked the position of the worker's right and left hands, and the position of the head.

Worker Camera- A Sony SNC-RZ30 pan-tilt-zoom camera was positioned on a tripod 1.1 meters in front of the worker's workspace. The camera was connected via analog coaxial cable to the video monitors mentioned above. All pan, tilt and zoom movements of this camera were logged with time-stamps for later analysis.

Displays – Two monitors were provided: a 13" NTSC video monitor that displayed the "worker camera" output, and a 17" LCD PC display that showed output from a webcam showing the helper's face.

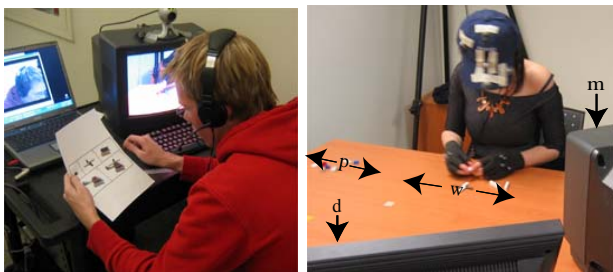


Figure 1. Photos of helper (left) and worker (right) setup for Task 1. The hat and gloves worn by the worker are used to track motion. The positions of pieces area (p), work area (w), monitor (m), and LCD display (d) are shown in the figure.

The helper's space consisted of a rolling table with a laptop PC, a Logitech Quickcam Pro 3000 USB webcam and a 13" NTSC video monitor. On the monitor the helper could see the output from the "worker camera." On the laptop display, the helper could see the output from the webcam, which was fixed on the helper's face and could not be controlled.

In the Helper-Controlled Camera condition, the helper operated the camera via the numeric keypad on an external keyboard attached to the laptop. The keyboard was directly in front of the helper. To move the camera, the helper used '4' and '6' to pan left and right, respectively, and '2' and '8' to tilt up and down. The 'Q' and 'W' keys were used to zoom in and out, because these could easily be controlled by the left hand. This control interface was iteratively developed for this study based on feedback from pilot users of an earlier, mouse based interface similar to that used in Liu, et al. [21]. Our experience and user comments suggested

that a keyboard interface was preferred due to similarity to other remote-control based camera interfaces (e.g. Polycom), the ability to operate it without looking at the control interface, and the speed of keypress input as compared with mouse movement [3].

The same interface was used by the dedicated camera operator in Operator-Controlled Camera condition, though in this case the PC used for control was located close to the worker's work desk.

All sessions were recorded for later analysis using mini-DV camcorders in both the helper and worker areas.

3.5 Materials

Tasks – One set of Lego plastic pieces was used in each task (see Figure 2). The sets varied in complexity and time required for completion, though participants were limited to no more than 30 minutes per task. Complexity varied in terms of the number of steps, number of pieces, the level of detail of the pieces, and the number of unique difficult-to-describe pieces (see Table 1). Difficulty of description was determined based on our own experience and that of pilot study participants.

Instructions – Nonverbal, picture-based, step-by-step directions were printed in color and provided to the helper for each task (see Figure 3).

Questionnaires – Questionnaires were administered to all participants prior to and following the experiment. The pre-test collected basic demographic data, extent of recent experience with videoconferencing, extent of experience with Lego toys, and included a color blindness test. The post-test included questions about the collaborative activity.

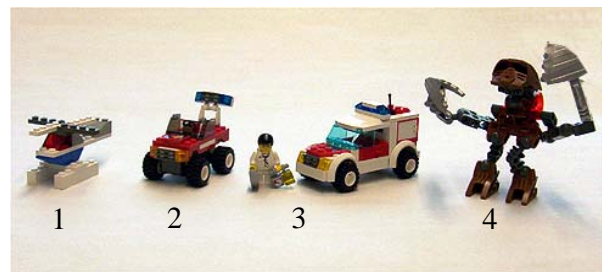


Figure 2. Lego objects in order (1-4) from left to right

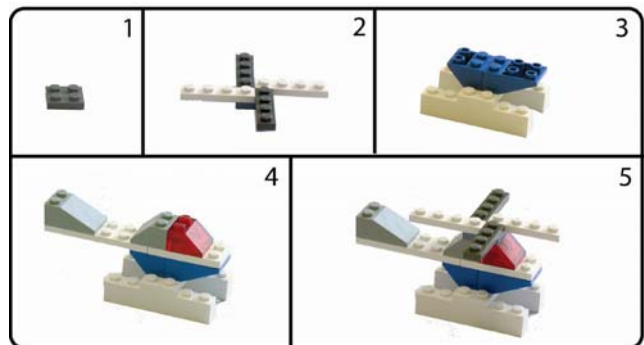


Figure 3. Sample instructions from Task 1 (helicopter).

Table 1 . Experiment Task Summary

Task	Model	Total Pieces	Unique Complex Pieces	Difficulty
1	Helicopter	15	5	Easy
2	Car	36	14	Moderate
3	Ambulance	78	27	Difficult
4	Robot	21	11	Moderate

3.6 Procedure

Once they had been randomly assigned to be “worker” or “helper,” participants were shown to their separate workspaces and the task was explained to them. Participants were then told that their goal was to, as efficiently and accurately as possible, build four objects according to instructions held by the helper.

Workers then put on the hat and gloves, and were given an opportunity to get comfortable in these. They were shown where the “pieces area” of the desk was, where the construction would take place, and what could be seen on each of the monitors. Workers were told they could not move more than four unattached pieces into the work area at once.

Depending on the camera control condition, the helper was told how to control the camera, or that they could not control the camera. In the Operator-Controlled Camera condition, they were told there was a human camera operator and that the operator could hear them and would choose appropriate camera shots throughout the task but would also respond to specific shot requests. In the Helper-Controlled Camera condition, they were given a chance to practice controlling the camera for 2-3 minutes.

Each participant then put on a headset and were asked if they could hear each other clearly. If this was true, they proceeded with the construction tasks. The order of the tasks was randomized over all of the participant pairs, and the printed instructions were given to the helper immediately prior to the start of each task.

3.7 Analysis

3.7.1 Video Analysis

Video of each session was screened to record precise task completion times, and to identify and count negotiations and critical errors. Negotiations were defined as any instance where there was back-and-forth dialogue between the participants about a piece that was difficult to place on the object, difficult to locate, or difficult for the helper to describe.

Critical errors were defined as errors by the worker that had to be corrected before certain future steps could be successfully completed. For instance, one of the tasks required a particular placement of a car steering wheel. If this was not placed properly, the windshield would not fit.

Videos were also used to transcribe specific episodes of interest for preliminary conversation analysis, and to provide validation of the Vicon motion data where details were unclear.

3.7.2 Motion Capture Data Analysis

The raw Vicon motion capture data consisted of time-stamped 3D coordinates for both of the worker’s hands and his/her head, in

addition to the position/orientation of the work desk, video monitor, and the camera. We captured these data points once per second for the duration of the four tasks. For analysis, we extracted the camera’s view vector using its position and pan-tilt-zoom values. Using the view vector we marked each time instant as “camera pointing to the pieces area”, “camera pointing to the work area” or “camera pointing to intermediate area”. Similarly, using the position of the hands we marked each time instant as “hand in pieces area”, “hand in the work area” or “hand in the intermediate area”.

The motion tracking data for one of the workers could not be captured correctly due to technical problems. Therefore, we did not include that pair’s data in the motion capture data analysis.

We were concerned that one of our “worker” participants was left-handed and that this would result in substantively different behavior that could bias our results. We closely examined the motion capture data and video data, however, and found no evidence to suggest that behavior or performance was different. As with the other workers, this participant reached into the pieces area with his right hand, and did assembly in the work area with both hands.

3.7.3 Validating Operator Consistency

To be sure that our camera operator’s behavior was consistent across all pairs of participants in the Operator-Controlled Camera condition, we examined the motion capture and video data. In doing so, we found no evidence to suggest systematic inconsistencies in operator behavior. While it might be expected that the operator would either get better or more complacent as the experiment wore on, this did not seem to occur, likely due to the operator being paid for the task as well as having received significant prior training during pilot studies.

4. RESULTS

In this section, we first examine the performance-related data, then we look at worker hand movement, and finally explore the nature of camera movement. Although our primary goal in tracking head movement was to get a sense of direction in which the worker was looking, we did not find any distinct pattern in the head movement for different conditions. One possible reason for this is that the coarse level at which we were tracking head movement was not always a good indicator of gaze direction. Therefore, we exclude any analysis of head movement in the following sections.

4.1 Performance

To measure the quality and efficiency of the participants’ performance in the four tasks, we used three measures.

First, we focused on task completion time. We hypothesized that the Operator-Controlled Camera condition would be faster than the Fixed-Scene Camera or Helper-Controlled Camera conditions. As can be seen in Table 2, however, the data do not support this hypothesis. While the mean task completion time is lower for the Operator-Controlled Camera than the other two conditions, this difference is not statistically significant in an ANOVA analysis ($F[2, 19] = 0.15, p = 0.86$). We also compared performance time across the three conditions for each task and, despite similar trends, were unable to find statistically significant differences. We suspect this is due in part to the exploratory nature of this work and the relatively small number of participants.

Second, we focused on the number of critical errors made by participants in each condition. Recall from Section 3.7.1 that a critical error was defined for our purposes as an error that impacted the successful completion of additional steps. We hypothesized that increased detail facilitated by camera control in the Helper- and Operator-Controlled Camera conditions would reduce the number of critical errors below that found in the Fixed-Scene Camera condition. As can be seen in Table 2, there was mixed support for this hypothesis. An ANOVA analysis does indicate a statistically significant main effect for camera condition ($F[2, 19] = 3.92, p < 0.05$), but testing the contrasts between conditions reveals no significant difference between the Fixed-Scene Camera condition and the Operator-Controlled Camera condition. Rather, there are significantly fewer errors in the Operator-Controlled Camera condition than in the Helper-Controlled Camera condition.

On the one hand, the lack of difference between the Helper-Controlled Camera and Fixed-Scene Camera conditions is not surprising. As we will show below, we did not see helpers make extensive use of the pan-tilt-zoom functionality of the camera, so this meant that the Helper-Controlled and Fixed-Scene Camera conditions were not substantially different for many pairs of participants during much of the time. At the same time, though, this is counterintuitive in that one might expect helpers in the Helper-Controlled Camera condition to move the camera to verify that steps were being completed correctly. In reviewing the videos, however, we found that they generally did not do so. In the Operator-Controlled Camera condition, on the other hand, shots were consistently tighter and more closely tracked the worker's hands (see below). Thus, monitoring detailed aspects of the task required less effort on the part of the helper, and this ease appears to have resulted in fewer errors. This suggests both the value of automated camera control and possible hazards from user control in mission critical situations.

Table 2 Mean Values By Condition for Performance Time, Critical Errors, and Self-Reported Effectiveness

	Fixed-Scene Camera (n=7)		User-Controlled Camera (n=8)		Operator-Controlled Camera (n=7)	
	Mean	SD	Mean	SD	Mean	SD
Total Time (min.)	48.5	10.31	51.24	13.90	48.38	9.58
Critical Errors	2.57 _a	.53	3.75 _a	1.28	2.00 _b	1.63
Effectiveness	4.29 _a	.49	4.63 _a	.74	3.57 _b	.79

Note: Means in the same row that do not share a subscript differ at $p < .05$ in contrast tests performed within an ANOVA analysis.

On the post-test questionnaire, we asked the helpers to evaluate, using a 5-point Likert Scale (anchored by strongly disagree and strongly agree), how effective the pair was at completing the tasks overall. Corresponding with our other performance hypotheses, we expected self-reported effectiveness to be highest in the Operator-Controlled Camera condition. As Table 2 shows, however, ANOVA results do show a statistically significant main

effect for camera condition ($F[2,19] = 4.48, p < 0.05$), but the difference is not in the expected direction. Rather, testing contrasts reveals that helpers felt they were most effective in the Helper-Controlled Camera condition. Given that this was also the condition in which performance time was slowest (even if not by a statistically significant margin) and error rate was highest, this is a potentially interesting finding. It becomes even more interesting in light of the fact that, as we shall demonstrate below, participants in this condition did not take advantage of camera control very often. Here we note that Rui, et al. [29] observed a split between participants who like to control the camera and participants who prefer to let the computer do the work for meeting video-archive viewing task.

4.2 Hand Movement and Camera Shot

We were interested in the extent to which worker hand movement correlated with camera movement. While some prior evidence from screen-based puzzle tasks [27] suggests that the helper is interested in seeing what the worker is doing, that was in an environment where worker motion was very easy to see. In our task, worker motion could easily be outside the camera's field of view. Thus, we were interested in how often camera movement paralleled hand movement in the two controllable camera conditions. In the Operator-Controlled Camera condition, we found that the correlation of camera view with hand position was moderate ($r = 0.54, p < 0.01$) with the right hand and weak ($r = 0.39, p < 0.01$) with the left hand. In the Helper-Controlled Camera condition, we found that this correlation was weak ($r_{right} = 0.22, r_{left} = 0.24, p < 0.01$) for both hands. The weakness of this correlation in the Helper-Controlled condition likely reflects the fact that most helpers kept the camera focused on the "work" area, while the worker's hand frequently moved back and forth between the "work" and "pieces" areas (see below).

In the Operator-Controlled Camera condition, on the other hand, right hand position was clearly followed more closely by the camera, but this correlation was still far from perfect. Given both our interest in using hand tracking to drive camera movement and our desire to claim that our operator was competent, this imperfection was of significant interest. We looked carefully at the motion capture data for the Operator- and Helper-Controlled Cameras, and identified a total of 510 discrete episodes where the worker's hand was outside of the camera shot.

Of these, the vast majority were cases where the worker's hand was outside of the camera shot for only a short period of time, and it was not possible or necessary to follow it with the camera. There were a total of 427 instances of this type (89 in the Operator-Controlled Camera condition and 338 in the Helper-Controlled Camera condition). In these cases, the worker's hand left the shot for a mean of 4.2 seconds ($SD=2.8$) before returning.

The remaining 83 cases (75 in the Operator-Controlled Camera condition and 8 in the Helper-Controlled Camera condition) were anticipatory or directive in nature. In some cases, these moves were (generally by the Helper) to direct worker focus towards a specific area or to identify a specific piece. In others, the move was anticipating a hand movement to a particular area, such as moving to the pieces area after a piece had been attached in the work area. There were also a small number of errors.

Figure 4 illustrates these types of camera moves with an approximately 500 second snapshot of camera and hand

movement to and away from the pieces area for a pair of participants in the Operator-Controlled Camera condition. In this plot, a rise indicates a move to the pieces area and a drop indicates a move to the work area. Note first that there are 3 very brief hand movements (labeled 'a') that do not have accompanying camera moves. These represent the first class of hand/camera misalignments discussed above. The second type is illustrated by the first camera move from the left (labeled 'b'). In this move, we see that the camera operator did not follow a brief movement to the pieces area, but then anticipated a hand movement back to the pieces area while the hand was in the work area.

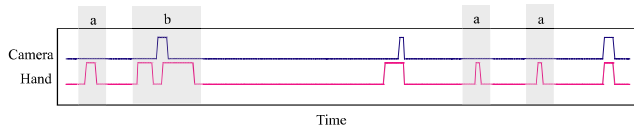


Figure 4. A 500 second snapshot of hand and camera movement in the Operator-Controlled Camera condition. A rise in the plot indicates move to the pieces area and a drop indicates move to the work area.

4.3 Worker Behavior Modification

We were also interested in the extent to which worker behavior was different across the three camera control conditions. On the one hand, support for hypothesized differences in behavior would suggest the utility of automated camera control. On the other hand, consistency across conditions could indicate patterns in worker behavior that might be useful in camera control.

As hand tracking seemed to be a promising indicator of worker activity location, we were interested in the extent to which workers made use of the entire workspace for completing the task. We hypothesized that adding pan-tilt-zoom functionality to the camera would result in less constrained movements of hand. When we looked at the hand movement on the desk between the pieces area and the work area we found an interesting pattern. The worker's hand position was largely restricted to these two areas in the Fixed-Scene Camera condition, but the distinction between these areas blurs in the Helper-Controlled Camera and Operator-Controlled Camera conditions. In other words, workers seemed to utilize a greater range of the available space when the camera could be moved to track them. We observed a marginally significant effect ($F[2, 18] = 3.13, p = 0.06$) of camera condition on the fraction of time the worker's hand spent in the intermediate area. Mean values were 0.05 ($SD = 0.03$), 0.08 ($SD = 0.13$), and 0.28 ($SD = 0.29$) per second for Fixed-Scene, Helper-Controlled and Operator-Controlled Camera conditions, respectively.

Figure 5 shows a continuous plot of how long the worker's right hand spent in different areas on the desk over the entire duration of the experiment for all participants, under different conditions. We can see that in the Operator-Controlled Camera condition the worker visits the intermediate region more frequently than in the other two conditions. This difference suggests that workers felt less need to constrain their movement in the Operator-Controlled Camera condition.

To further explore the effect of camera conditions on the user behavior, we analyzed the worker's hand movements towards the camera. We divided the work desk into two halves: towards the camera, and away from the camera. Figure 6 shows three top

views of the work surface for all participants, over the entire duration of the experiment, with one view for each camera control condition and the position of the worker's left hand positions indicated as black circles. The shaded area in the figure shows the desk half towards the camera. It can be seen that in the Fixed-Scene Camera and the Helper-Controlled Camera conditions the workers' hands moved to the half closer to the camera more often than in the Operator-Controlled Camera condition. Means of number of moves per minute are 2.69 ($SD = 5.16$), 2.23 ($SD = 4.80$), and 0.02 ($SD = 0.03$) for Fixed-Scene, Helper-Controlled, and Operator-Controlled Camera conditions respectively.

We did not find any such significant effect on the movements for the right hand. One possible reason for this is that the pieces area was to the worker's right, and thus closer to the right hand. Therefore, this hand was used to carry pieces back and forth, and was not extended towards the camera as much as the left hand.

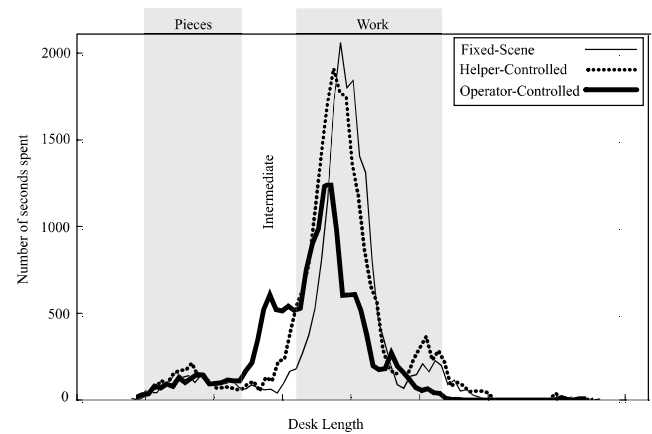


Figure 5. A plot of the number of seconds spent by the worker's right hand in the workspace, for all three conditions

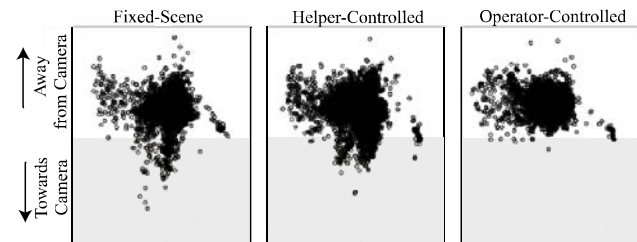


Figure 6. Three top views of workers' desk with left hand's positions indicated for the entire duration of the experiment, for all participants. Shaded area indicates the desk half closer to the camera.

This result suggests preliminarily that workers were more conscious of what was in the camera shot in the Fixed-Scene and Helper-Controlled Camera conditions, and modified their behavior accordingly. This provides some evidence in support of our hypothesis that adding pan-tilt-zoom functionality to the camera eases the establishment of a joint focus of attention. The fact that participants made less movement towards the camera in the Operator-Controlled Camera condition suggests that there was less need for workers to move closer to the camera to distinguish them from the rest of the shot, because the camera was already focused on them.

With regard to worker hand movement above the work surface, behavior appeared to be consistent across conditions. It is clear in Figure 7 that most worker action was conducted within 20 centimeters of the desk, but in all three conditions, we see some movement in vertical space. Interestingly, a clear line begins to emerge in the figure at about 30 centimeters above the desk which is just below the physical height of the camera above the work surface.

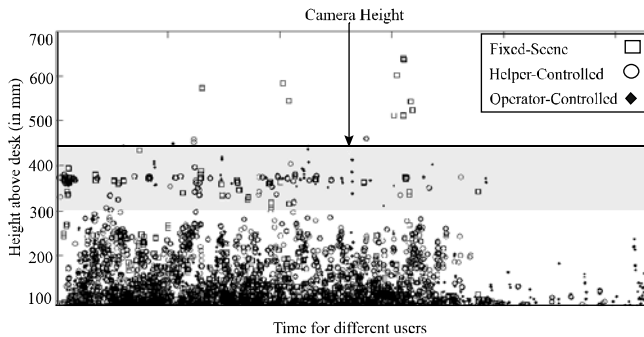


Figure 7. A plot of workers’ left hand height from 10cm above the desk height versus time for different users. The shaded region, just below the camera’s physical height, shows the region workers created to show the objects to helpers.

4.4 Understanding Camera Movement

Given these results suggesting strong, but far from perfect, relationships between movement and camera control, we wanted to preliminarily develop a better sense of the “threshold for movement.” In other words, what was different about hand movements that didn’t result in a camera move from those that did?

4.4.1 Why does the camera move?

We first wanted to characterize the nature of camera moves. As was pointed out earlier and in prior work [11], there are several potential uses for a shared visual space. Of these, camera movement and zooming are particularly well suited to both establishing a joint focus of attention, and monitoring the progress of specific portions of the task. We were interested in which of these, in the Helper-Controlled Camera condition, the camera was being used for, as this would provide some indication of when additional information that can be obtained via camera moves is useful to the user, in that they took the time to move the camera..

To investigate, we selected Task 3, which was the most complex of the four and the one with the most camera movement. Using the video data, we then coded all of the camera moves during this task for the 8 pairs of participants in the Helper-Controlled Camera condition, for a total of 52 camera moves. We coded them according to whether users were zooming in to identify a specific Lego piece (establishing a detailed joint focus of attention), panning to follow the worker’s movement to the pieces area or back to the work area (establishing a higher level joint focus of attention), zooming in to see a detailed aspect of the task (monitoring detailed task progress), or zooming out to get a more general overview (monitoring higher-level task progress) We found the results to be distributed reasonably evenly across these categories, though there were some differences. About half (52%) of the camera moves were to establish a joint focus of attention,

with 55% of these at a high level (moving between the pieces area and the work area) and 45% at a finer level of detail (zooming in to identify pieces). The other half of the moves (48%) were to monitor the task, with about 57% of these being detailed in nature (zooming in for detail) and the remaining 43% zooming out for an overview of the process. Note that there were no camera moves to see the worker’s face or otherwise monitor comprehension.

This suggests, at least preliminarily, that allowing for camera movement can serve as an aid in establishing a joint focus of attention or in monitoring a detailed task, while still maintaining the ability to have an overview without having to monitor multiple video sources simultaneously or be constrained to the worker’s field of view (as with a head-mounted camera).

4.4.2 When does the camera move?

We hypothesized that there would be more camera movement in the Helper-Controlled Camera condition when tasks were more complex, and in the early part of each task, when ambiguity was highest in terms of piece selection and construction. Support for these hypotheses was mixed.

With regard to task complexity, there does not seem to be an effect on camera movement frequency. We counted the number of camera moves and divided by the number of minutes for each pair, and compared these across the four tasks. The differences were not statistically significant ($F[3, 28] = .60, p = .62$).

As for when movement occurs within a task, however, this does seem to impact the amount of zooming that is done by helpers. When we compared the number of changes in camera zoom state per minute between the first and last thirds of each task for all pairs in condition 2, we see a statistically significant main effect in an ANOVA analysis ($F[1, 62] = 9.68, p < 0.01$). While there were a mean of 4.2 ($SD=4.3$, Median = 2.9) zoom changes per minute in the first third of each task, there were only .6 per minute ($SD=1.2$, Median = 0.0) in the final third. While an alternative explanation would hold that participants simply tired of zooming in and out and gave up as time went on, the fact that this result holds across all tasks (and that the tasks were performed in random order) suggests that reduced ambiguity at the end of the task lessened the need for zooming.

5. DISCUSSION and CONCLUSIONS

5.1 Theoretical Implications

Our goals in this study were to explore the benefits of having a designated or, potentially, automated camera operator as compared with user operation or a fixed-view camera, and to better understand how worker action relates to the information desired by a helper at any given moment.

We first hypothesized that there would be performance benefits, in terms of time and critical errors, to the Operator-Controlled Camera condition. While we could not show a statistically significant difference in performance time for this condition in the present work, there is a slight trend in the hypothesized direction and a larger study is needed to explore this result further.

There was, however, a statistically significant difference in the number of critical errors made by our participants in the three conditions. Somewhat surprisingly, participants who were permitted to control the camera had the largest number of critical errors, while those who were in the Operator-Controlled Camera

condition had the smallest number. This suggests that having to control the camera may have distracted these participants or that they were unwilling to take the time to move the camera, even when it would have been beneficial for them to do so. At the same time, this also suggests the advantage of an automated control system in allowing for relatively low-effort monitoring of detailed portions of the task where errors were likely to occur.

Despite their poor performance in terms of critical errors and relatively low numbers of camera moves, though, participants in the Helper-Controlled Camera condition self-reported their perceived effectiveness to be higher than participants in the other two conditions. This is somewhat puzzling and suggests preliminarily that there may be some psychological value in providing participants with a “manual override” in an automated setting that could boost feelings of control and, potentially, their perceived effectiveness.

We also hypothesized that camera movement would correlate with hand movement. While there was a moderate correlation in the Operator-Controlled Camera condition, this was not the case in the Helper-Controlled Camera condition. Rather, users in the Helper-Controlled Camera condition seemed to move the camera only when uncertainty about identifying a Lego piece (establishing a joint focus of attention) or alignment of detailed parts (monitoring task progress) forced them to do so. Given that monitoring task progress generally requires that the worker’s hands be present, whereas establishing a joint focus of attention does not (e.g., if the helper pans over to the pieces pile to zoom in on a desired piece and show it to the worker), this suggests that the utility of using worker motion to predict desired visual information may vary with the desired function of the visual information. While such cues may be difficult to obtain exclusively from motion tracking, such technology may have significant value in combination with speech-parsing technologies that may eventually be able to identify the desired function.

Finally, the behavior modification that we observed between conditions has several important implications. First, it suggests the potential value of automated camera control in ways that will be discussed below. Second, it suggests that providing the helper with “optimal” visual information at any given moment is a somewhat slippery optimization problem in that workers seem to modify their behavior based on what they knew the helper could or could not see. Thus, determining what the helper needs to see at any moment becomes a function, in part, of what the helper can see at any moment. This adaptation to technology is consistent with a broad range of field observations [24], and the mutual adaptation of users, technology and the environment is reminiscent of design scenarios described by Furnas [8].

5.2 Practical Implications

One practical implication of these results is that tracking hand motion appears to be different in important ways from the head-mounted camera used in prior studies. While both hand location and the head-mounted camera provide an indicator of the worker’s likely center of activity, tracking hand motion has the advantages of being less obtrusive, in that the worker need not wear a camera, and of not constraining the helper’s field of view to that of the worker. We saw in these results that there were several instances where the helper either did not need to see that the worker’s focus had momentarily shifted, or where helpers (or

the camera operator) moved the camera to a specific piece to redirect worker focus.

Another key implication is that we observed substantial behavior modification across conditions. Workers made use of space differently across the three conditions, depending on the extent to which their movement was being followed closely by the camera. The mobile nature of this construction task facilitated this sort of adjustment, however, in that workers could easily move pieces and objects around. This could be different in a setting where objects are less mobile (such as jet engine repair), and suggests that camera control may be more valuable in such settings. More work is needed in order to fully substantiate this claim, however.

Finally, it must be noted that tracking motion in 3D remains difficult and expensive, but the technology is becoming increasingly accessible. Although we use a commercial motion tracking system with reflective markers in this study, research in computer vision is approaching robust, real time tracking of bare hand postures and movement in 3D space [23].

5.3 Limitations and Future Work

In considering these results, there are several limitations that must be kept in mind.

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. At the same time, 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. At the same time, however, this limitation is common to many experimental studies in this area.

A second potential limitation of these results is the reversed orientation of the camera, as compared with prior work. We did not expect this to be a significant problem, and we found no substantial evidence to suggest otherwise. Though pairs made occasional errors in references, these were generally corrected very quickly (e.g., “No, the other left.”). More common, though, was a shift from a global coordinate space to an object-based coordinate space. In construction, most helpers instructed workers to place objects, for example, “on the left side of the car” as opposed to “on your right.” This was not possible, however, when establishing a joint focus of attention away from the object being constructed. In those cases, helpers generally specified directions in the worker’s reference frame (e.g., “on your right.”).

We plan to conduct future work on three specific problems. First, we will use the data gathered here to develop a very preliminary predictive model of desired visual information, and begin the iterative refinement of this model. Second, we will continue to refine the tasks used in this study and persist in our efforts to demonstrate performance differences between camera conditions. Third, we will explore the potential benefits of combining other information sources, such as speech [13], with tracking data in predicting what the helper wishes to see.

6. ACKNOWLEDGEMENTS

Removed for blind review.

7. REFERENCES

- [1] Vicon Web Site. <http://www.vicon.com>
- [2] Ballantyne, G.H. Robotic surgery, telerobotic surgery, telepresence, and tementoring. *Surg Endosc* 16, (2002), 1389-1402.
- [3] S. Card, Moran, T. and Newell, A. *The Psychology of Human Computer Interaction*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1983.
- [4] H.H. Clark. *Using language*. Cambridge U. Press, New York, 1996.
- [5] H.H. Clark. *Arenas of Language Use*. University of Chicago Press, Chicago, IL, 1992.
- [6] Clark, H.H. and Brennan, S.E. Grounding in Communication. In *Perspectives on Socially Shared Cognition*, L.B. Resnick, R.M. Levine and S.D. Teasley Eds. American Psychological Association, Washington, DC. 1991, 127-149.
- [7] Finholt, T.A. Collaboratories as a new form of scientific organization. *Economics of Innovation and New Technologies* 12, 1 (2003), 5-25.
- [8] Furnas, G. Future design mindful of the MoRAS. *Human-Computer Interaction* 15, (2000), 207-263.
- [9] Fussell, S.R., Setlock, L.D., Yang, J., Ou, J., Mauer, E. and Kramer, A.D.I. Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction* 19, (2004), 273-309.
- [10] Fussell, S.R., Setlock, L.D. and Kraut, R.E. Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In *Proc. CHI*, (Ft. Lauderdale, FL, April 5-10, 2003). ACM Press, 513-520.
- [11] Fussell, S.R., Kraut, R. and Siegel, J. Coordination of communication: Effects of shared visual context on collaborative work. In *Proc. CSCW*, (Philadelphia, PA, 2000). ACM Press, 21-30.
- [12] Gaver, W., Sellen, A., Heath, C. and Luff, P. One is not enough: Multiple views in a media space. In *Proc. InterCHI*, (April 24-29, 1993). ACM Press, New York, 335-341.
- [13] Gergle, D., Kraut, R.E. and Fussell, S.R. Action as language in a shared visual space. In *Proc. CSCW*, (Chicago, IL, Nov. 4-6, 2004). ACM Press, New York, 487-496.
- [14] Gutwin, C. and Greenberg, S. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work* 11, (2002), 411-446.
- [15] Inoue, T., Okada, K. and Matsushita, Y. Learning from TV programs: Application of TV presentation to a videoconferencing system. In *Proc. UIST*, (Pittsburgh, PA, November 15-17, 1995). ACM Press, New York, 147-154.
- [16] Jouppi, N. First steps towards mutually-immersive mobile telepresence. In *Proc. CSCW*, (New Orleans, LA, Nov. 16-20, 2002). ACM Press, New York. 354-363.
- [17] Karsenty, L. Cooperative work and shared visual context: An empirical study of comprehension problems in side-by-side and remote help dialogues. *Human-Computer Interaction* 14, (1999), 283-315.
- [18] Kouzes, R., Myers and Wulf, W. Collaboratories: Doing science on the internet. *IEEE Computer* 29, 8 (1996), 40-46.
- [19] Kraut, R.E., Fussell, S.R. and Siegel, J. Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction* 18, (2003), 13-49.
- [20] Kuzuoka, H. Spatial workspace collaboration: A sharedview video support system for remote collaboration capability. In *Proc. CHI*, (Monterey, CA, May 3-7, 1992). ACM Press, 533-540.
- [21] Liu, Q., Kimber, D., Foote, J., Wilcox, L. and Boreczky, J. FLYSPEC: A multi-user video camera system with hybrid human and automatic control. In *Proceedings of ACM Multimedia*, (Juan-les-Pins, France, December 1-6, 2002). ACM Press, New York, 484-492.
- [22] Nardi, B., Schwarz, H., Kuchinsky, A., Leichner, R., Whittaker, S. and Scabassi, R. Turning away from talking heads: The use of video-as-data in neurosurgery. In *Proc. InterCHI*, (Amsterdam, 1993). ACM Press, New York, 327-334.
- [23] Nickel, K. and Stiefelwagen, R. Pointing gesture recognition based on 3-D tracking of face, hands and head orientation. In *Proc. ICMI*, (Vancouver, BC, Nov. 5-7, 2003). ACM Press, New York, 140-146.
- [24] Olson, G.M. and Olson, J.S. Distance matters. *Human-Computer Interaction* 15, (2001), 139-179.
- [25] O'Malley, C., Langton, S., Anderson, A., Doherty-Sneddon, G. and Bruce, V. Comparison of face-to-face and video-mediated interaction. *Interacting with Computers* 8, 2 (1996), 177-192.
- [26] Ou, J., Oh, L.M., Fussell, S.R., Blum, T. and Yang, J. Analyzing and predicting focus of attention in remote collaborative tasks. In *Proc. ICMI*, (Trento, Italy, October 4 - 6, 2005). ACM Press, 116-123.
- [27] Ou, J., Oh, L.M., Yang, J. and Fussell, S.R. Effects of task properties, partner actions, and message content on eye gaze patterns in a collaborative task. In *Proc. CHI*, (Portland, OR, April 2-7, 2005). ACM Press, 231-240.
- [28] Paolos, E. and Canny, J. PRoP: Personal roving presence. In *Proc. CHI*, (Los Angeles, CA, Apr. 18-23, 1998). ACM Press, New York, 296-303.
- [29] Rui, Y., Gupta, A. and Cadiz, J.J. Viewing meeting captured by an omni-directional camera. In *Proc. CHI*, (Seattle, WA, March 31 - April 5, 2001). ACM Press, New York, 450-457.
- [30] Rui, Y., Gupta, A. and Grudin, J. Videography for telepresentations. In *Proc. CHI*, (Ft. Lauderdale, FL, April 5-10, 2003). 457-464.
- [31] Schober, M.F. Spatial perspective-taking in conversation. *Cognition* 47, 1 (1993), 1-24.
- [32] Veinott, E., Olson, J.S., Olson, G.M. and Fu, X. Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In *Proc. CHI*, (Pittsburgh, PA, May 15-20, 1999). Pittsburgh, PA